Post-glacial vegetation dynamics: Bayesian inference for tree abundances using the fossil pollen record

Chris Paciorek
Department of Biostatistics
Harvard School of Public Health

Jason McLachlan
Center for Population Biology
UC-Davis

September 16, 2005

www.biostat.harvard.edu/~paciorek

LyX - FoilTEX - pdfLATEX

Scientific Setting

Trees release pollen that accumulates in lake sediments over time.

Ecologists use fossil pollen data collected from lake sediment cores and identified to genus to assess tree species abundances over time.

The pollen record is a biased, noisy reflection of the true tree vegetation.

Current analysis methods focus on time series plots of individual pond records.

A spatio-temporal model can help to estimate spatial maps of tree species compositions at multiple time points over several millenia by understanding the relationship between the pollen record and vegetation.

The model needs at least one time point of concurrent pollen and ground-truth vegetation data to estimate the relationship between pollen and actual vegetation.

Central New England pollen and vegetation data

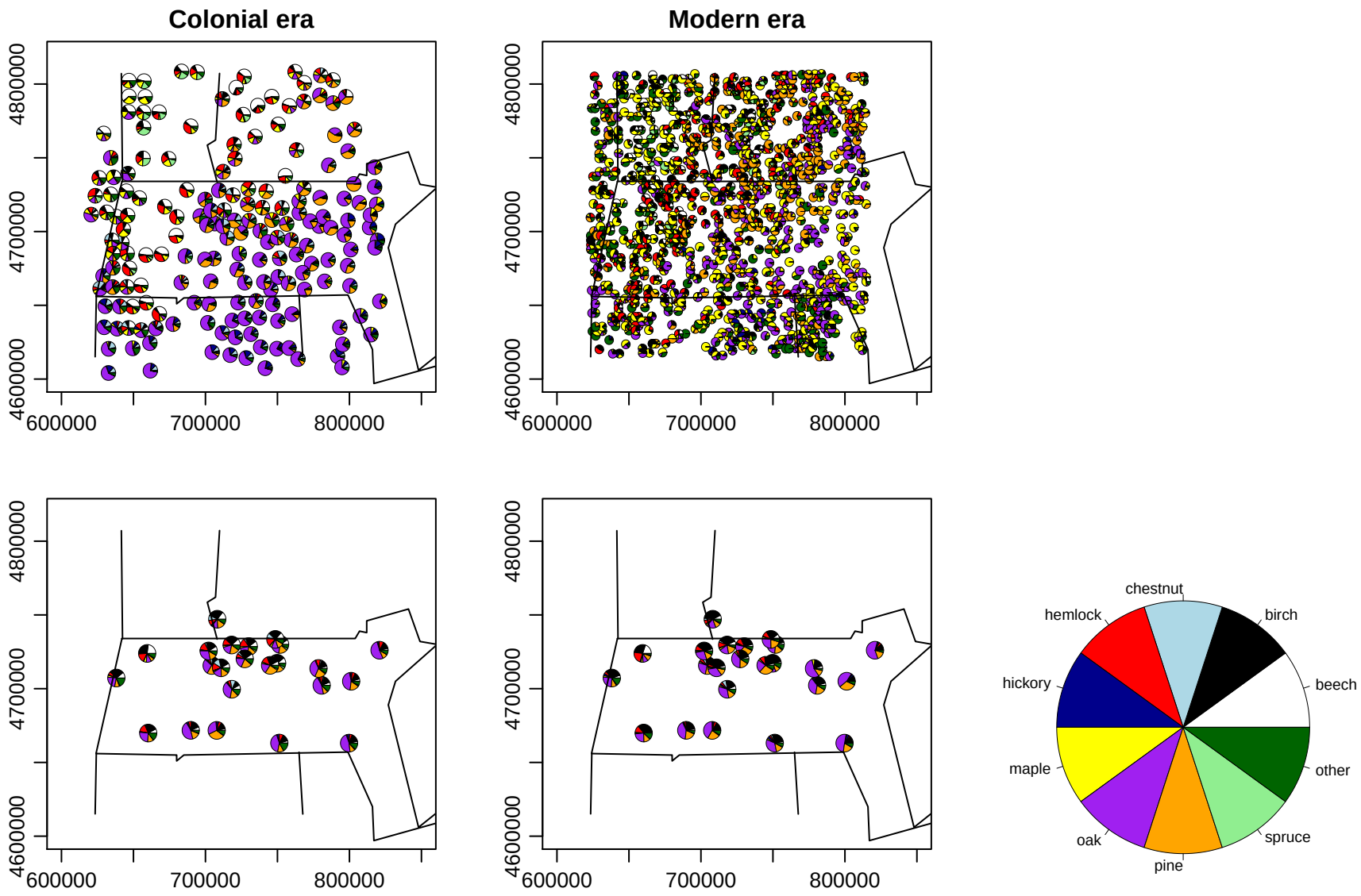


Figure 1: Colonial era witness tree data (top left) and modern plot data (top right); colonial (bottom left) and modern (bottom right) pollen data

Goals

- Understand the spatial relationship between pollen and vegetation. At what resolution do ponds predict vegetation? How far and in what quantity does pollen disperse?

- Estimate and compare spatial patterns in tree abundances for the colonial and modern eras based on relevant vegetation data.

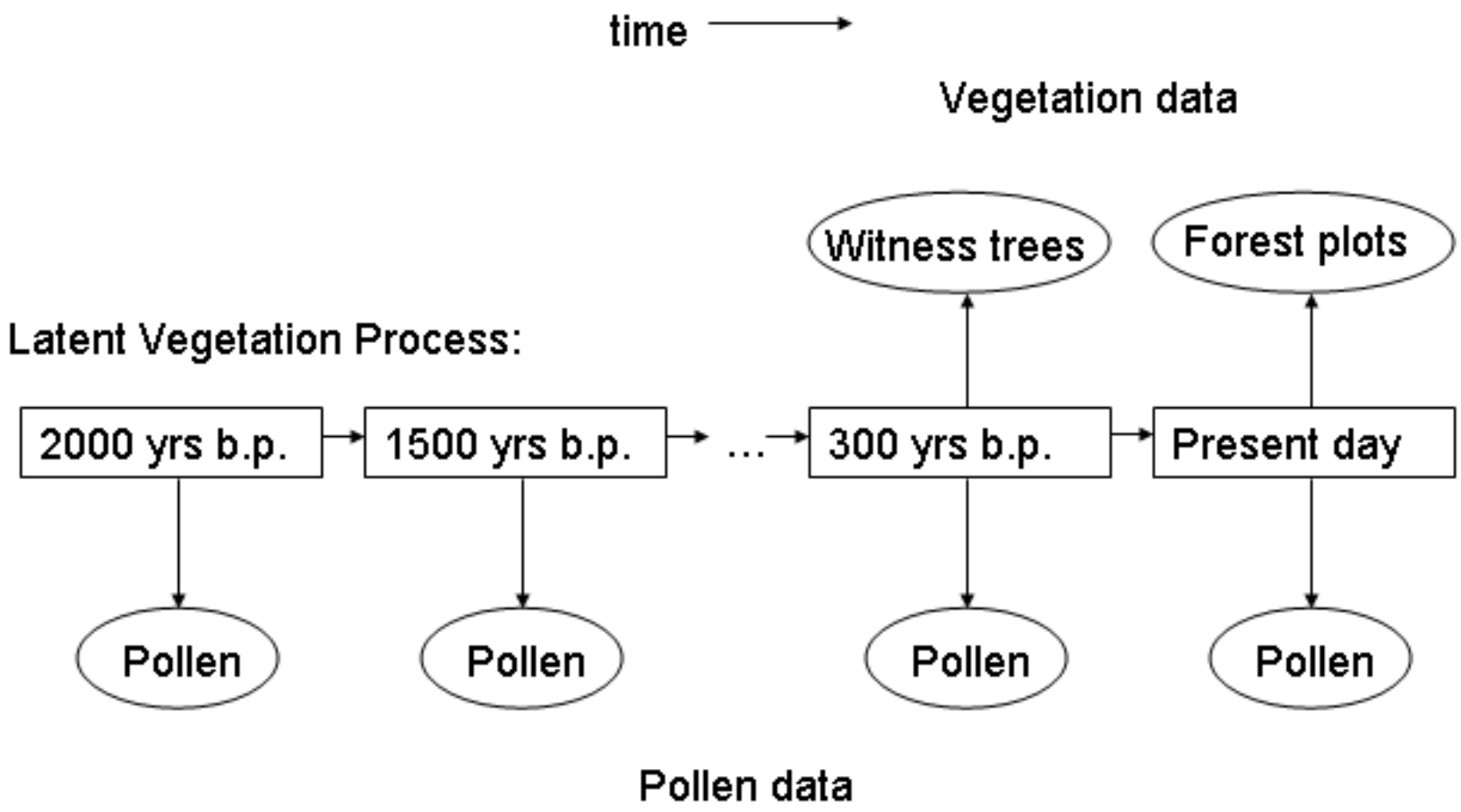
- Predict spatial patterns in tree abundances based on pollen data over the past few thousand years, with uncertainty estimates.

- Assess the predictions to understand tree community dynamics: changing abundance and ranges of tree taxa over time.

- Assess the degree of certainty reasonable in making inference about tree abundances based on pollen records. At what spatial scale can the pollen record distinguish spatial heterogeneity in tree abundances?

- Use the model to integrate genetic data and understand genetic heterogeneity.

Basic model and data structure



Model (1): Latent spatial processes

For each time t, 9 latent Gaussian spatial processes represent spatial taxa compositions for the 9 taxa of primary interest:

$$g_p(\cdot) \sim \mathrm{GP}(\mu_p \mathbf{1}, \sigma_p R(\rho, \nu))$$

Proportion of taxa p at location s, $r_p(s)$, is calculated using the additive log-ratio transformation (Aitchison 1985), with the reference group being all 'other' trees:

$$r_p(s) = \frac{\exp(\mu_p + \sigma_p g_p(s))}{1 + \sum_{k=1}^{9} \exp(\mu_p + \sigma_p g_k(s))},$$

Processes efficiently represented on a 16 by 16 grid:

$$\boldsymbol{g}_p \quad = \quad \mu_p \mathbf{1} + \sigma_p \Psi \boldsymbol{\alpha}_p; \qquad \boldsymbol{\alpha}_p \sim \mathrm{N}(\mathbf{0}, V(\rho, \nu)),$$

where Ψ is the Fourier basis matrix (Wikle 2002, Paciorek & Ryan 2005) and $V(\rho, \nu)$ is a diagonal variance matrix based on the spectral density of the Matern (ρ, ν) correlation function

Model (2): Likelihood terms

- Modern plot data (tree counts), $s = 1, \ldots, 1161$:
  - $\boldsymbol{f}_s \sim \text{Dir-multi}(n_{f,s}, \alpha_t \boldsymbol{r}(\boldsymbol{s}))$
  - α_f is extra-multinomial heterogeneity
    $\boldsymbol{r}(\boldsymbol{s})$ is composition vector $(r_1(s), \ldots, r_{10}(s))$
- Colonial surveys (witness tree counts in townships), $s = 1, \ldots, 183$:
  - $\boldsymbol{w}_s \sim \text{Dir-multi}(n_{w,s}, \alpha_w \overline{\boldsymbol{r}(\boldsymbol{s})})$
  - $\overline{\boldsymbol{r}(\boldsymbol{s})}$ is the weighted composition based on town-gridbox overlap
- Pollen data (pollen grain counts from 22 ponds at a fixed time), $s = 1, \ldots, 22$:
  - $\boldsymbol{c}_s \sim \text{Dir-multi}(n_{c,s}, \alpha_c \boldsymbol{\phi} \cdot \boldsymbol{r}(\boldsymbol{s}))$
  - ϕ scales compositions to account for taxa heterogeneity in pollen production and dispersal (element-wise multiplication)

Initial results (1): Colonial and modern vegetation

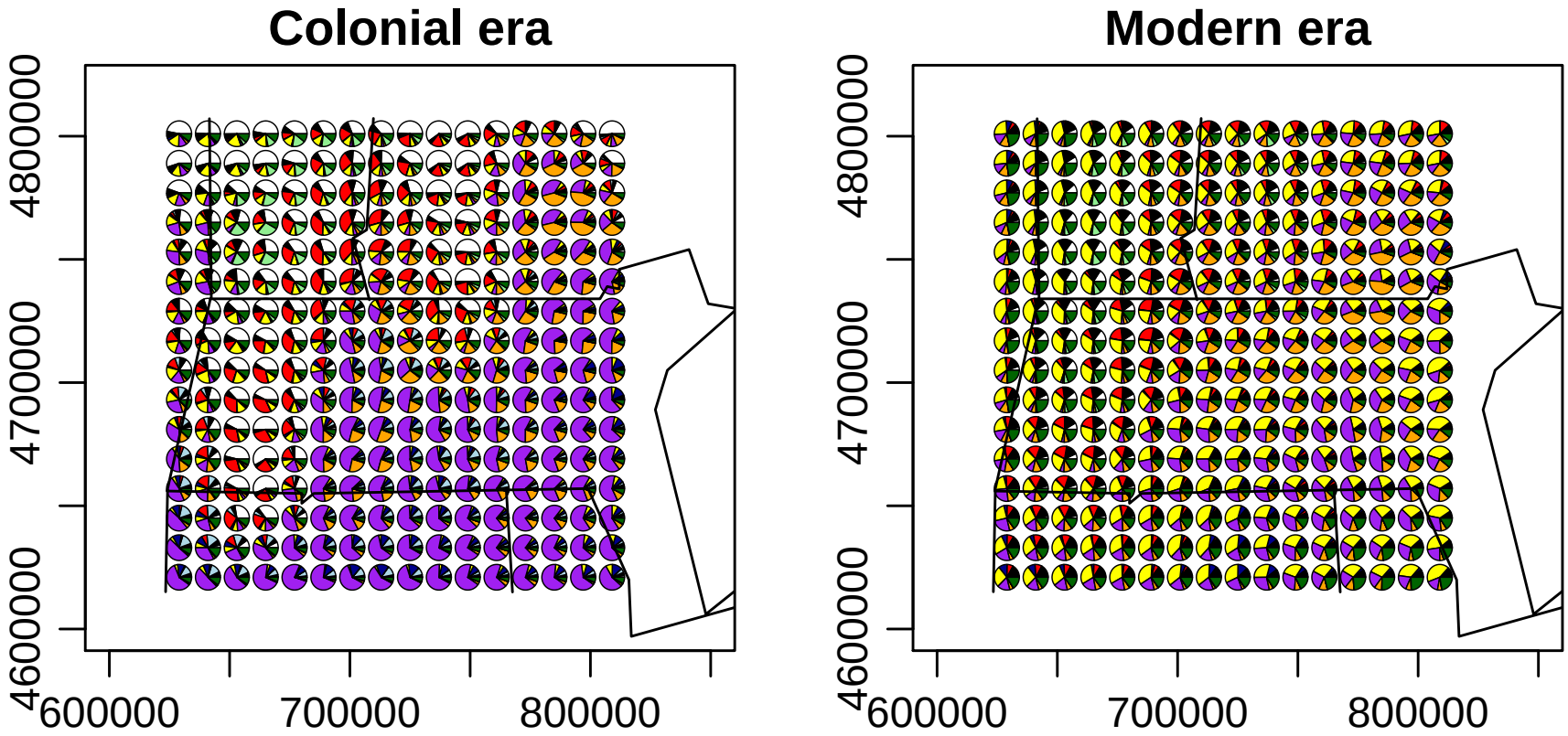


Figure 2: Tree composition predictions using ground-truth vegetation data.

Initial results (2): Predictions based on pollen

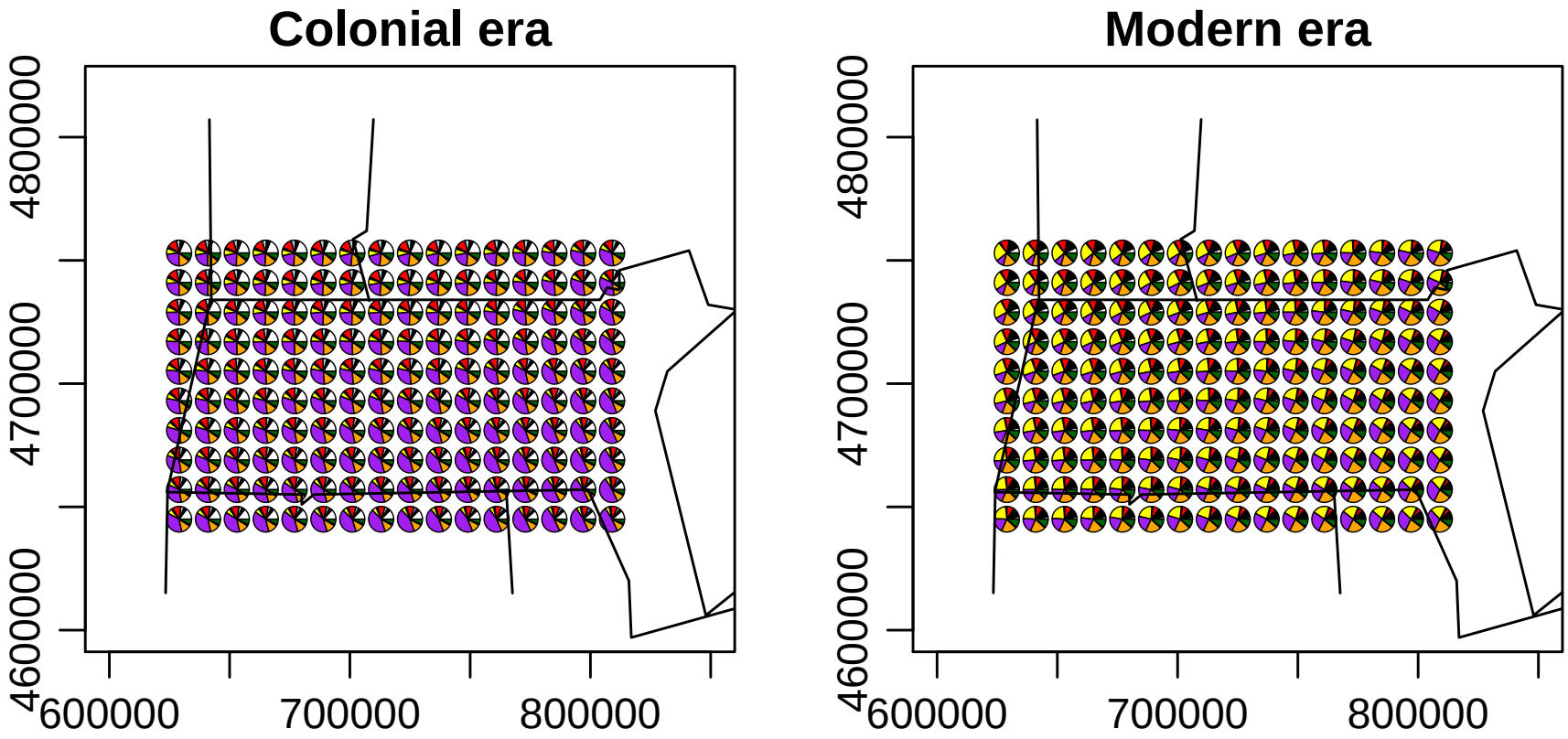
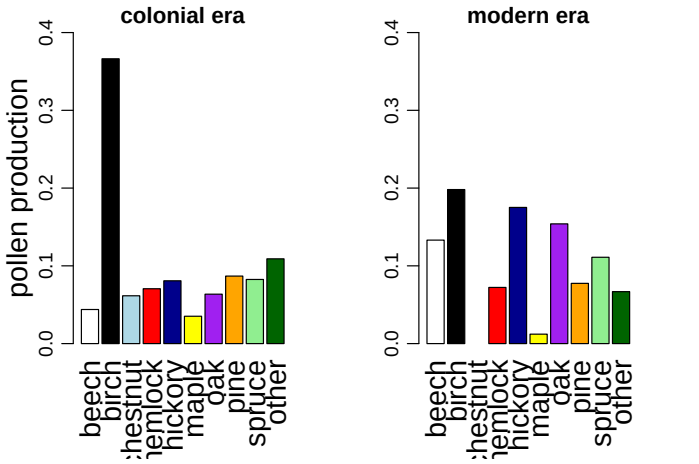
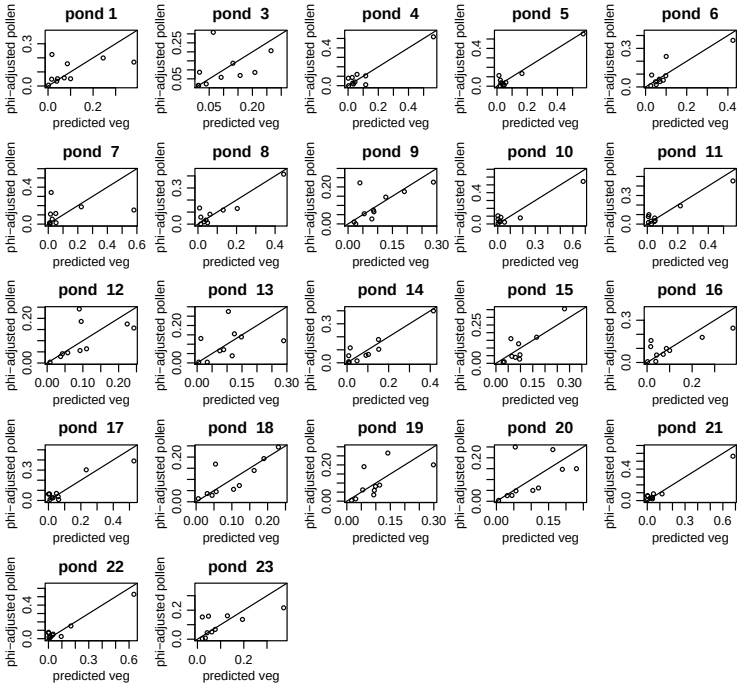


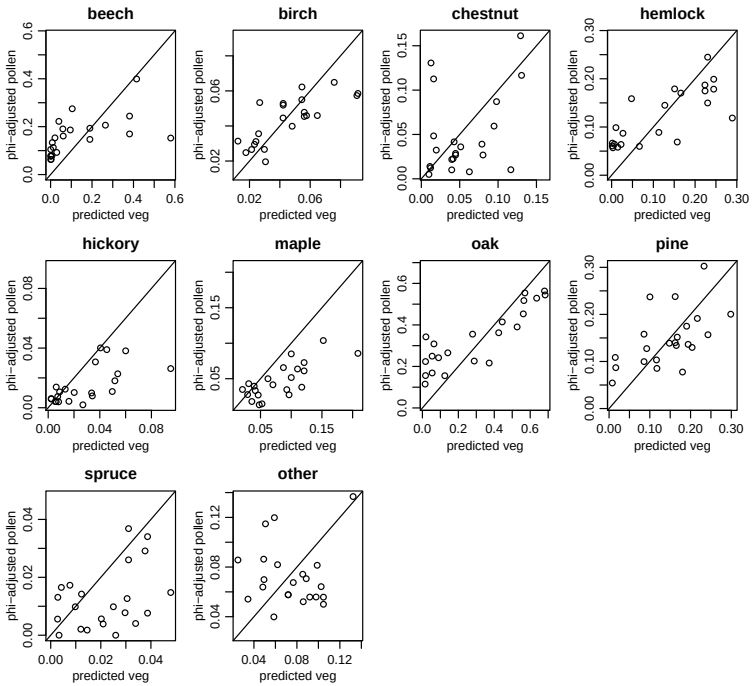
Figure 3: Tree composition predictions using only pollen data.

Pollen-vegetation mismatch





ϕ parameters (see left) adjust for differential pollen production and dispersal, but even after adjusting for $\hat{\phi}$, we see more homogeneity in the pollen data than in the smoothed ground-truth vegetation. This appears to make it difficult to infer vegetation from pollen, based on diagnostic plots for the colonial era comparing vegetation composition to pollen composition, by pond (below left) and by taxa (below right).



Current challenges

- Some ponds poorly reflect ground-truth vegetation - why?
  - Can we use relevant covariates to explain pond anomalies?
- Predictions from pollen are very smooth - is the model oversmoothing?
  - Would different model structure better identify heterogeneity or is our spatial scale too small?
- MCMC mixing is rather slow - would a CAR representation help?

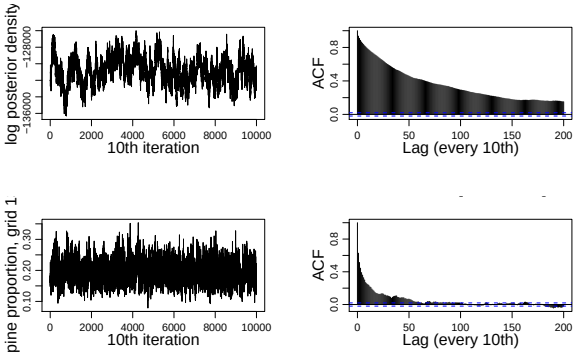


Figure 4: Mixing of log posterior and an example taxa proportion from the colonial era model

- How should we assess joint uncertainty, across multiple locations and taxa?

Future work

- Continued model exploration and selection
- Addition of ponds from north (NH/VT) and south (CT) parts of study region to assess ability of pollen to resolve spatial variability
- Predictive inference over the past 4000 years, including modelling temporal autocorrelation in the latent processes if necessary
- More explicit pollen dispersal modelling, perhaps with a long-distance component
- Expansion to the northeastern United States + southeastern Canada to get to a scale where pollen can resolve spatial hetero-geneity and assess tree migration
- Integration of the modelling with genetic data to better understand migration and genetic heterogeneity