Statistical Inference in Paleoecology, with a Focus on Bayesian Hierarchical Modeling

Chris Paciorek
Department of Statistics
University of California, Berkeley
www.stat.berkeley.edu/~paciorek

Joint work with Jason McLachlan, Notre Dame Biology, and the PalEON Project (PIs: J. McLachlan, M. Dietze, S. Jackson, C. Paciorek, J. Williams)

(Some) Paleoecological Data Sources

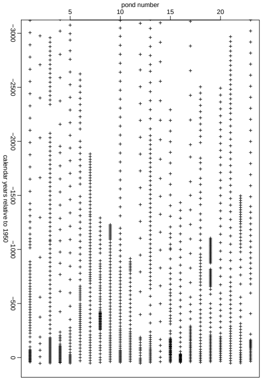
- Counts of pollen grains from sediment cores in lakes and other depositional environments
    - Inference: vegetation composition, vegetation types, ecosystem boundaries
- Counts of charcoal particles from sediment cores
    - Inference: fire frequency and severity
- Ring widths from tree cores
    - Inference: growth, biomass and carbon balance
- Fire scar data from tree cores
    - Inference: fire frequency and severity

(Some) Goals of Paleoecology

- Understand past distributions of vegetation and changes in those distributions
- Use long-term data to understand the nature of vegetation dynamics:
  - competition
  - species dispersal/spread
  - species declines and causes of those declines – impacts of disturbance, disease, herbivory, climate
  - stability of species assemblages
- Understand patterns and rates of large-scale disturbance

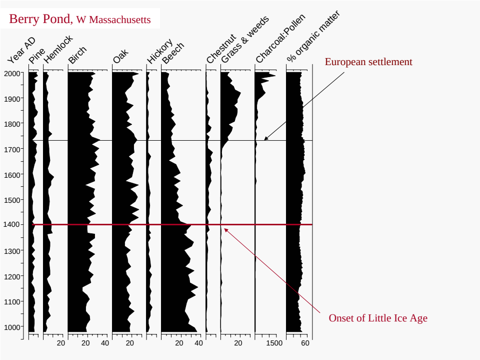
Challenges of Paleoecological Data Sources

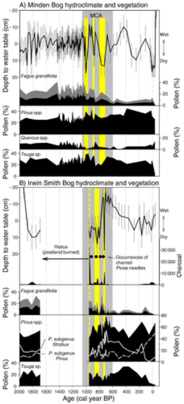
- Sparsity and irregularity in space and time
    - Certain proxies are only available in certain regions
    - Many records of limited duration
- Lack of replication
- Proxies are not direct measurements of the quantities we care about
- Calibration data are scarce
- Calibration against modern data may be less relevant for periods in the past (the no analog problem)
- Many of the quantities of interest do not have paleodata proxies
- Dating is uncertain and dating methods are expensive



Temporal sampling density for 23 ponds in central New England

Analysis of Pollen Diagrams





Booth et al. (2012) Ecology 93:219

Pollen in a Spatial Context

690 KERRY D. WOODS AND MARGARET B. DAVIS Ecology, Vol. 70, No. 3

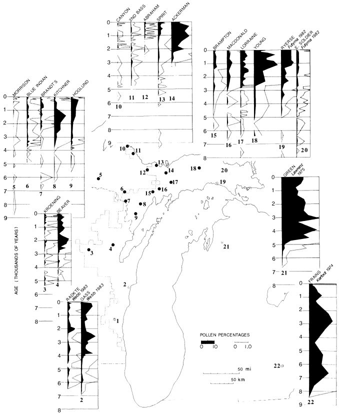
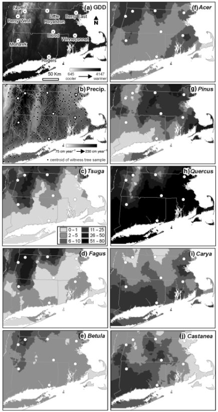
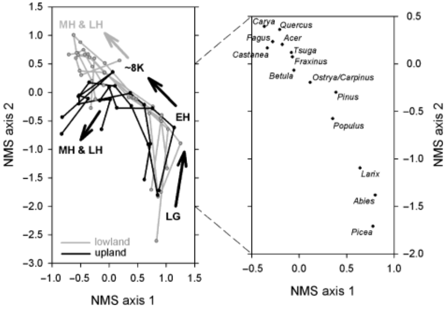


FIG. 5. Curves showing beech pollen percentages plotted against age in thousands of radiocarbon years for selected sites. Beech range limit, generalized to township, is shown by the dotted line. ◎ sites from the literature. Curves are shown for all unpublished sites (●) used. Site numbers are given at the bottom of each curve. Vertical scales are in thousands of years before present. Dotted curves (unshaded) are pollen percentages multiplied by 10.

Dimension Reduction





Oswald et al. (2007); J of Biogeography 34:900

Calibrating Pollen to Vegetation

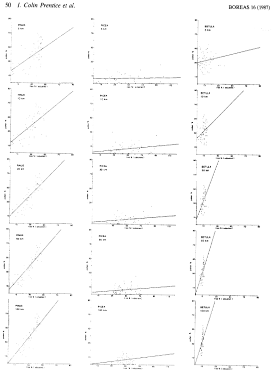
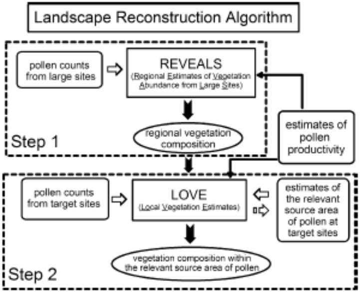

50 *I. Colin Prentice et al.* BOREAS 16 (1987)

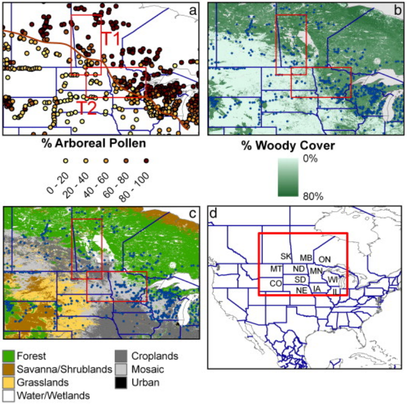
Fig. 4A. Scatterplots showing relationships between surface pollen percentages and tree volume percentages within various distances of the pollen sampling sites. The tree percentages have been adjusted by multiplying by site factors (equation (3)). This adjustment is designed to eliminate the effects of the percentage constraint. The straight lines fitted to the data have slopes α_i and intercepts z_i (equation (2)), as estimated by the method of Parsons & Prentice (1981) and Prentice & Webb (1986). The goodness-of-fit to these lines indicates goodness-of-fit to the extended R-value model of the pollen-vegetation relationship.

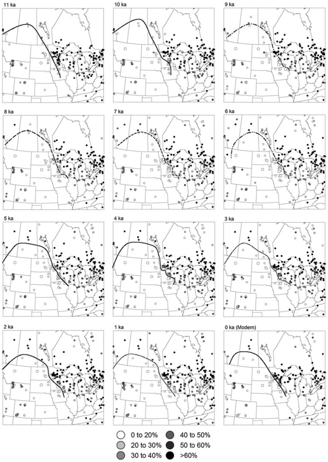
Prentice (1987); Boreas 16:43



Sugita (2007); The Holocene 17:243

Ecosystem Boundary Reconstruction





Williams et al. (2009); Global and Planetary Change 66:195

Fire History Reconstruction

Fire return intervals from peak detection

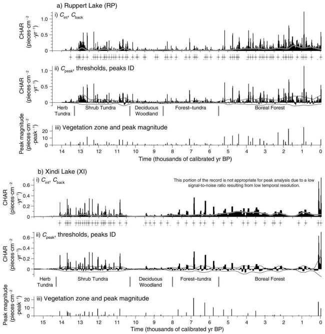


Fig. 4. Charcoal records for (a) Ruppert, (b) Xindi, (c) Code, and (d) Wild Tussock lakes. (i) Interpolated charcoal accumulation rates (CHAR), C_{int} (black), and background CHAR, C_{back} (gray); (ii) Peak CHAR, C_{peak}, with the values identifying noise-related variability (positive and negative gray lines) and peaks identified with each threshold criterion. The 99th percentile criterion used for interpretation is represented with +, and the 95th and 99.9th percentile results are represented with gray dots. (iii) Pollen-inferred vegetation zone and peak magnitude for all charcoal accumulation rate (CHAR) values exceeding the positive threshold value in panels ii. Note peak magnitude values not corresponding to + symbols in panels ii are those that did not pass the minimum-count screening (see *Methods* for details), and + symbols in panels ii with no apparent peak magnitude value correspond to very small peak magnitudes.

Discussion

Interpreting sediment charcoal records and detecting changes in fire regimes

We introduce three general tools that facilitate the interpretation of fire history from sediment-charcoal records. First, the signal-to-noise index provides a semi-objective way to judge if a record is appropriate for peak analysis. For example, while >0.8 in most records, SNI values were consistently <0.5 for the 8000–0 yr BP in the Xindi Lake record (data not shown), indicating that this section was not suitable for peak identification. Second, our use of a Gaussian mixture model to determine threshold values for peak identification allowed us to treat all charcoal records with one set of semi-objective criteria. These criteria are consistent with a mechanistic

Local area burned from background charcoal

Table 3. Alternative regression models relating charcoal accumulation in a composite record to area burned from AD 1675 to 1960 ($n = 19$). Positive reduction of error (RE) values indicate that the model is a better predictor of area burned than the mean of the series alone (i.e. the model has skill). Cross-validation involved constructing 5000 models based on a random subset of data points (53%) and then calculating the RE statistic for predictions using data excluded from the model (see methods)

Sites contributing	Model: $y = ax^b$	F-stat	p	r^2	r^2_{adj}	RE	Cross-validation RE_{median}
DU, DR, MA, WT	$a = 60202; b = 2.250$	93.31	0.0000	0.80	0.79	0.78	0.67
MA, WT	$a = 25950; b = 1.711$	40.55	0.0000	0.61	0.59	0.46	0.52
WT	$a = 70780; b = 4.936$	36.50	0.0000	0.66	0.64	0.53	0.65

Site codes are listed in Table 1.

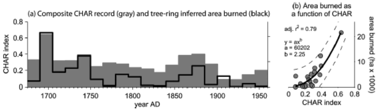


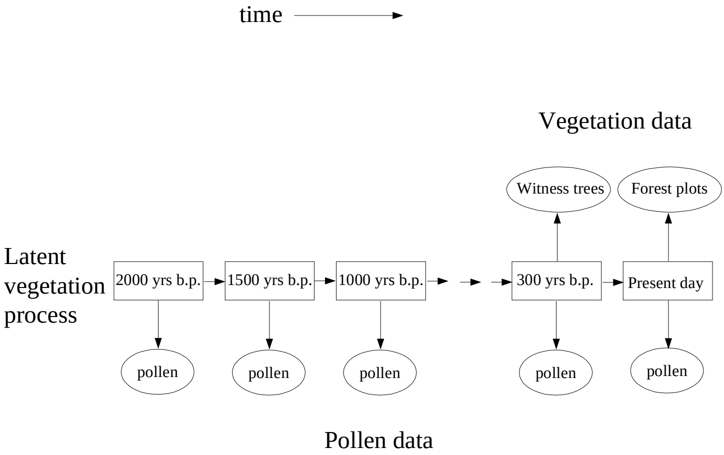
Figure 4. Comparison between the four-site composite charcoal record and area burned within the entire study. (a) Composite charcoal record, expressed as a charcoal accumulation rate (CHAR) index (gray bars, left y-axis), and area burned (thick line, far right y-axis) for the AD 1675–1960 calibration period. (b) Scatter plot of area burned as a function of CHAR from the two series in (a) with the best-fit power model and adjusted r^2 statistic. Dashed lines represent 90% confidence intervals for new predictions

Higuera et al. (2011) Ecological Applications 21:3211

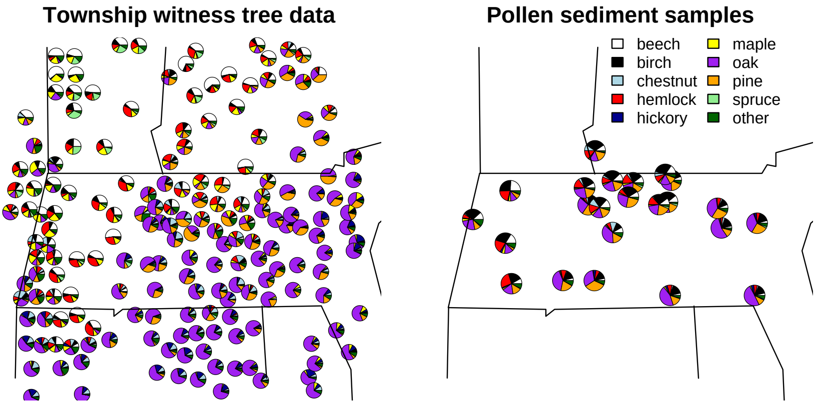
Overview

- Hierarchical statistical models build a (possibly complicated) statistical model that relates data to unknown quantities of interest in (relatively simple) stages.
    - Measurement model: Data are related to a latent process (often a space-time process representing a relevant field)
    - Process model: Latent process is modeled stochastically (potentially with deterministic components) that build in appropriate dependencies
    - Parameter model: Additional 'tuning' parameters govern the behavior of the latent process.
- The goal is to make inference (including uncertainty assessment) about the the key quantities of interest, which may be the latent process or parameters or functionals of those.
- Given a model, there are standard (but sometimes inadequate) computational approaches to computing the inferences

Example: STEPPS Model for Vegetation Reconstruction

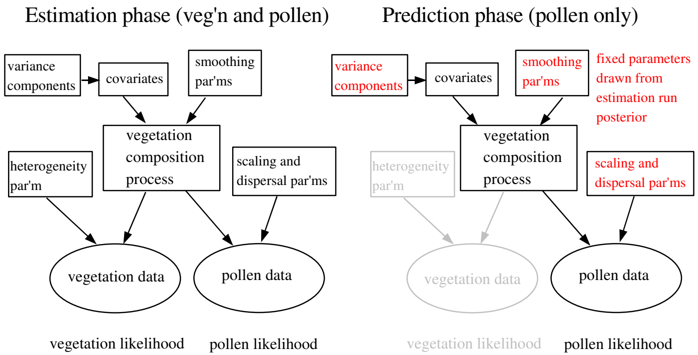


Calibration Data

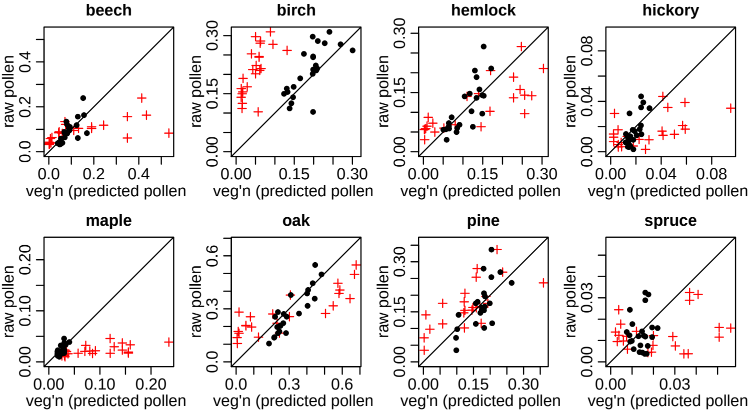


183 towns, 26-3149 trees per town   23 ponds, 500 grains per pond

A Cartoon of the Model



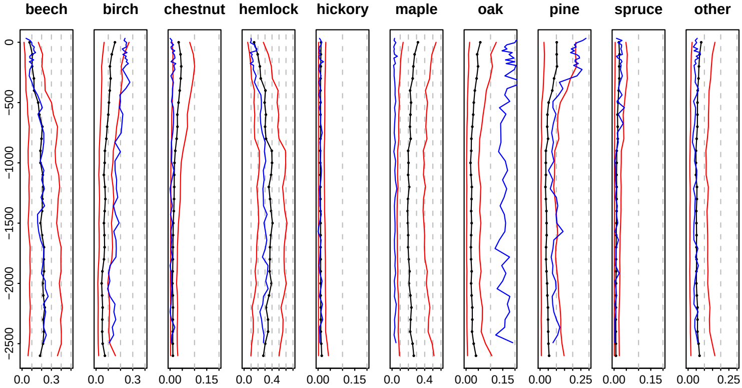
Calibration of Pollen to Vegetation



+ = raw pollen vs. spatially-smoothed vegetation

• = raw pollen vs. model-predicted pollen based on vegetation, accounting for species-specific pollen production and for long-distance pollen dispersal

Inference: Time

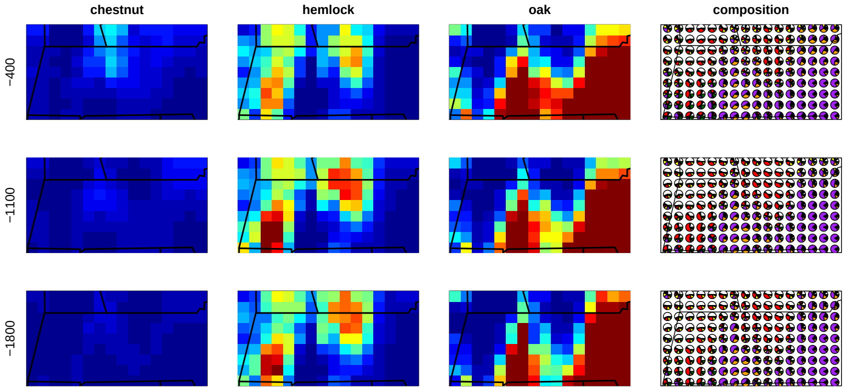


– – – = raw pollen proportions

– • – = model-estimated vegetation proportions

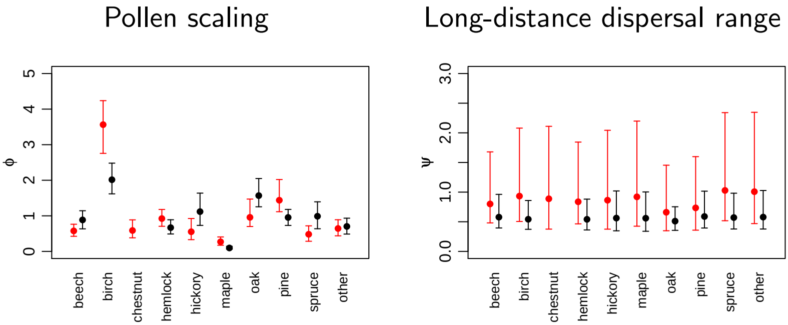
– – – = uncertainty estimates

Inference: Space



We can also present spatial predictions in the context of uncertainty, in particular assessing our confidence in changes over time and differences across space.

Inference: model parameters



red = colonial estimates

black = modern estimates

Key Aspects of Hierarchical Approach

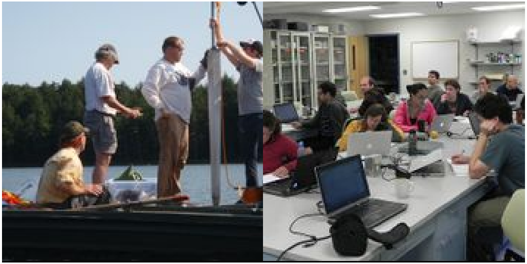
- *Sparsity and irregularity in space and time:* Borrow strength and smooth in space & time
- *Certain proxies are only available in certain regions*
    - *Many records of limited duration*
- *Lack of replication:* Smoothing in space accounts for a form of replication
- *Proxies are not direct measurements of the quantities we care about:* Calibrate to direct measurements
- *Calibration data are scarce*
- *Calibration against modern data may be less relevant for periods in the past (the no analog problem)*
- *Many of the quantities of interest do not have paleodata proxies*
- *Dating is uncertain and dating methods are expensive:* Include dating uncertainty in statistical model, e.g., the BACON model (Blaauw & Christen 2011)

Open Issues

- Data sparsity
- Data dropout tends to have large effects – e.g., losing an observation far from other observations can cause large changes in predictions
- To what extent can we interpret parameter estimates as physically meaningful?
- Computation can be difficult

PalEON: A PaleoEcological Observatory Network

- Multi-institution collaboration of paleoecologists, statisticians, and ecosystem modelers



- Overarching goal: Use paleodata to help understand global change
- Current focus on northeastern/midwestern US over the past 3000 years

Motivation for PalEON

- Paleoecological data have not been used extensively in considering global change, even though they are the only data on long-term changes
- Proxies are often not directly related to quantities of interest for global change and are not in a form directly useful for quantitative analysis
- Terrestrial ecosystem models and paleodata are at different spatial and temporal scales

PalEON Goals

Develop networks of paleodata, synthesized statistically, to inform ecosystem models:

- Assess models against paleodata
- Initialize models based on paleodata
- Assimilate paleodata into models
- Improve model formulations
- Prioritize new data collection