Geostatistical modeling with graphs

w/ many of my invaluable co-authors

Bora Jin

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

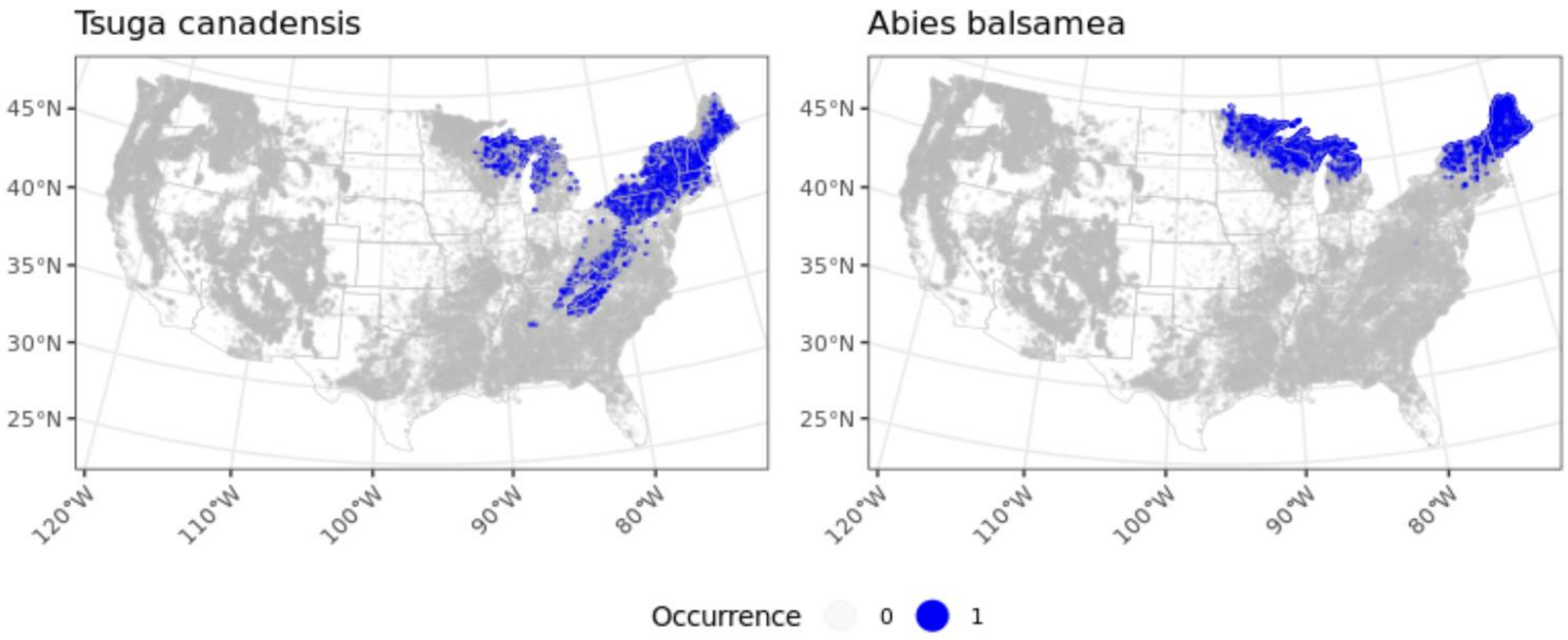
At JSM on August 6, 2025

Spanning tree-based multivariate spatial model

Bora Jin, Andrew Finley, Abhirup Datta

Motivation

- Spatial distribution of particle size curves (88 particle sizes, 3340 samples)
- Tree species co-occurrence analysis (96 species, 78804 samples)
- Air quality monitoring, spatial transcriptomics, spatial proteomics, etc.
- We want a **scalable** and **interpretable** method for highly multivariate ($q > 30$) and large ($n > 10000$) geostatistical data.

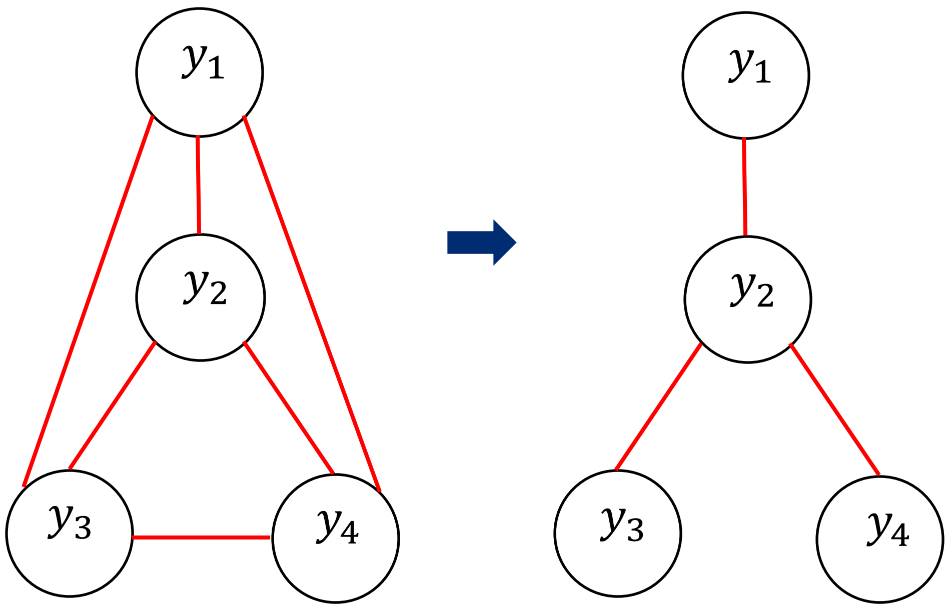


Existing multivariate methods

- Spatial factor model (+ NNGP)
  - Finley et al. (2015), Tikhonov et al. (2020), Doser et al. (2022)
  - Difficult to interpret (a latent process w is a linear combination of independent latent processes...🤯)
  - Difficult to fix or assign priors to hyperparameters
  - Choosing the ‘right’ number of factors is another area of research
- Parsimonious cross-covariance matrix function (+ NNGP)
  - Bevilacqua et al. (2015), Peruzzi (2024)
  - Computational burden and dimension increases quadratically (or more) with q
- “treed” DAG
  - Peruzzi and Dunson (2022)
  - Multiresolution/ recursive scheme scales poorly with q
- Process-level conditional independence using a graphical model
  - Dey et al. (2022)
  - Exhaustive stochastic exploration over sparse graph space is infeasible

Spanning tree-based approach

- Construct a multivariate process exploiting variable-level conditional independence relationships implied by a data generating inter-variable graph.
- Consider a minimum spanning tree as the backbone of the inter-variable graph.
- **Spanning trees** are economical to handle
    - Span all q variables
    - Include only $q-1$ edges

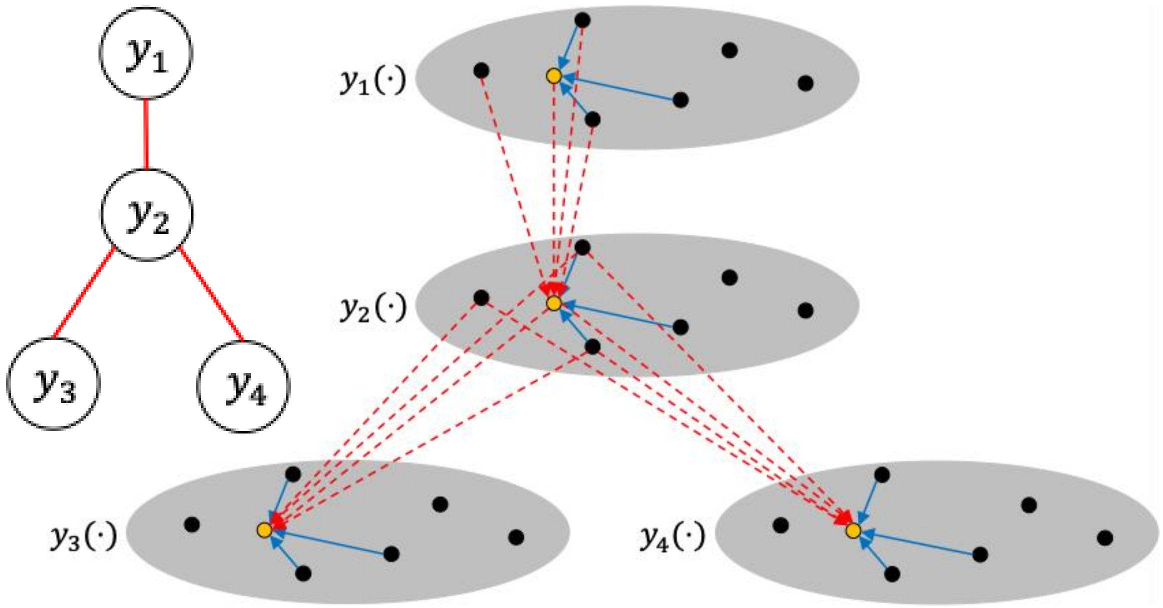


Spanning tree-based approach

- The spanning tree T on variables + a sparse DAG on locations
- For $\boldsymbol{y}(s) = \left(y_1(s), \dots, y_q(s)\right)^T$,

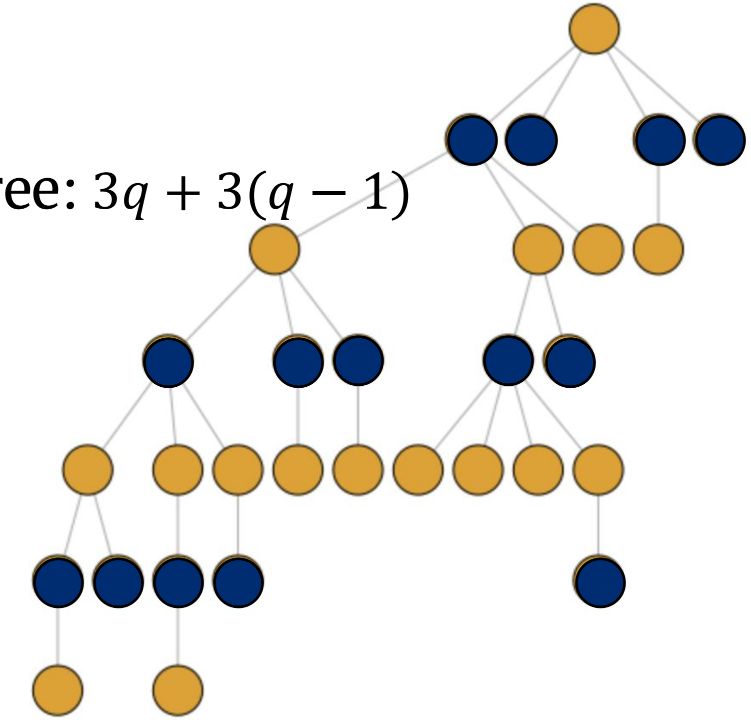
$$\tilde{f}(\boldsymbol{y}(s)) = f\left(y_1(s) \middle| y_1(N(s))\right) \prod_{(j,k) \in E_T} f\left(y_k(s) \middle| y_k(N(s)), y_j(s), y_j(N_u(s))\right)$$

  - Variable k is independent to other variables conditional on variable j connected by T
  - Different sets of neighbors from the parent variable
  - Useful when data are misaligned



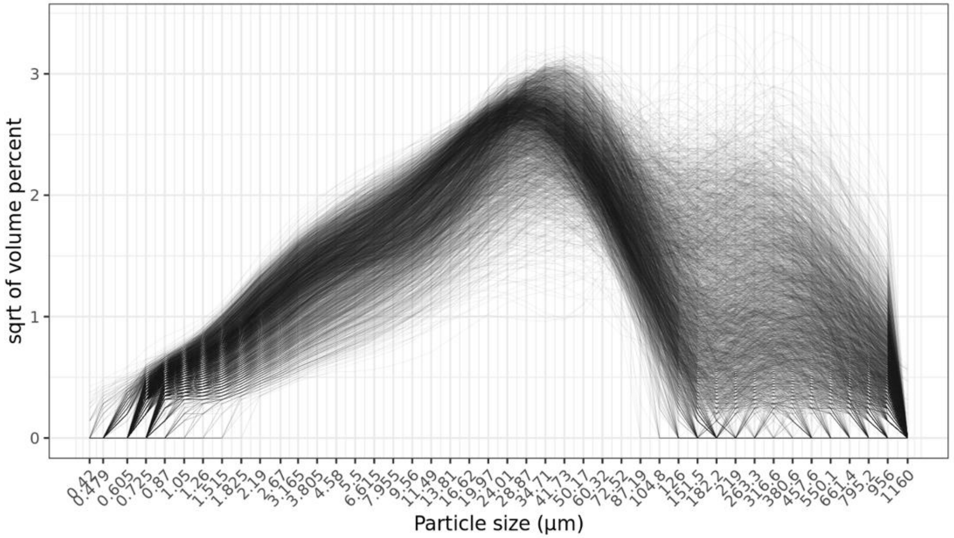
Properties

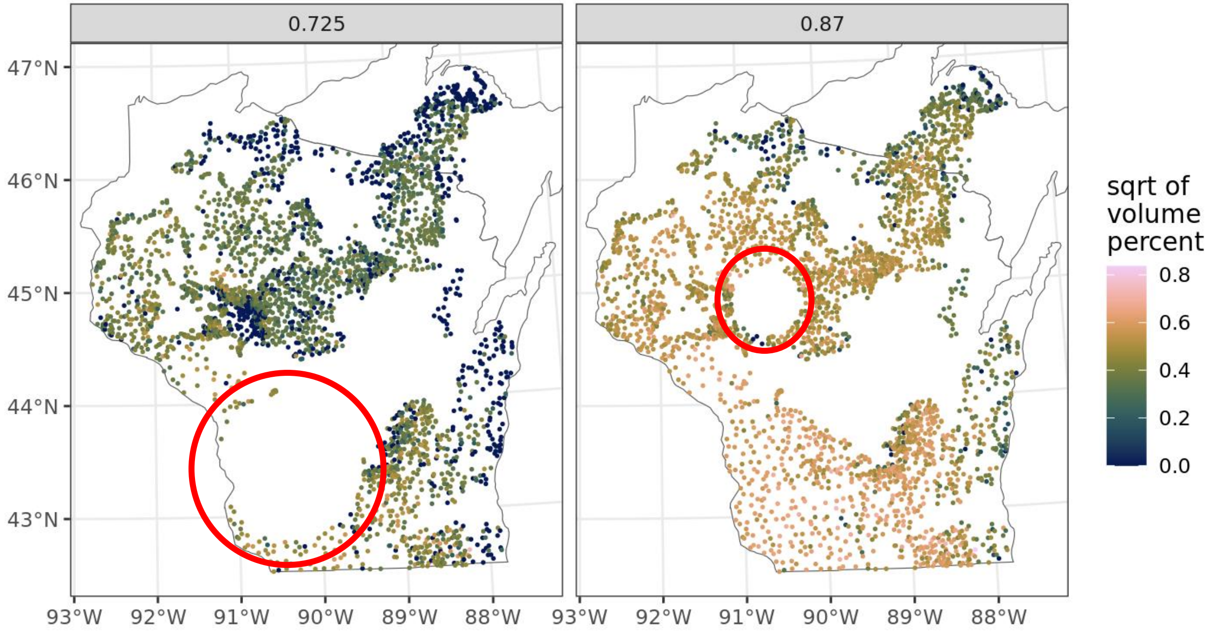
- Any variable can serve as the root; No arbitrary variable ordering.
- Resulting multivariate process preserves process-level conditional independence specified by T.
- Sufficient to ensure validity of a **bivariate** cross-covariance function for each pair of variables connected by T.
    - Substantial dimension reduction when q is large.
    - Multivariate Matérn: $3q + 3q(q-1)/2$ vs. Spanning tree: $3q + 3(q-1)$
- Parallelization using graph coloring
    - Only 2 colors because of a tree structure
    - All variables in yellow updated in parallel,
    - then all variables in blue in parallel



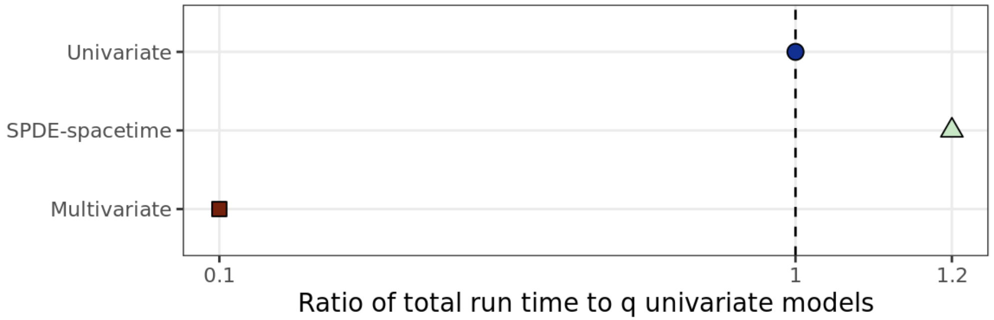
Particle size curves

- $n = 3340$ soil samples around Wisconsin and Michigan
- A curve representing sqrt of the volume of particles across $q = 44$ different sizes at each location
- Misalignment at every other particle size
- Aim to predict a curve at a new location
- Choice of T: path graph

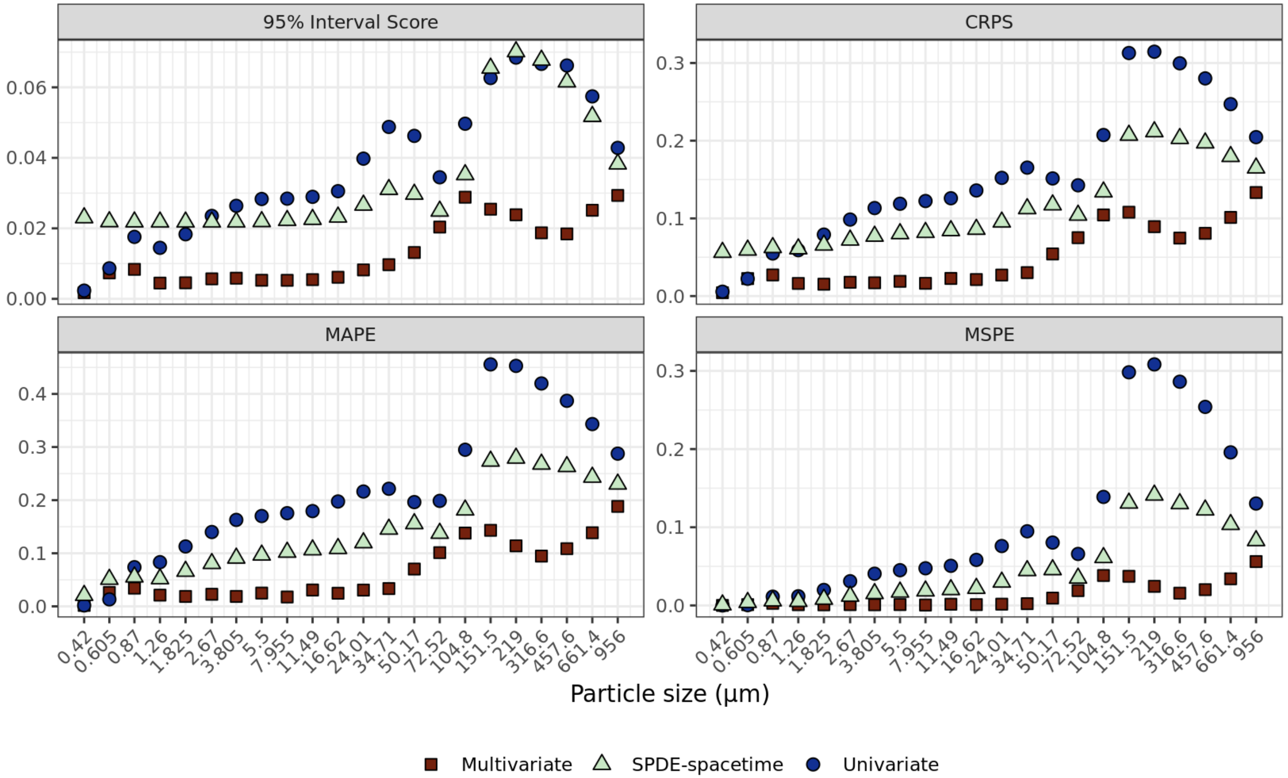


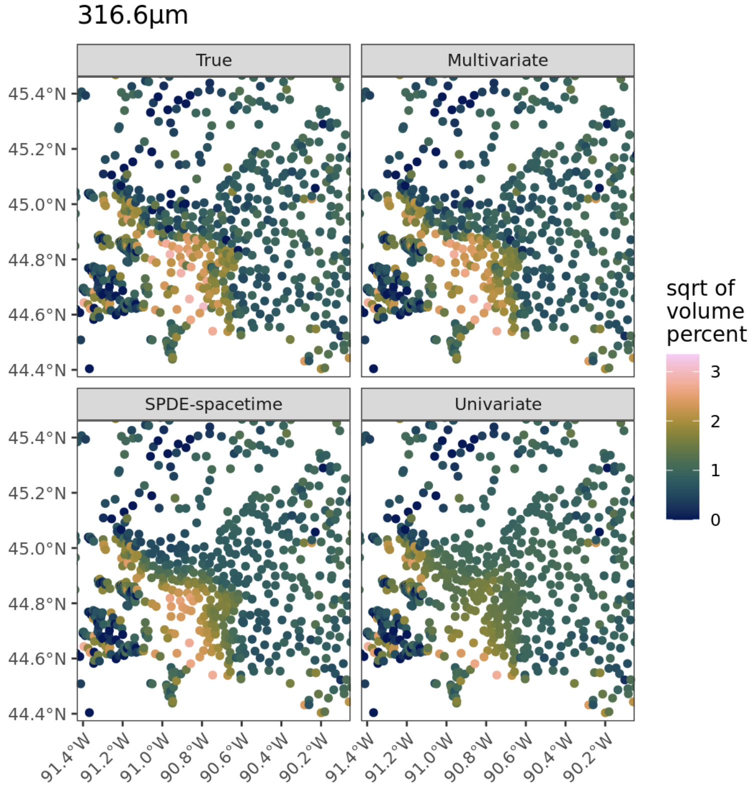


Particle size curves



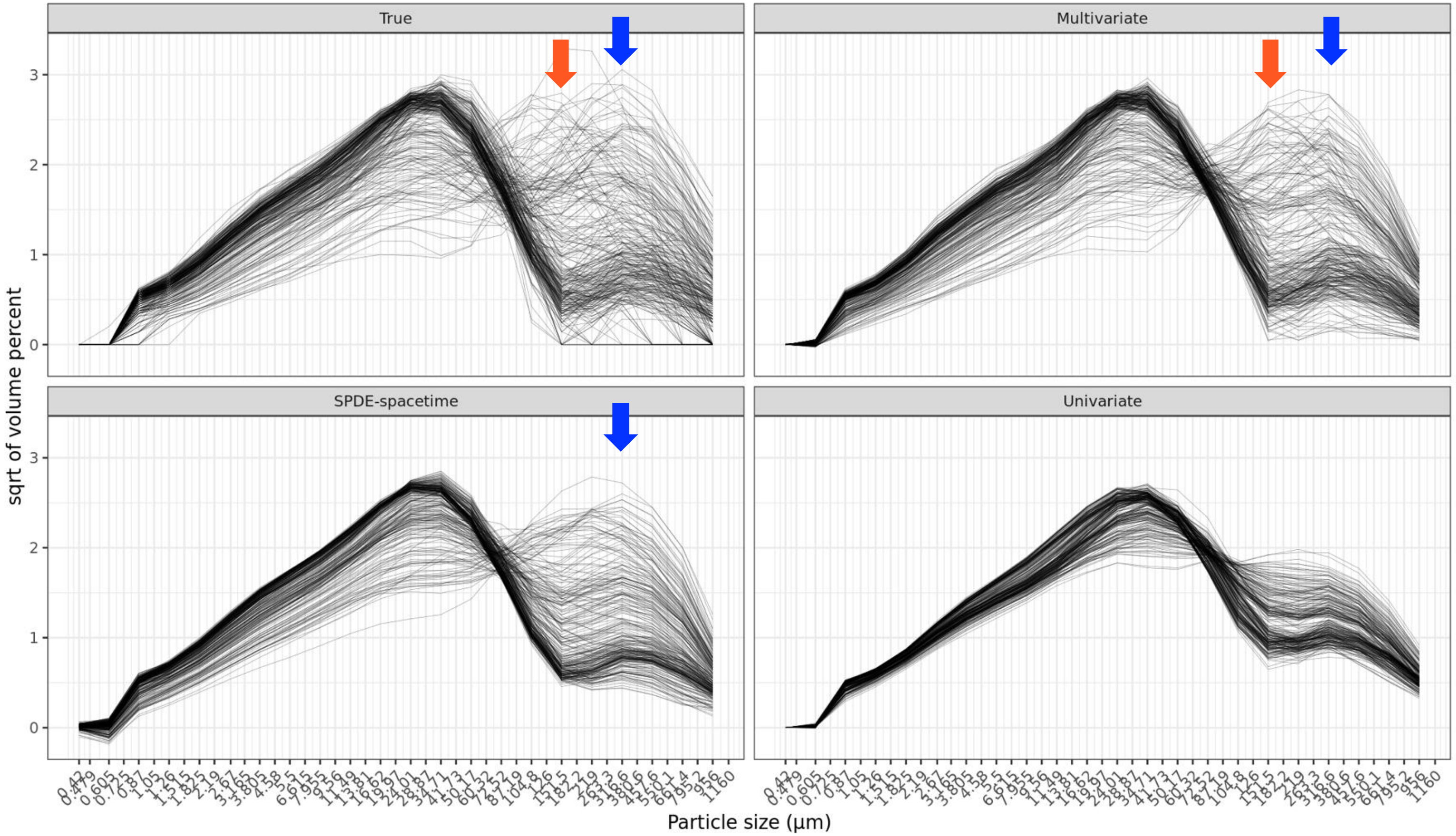
- Reduction in computation time
- Gain in prediction accuracy for misaligned locations
  - Silt dominating area; competitors underestimate volume of coarse particles





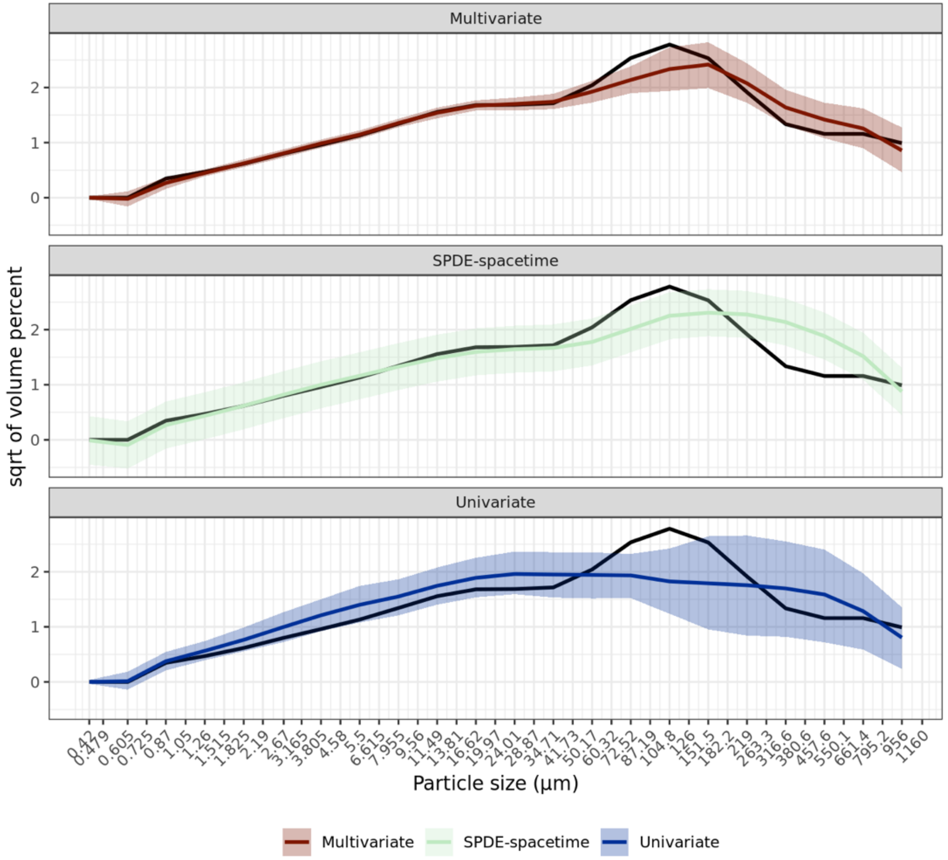
Particle size curves

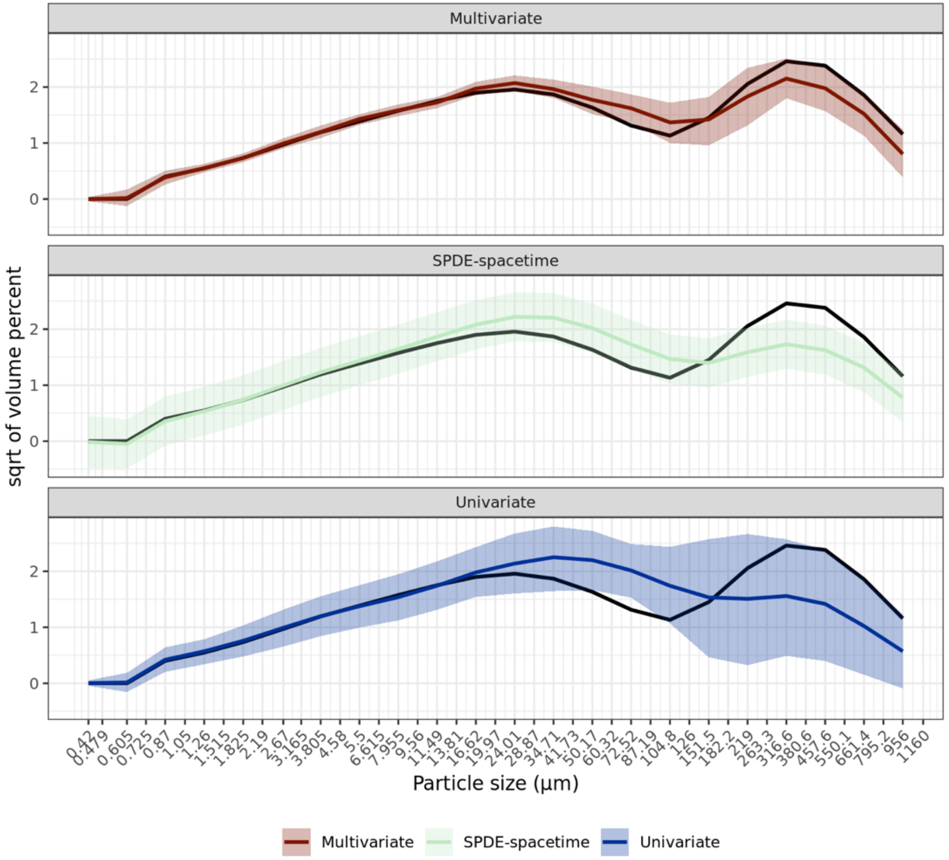
- Separable spacetime model and independent univariate model struggle to find hotspots rich with medium or large sand.



Particle size curves

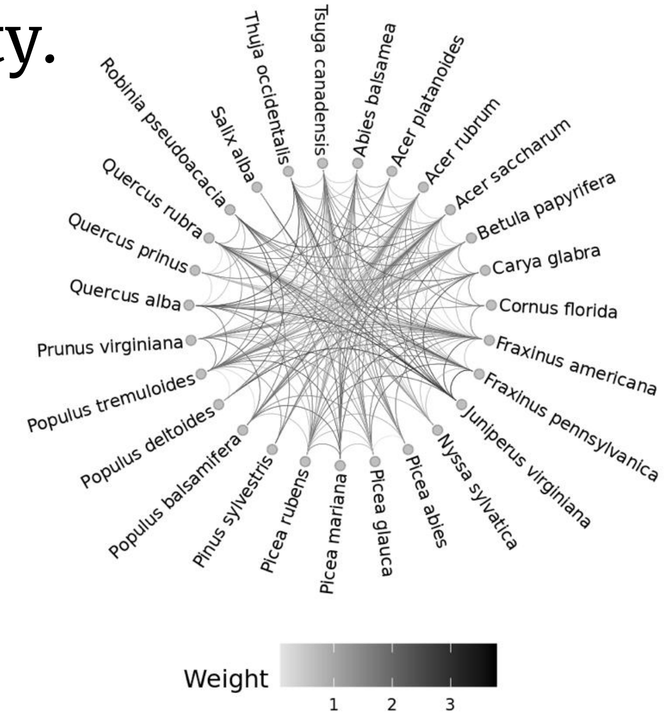
- Separable spacetime model and independent univariate model struggle to find hotspots rich with medium or large sand.

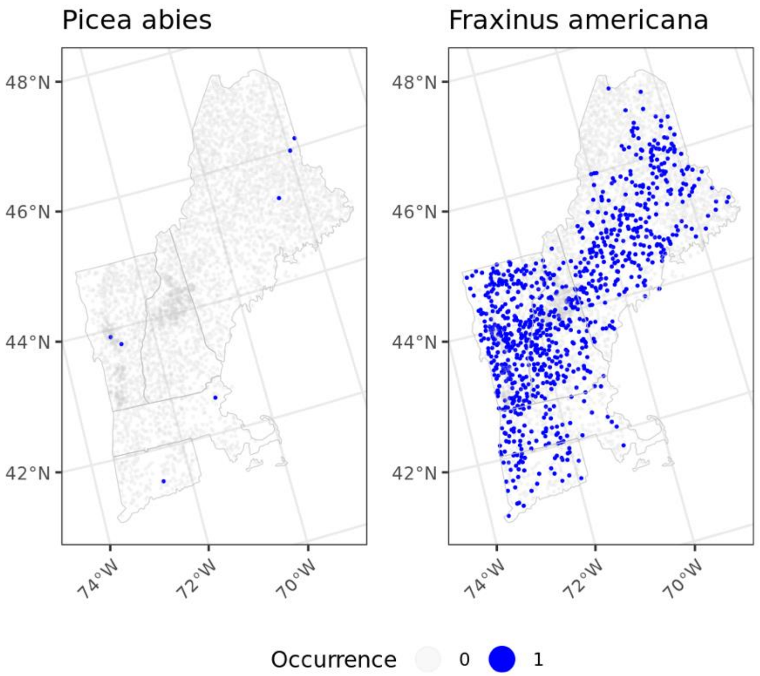




Tree species co-occurrence

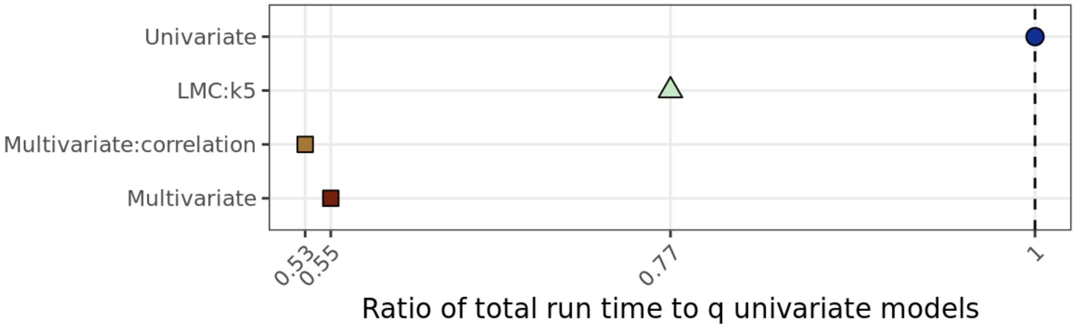
- $q = 27$ tree species occurrence data at $n = 3663$ locations around New England
- Inter-variable graph created based on field knowledge whose weights are defined by closeness in a space of trees' resistance to drought and wood density.

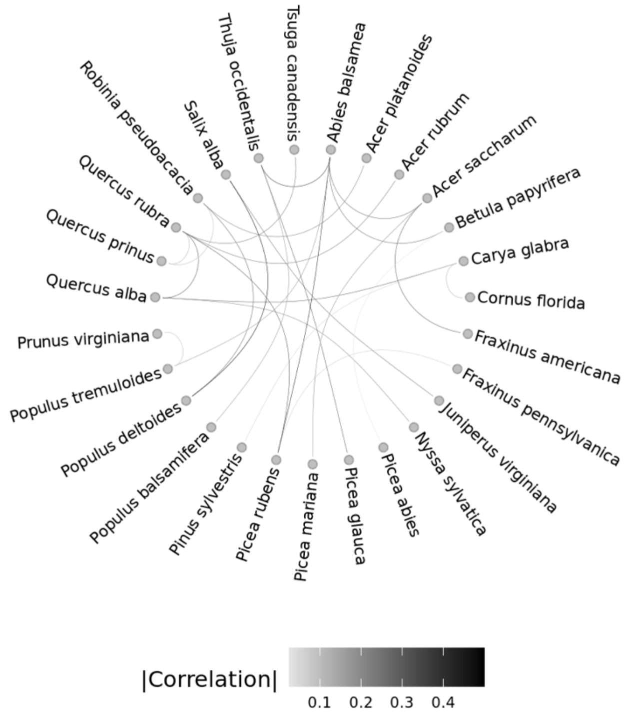




Tree species co-occurrence

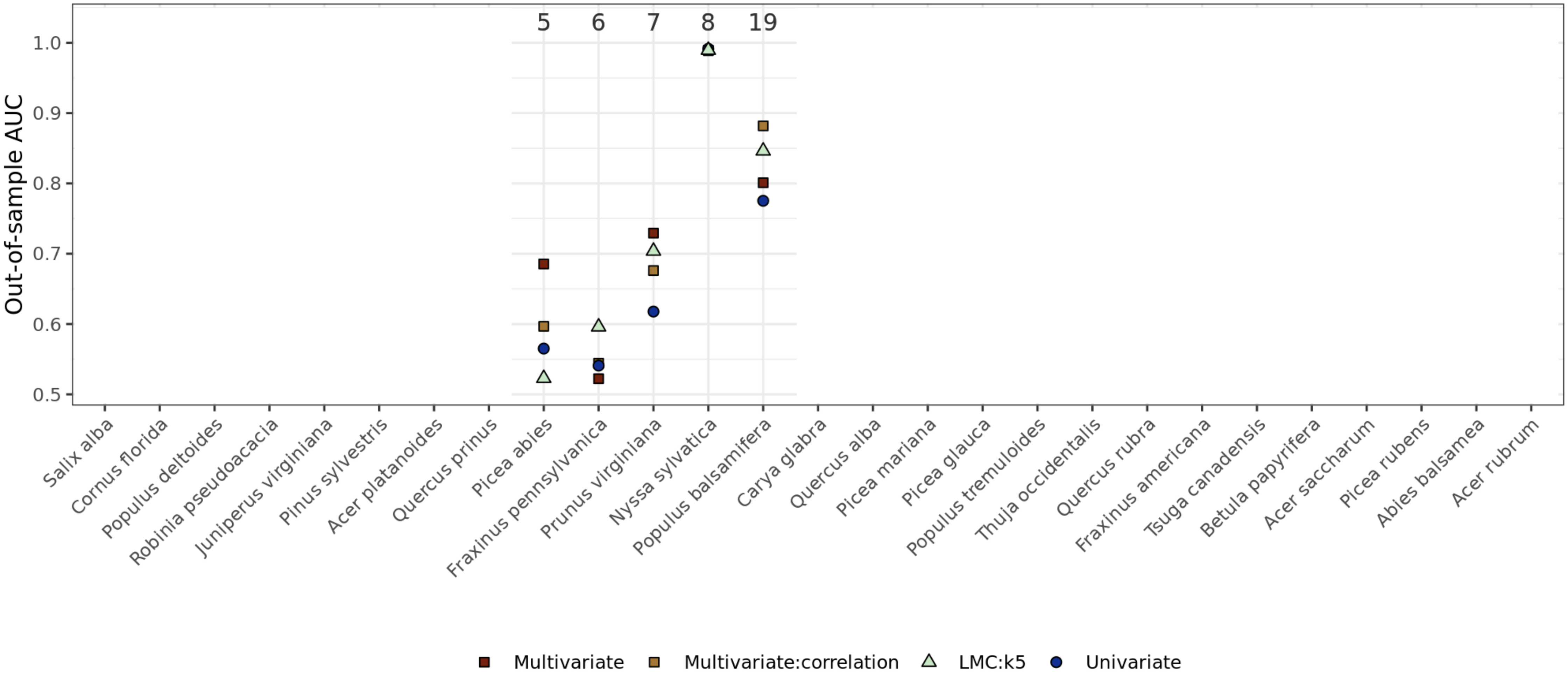
- Choice of T: minimum spanning tree with
  - negative weights
  - negative absolute correlations
- Reduction in computation time





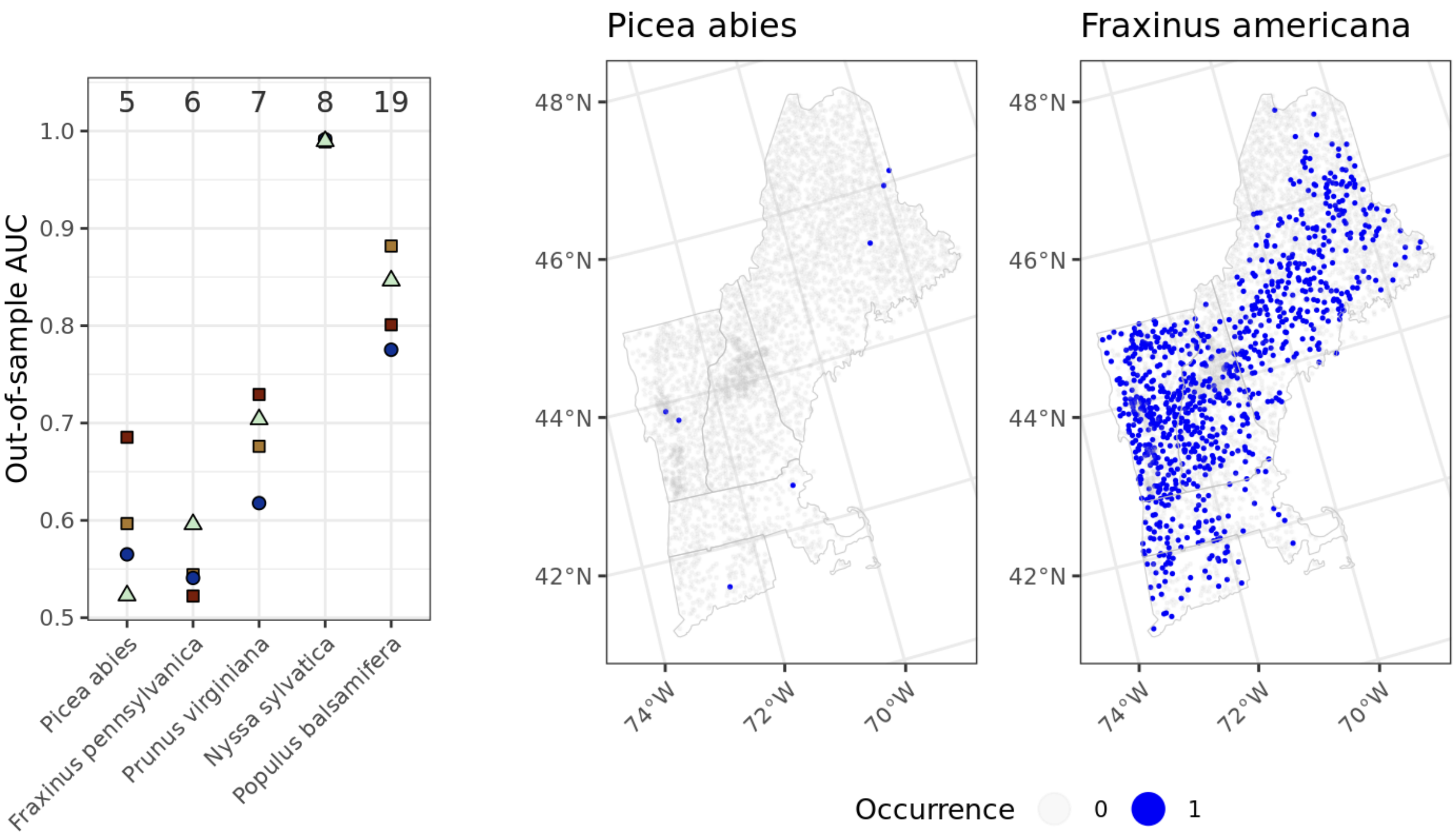
Tree species co-occurrence

- Gain in prediction accuracy for moderately rare species



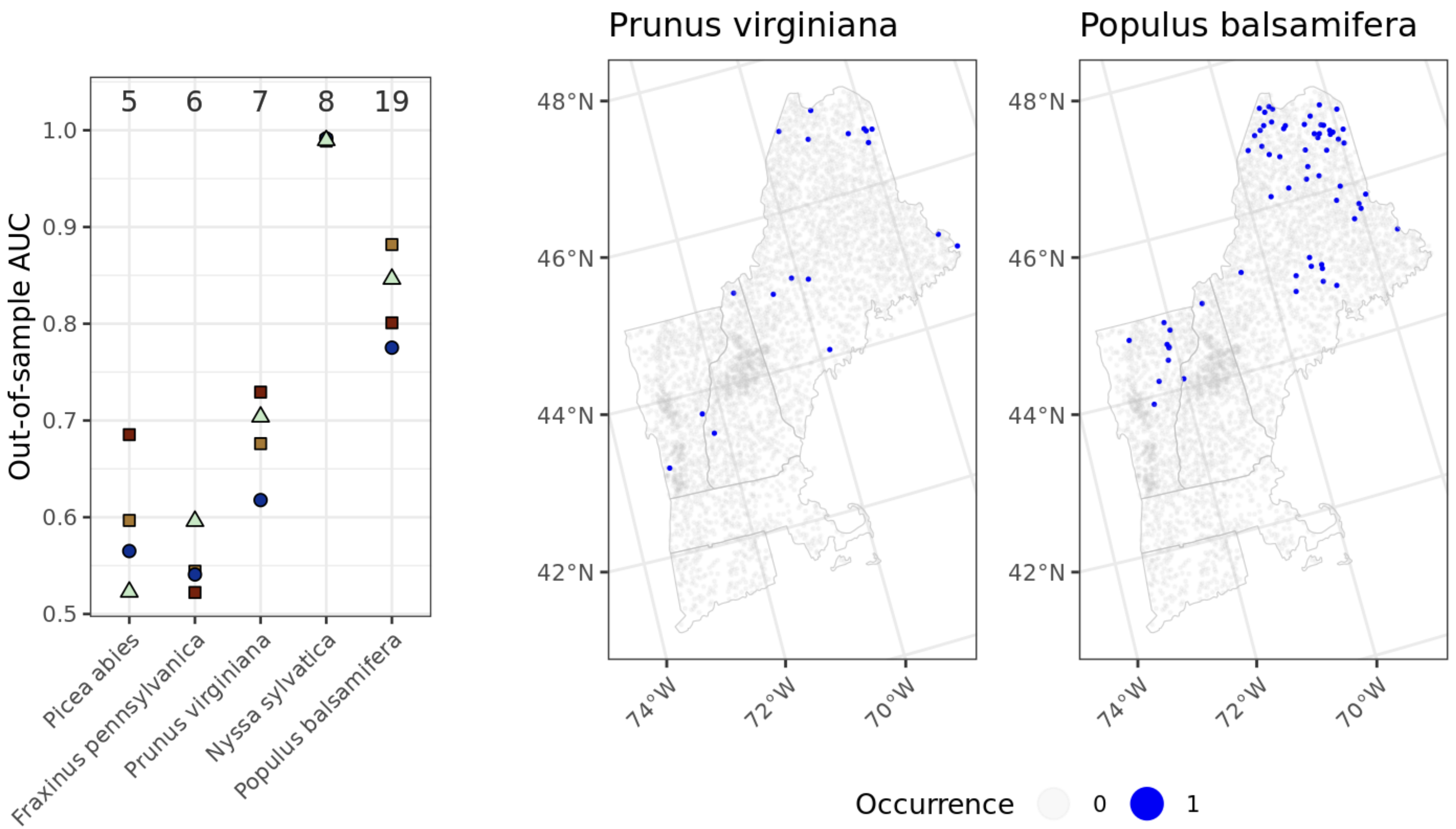
Tree species co-occurrence

- Gain in prediction accuracy for moderately rare species



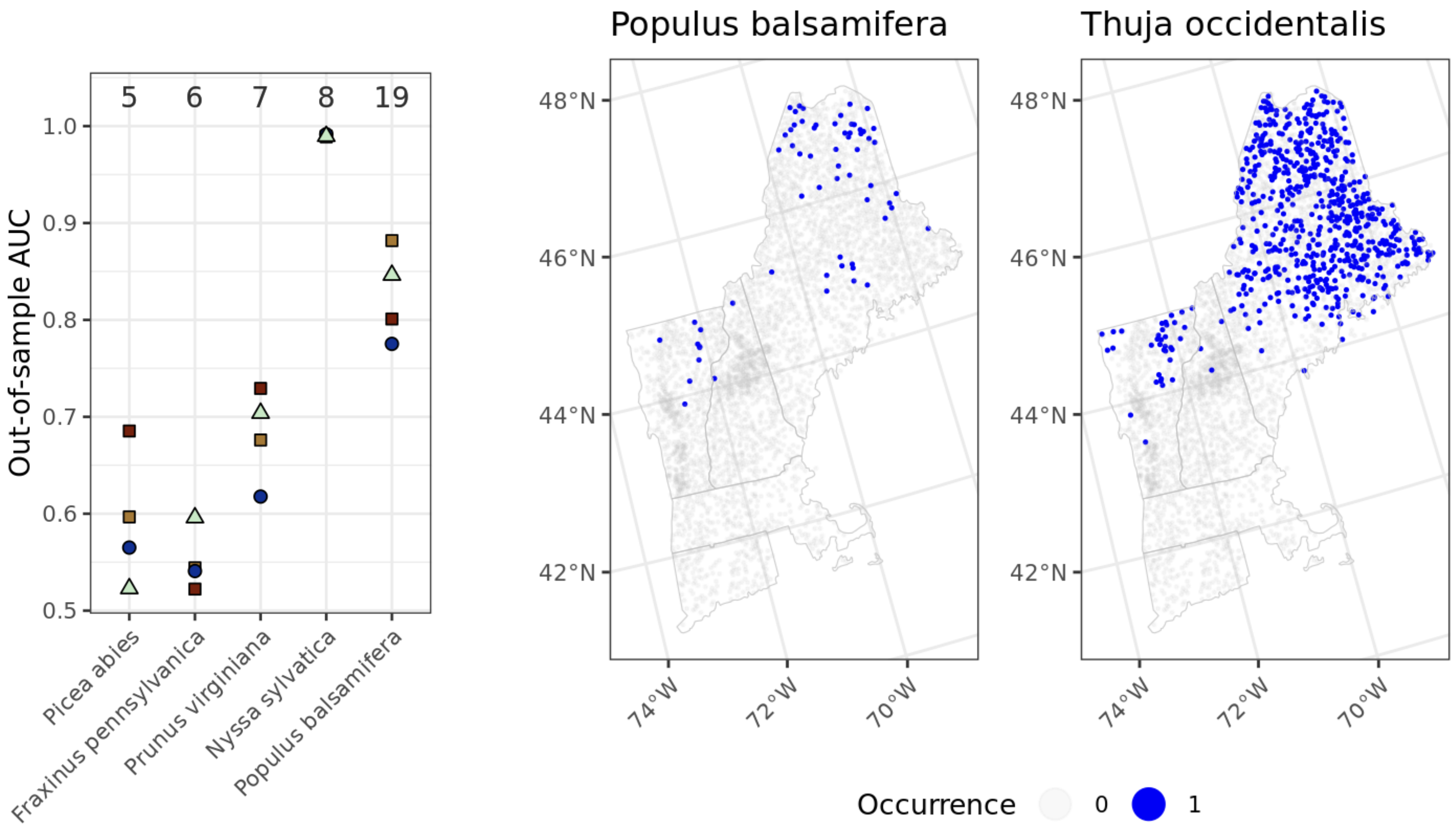
Tree species co-occurrence

- Gain in prediction accuracy for moderately rare species



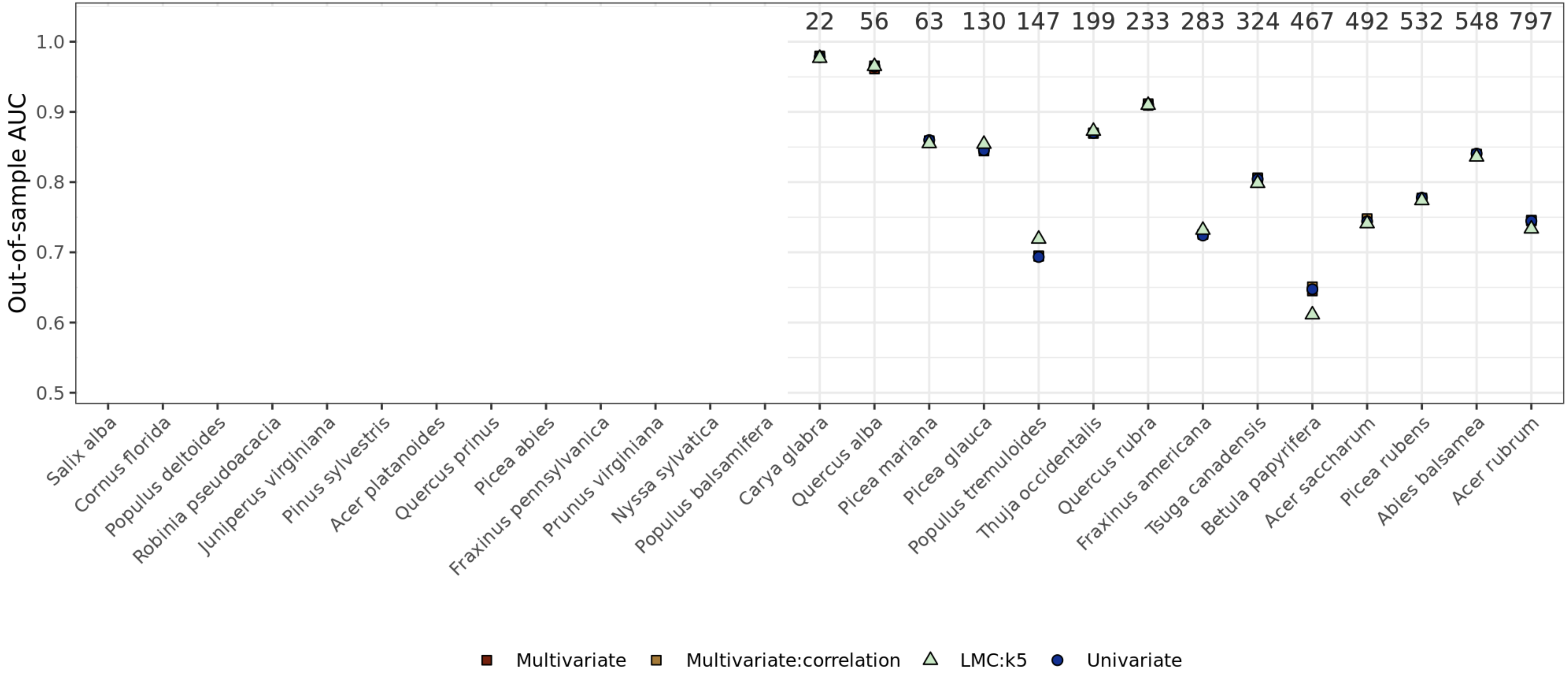
Tree species co-occurrence

- Gain in prediction accuracy for moderately rare species



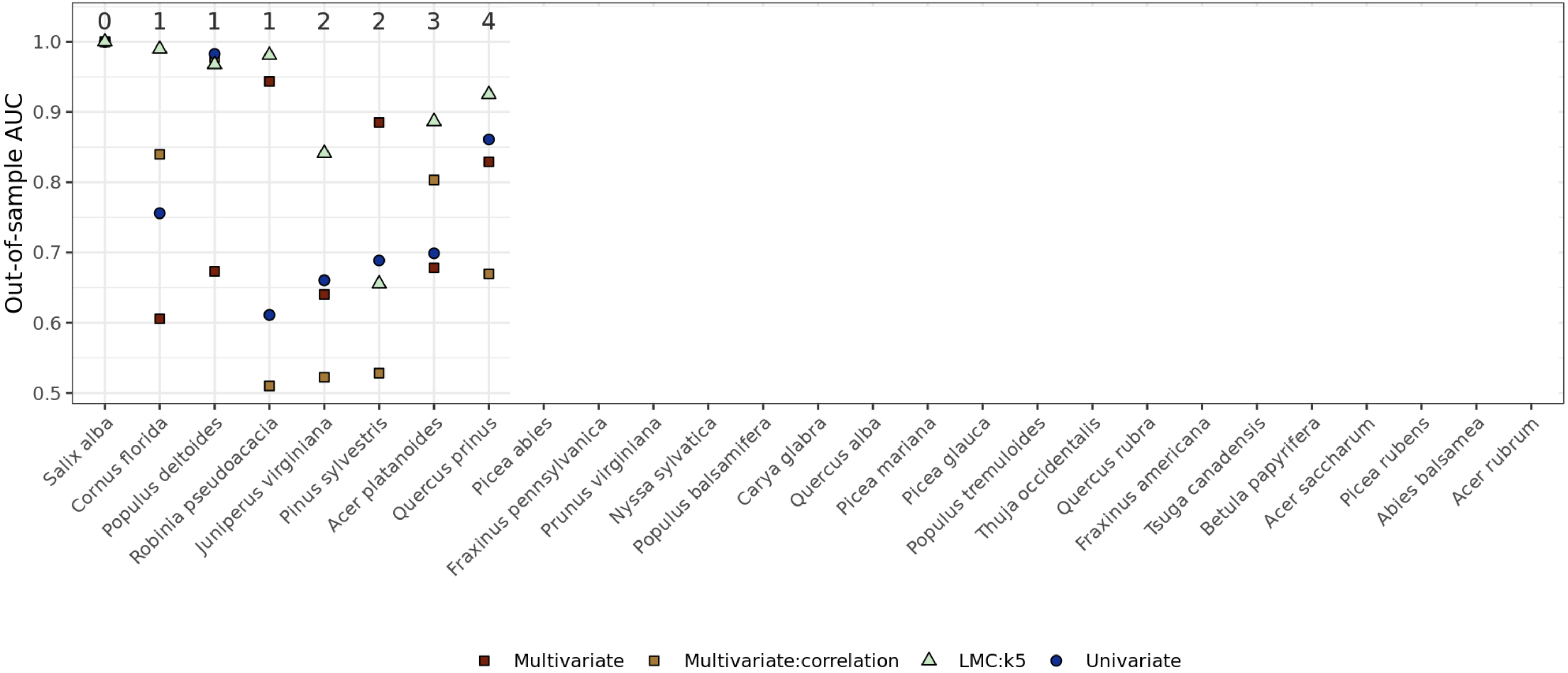
Tree species co-occurrence

- Similar prediction performance for frequently observed species



Tree species co-occurrence

- Factor model helps the most for extremely rare species



Future direction

- Using a minimum spanning tree can be too restrictive; combine results over multiple minimum spanning trees
- Choice of a minimum spanning tree can be arbitrary when graph structure is not intrinsic among variables; alternative ways to infer inter-variable relationships?
- With fixed covariance parameters, MCMC can be avoided (predictive stacking; Zhang et al. 2023).

Happy to hear your insight/suggestions/feedback!

bjin9@jh.edu

Reference

- Bevilacqua, M., Fassò, A., Gaetan, C. *et al.* Covariance tapering for multivariate Gaussian random fields estimation. *Stat Methods Appl* **25**, 21–37 (2016). https://doi.org/10.1007/s10260-015-0338-3
- Dey, D., Datta, A., Banerjee, S. (2022) Graphical Gaussian process models for highly multivariate spatial data, *Biometrika*, Volume 109, Issue 4, December 2022, Pages 993–1014, https://doi.org/10.1093/biomet/asab061
- Doser, J. W., Finley A. O., Kéry, M., & Zipkin E. F. (2022). spOccupancy: An R package for single-species, multi-species, and integrated spatial occupancy models Methods in Ecology and Evolution, 13, 1670-1678. https://doi.org/10.1111/2041-210X.13897
- Finley, A. O., Banerjee, S., & Gelfand, A. E. (2015). spBayes for Large Univariate and Multivariate Point-Referenced Spatio-Temporal Data Models. *Journal of Statistical Software*, *63*(13), 1–28. https://doi.org/10.18637/jss.v063.i13
- Peruzzi, M., & Dunson, D. B. (2022). Spatial Multivariate Trees for Big Data Bayesian Regression. *Journal of machine learning research : JMLR*, *23*, 17.
- Peruzzi, M. (2024). Inside-out cross-covariance for spatial multivariate data. arXiv preprint arXiv:2412.12407.
- Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M. J., Oksanen, J., & Ovaskainen, O. (2020). Joint species distribution modelling with the r-package Hmsc. *Methods in ecology and evolution*, *11*(3), 442–447. https://doi.org/10.1111/2041-210X.13345
- Zhang, L., Tang, W., Banerjee, S. (2023). Bayesian Geostatistics Using Predictive Stacking. arXiv preprint arXiv:2304.12414.