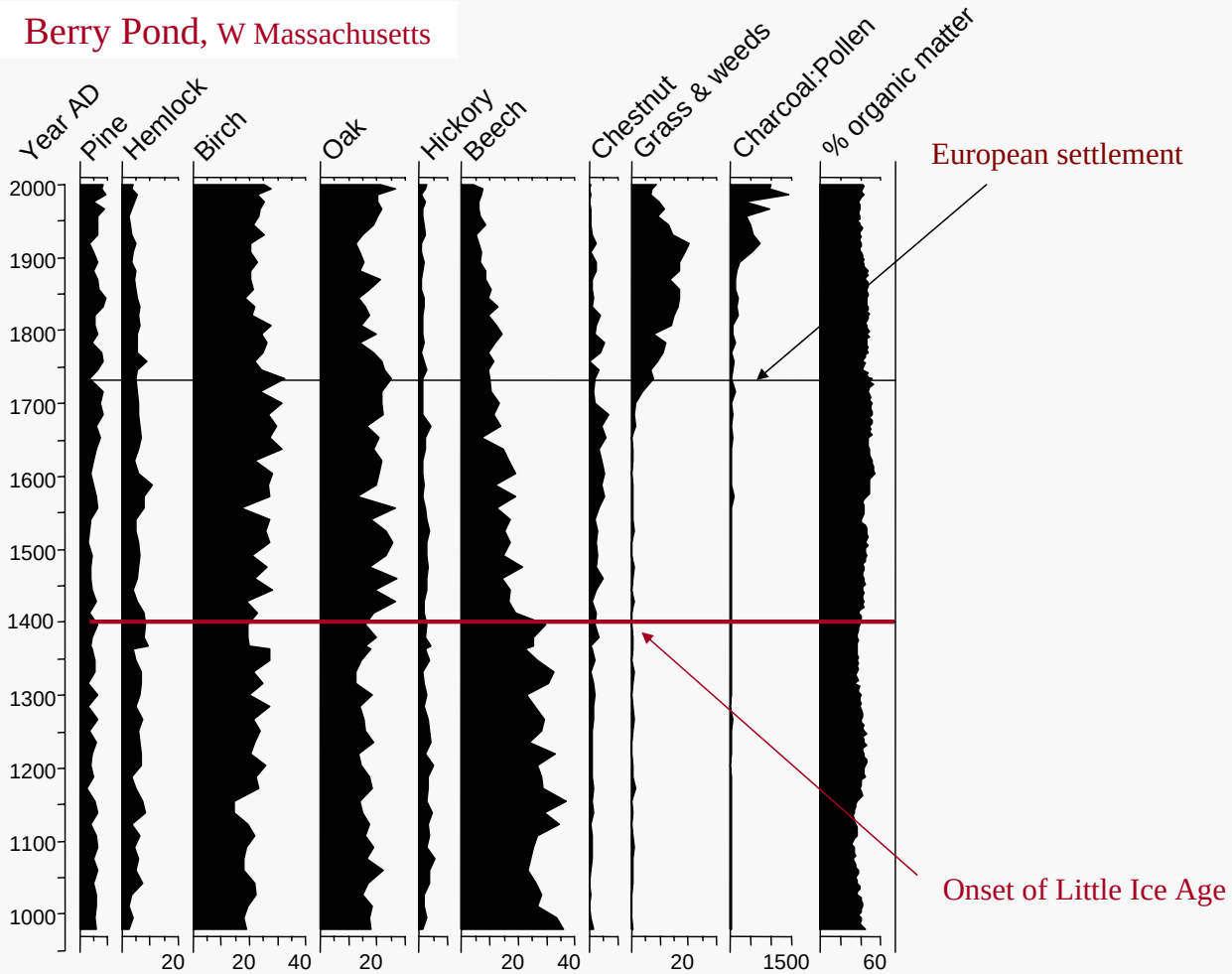
Mapping Ancient Forests: Bayesian Inference for Forest Composition Using the Fossil Pollen Proxy Record

Chris Paciorek
Department of Biostatistics
Harvard School of Public Health
www.biostat.harvard.edu/~paciorek

Joint work with Jason McLachlan, Notre Dame Biology

A pollen diagram



Scientific goals

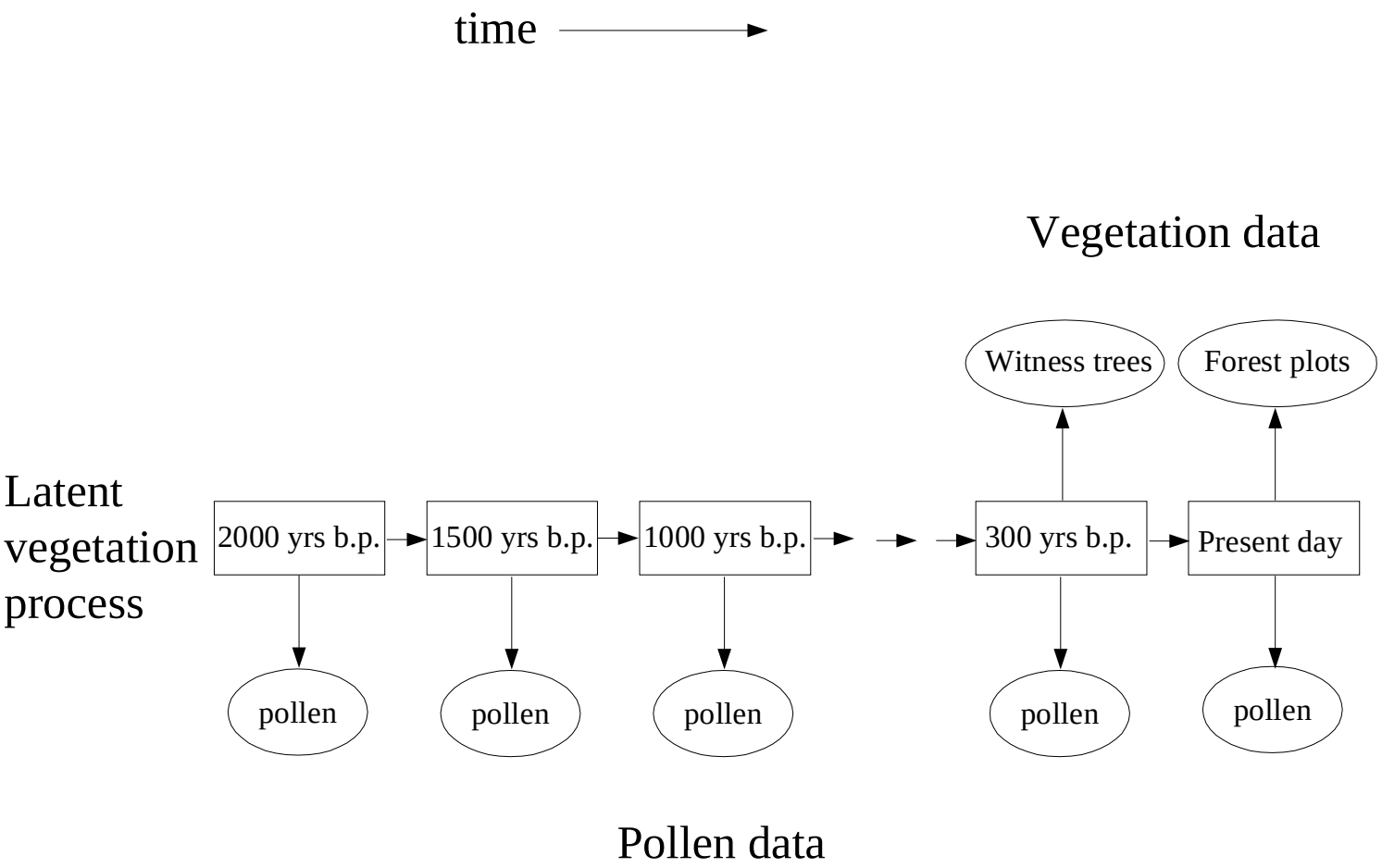
- Understand the relationship between pollen in ponds and trees on the landscape
  - Relationship is noisy because of dispersal and deposition processes.
- Map spatio-temporal patterns in tree populations
  - Relative abundances
  - Species ranges
- Understand vegetation dynamics, particularly vis-a-vis changing climate
  - Population growth and decline
  - Migration patterns

Bayesian hierarchical modeling

Key questions:

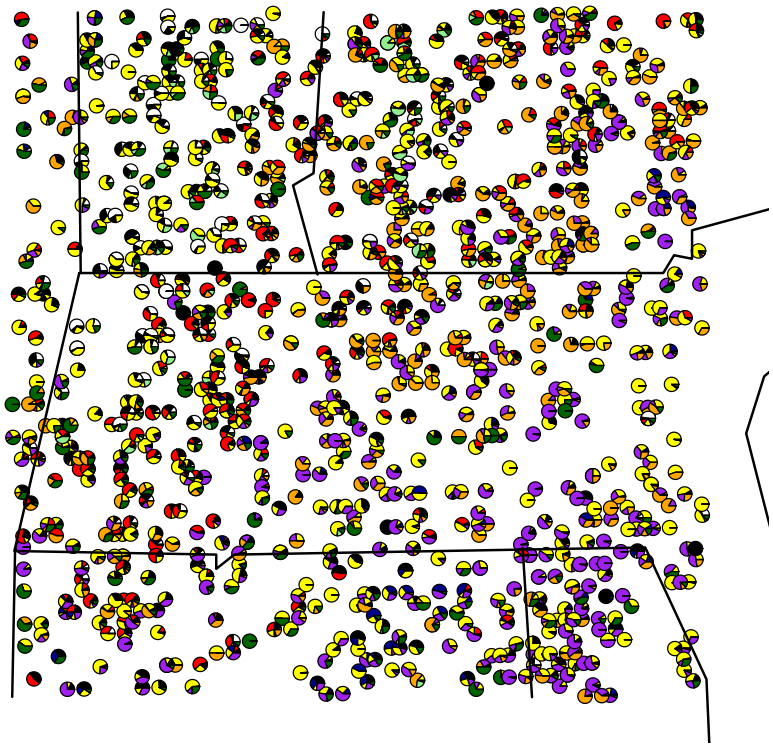
- How complex should a model be?
- How do we develop a hierarchical model?
- Are our parameter estimates interpretable?
- How do the data inform the various pieces of the model?
- Does the model appropriately synthesize information from disparate data sources?
- Is this art or science?

Basic problem structure



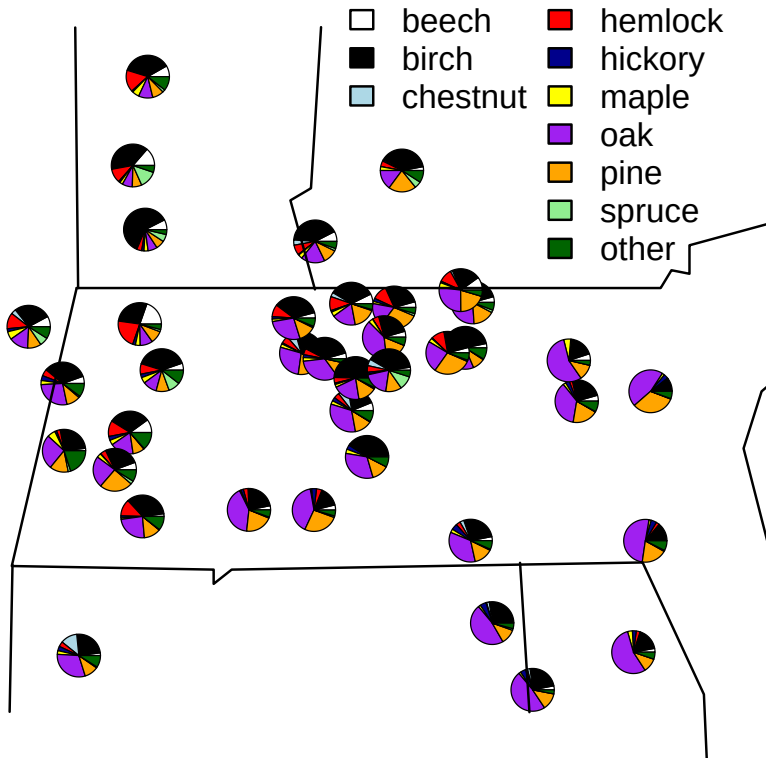
Modern data

Forest Service vegetation data



1161 plots, 1-115 trees per plot

Pollen sediment surface samples

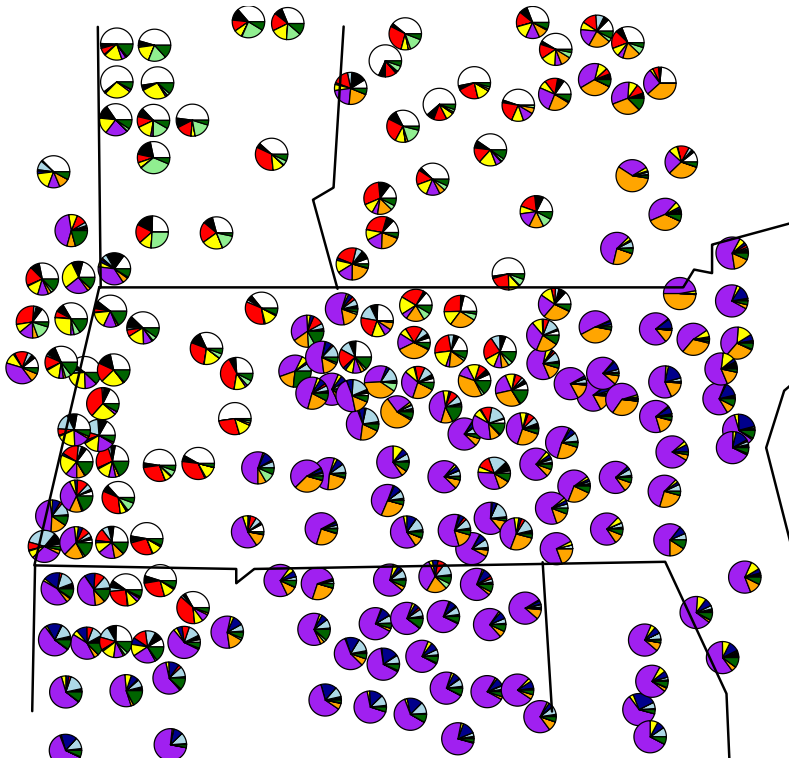


38 ponds, 500 grains per pond

R help: "Pie charts are a very bad way of displaying information."

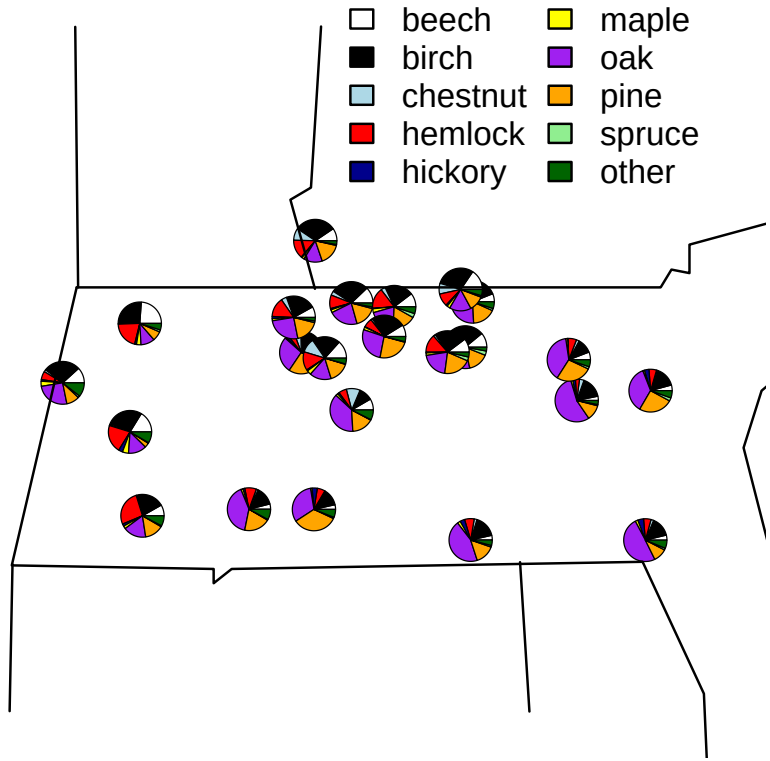
Colonial data

Township witness tree data



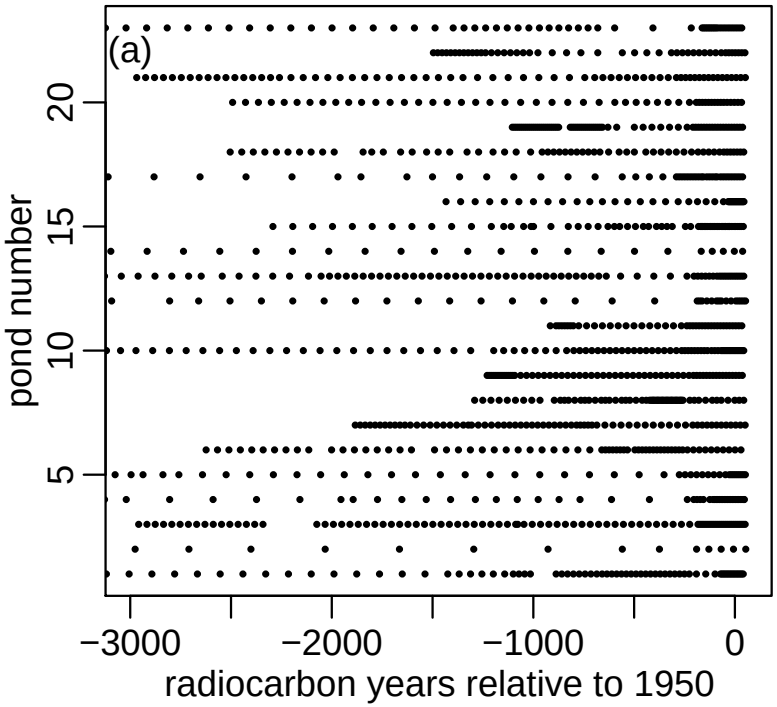
183 towns, 26-3149 trees per town

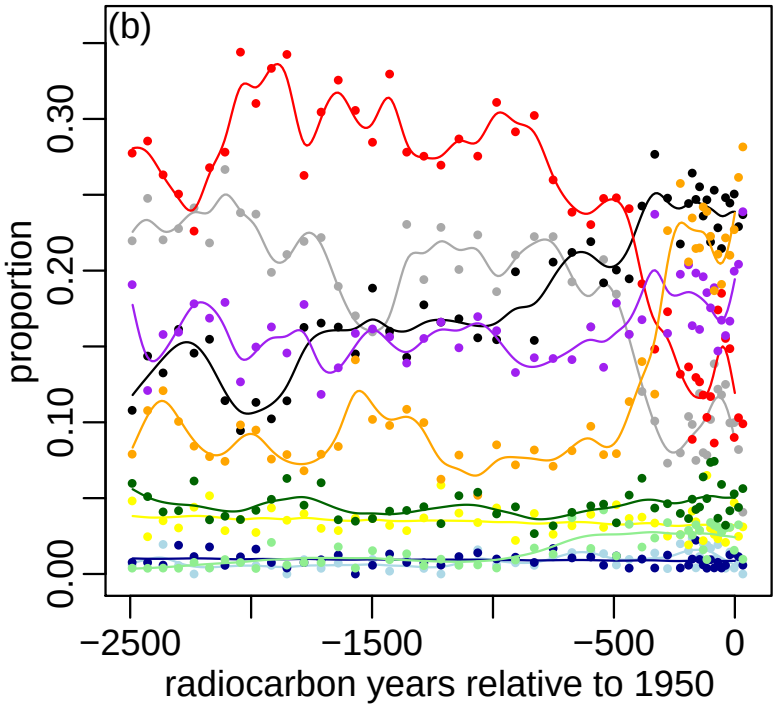
Pollen sediment samples



23 ponds, 500 grains per pond

Pollen time series

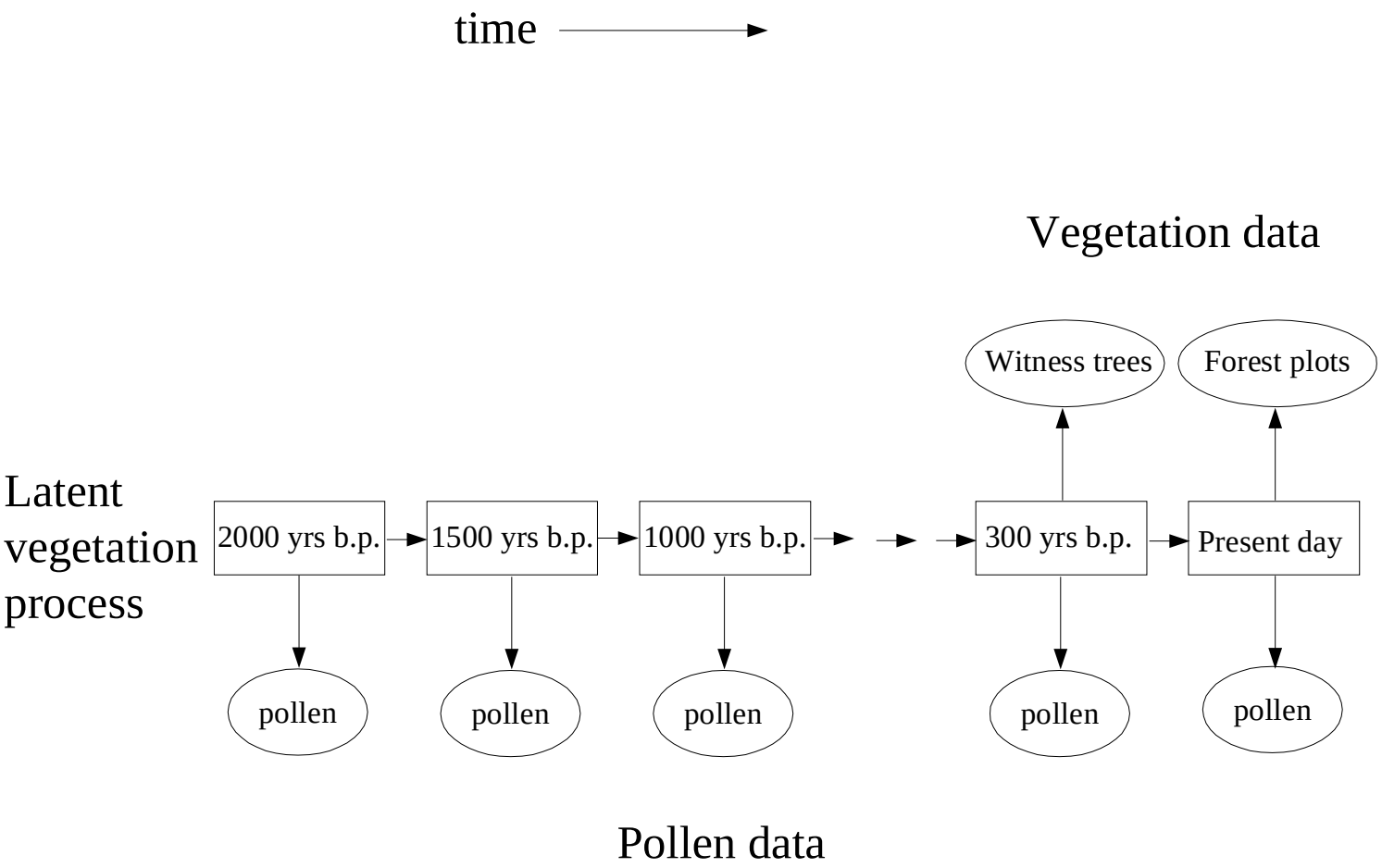




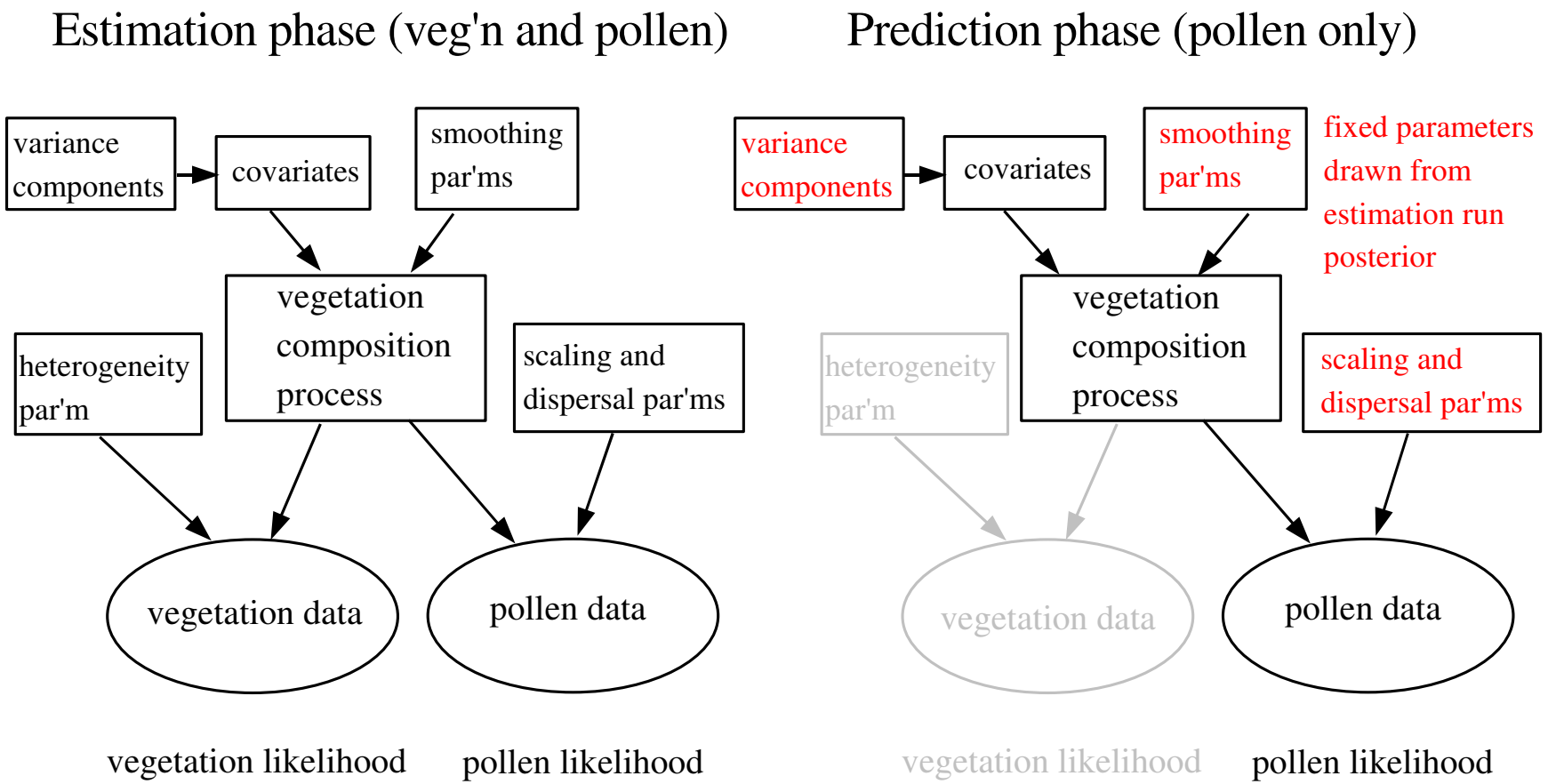
Modeling Goals

- Understand the spatial relationship between pollen and vegetation
  - At what resolution are ponds a good proxy for vegetation?
  - How far and in what quantity does pollen disperse?
- Predict spatial patterns in tree abundances over the past 2000 years
  - Provide uncertainty estimates to allow inference about spatio-temporal patterns
- Assess the predictions to understand vegetation dynamics: changing abundance and ranges of tree species over time.
- Use the model as an ongoing research framework.

Basic problem structure



Model structure



Data model

	types of data	data size	likelihood	relation to latent process
Modern plot data	tree counts in plots	1161	$\mathbf{v}_i^{(2)} \sim \mathcal{DM}(n_i^{v,2}, \alpha_v^2 \mathbf{r(s(i))})$	vegetation in grid cell of sample
Colonial witness data	tree counts in townships	183	$\mathbf{v}_i^{(1)} \sim \mathcal{DM}(n_i^{v,1}, \alpha_v^1 \overline{\mathbf{r(s(i))}})$	weighted average of vegetation in cells overlapped by township
Modern pollen	counts from surface samples	38	$\mathbf{c}_i^{(2)} \sim \mathcal{DM}(n_i^{c,2}, \phi \bullet \widetilde{\mathbf{r(s(i))}})$	weighted average of grid cell and distance-weighted vegetation in other cells
Colonial pollen	pollen counts from settlement horizon	23	$\mathbf{c}_i^{(1)} \sim \mathcal{DM}(n_i^{c,1}, \phi \bullet \widetilde{\mathbf{r(s(i))}})$	same

Process representation

- Proportion of species p at location s, $r_p(s)$, via additive log-ratio transformation (Aitchison 1985) of independent Gaussian processes:

$$r_p(s) = \frac{\exp(g_p(s))}{\sum_{k=1}^{10} \exp(g_k(s))}; \qquad \sum_p r_p(s) = 1$$

- For fixed time, $P = 10$ latent Gaussian spatial processes:

$$g_p(\cdot) \sim \mathcal{GP}(\mu_p 1(\cdot) + \beta_{1,p}\text{elevation}(\cdot) + \beta_{2,p}\text{latitude}(\cdot), \sigma^2 R(\rho, \nu))$$

Computational representation

Processes efficiently represented on a 16 by 16 grid:

$$\mathbf{g}_p = \mu_p \mathbf{1} + \beta_{1,p}\text{elev'n} + \beta_{2,p}\text{latitude} + \sigma\Psi\mathbf{u}_p; \quad \mathbf{u}_p \sim \mathcal{N}(\mathbf{0}, V(\rho,\nu))$$

- Ψ is the Fourier basis matrix
- $\mathrm{V}(\rho,\nu)$ is a diagonal variance matrix based on the spectral density of the Matern (ρ,ν) correlation function
- One ρ and one σ^2 common to all species seem sufficient when covariates included.
- Mixing issues: 'only' 256 locations, but 10 processes and no closed-form conditionals.

Borrowing strength across species and time

- $\{\beta_{1,p}\}$ and $\{\beta_{2,p}\}$ parameters may vary over time, we have limited information from pollen with ≤ 23 ponds.
- One solution is to assume exchangeability, $\beta_{k,p} \sim \mathcal{N}(0, s^2_{\beta_k})$ and estimate $s^2_{\beta_k}$ only in estimation runs.
  - $s^2_{\beta_k}$ stabilizes the posteriors for $\beta^t_{k,p}$ in the prediction runs.
- We also borrow strength across species to estimate the pollen dispersal distance.
- We assume vegetation process smoothness (i.e., σ^2 and ρ) doesn't vary in time.

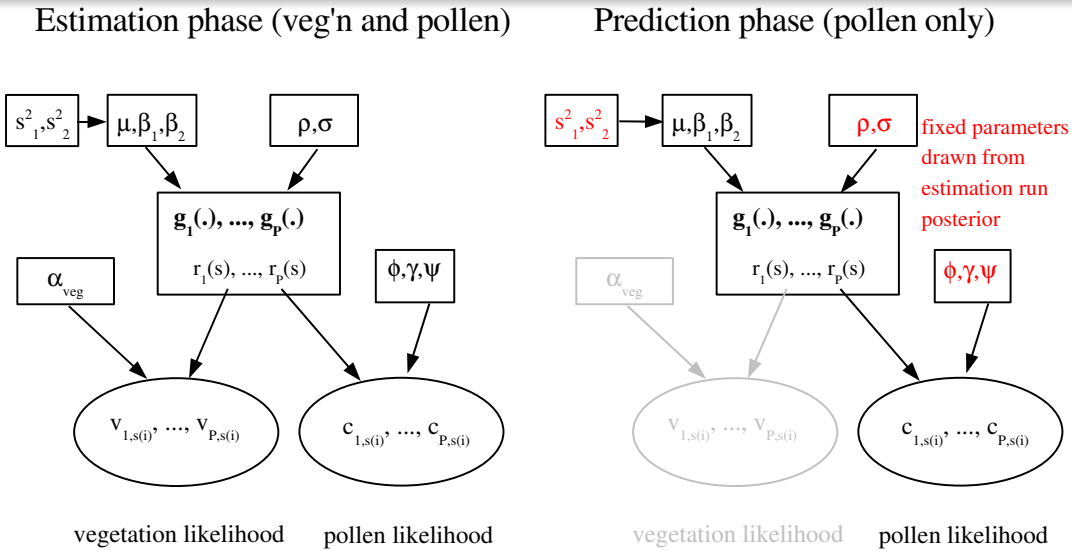
Model development

Build model in stages:

- Understand what features of the model are most important
- Assess how much we trust various aspects of model structure
- Assess different specifications as we go along
  - Assess the bias-variance tradeoff of single parameters vs. time/space/covariate-varying parameters
- Improve our understanding of how the model synthesizes information from multiple data sources
- Detect coding bugs
- Understand what impedes MCMC mixing

Think of the model as a complex system whose dynamics we need to understand.

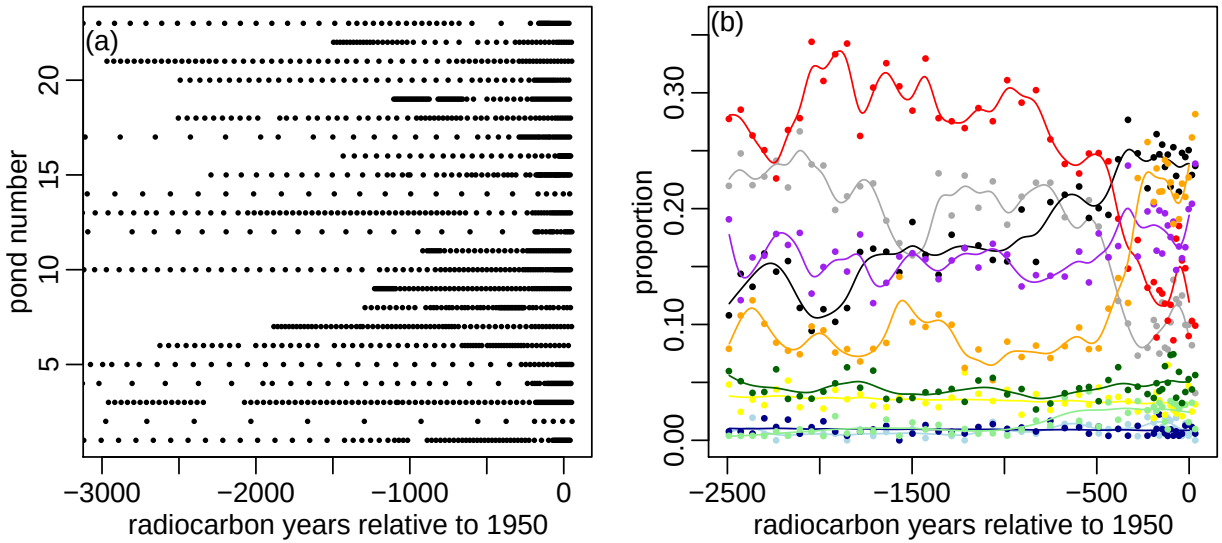
Using the model for prediction



We do not sample a variety of parameters in the prediction runs.

- Why? No information in the pollen-only time points about these parameters.
- Instead, average over samples of the key parameters from their estimation run posterior.
- Elevation and latitude covariates estimated at each time but stabilized by variance parameters from estimation runs.

Using the model for prediction

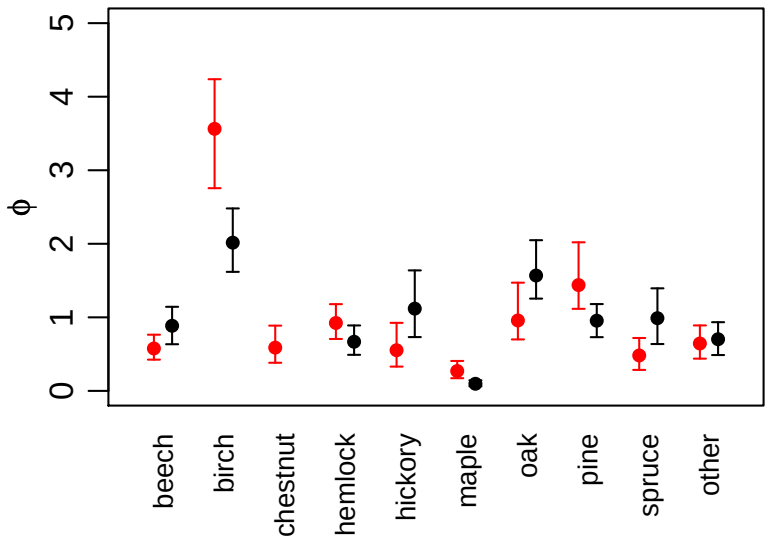


Likelihood for time-smoothed pollen data:

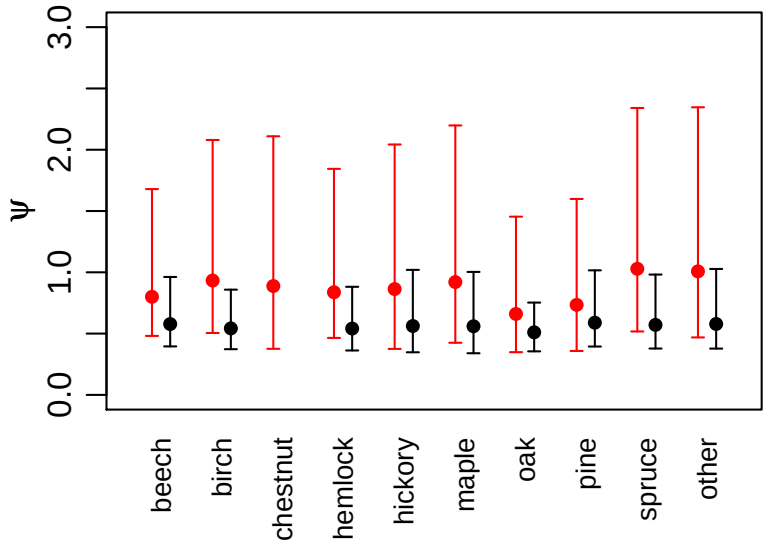
- Recall that we smooth counts over time using gam() at each pond to get predicted composition at fixed times, $\hat{\mathbf{p}}_{\mathbf{i,t}}, \quad t = 0, -100, \ldots - 2000$
- Likelihood for proportions $\hat{\mathbf{p}}_{\mathbf{i,t}} \sim \text{Dir}(\phi \bullet \widetilde{\mathbf{r_t(s(i))}})$

Heterogeneity in pollen production/dispersal

Pollen scaling



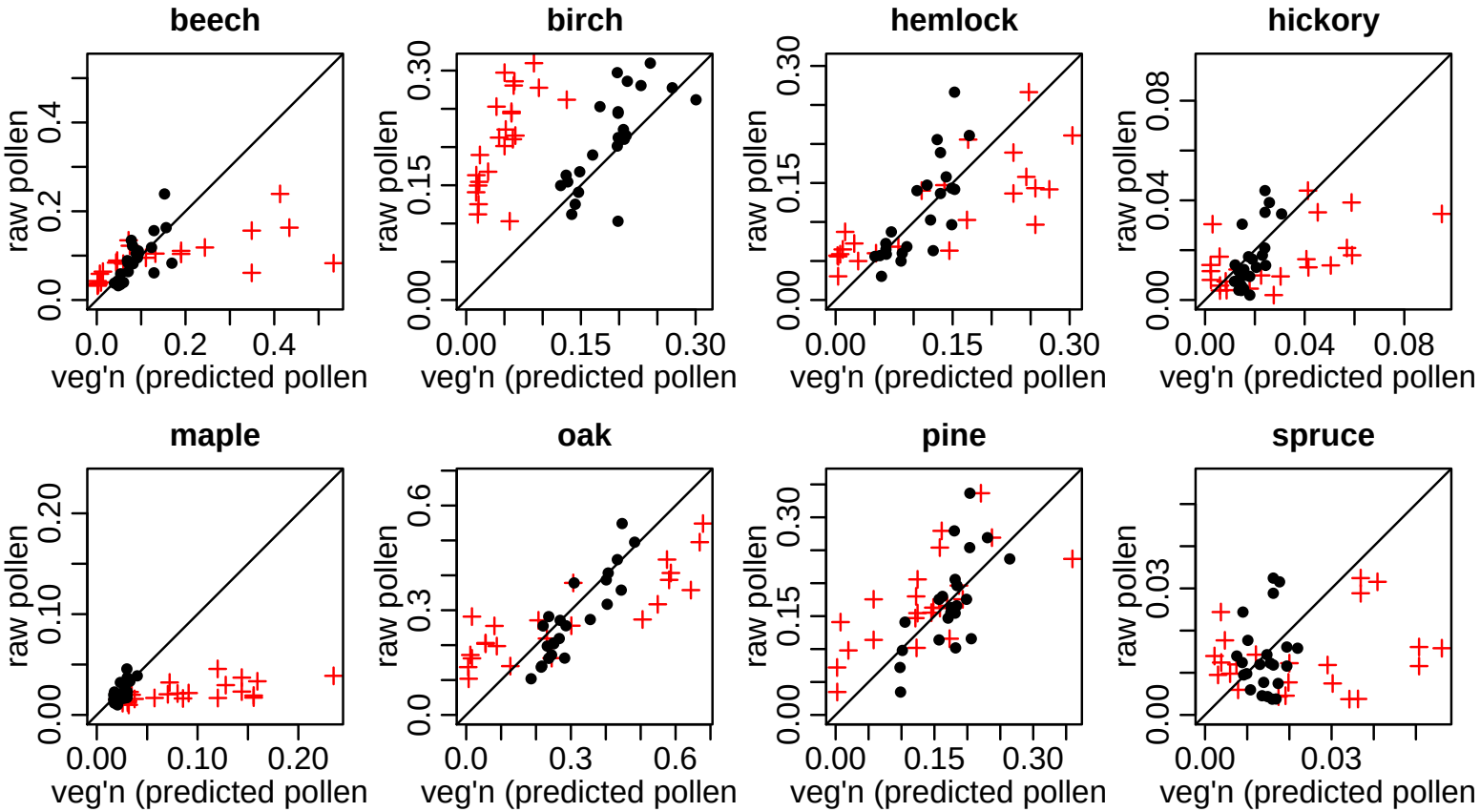
Long-distance dispersal range



red = colonial estimates

black = modern estimates

Pollen-vegetation relationship



+ = raw pollen vs. spatially-smoothed vegetation

• = raw pollen vs. model-predicted pollen based on vegetation

Parameter interpretability

- We estimate the proportion of local pollen (γ) is 10-40%.
  - Ecological knowledge suggests that true proportion is higher, but poorly known.
- γ is a statistical parameter.
  - Model is just optimizing with respect to likelihood and prior.
- How much we can trust the estimate as literally reflecting long-distance pollen contribution?
  - Is the model biased given the data available?
  - True parameter value may lie well outside the area of high posterior density.

Predictive assessment

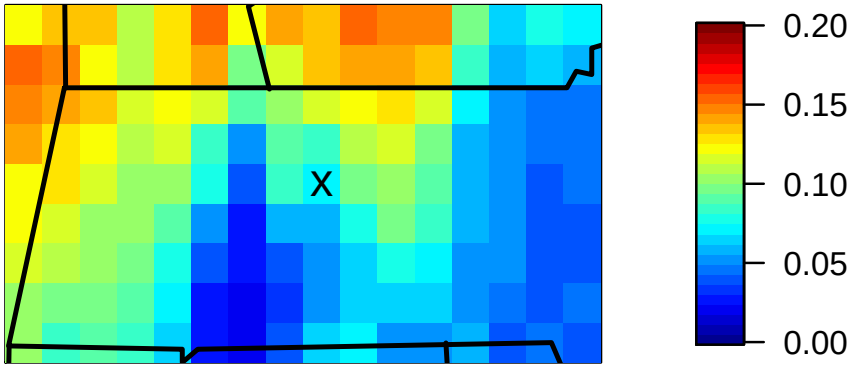
Cross-validation is tricky:

- only have gold standard (vegetation) in two time periods
  - modern parameter estimates + colonial pollen $\Rightarrow$ colonial predictions
  - colonial parameter estimates + modern pollen $\Rightarrow$ modern predictions
- interest lies in qualitative performance: patterns and trends more than absolute levels

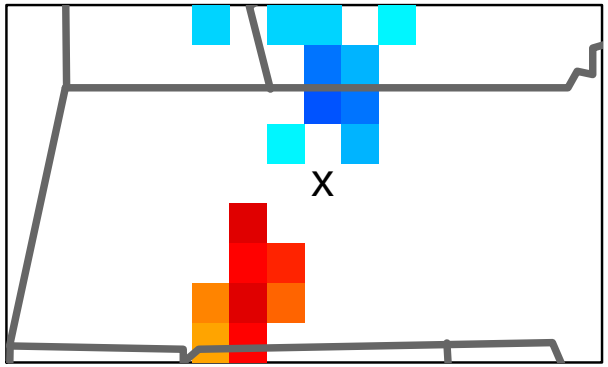
Good diagnostics/plots are critical.

Feature significance

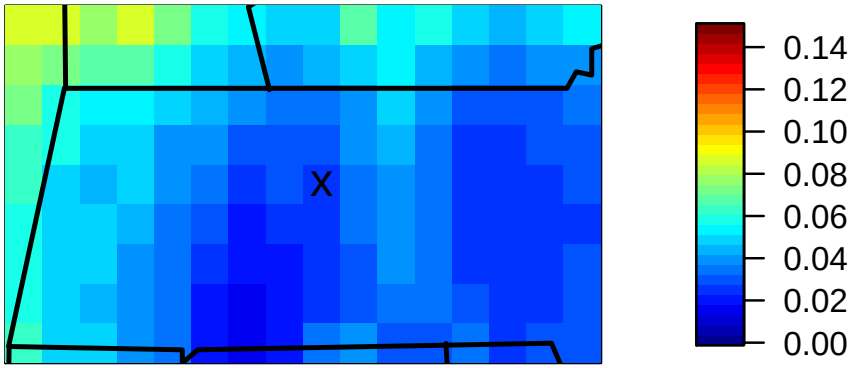
Posterior mean



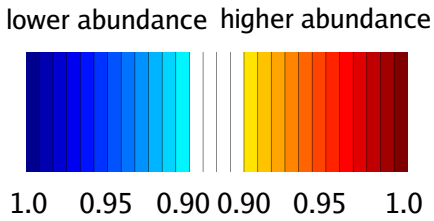
Focal cell feature significance



Posterior std. dev.

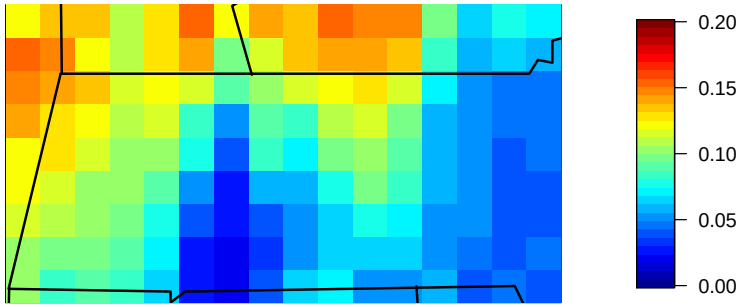


Posterior probability of

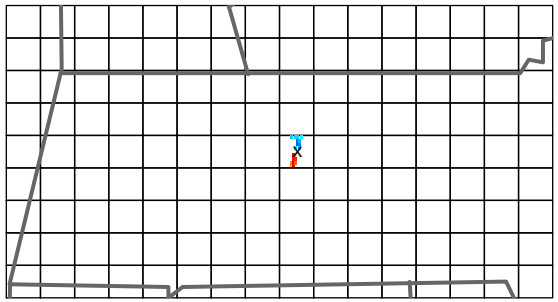


Feature significance (2)

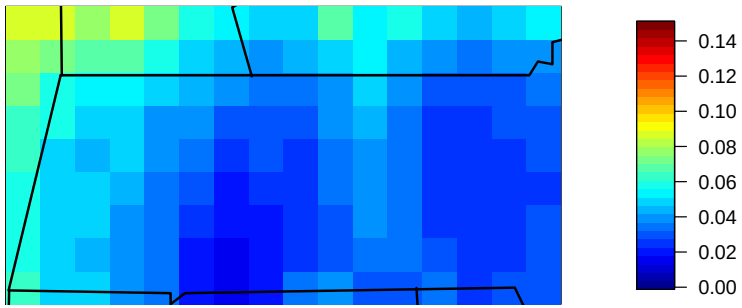
Posterior mean



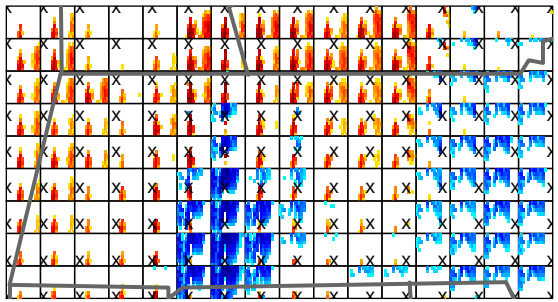
Focal cell feature significance



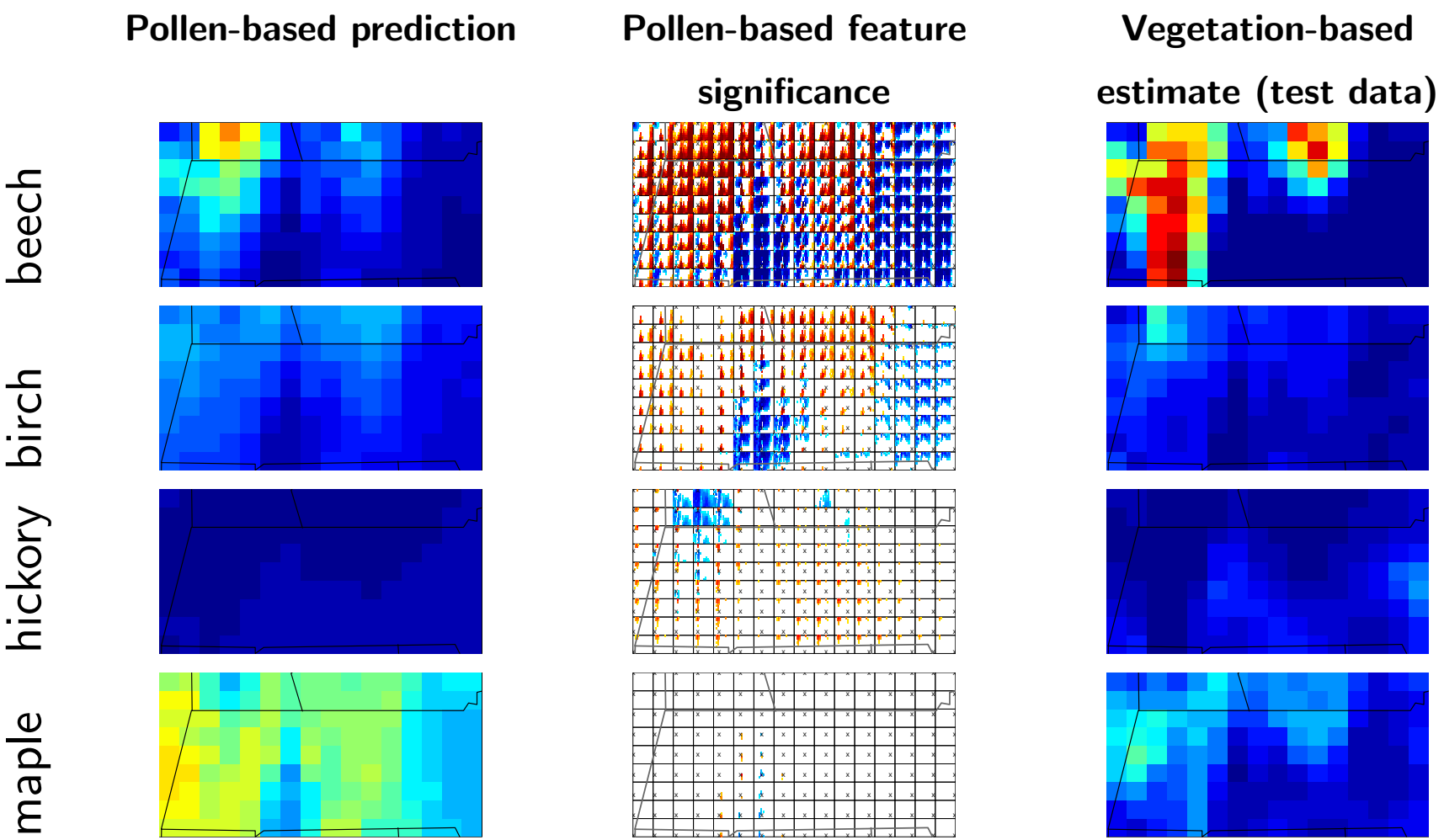
Posterior std. dev.



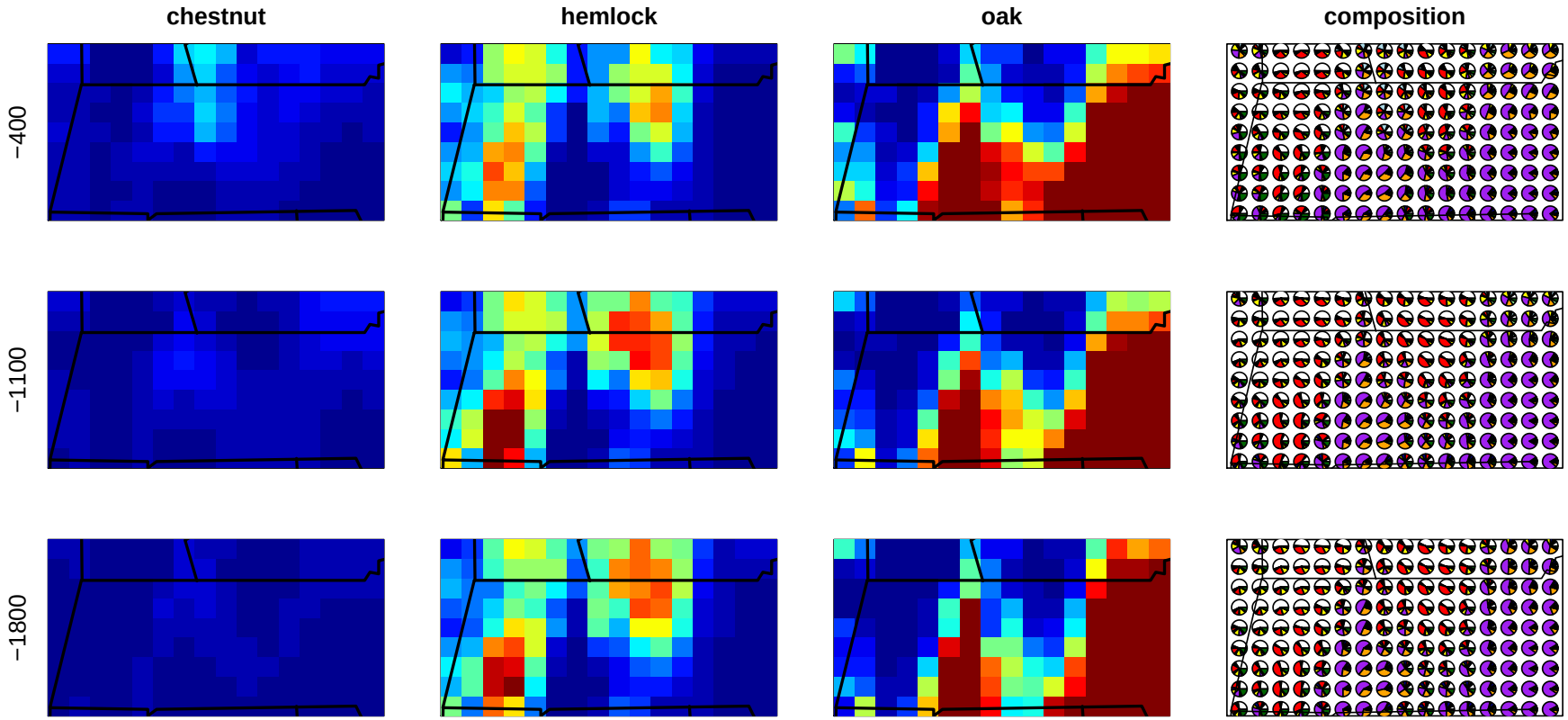
Full feature significance



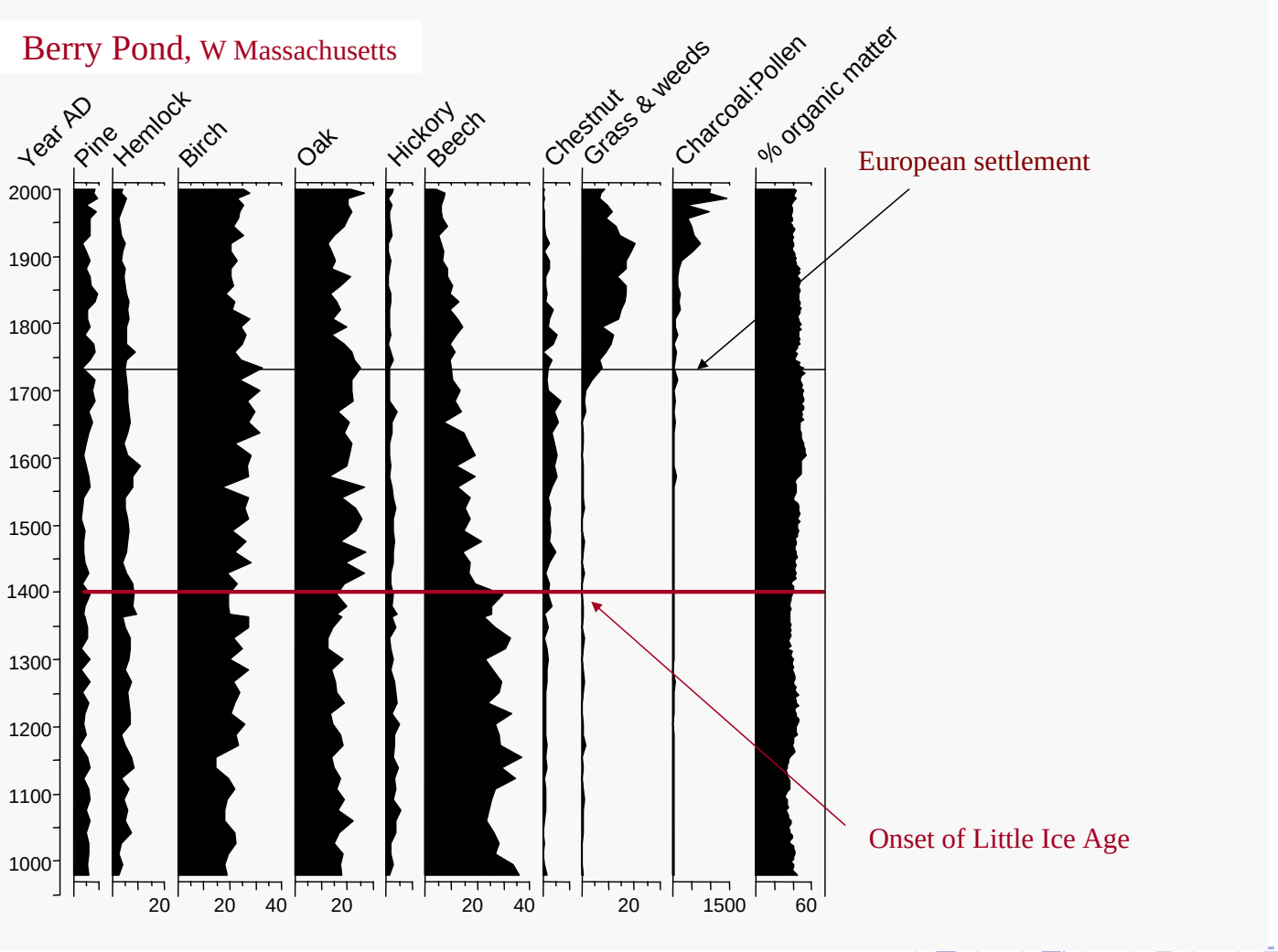
Cross-validation



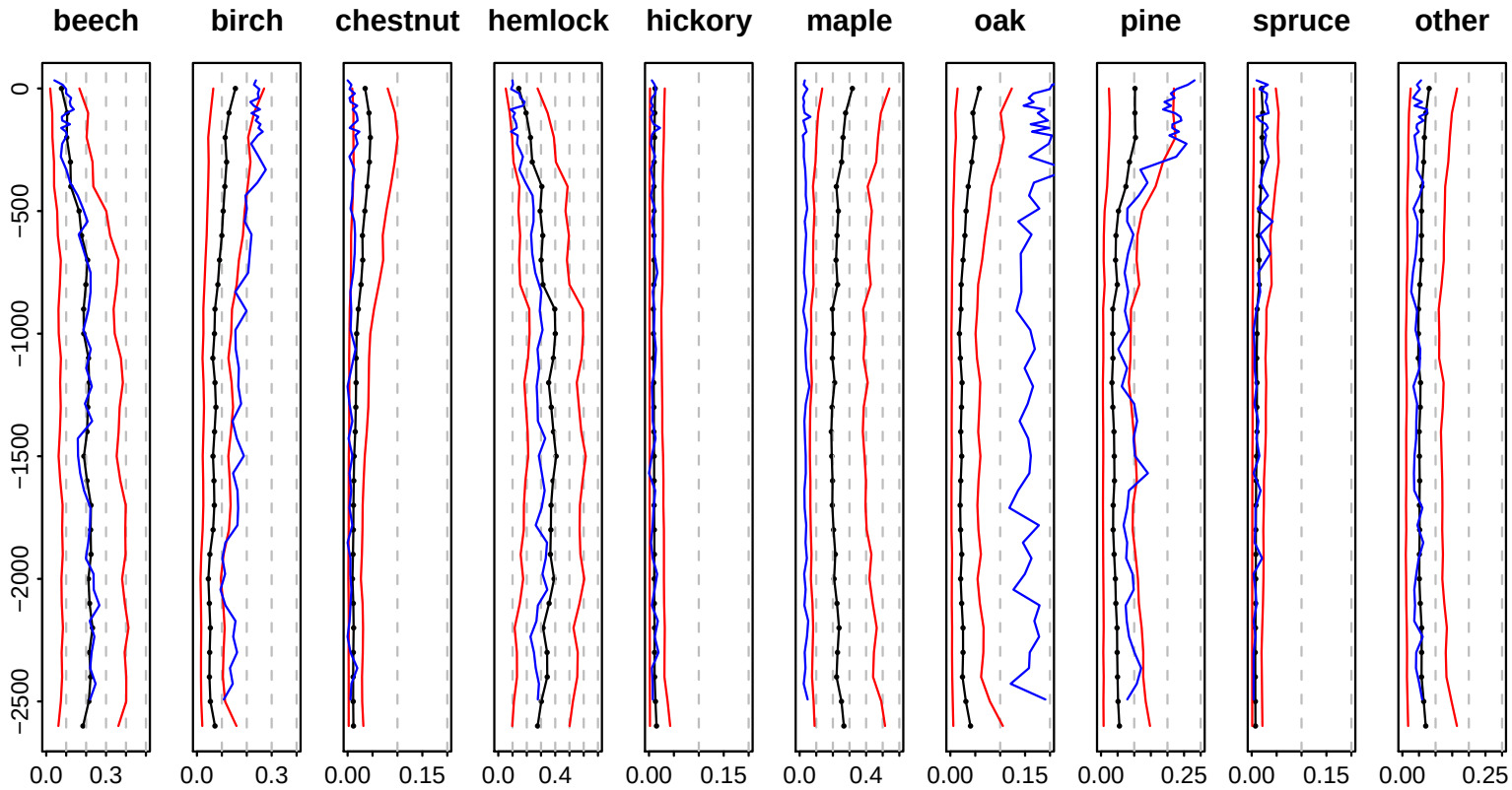
Mapping ancient forests



From pollen diagrams....



.... to statistical vegetation diagrams



Things we left out

- Uncertainty in the radiocarbon dating
  - Not thought to be a key source of uncertainty
- Correlation amongst species
  - Avoid assuming constant inter-species association across space
- Rudimentary approach to spatial and temporal misalignment
- Joint estimation and prediction model
  - Improved model robustness

Next steps

- Full space-time latent vegetation process in prediction phase
  - will give more confidence in temporal comparisons
- Collect additional vegetation data to inform parameters in estimation runs
- Use model in larger space-time domains with more signal
- Parameterize and infer population growth and space-time dynamics
  - Also link to climate variables
- Link with genetic data to infer migration process

Hierarchical Modeling: Art or science?

Science:

- Firm probabilistic foundation
- MCMC well-justified
- MCMC algorithm choices and tricks may become more systematized

Art:

- Model structure subjective
  - Wide variety of choices made in the modeling process
  - Would two statisticians come up with the same results, if not the same model?
- Lack of reproducibility
- Parameter interpretability
- Understanding how information flows through the model
- Choice of output

Hierarchical Modeling: Mathematics or Statistical Science?

Mathematical Statistics:

- Firm probabilistic foundation
- MCMC well-justified
- MCMC algorithm choices and tricks may become more systematized

Statistical Science:

- Model structure subjective
  - Wide variety of choices made in the modeling process
  - Would two statisticians come up with the same results, if not the same model?
- Lack of reproducibility
- Parameter interpretability
- Understanding how information flows through the model
- Choice of output

Bayesian hierarchical modeling in science

- Do we teach our collaborators to implement these models?
- What scientific problems justify the effort of constructing a complicated statistical model?
- In which problems are hierarchical models particularly helpful?