Model uncertainty in Airborne Laser Scanning assisted forest inventories

Philip Mundhenk[1] Christoph Kleinn[1] Thomas Kneib[2]
Liviu Ene[3] Mike Wulder[4] Steen Magnussen[4]

[1]Chair of Forest Inventory & Remote Sensing
RTG 1644 — *Scaling Problems in Statistics*
Georg–August University Göttingen, Germany

[2]Chairs of Statistics & Econometrics
Georg–August University Göttingen, Germany

[3]Department of Ecology and Natural Resource Management
Norwegian University of Life Sciences (UMB)
Ås, Norway

[4]Pacific Forestry Center
Canadian Forest Service
Victoria, BC, Canada

Biometry Workshop — Freising, November 7, 2013

Introduction

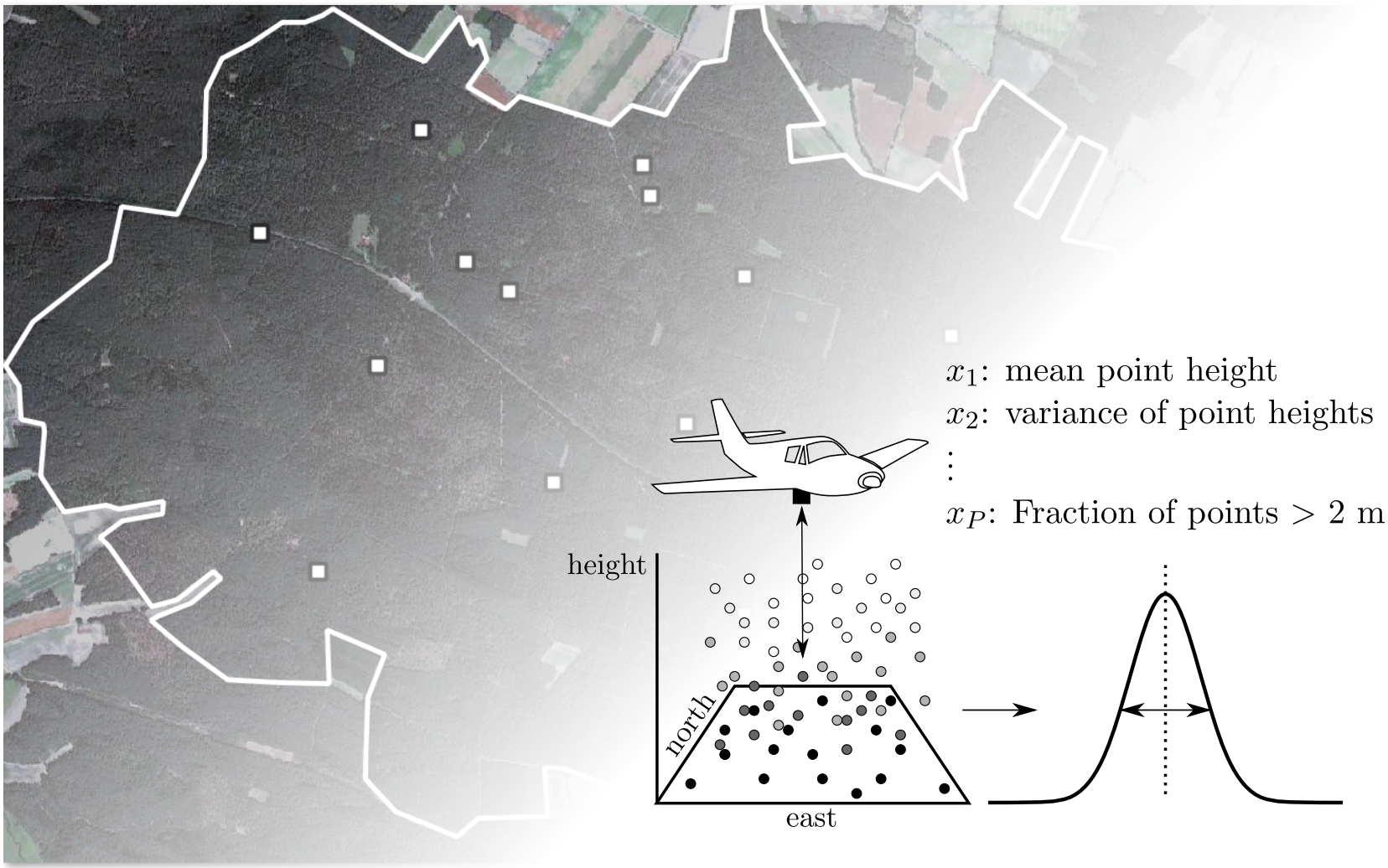
Target variables

Introduction

Field sample: observations/estimates of the target variable(s)

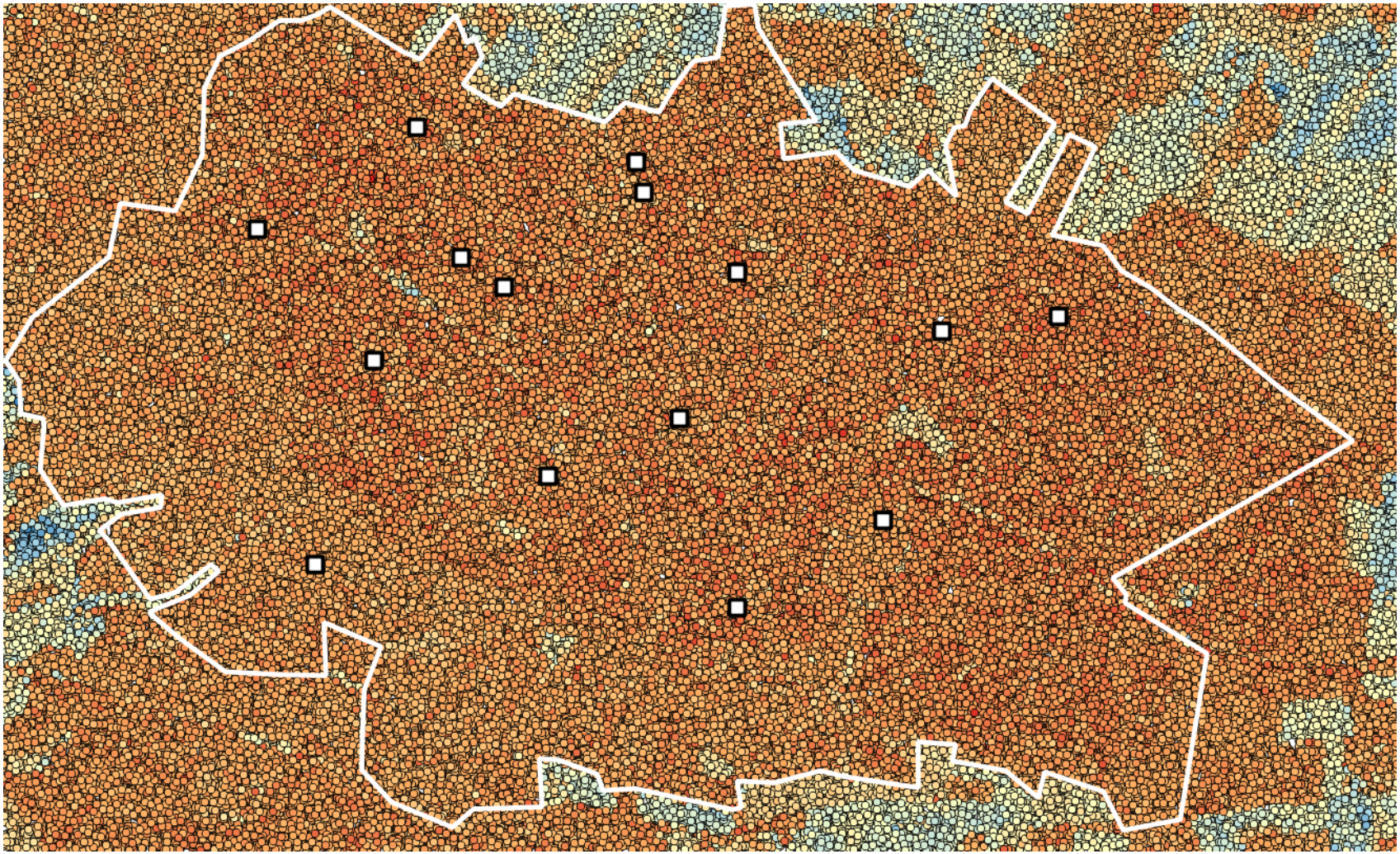
Introduction

Auxiliary information: Airborne Laser Scanning (ALS) data



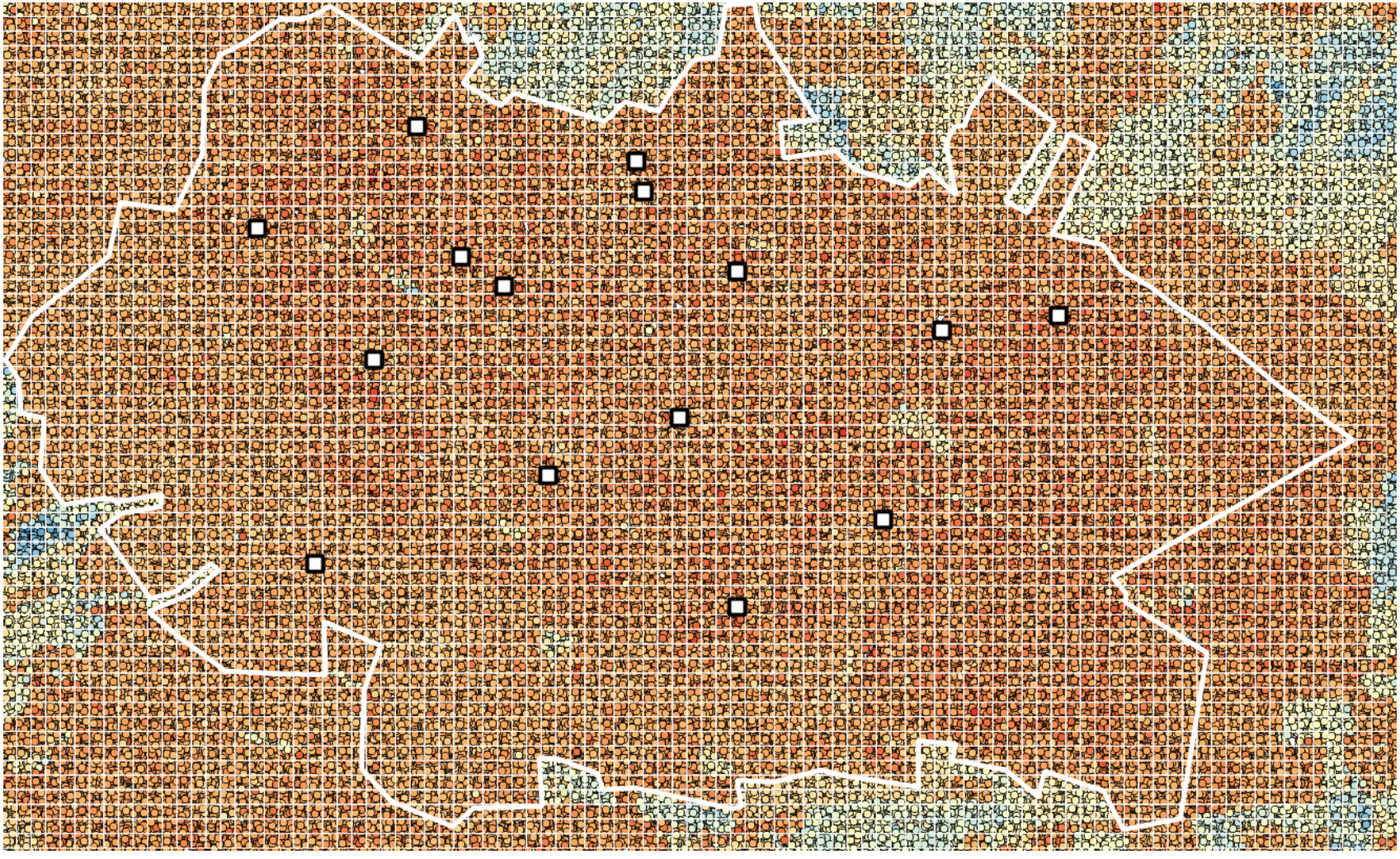
Introduction

A huge point cloud



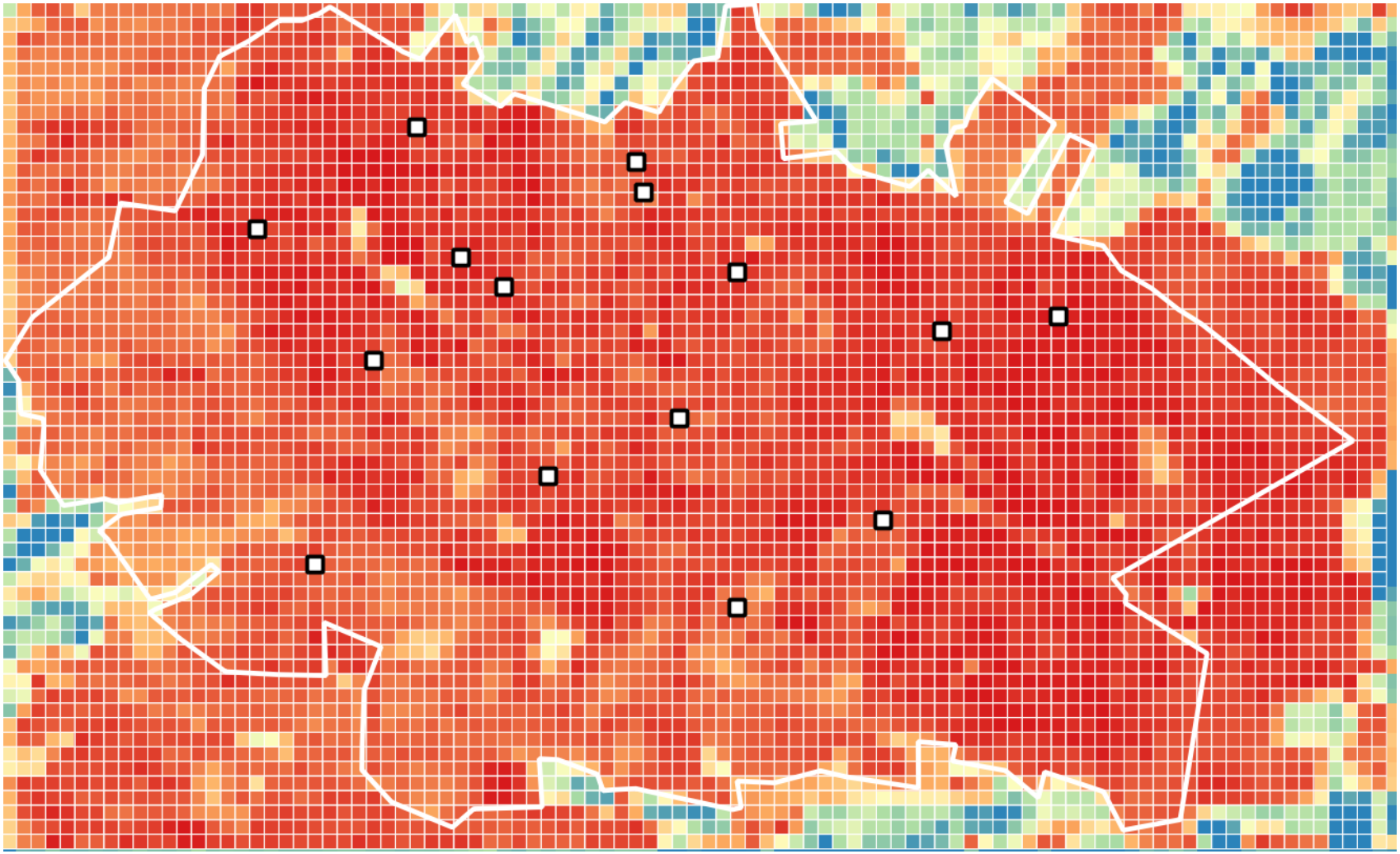
Introduction

"Gridify" the cloud



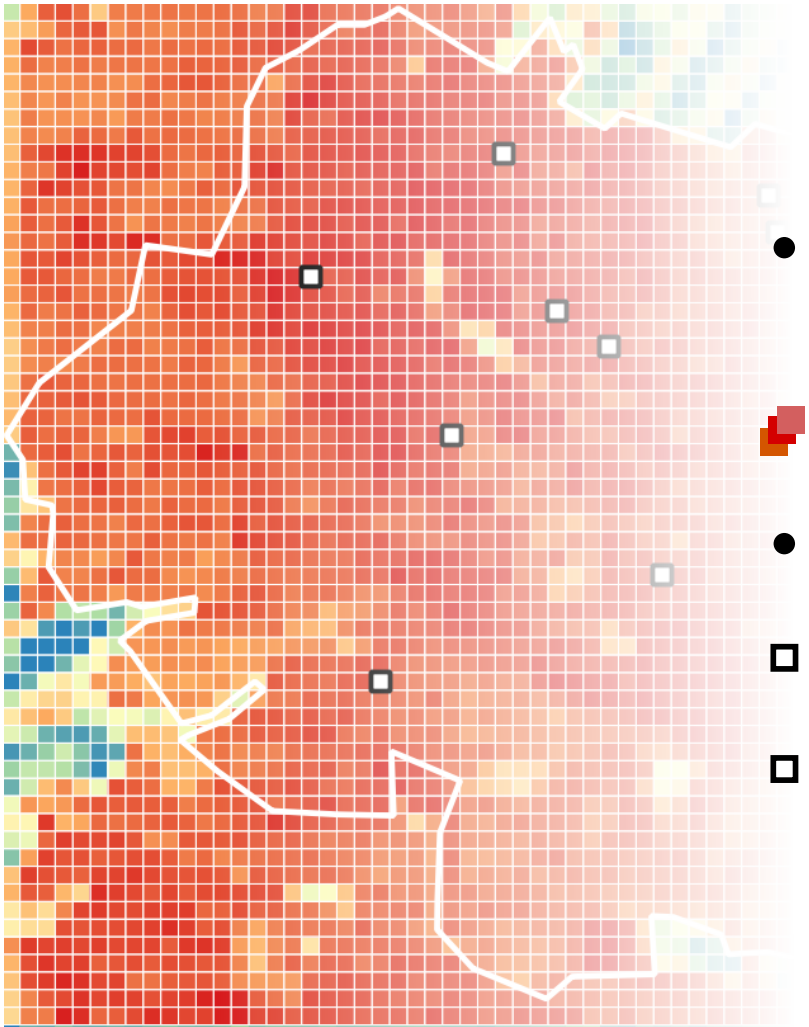
Introduction

ALS metrics for each grid cell



Introduction

Notation



- Finite population U of N grid cells $k = 1, 2, \ldots, N$
- $\boldsymbol{x}_k = (x_{k1}, \ldots, x_{kP})' \quad \forall k \in U$
- P: number of ALS metrics
- Sample s of size n (SRSwoR)
- For all $k \in s$: $(y_k, \boldsymbol{x}_k')$

Introduction

The "working–model"

- "Working–model":

$$y_k = \boldsymbol{x}_k^\top \boldsymbol{\beta} + \varepsilon_k \quad \forall k \in s$$

- Prediction:

$$\hat{\boldsymbol{\beta}}_s = (\boldsymbol{X}_s' \boldsymbol{X}_s)^{-1} \boldsymbol{X}_s' \boldsymbol{y}_s$$

$$\hat{y}_k = \boldsymbol{x}_k^\top \hat{\boldsymbol{\beta}} \quad \forall k \in U - s$$

- Simple approach:

$$\hat{t}_y = t_{y,\text{field plots}} + \hat{t}_{\hat{y},\text{model}}$$

- To obtain an unbiased estimate, the model has to be correctly specified

Introduction

The generalized regression (GREG) estimator (Särndal *et al.*, 1992)

- Alternative (design-based) approach:

$$\hat{t}_y = \hat{t}_{y\pi} + \hat{\boldsymbol{B}}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi})^\top$$

where

$$\hat{t}_{y,\pi} = \sum_s \frac{y_k}{\pi_k}, \quad \pi_k = \mathbb{P}(k \in s) \quad (= \frac{n}{N} \quad \text{under SRSwoR}),$$

$$\boldsymbol{t}_x = \sum_U \boldsymbol{x}_k, \quad \hat{\boldsymbol{t}}_x = \sum_s \frac{\boldsymbol{x}_k}{\pi_k}$$

and

$$\hat{\boldsymbol{B}} = (\hat{B}_1, \ldots, \hat{B}_J)' = \left(\sum_s \boldsymbol{x}_k \boldsymbol{x}_k' / \sigma_k^2 \pi_k \right)^{-1} \sum_s \boldsymbol{x}_k y_k / \sigma_k^2 \pi_k$$

$\rightarrow \hat{\boldsymbol{B}}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi})^\top$ should average to zero (on the long run...)

Introduction

Variance estimator for GREG

$$\hat{V}(\hat{t}_y^{GREG}) \doteq N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) \hat{s}_{\hat{e}}^2$$

where

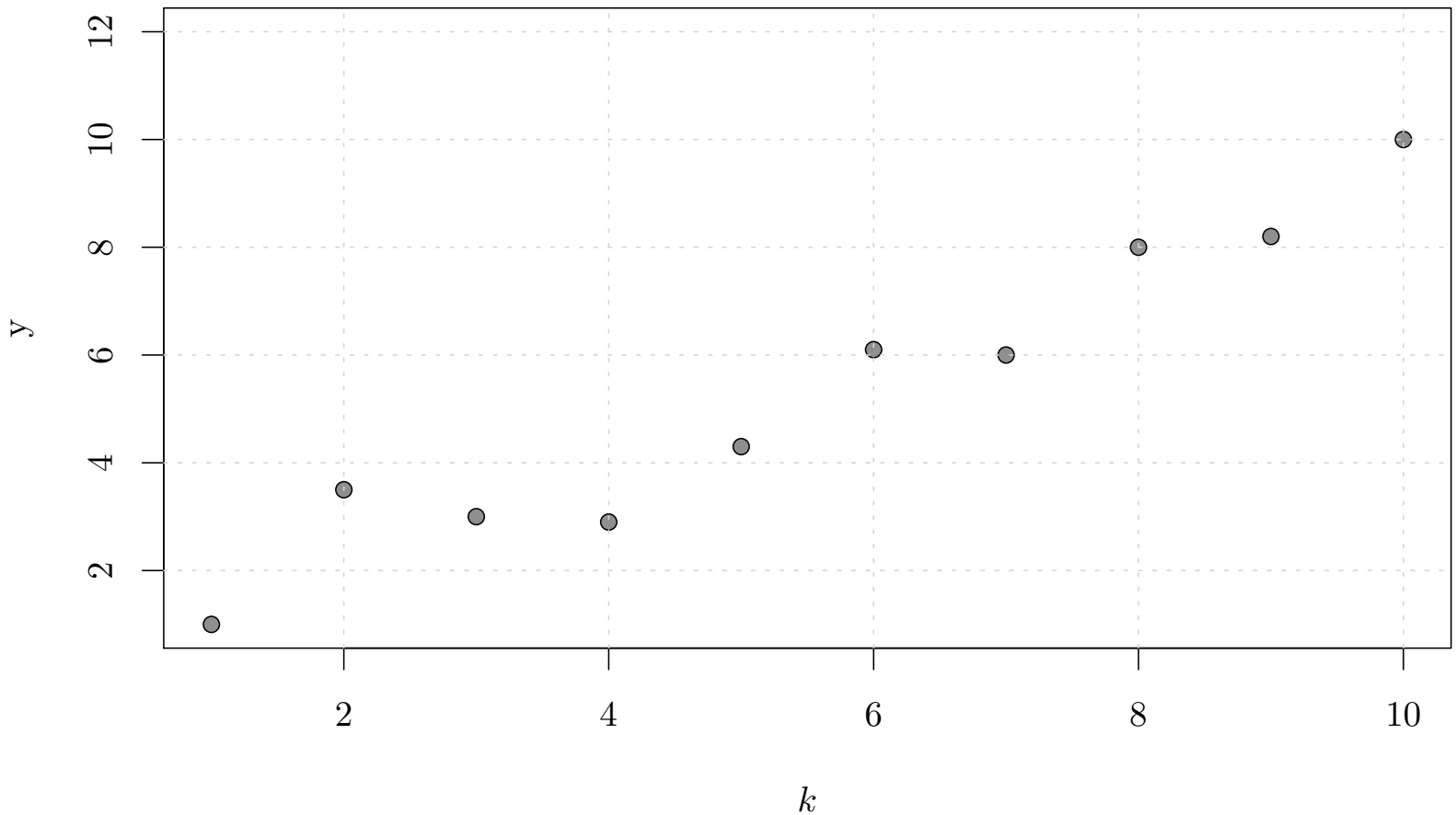
$$\hat{s}_{\hat{e}}^2 = \sum_s (\hat{e}_k - \bar{\hat{e}}_k)^2/(n-1), \quad \bar{\hat{e}} = \sum_s \hat{e}_k/n, \quad \hat{e} = y_{k\in s} - \hat{y}_{k\in s}$$

and

$$\hat{y}_{k\in s} = \boldsymbol{x}'_{k\in s}\hat{\boldsymbol{B}}$$

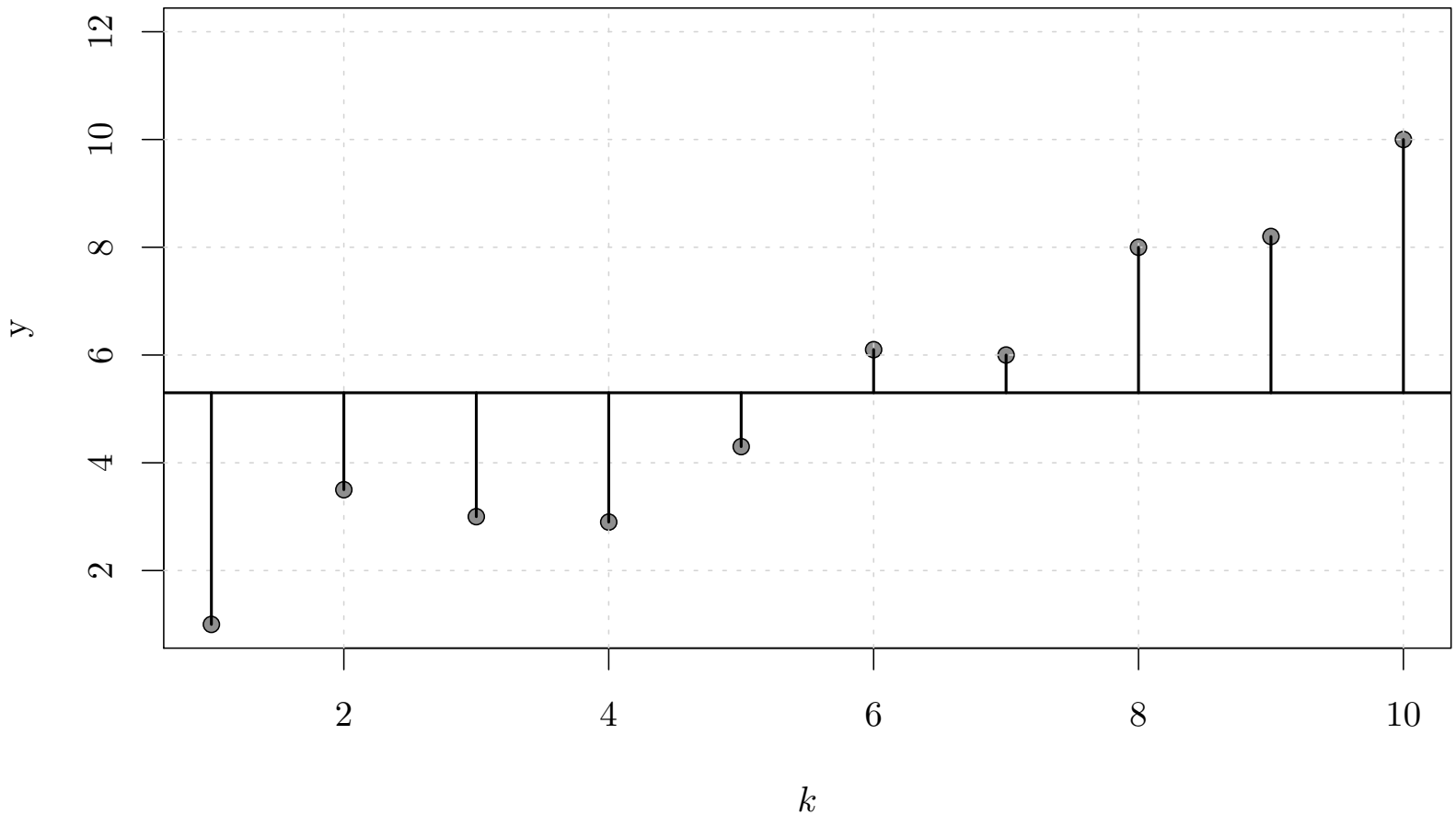
Introduction

A sample of y_k's



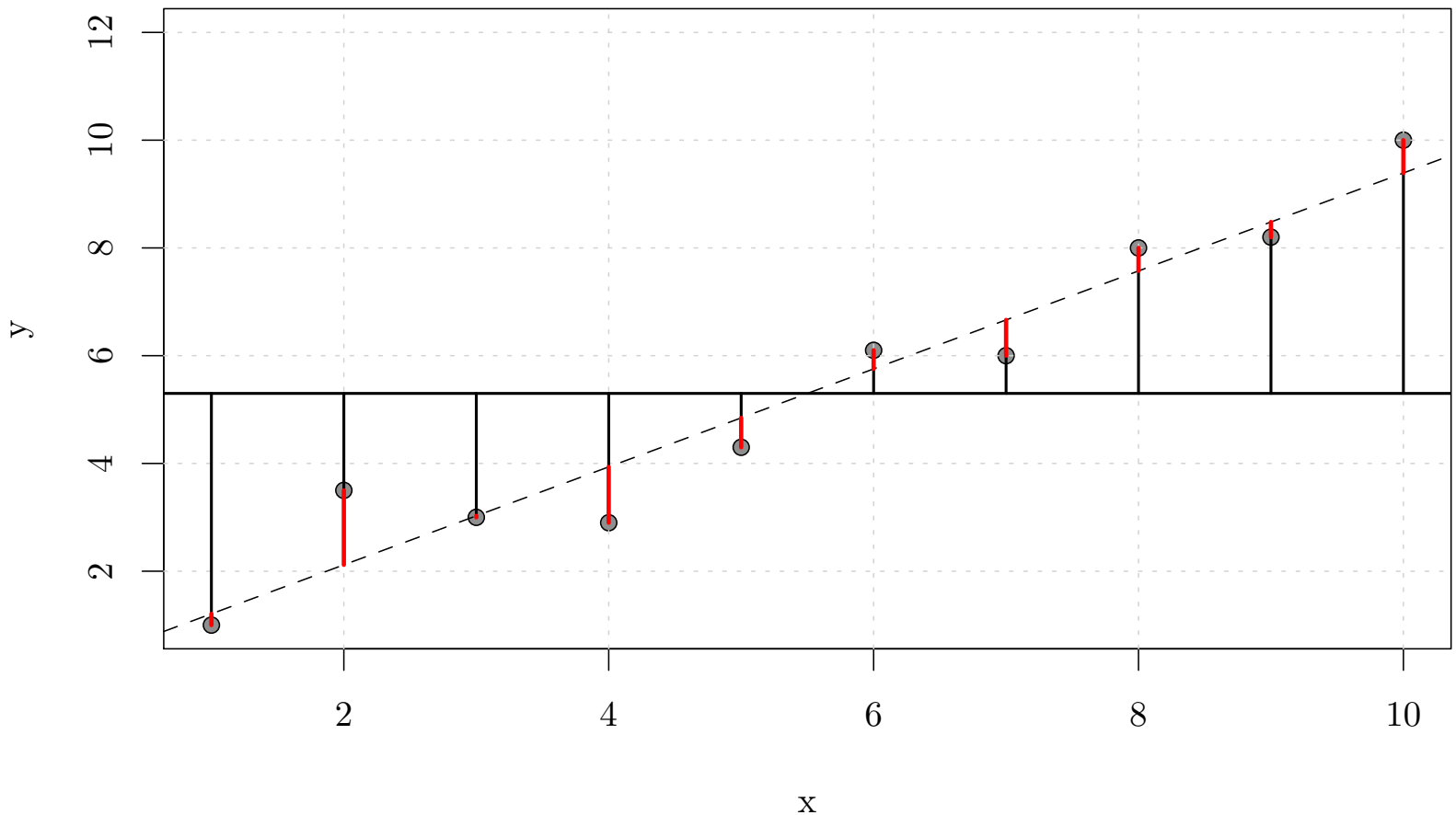
Introduction

"Global" mean & variance (no auxiliary information)



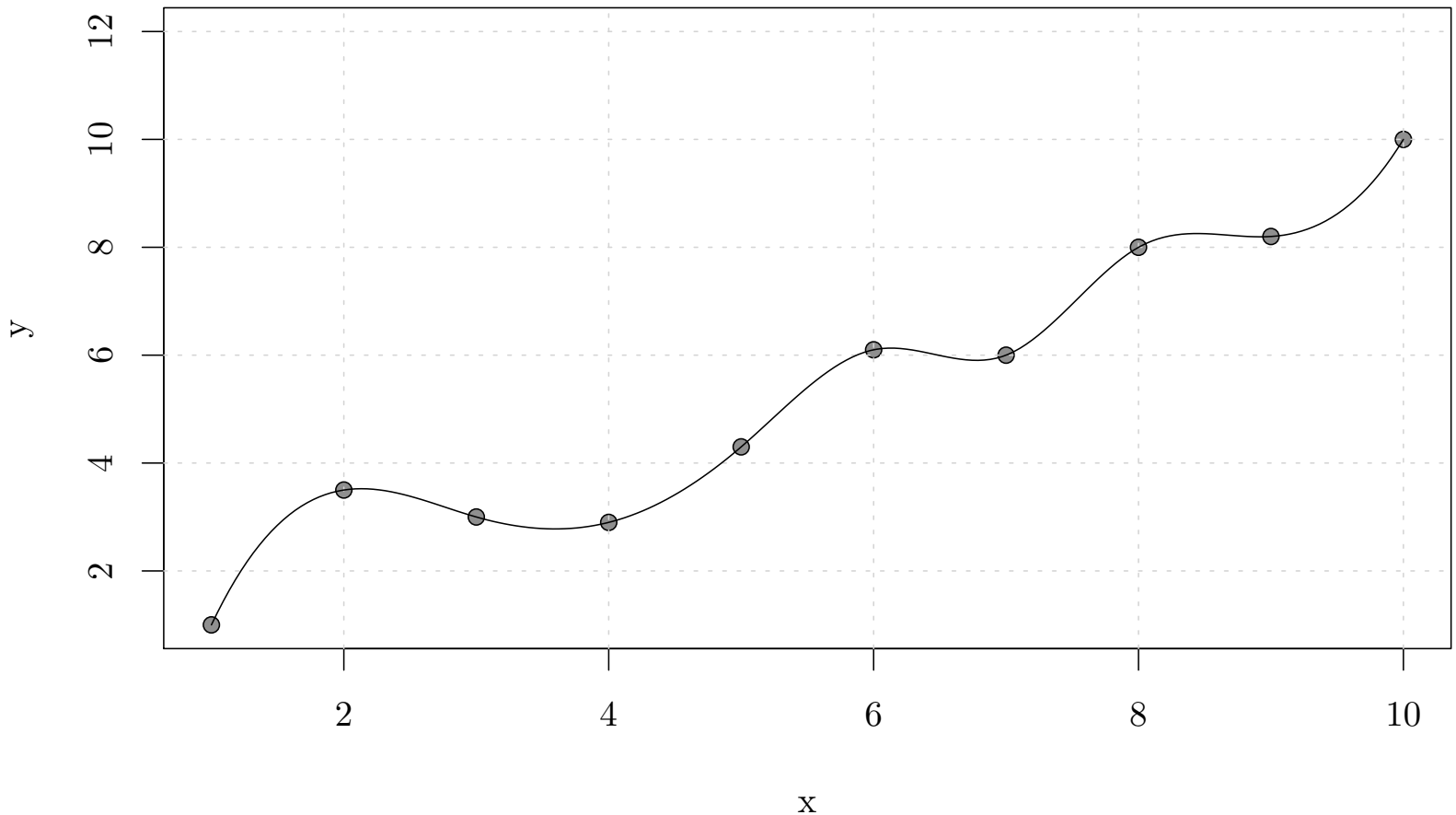
Introduction

Variance reduction when using auxiliary information: $\hat{V}(\hat{t}_y^{GREG}) = \hat{V}(\hat{t}_{y\pi}) \times (1 - R^2)$



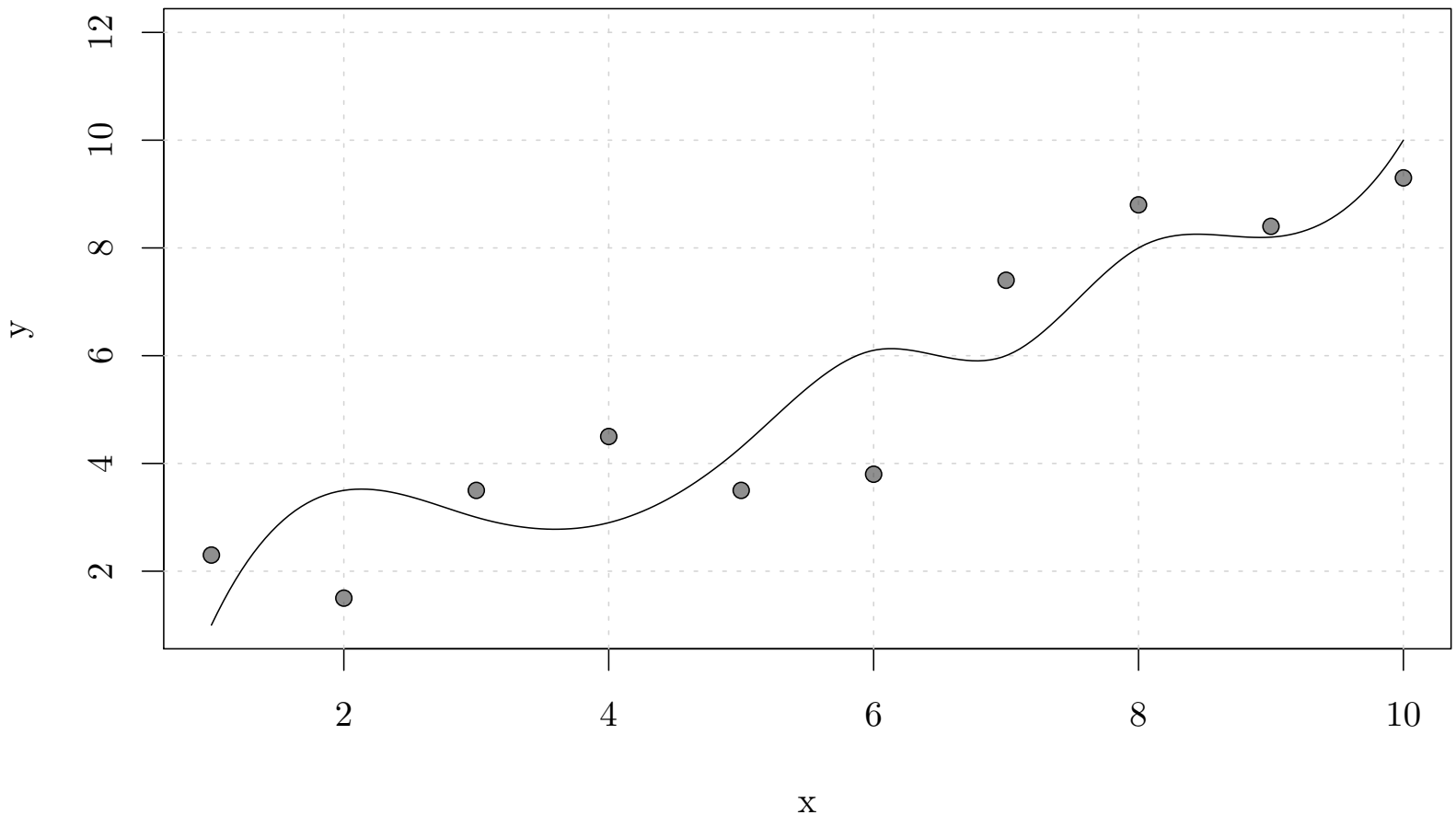
Introduction

A perfect fit: $\hat{V}(\hat{\mu}_y^{GREG}) = 0$



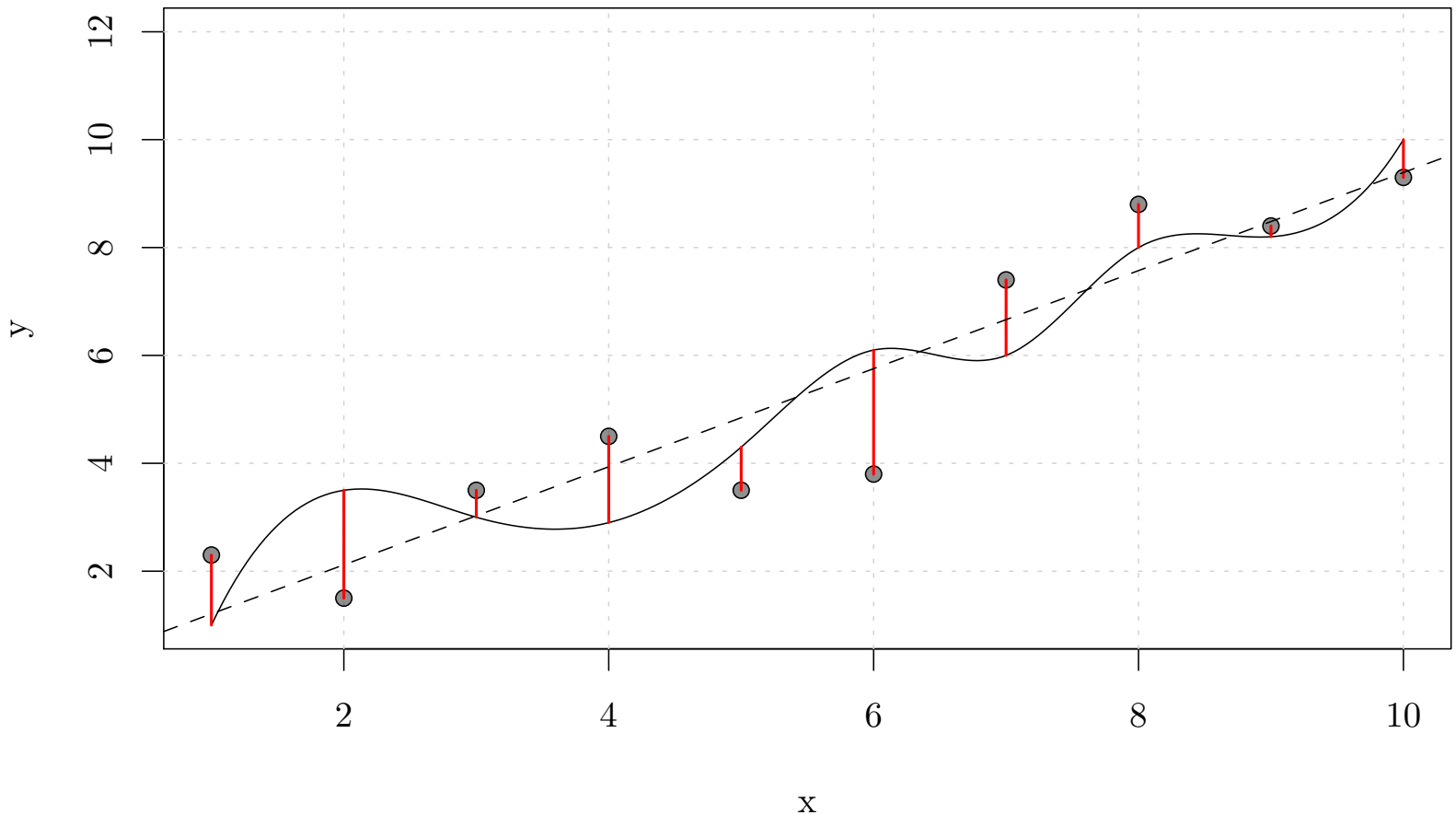
Introduction

"Perfect" fit to another sample: $\hat{V}(\hat{\mu}_y^{GREG}) \gg 0$



Introduction

Simple vs complex model



Introduction

ALS-assisted forest inventories

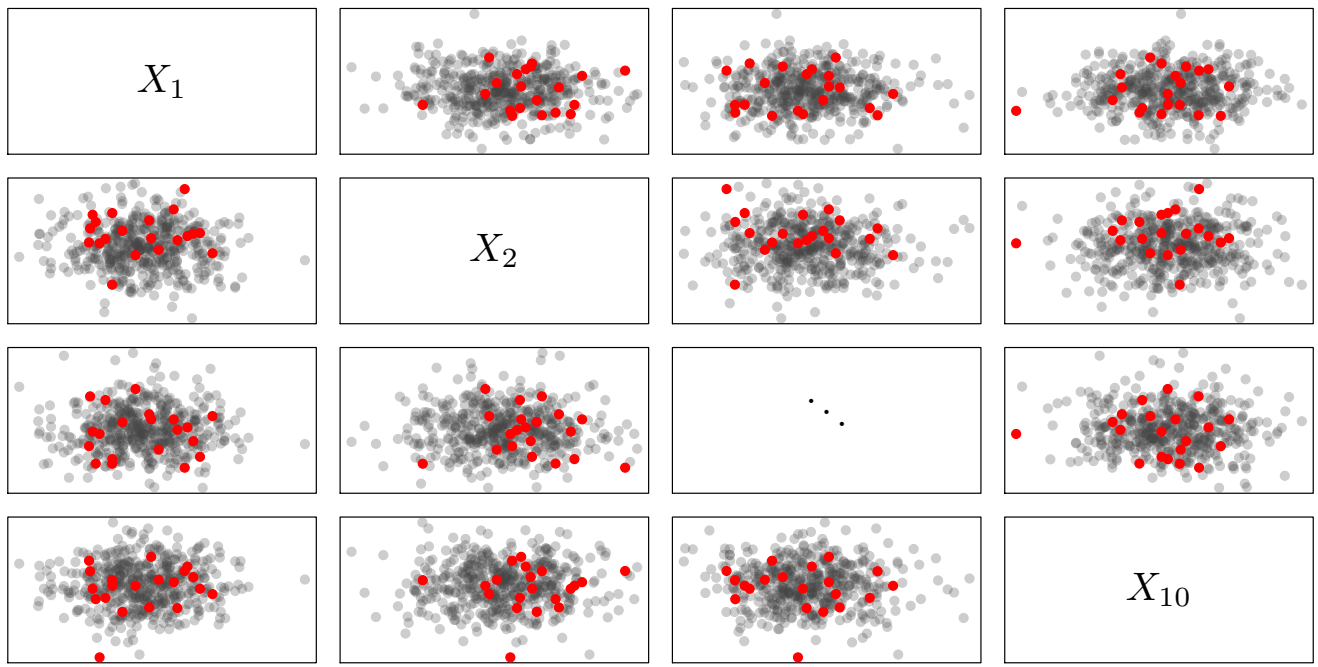
In ALS-assisted forest inventories...

- we use the **sample data** to formulate and fit the "working-model"
- the sample size is small relative to the population size: n $\ll$ N
- there are many potentially useful explanatory variables (ALS metrics): large P, small n
- how to "find" the "working" model? How to select variables?

Introduction

Example (taken from Chatfield, 1995)

- Suppose we have $p = 10$ random variables $X_1, X_2, \ldots, X_{10}$
- Each X_p follows $\mathcal{N}(0, 1)$
- Select a sample of size $n = 20$ from each X_p



Introduction

Example II

- Let $X_1 \sim f(X_2, \ldots, X_{10})$:

$$x_{k1} = \beta_0 + \beta_1 x_{k2} + \cdots + \beta_9 x_{k10} + \varepsilon_k$$

- The true model is the null model!
- Select four variables that "explain" X_1 best (using AIC)
- Simulation shows: $\bar{R}^2 = .42$ (max. .86)
- Thus, we explain on average 42% of the variability in X_1, although there should be nothing to explain at all!

"The moral is that subset selection can be dangerous..."
(Chatfield, 1995)

Introduction

Research questions

Since $\hat{V}(\hat{t}_y^{GREG}) = \hat{V}(\hat{t}_{y\pi}) \times (1 - R^2)$:

- . . . do we overestimate precision when we search for a model that fits the (single) sample data best?
- . . . are there differences among different statistical variable selection procedures?
- How does the sample size affect estimates of precision?

Methods

Methods

Data

- Two synthetic populations based on FI data from
  - Hinton Wood Products, western Alberta, Canada[1]
  - Hedmark County, south-eastern Norway[2]
- Airborne Laser Scanning (metrics) + field plot data
- Study variable: aboveground biomass in Mg ha^{-1}
- Auxiliary information: ALS metrics
  - 29 Hinton
  - 19 Hedmark
- Canonical vine copulas and kNN were used to simulate aboveground biomass, $y_{k \in U}$ (Ene *et al.*, 2013)
- Result: two populations with $N = 50,000$ where $(y_k, \boldsymbol{x}'_k)$ are known $\forall k \in U$

[1] Provided by J. White & M. Wulder

[2] Provided by E. Næsset, Terje Gobakken and L. Ene

Methods

Simulation study

- Select sample (SRSwoR) with $n = 100$
- Fit "working" model using different variable selection procedures (next slide)
- Estimate $\hat{t}_y^{GREG}$ and $\hat{V}(\hat{t}_y^{GREG})$
- Repeat 100,000 times for each variable selection procedure

After a couple of days...

- Get average $\sqrt{\hat{V}(\hat{t}_y^{GREG})} = \hat{\text{SE}}$ for each variable selection procedure
- Get standard deviation of $\hat{t}_y^{GREG}$ for each variable selection procedure (empirical standard error)

Methods

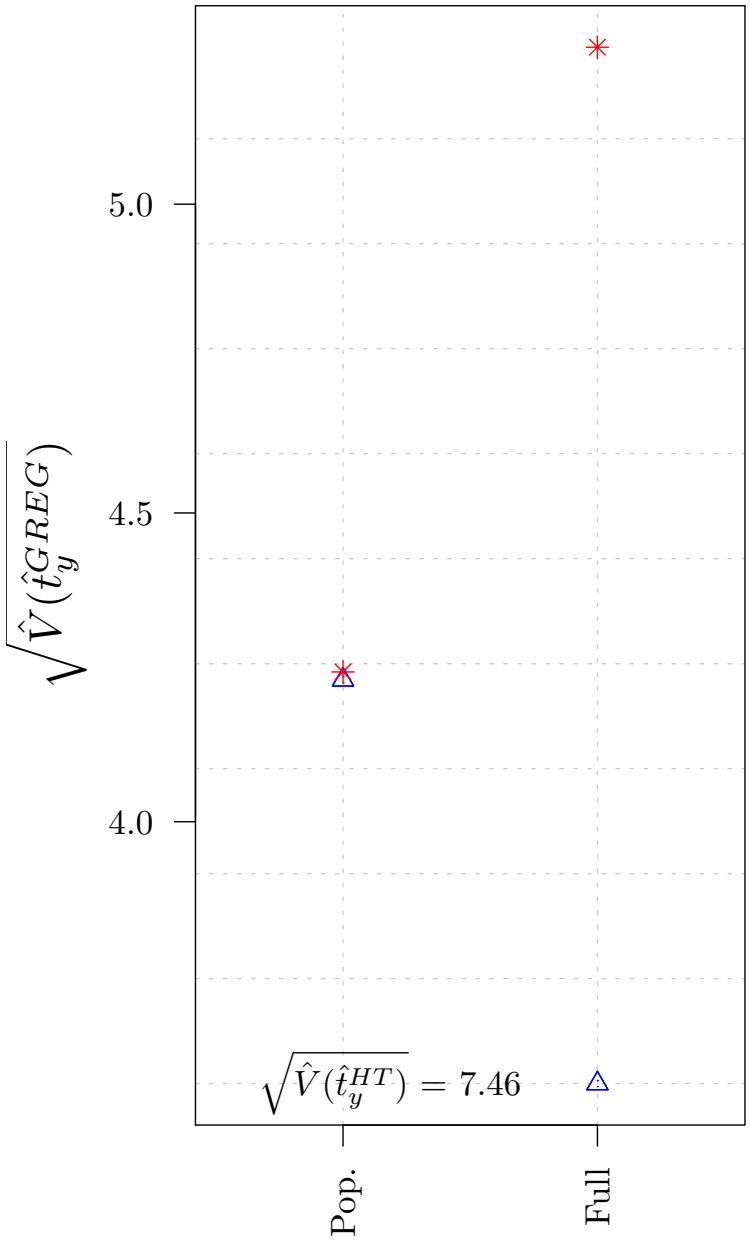
Variable "selection" procedures

- Full, "true" population model
- Full model (sample), i.e., no variable selection
- AIC and BIC (backward selection)
- Ridge regression
- Bayesian model averaging (BMA)
- . . . (e.g., Lasso, elastic net, PLS, PCR, etc.)

Results

Results

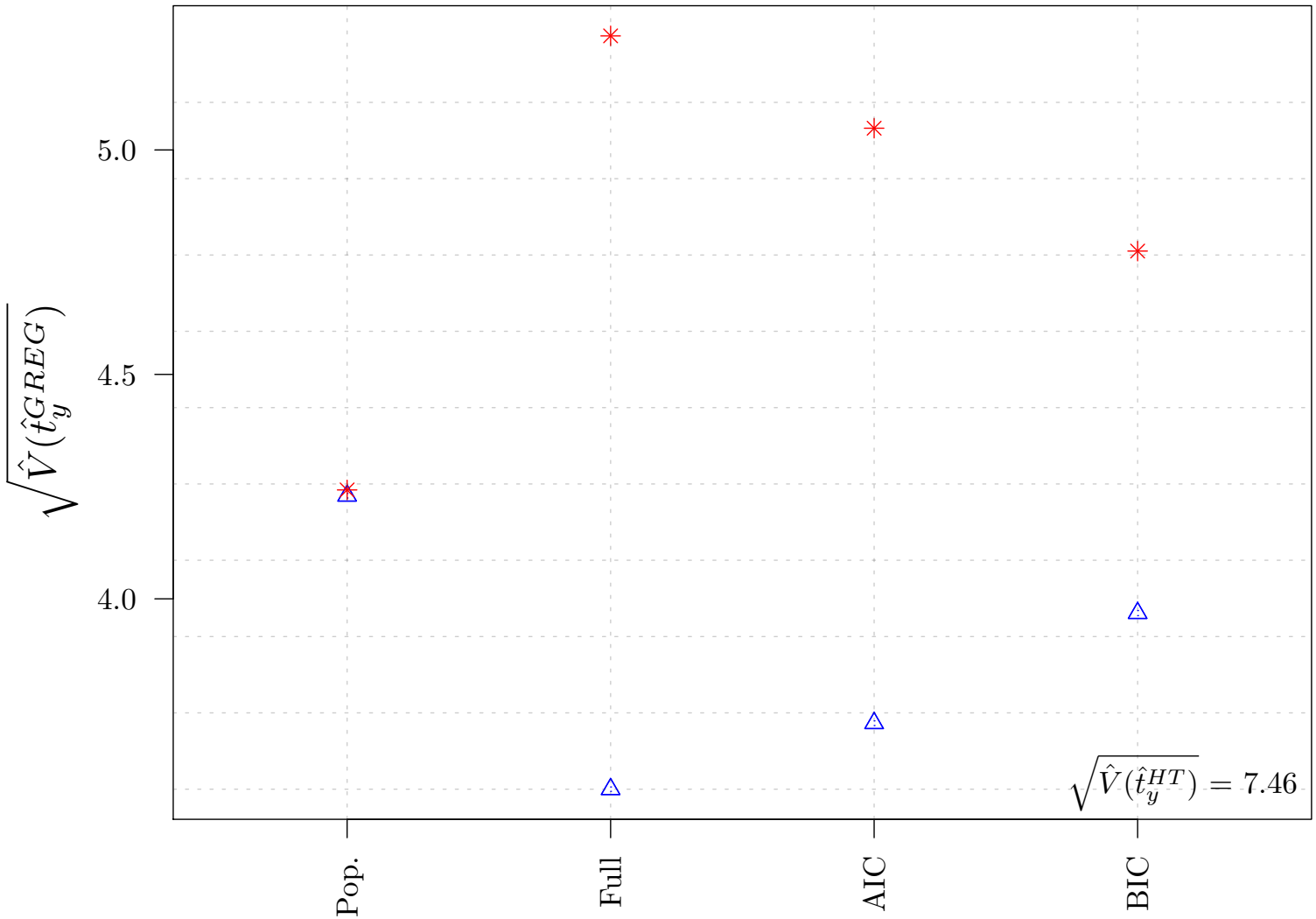
Pop.: model fitted to population data (for comparison), full: no variable selection



- Sample size: $n = 100$
- x-axis: variable "selection" procedure
- y-axis: estimated SE
- $\triangle$: mean estimated SE (after 100,000 runs)
- $\star$: SD of estimated means (empirical SE)
- For the full (sample based) model the SE is underestimated by $\sim 32\%$!

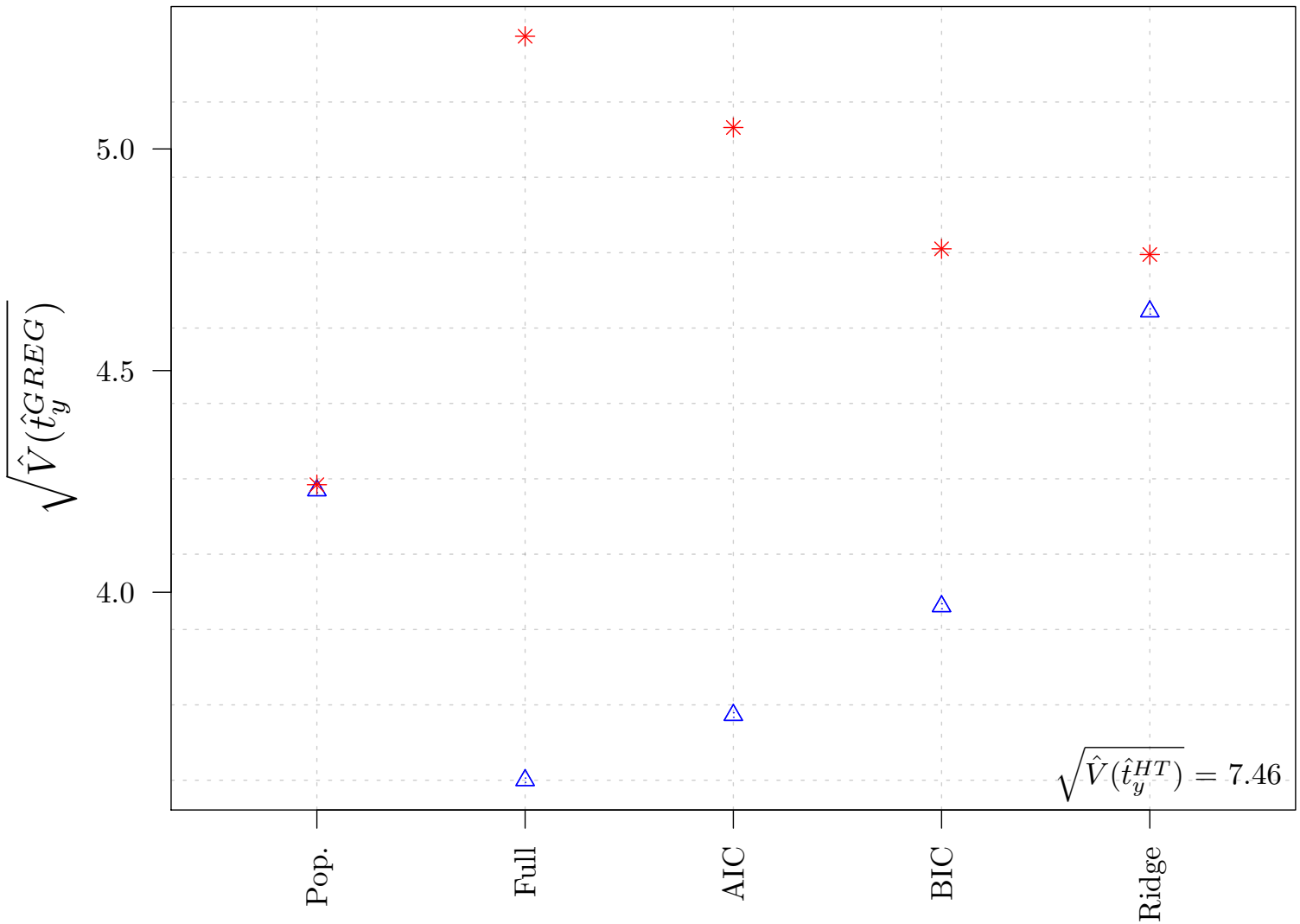
Results

Full model, AIC and BIC



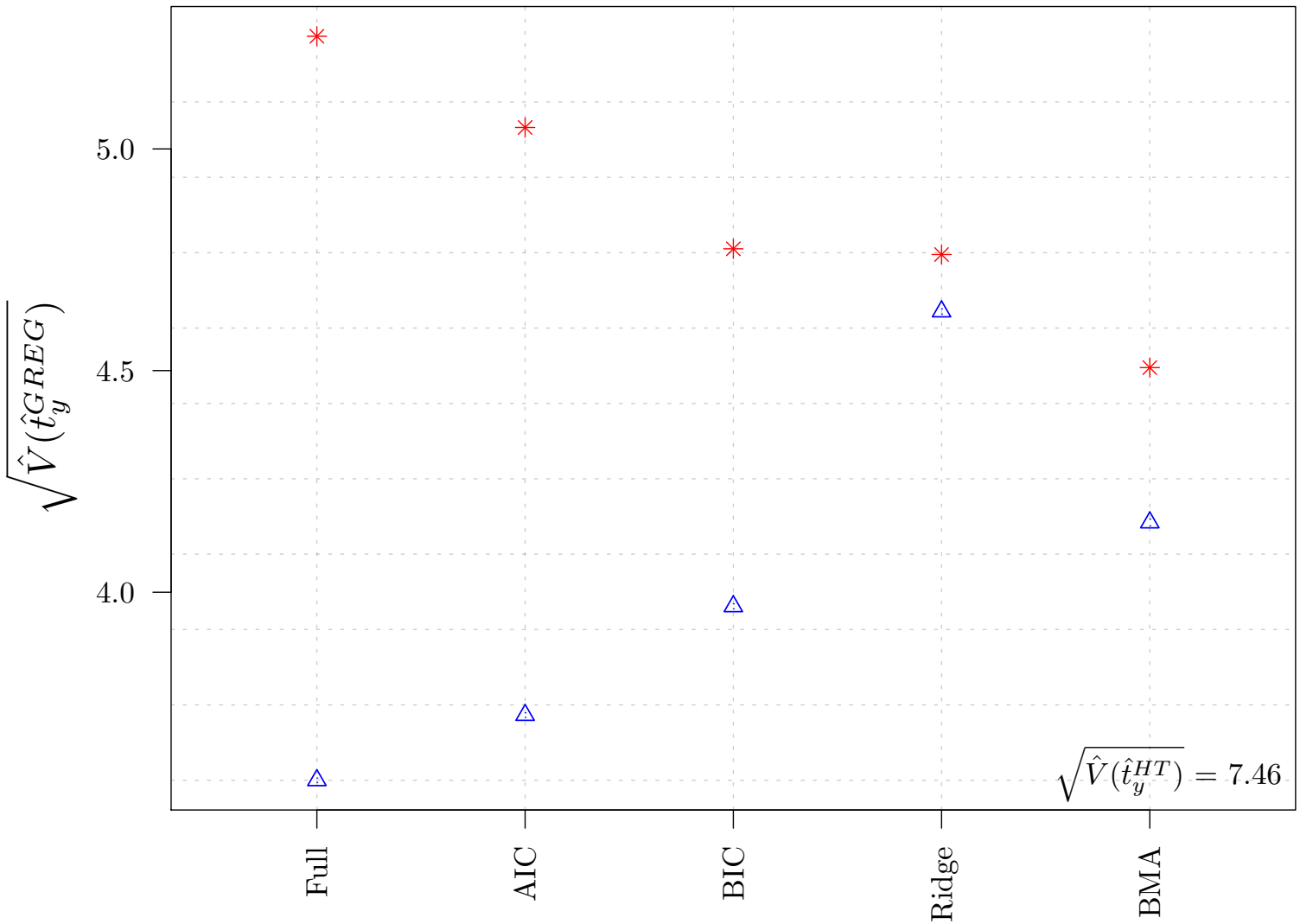
Results

Full model, AIC, BIC, Ridge regression



Results

Full model, AIC, BIC, Ridge and BMA



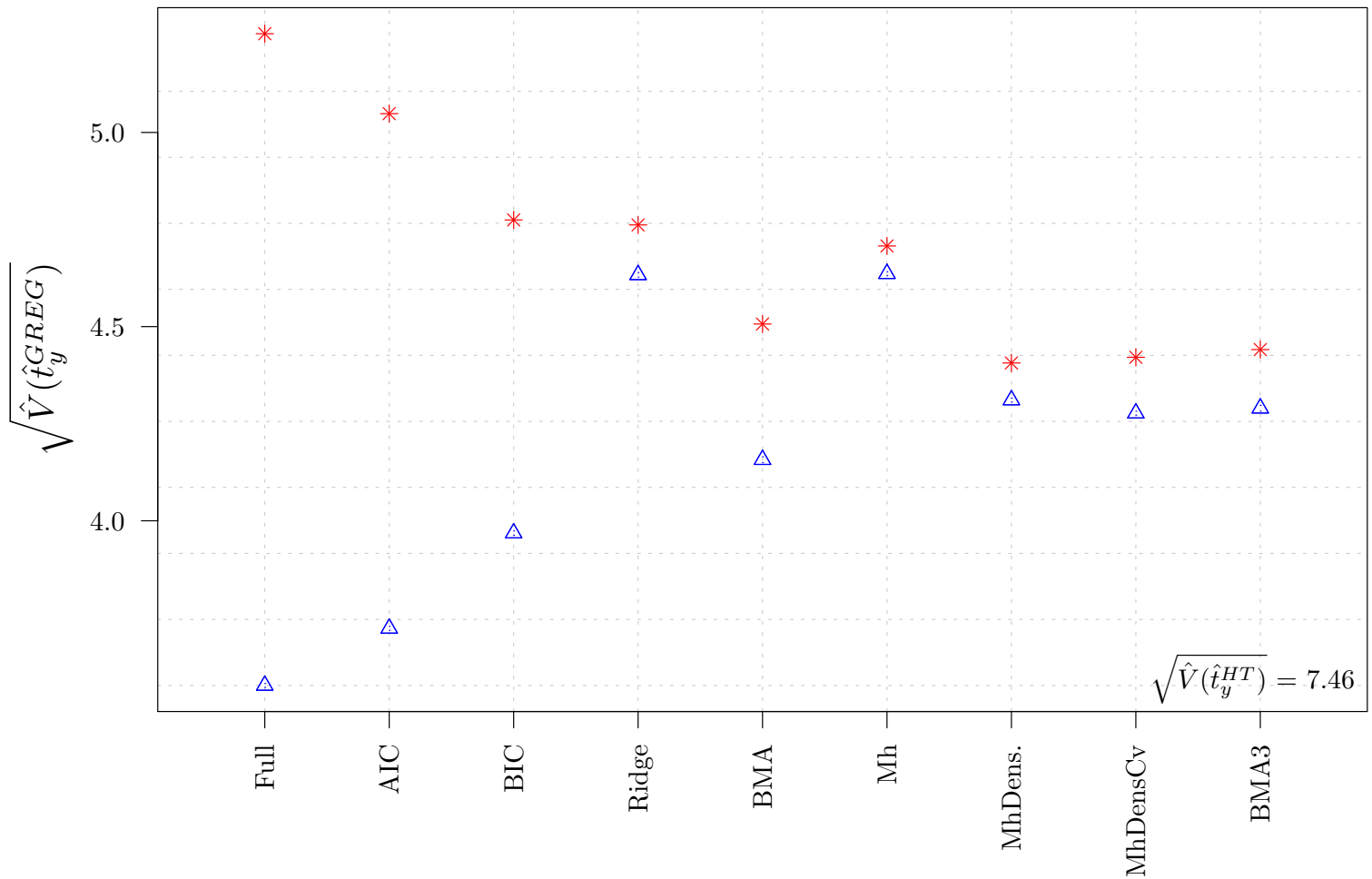
Results

Auxiliary variables that frequently appear in the literature

Mh: mean point height

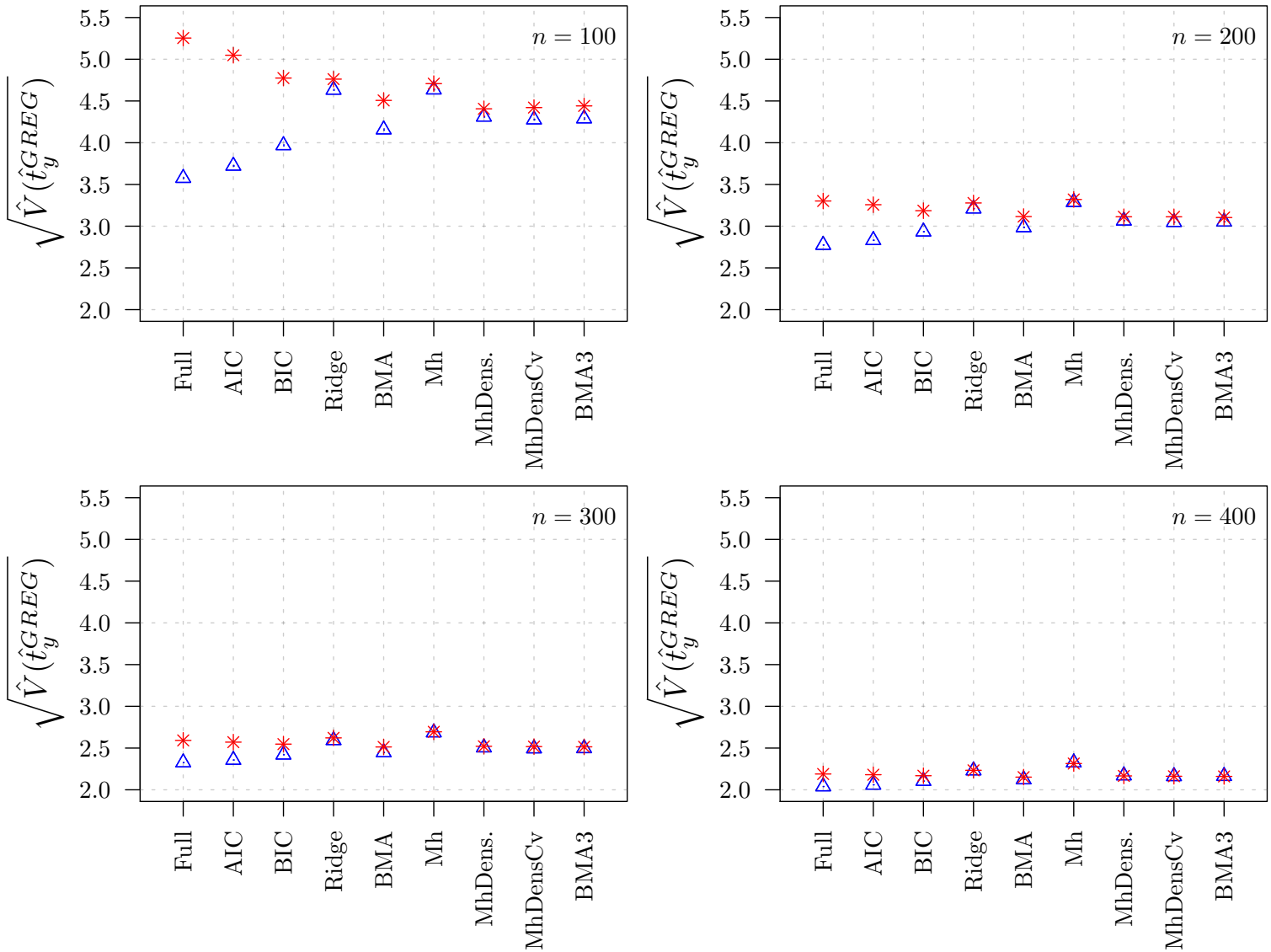
MhDens.: Mh + fraction of points above 2 meters (density)

MhDensCv: Mh + Dens. + coefficient of variation



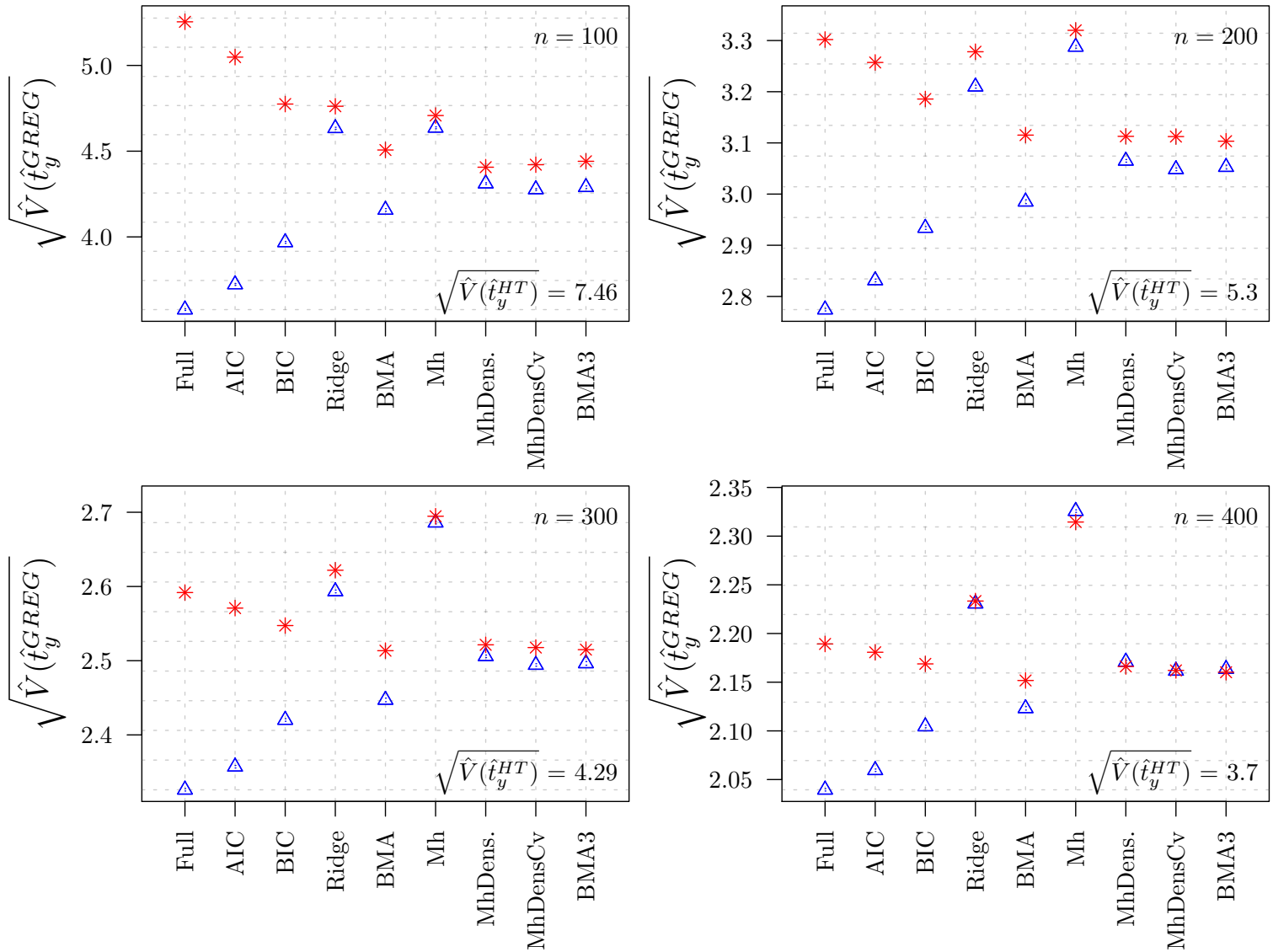
Results

Increasing the sample size: from $n = 100$ to $n = 400$ by 100



Results

Increasing the sample size: from $n = 100$ to $n = 400$ by 100



Conclusions I

- Incorporating ALS data in estimation makes sense (more efficient than using field data alone)
- If we would know the ("true") population model, we would estimate precision correctly
- No variable selection (large P) + small n: precision is overestimated (here $\sim 32\%$)
- Using AIC and BIC to select variables does not lead to correct estimates of precision $\rightarrow$ we underestimate the SE
- BMA performs slightly better

Conclusions II

- Ridge regression performs well (even for small sample sizes)
- Simple models (1 – 3 metrics/coefficients) perform best
- Conclusion: use simple models!
- Different study site, different variables?
- Alternative approaches: boosting, other Bayesian approaches (e.g., intrinsic priors), etc. (lots of possibilities!)

Acknowledgments

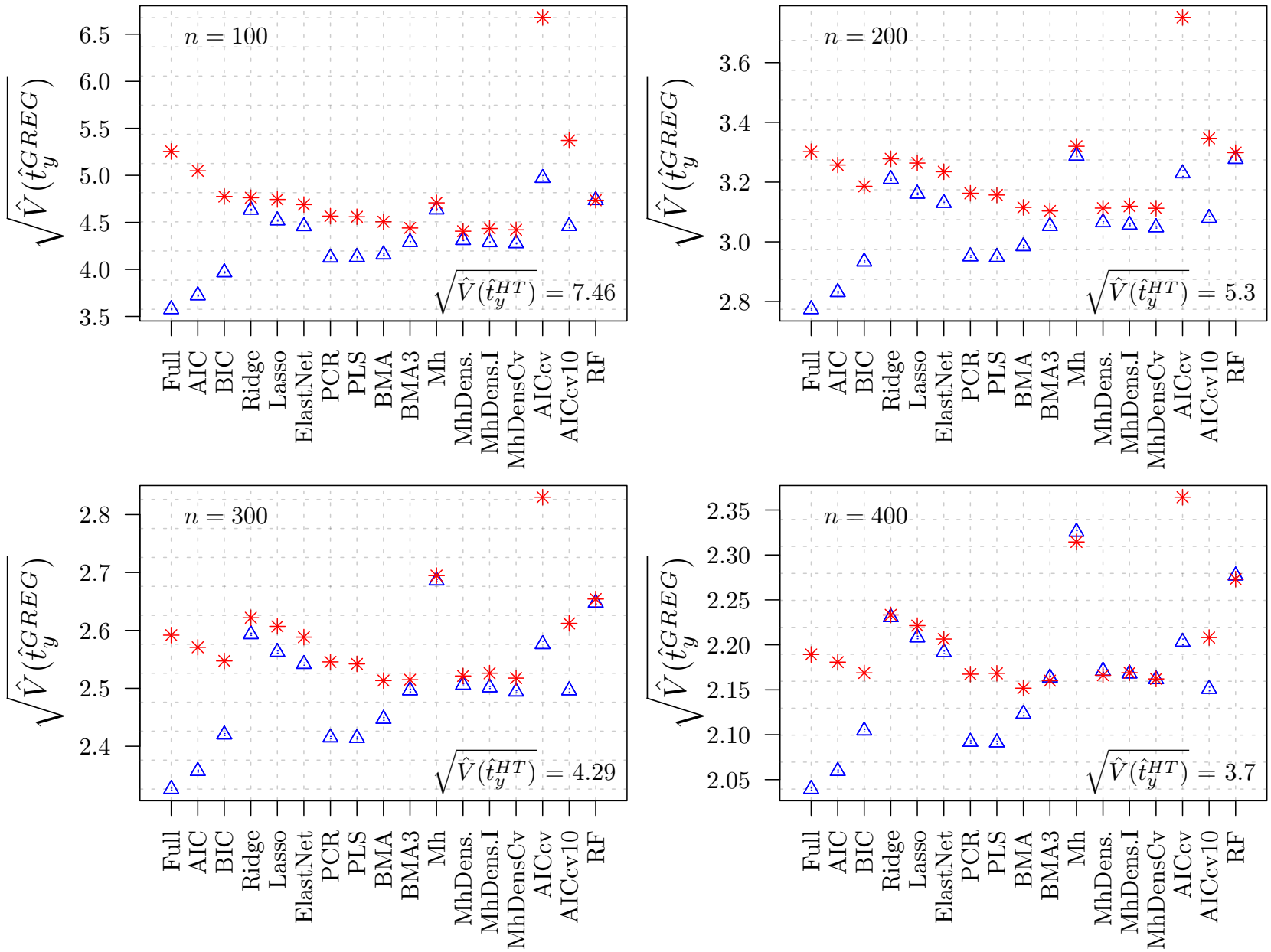
- Steen Magnussen, Joanne White and Mike Wulder (PFC, Victoria, BC, Canada)
- Christoph Kleinn and Thomas Kneib (University of Göttingen, Germany)
- Erik Næsset, Terje Gobakken and Liviu Ene (UMB, Ås, Norway)

Thank you for your attention!

References

- Särndal, C.-E., Swensson, B. & Wretman, J. 1992. *Model Assisted Survey Sampling*, 2nd edition, Springer
- Chatfield, C. 1995. Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society. Series A*, 158, 419-466
- Ene, LT, Næsset, E. & Gobakken, T. 2013. Model-based inference for k-nearest neighbours predictions using a canonical vine copula. *Scandinavian Journal of Forest Research*, 28:3, 266–281

Results: random forest, cross-validation



Results: using g-weights