Calibrating the pollen-vegetation relationship

Andria Dawson

October 23, 2014

1 Introduction

Understanding forest ecosystems of the past can provide us with valuable information about how ecosystems respond to biotic and abiotic factors. In particular, mapping forests back through time offers new information not only about the climate-forest relationship, but also forest-atmosphere interactions. In order to quantify forest ecosystem change through time, we need spatio-temporal data. There is no forest data that extends back through the last several millenia, but there is a wealth of paleo-data, including fossil pollen data, that serves as a proxy for surrounding vegetation. To make use of this data to estimate past forest composition relies on our ability to quantify the pollen-vegetation relationship.

The complexity of the pollen-vegetation relationship has made it a long-studied question in the paleoecological literature. Different taxa produce different amounts of pollen, and XXX Describe some of the research here....

Sampling to quantify pollen-vegetation relationships usually involves collecting surface sediment samples to obtain pollen counts at the time of sampling, as well as a survey of vegetation composition and abundance of the surrounding forests. Inference can then be made about processes that affect pollen production, dispersal, and deposition, but only for the time of sampling. Due to widespread land-use change, it is likely that these relationships have not remained static back through time.

Here we use the awesome PLS data set...

To calibration the pollen-veg relationship against the PLS forest composition data set requires that we identify pollen samples that date back to pre-settlement. Raw fossil pollen data records depths and counts. Usually there are additional radiocarbon dates for some number of macrofossils scattered along the cores, not necessarily aligned with sampled depths. These radiocarbon dates constrain age estimates, and although radiocarbon dating facilities do assign dating errors, it is important to keep in mind that macrofossil dates may not be exactly aligned with dates of sediment pollen from the same depth **?**. To infer age as a function of depth requires an age-depth model, although there is no consensus on the most appropriate model. Recently, the community has recognized that importance of estimating uncertainty, and a Bayesian age-depth model coined Bacon developed by **?**which does just that has gained momentum.

In addition to radiocarbon dates, other constraining geological markers can often be identified from looking at stratigraphic plots of pollen proportions.

In the case where samples are taken at multiple sites, pollen counts are rarely modelled in a spatial context. Here we use a spatial Bayesian hierarchical model of pollen counts at a network of sites developed by **?**.

2 Data

2.1 Spatial domain

Our study area is the upper Midwestern US, and includes Minnesota, Wisconsin, and the upper peninsula of Michigan. The lower peninsula of Michigan was not included because: 1) it is spatially disjoint from the rest of the domain; and 2) the public land survey forest data is still in the process of being digitized, so the composition data available is incomplete.

2.2 Tree taxa

We focus on a subset of taxa that are of particular interest, inlcuding the most abundant taxa and any taxa that are of specific ecological importance. Our modelled taxa includes: Ash, Beech, Birch, Elm, Hemlock, Maple, Oak, Pine, Spruce, Larch, as well as Other Conifer and Other Hardwood groups which include those respective tree types not explicitly included in the aforementioned list of taxa. This separation of other hardwood and conifers as opposed to having a single other group was motivated by ecological modellers who are often interested in grouping taxa as deciduous, conifer, and deciduous conifers. Additionally, the separation of other hardwood and conifers allows the model to treat each group seperately and tease out inherent differences between conifer and deciduous seed production, although we recognize that the variability in production and dispersal within each of these groups is still large.

2.3 Public Land Survey (PLS) data

Prior to major European settlement, the US General Land Office conducted a Public Land Survey (PLS) throughout much of the United States in order to simplify the sale of federal lands. Surveyors documented section location using trees as landmarks, and recorded genus or species, diameter, and location (azimuth and distance from corner). This data set provides a systematic survey of the forest before settlement, and has been used by foresters, ecologists, and historians to understand ecosystem and land-use change through time. In the Upper Midwest, the survey was conducted during XXXX-XXXX. Due to the slow-growing nature of temperate forests, including those in the Upper Midwest, we can think of the PLS data as a snapshot of forest composition in time.

Survey data for the Upper Midwest has been recently digitized, and aggregated to an 8km square grid **?**. However, due to its sparse nature, sampling methodology, and surveyor bias the PLS data set contains inherent variability. Here we work with a smoothed version of the PLS data, based on a Bayesian spatial multinomial model **?**.

2.4 Pollen data

With the push for robust and reproducible research from the scientific community, paleoecoinformatics has responded with the development of tools that make accessing and using large datasets possible. One such tool that has made this work possible is the Neotoma database (`neotomadb.org`; **?**), which stores a variety of types of paleoecological data, including pollen data. Accessing this data can be done using the Neotoma API (), or using the R neotoma package **?**. Using these tools granted us access to 176 fossil pollen cores falling within our domain.

In addition to the data obtained from Neotoma, we also had access to a data set belonging to Calcote, Hotckiss, XXX. This data set included 57 cores in our domain. Of these 57, 9 were long cores (analogous to those from neotoma), while the remaining 48 had only core top and pre-settlement samples (at least in the data file we had access to).

Associated with each of the cores is a table containing counts by taxon for a series of depths. For each pollen core we are interested in the pre-settlement sample that is closest in time to the PLS data. Typically, age-depth models are used to assign ages to sample depths. However, there are many different types of age-depth models, each with its own set of benefits and shortcomings. Instead, we rely on a panel of experts to interpret patterns in the pollen count data to identify pre-settlement sample estimates. All long cores were suitable for this exercise.

2.5 Expert elicitation of pre-settlement depth

Widespread land clearance that occurred during European settlement provided habitat for certain non-arboreal colonizers, resulting in increases in non-arboreal pollen in the sediment. In the Upper Midwest, significant increases in Ambrosia, Rumex, or Poaceae are typically coincident with this settlement horizon. When these increases can be identified based on pollen count data, we can identify this settlement horizon, and in particular what we call the pre-settlement sample - the sample that falls immediately before these increases in agricultural indicator species. In practice, identifying increases in agricultural indicators is often difficult, when possible, and can be subjective.

For this study, in the interest of reducing uncertainty, we want to: 1) identify pre-settlement samples using consistent methodology, and 2) assess the variability in assignment of pre-settlement among analysts. To address these questions, we asked a team of expert palynologists to identify the pre-settlement sample for 185 pollen records (176 from Neotoma and 9 from the Calcote data set). Experts were provided with pollen diagrams depicting proportional changes through time as a function of depth for key indicator species and the ten most abundant arboreal taxa, and were prohibited from relying on stratigraphic dates (radiocarbon or other) or age-depth model estimates of sample age. In the case that there was no distinguishable pre-settlement sample, experts were instructed to report NA. In the case that experts were uncertain about their pre-settlement sample assignment, they were instructed to note this, with or without justification.

Results from this exercise will define depths associated with pre-settlement samples. Pollen counts associated with these depths will then become part of our calibration data set.

3 Calibration model

Here we describe the Bayesian hierarchical calibration model used to quantify the pollen-vegetation relationship.

We treat space as a regular grid composed of 8 km square grid cells, which is the resolution defined by the PLS data. The grid is composed of discrete cells, but the underlying vegetation composition and dispersal spatial processes are assumed to be smooth. Spatial cells are indexed by $s = 1, \ldots, S$, where $S = 8013$.

3.1 Model description

Trees are sources of pollen - they produce and distribute pollen across the landscape. Here, the gridded PLS data provides us with a representation of how these trees are distributed throughout the Upper Midwest domain. We can then think of grid cells as being producers of pollen, and the amount of pollen produced by a given grid cell depends on the compositional makeup of that cell (among other things).

Pollen produced by vegetation within each grid cell can be deposited locally within that same grid cell, or can be dispersed into the neighborhood around that grid cell. cells. For a focal grid cell s_i, the pollen produced by taxon p within that cell that remains local is described by

$$\gamma \phi_p r_p(s_i) \tag{1}$$

where γ is the proportion of pollen produced in s_i that is deposited locally, ϕ_p is the scaling factor that accounts for differential production, and $r_p(s_i)$ is the proportional abundance of vegetation in s_i.

The remaining proportion $(1 - \gamma)$ of pollen produced in s_i is dispersed to other grid cells according to an isotropic dispersal kernel centered at s_i. The dispersal kernel weights all pollen dispersing away from the focal cell as a function of the distance from s_i to any neighboring cell s_k by $w(s_i, s_k)$. Here $w(s_i, s_k)$, is defined to be an unnormalized gaussian dispersal kernel written as

$$w(s_i, s_k) = \exp\left(-\frac{d(s_i, s_k)^2}{\psi^2}\right), \tag{2}$$

where $d(s_i, s_k)$ defines the distance between cells s_i and s_k and ψ is a parameter that describes the spread of the kernel. As expected, the weight assigned by this kernel is a decreasing function of distance - less pollen is distributed farther away.

We can then define the pollen produced by taxon p dispersing from s_i to s_k by

$$\frac{1}{C}(1 - \gamma)\phi_p r_p(s_i) w(s_i, s_k), \tag{3}$$

where C is a normalizing constant equal to the sum of the weights of all the cells to which pollen can be dispersed, defined be a rectangular region that covers and extends beyond the limits of the domain.

So far we have described the model from a source-based perspective, describing how pollen produced in grid cells is dispersed. What we really want is to model the pollen arriving at the grid cell in which a pond lies $s(i)$ - we have counts of pollen that arrived at lakes. To do this, we simply sum all the contributions that have been dispersed to $s(i)$, which includes both the locally deposited pollen plus the pollen dispersed to $s(i)$ from all other grid cells in the domain. Therefore the pollen from taxon p arriving at $s(i)$ is given by

$$\gamma\phi_p r(s(i)) + \frac{1}{C}(1-\gamma)\phi \sum_{s_k \neq s(i)} r(s_k) w(s(i), s_k). \tag{4}$$

Finally, pollen counts at pond i, denoted by $\boldsymbol{y}_i$, are modelled using the dirichlet-multinomial to account overdispersion resulting from the placement of lakes within grid cells. The pollen counts observed at a lake are not equivalent to the averaged counts that we predict for a grid cell. We have

$$\boldsymbol{y}_i \sim DM(n_i, \alpha_i) \tag{5}$$

where the precision parameter α_i is equal to the sum of 4 over all taxa

$$\alpha_i = \sum_{p=1}^{K} \gamma\phi_p r(s(i)) + \frac{1}{C}(1-\gamma)\phi \sum_{s_k \neq s(i)} r(s_k) w(s(i), s_k). \tag{6}$$

This precision parameter α is also affected by how close a lake is to the domain boundary. Since we do not have data outside of our domain, for any lake that falls close to a boundary we will miss including some of the pollen contributions from the vegetation growing outside the domain. The repercussion of this is that the local pollen is up-weighted because the sum of the weights of the grid cells contributing to a grid cell near the boundary is much less than the sum of the weights of the total potential neighborhood defined by C (see 6.

3.2 Numerical implementation

Due to the non-conjugate nature of the multivariate likelihood terms, we are relegated to using MCMC methods. With the goal of achieving more efficient sampling with respect to the effective sample size per unit time, we used the Stan statistical modeling software to estimate parameters. Stan implements a variant of the Hamiltonian Monte Carlo method called the No-U-Turn Sampler, which is a gradient based sampling method that uses these directional derivatives to make informed decisions about how to move along (and sample from) the joint posterior surface.

4 Results

4.1 Expert elicitation of pre-settlement depth and site suitability

Four experts participated in the elicitation exercise. For 59 out of 185 sites, the experts were in total agreement: they all identified the same pre-settlement sample (55 cases) or

were in agreement that no such sample could be identifed (4 cases). The remaining sites varied in level of disagreement - in 79 cases, experts identified two pre-settlement samples; in 38 cases there were 3 pre-settlement samples identified; and in 4 cases there was no agreement. Without further analysis, these results confirm our assumption that identification of biostratigraphic markers is subject to variability between analysts.

Of the 4 sites that experts agreed had no discernible settlement signal, two of these sites were at Rice Lake, which is an anomalous site overwhelmed by a Zizania and Poaceae signal, and does not show any indication of settlement.

Based on these results, we were subsequently faced with the challenge of establishing site suitability criteria to determine which sites to include in the calibration data set. Ideally, the now quantified pre-settlement sample uncertainty would be included into the modelling framework. However, this elicitation exercise varies from those that are typically conducted to construct priors; here we are uncertain about our data, not our parameters. Instead of increasing the complexity of the statistical model, we filter suitable sites and assign pre-settlement depths in a systematic way based on the expert determinations.

Sites were considered unsuitable if: 1) the majority chose not to assign a pre-settlement sample, or 2) if half of the experts chose not to assign a pre-settlement sample and the remaining half identified pre-settlement samples whose modelled ages were far from the approximate time of settlement. There were 11 sites for which a majority of experts (3 or 4) did not assign a pre-settlement sample, indicating no consensus regarding the existence of a land-clearance signal. As such, these sites were deemed unsuitable for calibration. For our second criterion, we looked to existing age models associated with our pollen cores. Modelled ages for the identified pre-settlement depths were compared to 1850, which serves as an approximate year of settlement in the upper midwest. Uncertainty associated with age estimates from age-depth models can be large, but there is also a tendency for us identify patterns when in fact there are none **?**. This criterion is a crude way to try to account for the phenomenon. There were seven sites for which half of the experts did not assign a pre-settlement sample and the other half identified depths whose estimated ages were more than 500 years away from 1850. All unsuitable sites were excluded from further analysis.

After completion of the elicitation exercise, a more thorough examination of the stratigraphic data used to calibrate age models, we noticed that several cores had core tops with dates much older than expected. Further investigation revealed that three of the cores included in our analysis were missing core tops. Core tops for both Lake Mary and Green Lake pre-date settlement, and therefore were discarded. The third core for Lake Kotiranta had a core top corresponding to the pre-settlement sample, and was retained for inclusion in the calibration data set. Interestingly, for each of these cores, 1-3 experts identified a pre-settlement sample (although to be fair, they likely made the assumption that the uppermost sample was in fact a surface sample).

The additional 48 short cores in the Calcote data set were not candidates for the elicitation exercise because they had only a core top and a pre-settlement sample recorded. We comtemplated their inclusion in our calibration data set because their pre-settlement samples are identified using different methodology (and analyst) than the remaining sites. We opted to include these samples, based on our confidence in the data set/analyst and the recognition that including these sites would result in a substantive increase in sample size.

After suitability screening, we were left with 165 long cores plus 48 additional short cores,

for a total of 213 calibration.

XXX: How did we choose the depths?

4.2 Exploratory data analysis

To visually assess the relationship between sediment pollen and tree taxa, we compare pie maps which depict the relative proportions of taxa across space (Figure 1). If the patterns were identical, then the relationship between vegetation and pollen would be 1:1, i.e. the pollen and vegetation proportions would be identical to each other at each location. We know this is not the case, but we need to assess how they are different from this 1:1 ideal. In the pie maps the patterns are consistent with each other, although there are some striking differences. First, we see that pine dominates the pollen records for most of the northern half of the domain. Second, we see areas that have higher relative abundances of Hemlock, Tamarack, and Maple. There are more subtle differences in the less abundant taxa. As expected, these pictures confirm that the relationship between sediment pollen and vegetation is complex. Note that the PLS pie map represents an aggregated version of the 8km gridded dataset, and that this coarsened data set is used for exploratory purposes only (pies would not be visible if the original scale were used).

To better assess spatial distributions of pollen versus PLS data, we can plot the data as heat maps by taxon. Differences in extent of these distributions are indicative of successful pollen dispersal. For example, for both Birch and Pine, we see that the distributions of sediment pollen extend well beyond the boundaries of those shown in the PLS data (Figures **??** & **??**).

4.3 Modelling results

Three chains were run for with a warmp-up of 250 iterations, followed by a sampling period of 10,000 iterations. Warm-up iterations were not used in further analysis. For a total of 30,000 iterations, the effective sample size of the joint log posterior was XXX. Trace plots for the joint log posterior show the efficient mixing achieved by the sampler **??**.

Parameter estimates from the calibration model run allow us to quantify the relationship between the sediment pollen and the vegetation on the landscape. In particular, we are interested in learning whether sediment pollen from a network of sites can be used as a proxy for vegetation composition at large spatial scale. An intuitive way to think about the pollen-vegetation relationship is to plot the proportion of sediment pollen against the proportion of vegetation, by taxon. We know this relationship is complicated, and the hope is that the calibration model allows us to better predict sediment pollen than if we tried to predict based on data alone.

In Figure 4, we plot the raw pollen proportions against the vegetation proportions for each grid cell. The relationship between the pollen and vegetation is clearly not 1:1. In particular, we see that some taxa, such as beech, maple, other conifer, and tamarack, are not prolific pollen producers - they can appear in large proportions on the landscape, but do never appear in large proportion in the pollen record. (Note that it could also have been the case that the produced pollen may have been abundant, but travelled elsewhere.) We also point out that some relationships are difficult to identify, for example pine, where it varies

from being sparse to abundant on the local landscape, but this provides little indication about the relative abundance of pine pollen.

Differential pollen production is one of the reasons these relationships are complicated - in the model, this is accounted for by the scaling parameter ϕ. After scaling the vegetation in the pond cell by ϕ, we can compare the raw pollen proportions with these pollen predictions (4). Scaling by ϕ does seem to improve our pollen estimates; for many taxa, the black points have shifted towards the 1:1 line. However, the improvement is minimal, indicatins that something is still missing.

In Figure 5, we again plot raw pollen against the vegetation or predicted pollen by taxon, but here the predicted pollen is based on the local plus non-local predictions obtained from the full calibration model. Now that dispersal has been accounted for, we see that the raw versus predicted pollen points fall more closely along the 1:1 line. This indicates that in this region, dispersal processes account for much of the pollen that we see - scaling the local vegetation to account for differential production was not sufficient to predict deposited pollen. The model is less succesful at predicting Hemlock pollen. In Figure 5. we see that the model underpredicts the pollen at sites where there is a high proportion of Hemlock. This suggests that pollen dispersing to these sites is estimated to make a larger contribution to the sediment pollen than it should for this taxon.

XXX There are several cases where the model does not do a good job

Taxon-specific estimates for the production/dispersal parameter ϕ form two distinct groups in ϕ parameter space: low to intermediate, and high ϕ or production/dispersal (Figure 6). The group of taxa with lowest production/dispersal in decreasing order are Hemlock, Beech, Maple, Other Conifer, and Tamarack. This low production/dispersal pattern is evident in Figure 5, where the representation of these taxa in the pollen records is consistently less than their representation on the landscape (except at a few anomalous ponds). The high production/dispersal group includes pine and birch, which can also be seen in Figure 5 by their propensity to be over-represented in the pollen record relative to the landscape. All other taxa fall into the intermediate production/dispersal region of the spectrum. These results agree in generel with estimates of the pollen-vegetation relationship based on sites in Wisconsin and the Upper Peninsula of Michigan found by Prentice and Webb [1986]. In that work, Pine and Birch had the largest slopes (top producers/dispersers), and maple and tamarack as limited producers/dispersers. Jackson [1990] also found that Pine and Birch were good dispersers with effective source areas of 1000m, while Maple was found to have a much smaller source area indicating limited dispersal.

The local versus non-local weight γ was estimated to have a mean of $0.20\,(0.19, 0.22)$. This indicates that 20% of the pollen produced by vegetation in a focal grid cell is deposited in that grid cell, while the remaining 80% disperses elsewhere in the domain. The dispersal kernel specifies the weights that determines where the pollen goes, and depends on the spread parameter ψ which was estimated to be $210\,(200, 220)$ km. This implies that 50% of the pollen produced in a focal cell is deposited within 140 km of the focal cell, and 90% is deposited within 296 km of the focal cell 7.

To assess the assumption that all taxa share a dispersal kernel with spread controlled by the scalar ψ, we re-ran the model letting ψ vary by taxon. Five of the taxon-specific 95% credible intervals had some overlap with the credible intervals for the single-ψ case, while the remaining taxa had estimated ψ value that were larger (Figure 8). Elm, Tama-

rack, Maple, and Ash had the largest mean value of ψ, although the uncertainty of these estimates was large - for example, for Tamarack the mean value and credible interval of ψ was $0.73(0.10, 1.9)$. When we let ψ vary by taxon, there was little effect on the values of ϕ. The ordering of the taxa remained identical, except for a minor switch in position of Oak with Elm (Figure 9). XXX: I don't get the dispersal pars for elm, tam, maple, ash - why are these so big?

To visualize the dispersal patterns predicted by our model, we can estimate deposited pollen for each grid cell in the domain (previously we were estimating deposited pollen only at lakes for which we had pollen data) (Figure 11). These spatial maps help us understand how the model treats dispersal. We can compare the maps of predicted relative pollen abundance with those depicting the relative tree abundance (Figure **??**), remembering that the pollen predictions have been scaled for differential production. This scaling for differential production is particularly apparent for Tamarack. We know from the data that Tamarack pollen proportions are always lower than the vegetation proportions, and as a result, we expect that the model will predict low amounts of Tamarack pollen, which it does. To see how the model predicts dispersal by taxon, based on parameters estimated from that data, we can compare the vegetation and pollen ranges. The effects of dispersal are especially clear in the Pine and Oak cases, where we see pollen dispersing well beyond the range limits of the vegetation range boundaries.

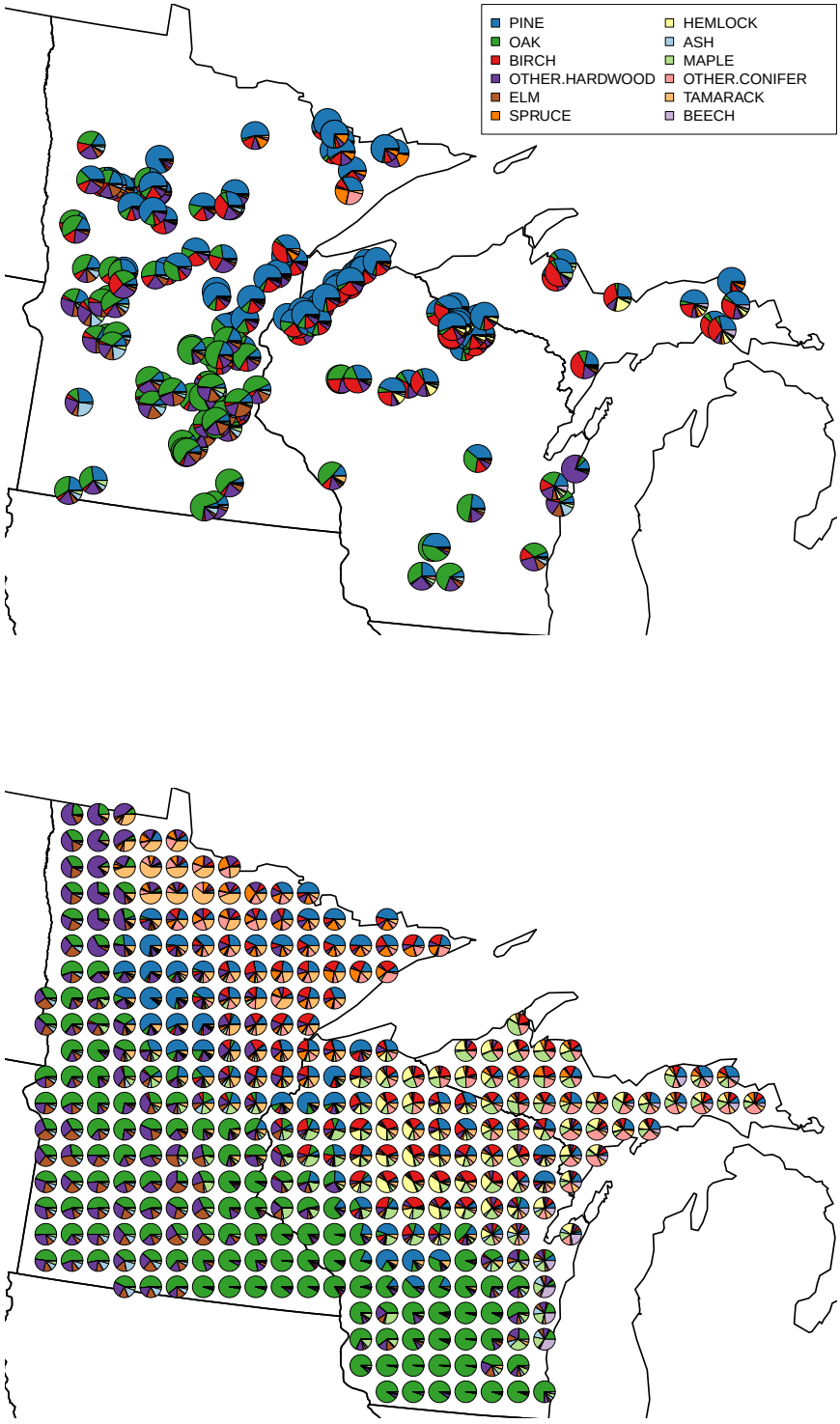

Figure 1: Pie maps depicting the relative composition of pollen (top) and PLS vegetation (bottom) from the data. Note that the PLS data has been aggregated to a coarser resolution for illustrative purposes.

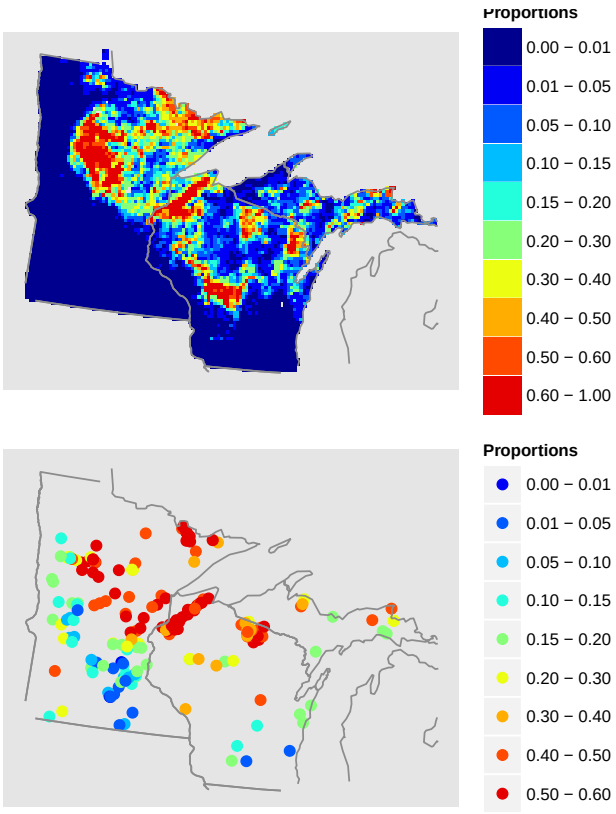

Figure 2: Heat maps showing the range limits of Pine in the PLS composition data (top) and the sediment pollen (bottom).

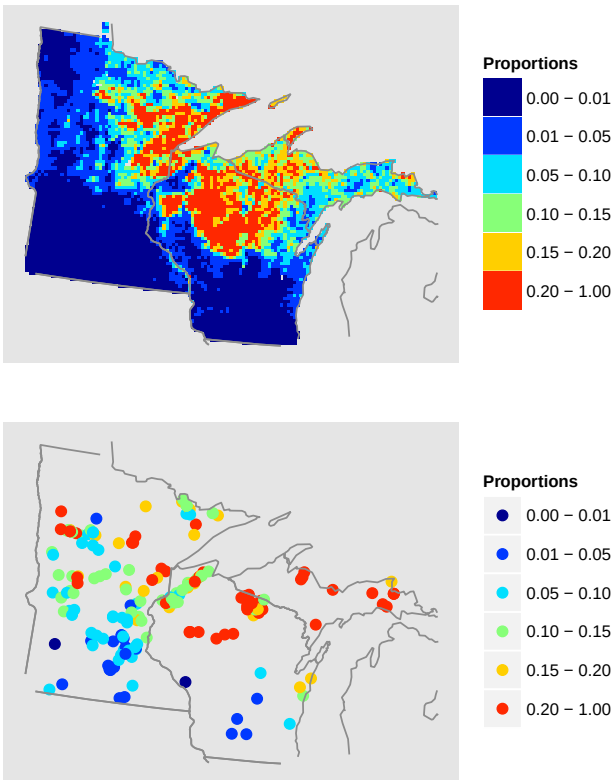

Figure 3: Heat maps showing the range limits of Birch in the PLS composition data (top) and the sediment pollen (bottom)

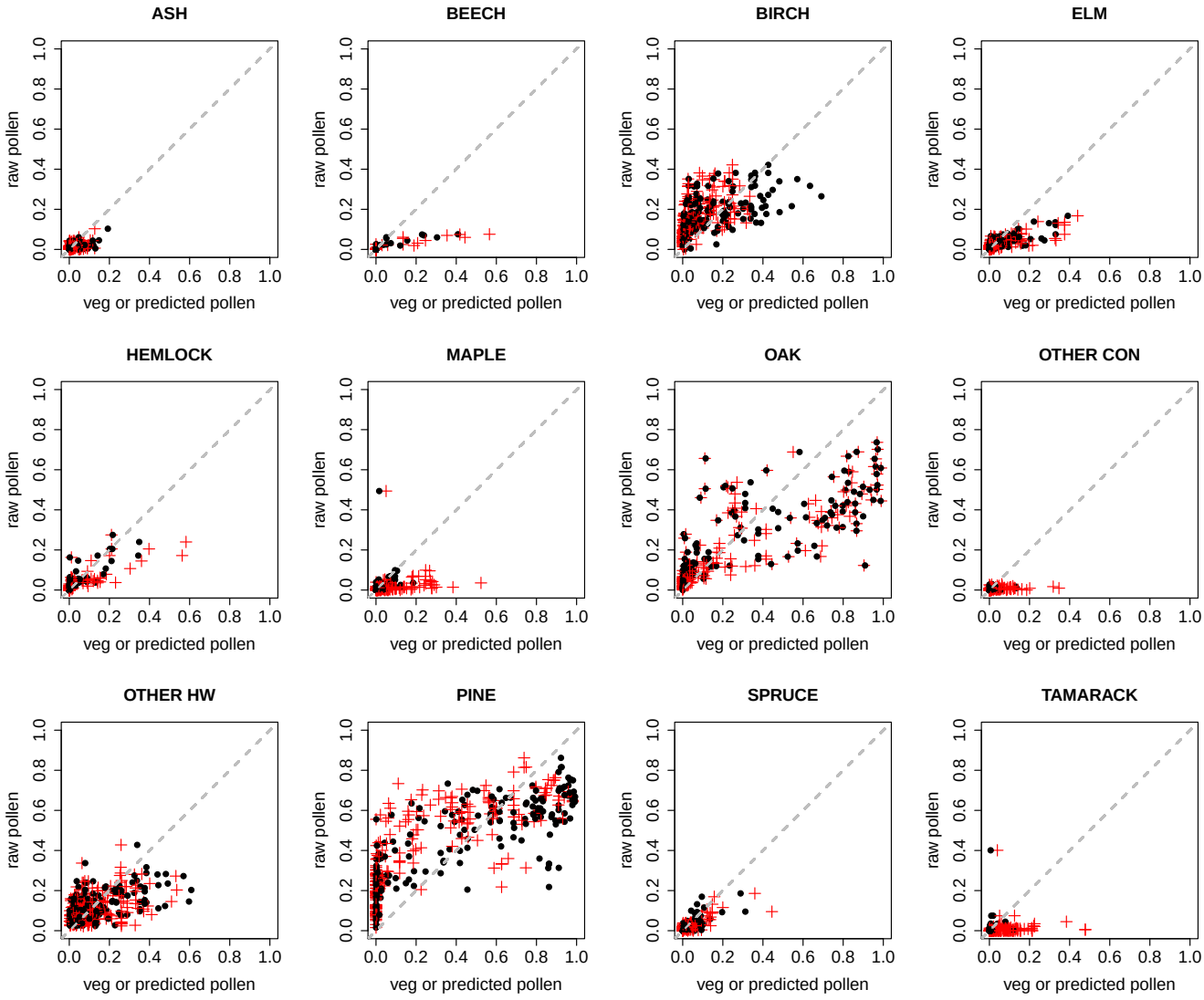

Figure 4: Pollen proportions plotted against local vegetation proportions (red crosses) or local vegetation proportion scaled by ϕ (black dots), by taxon.

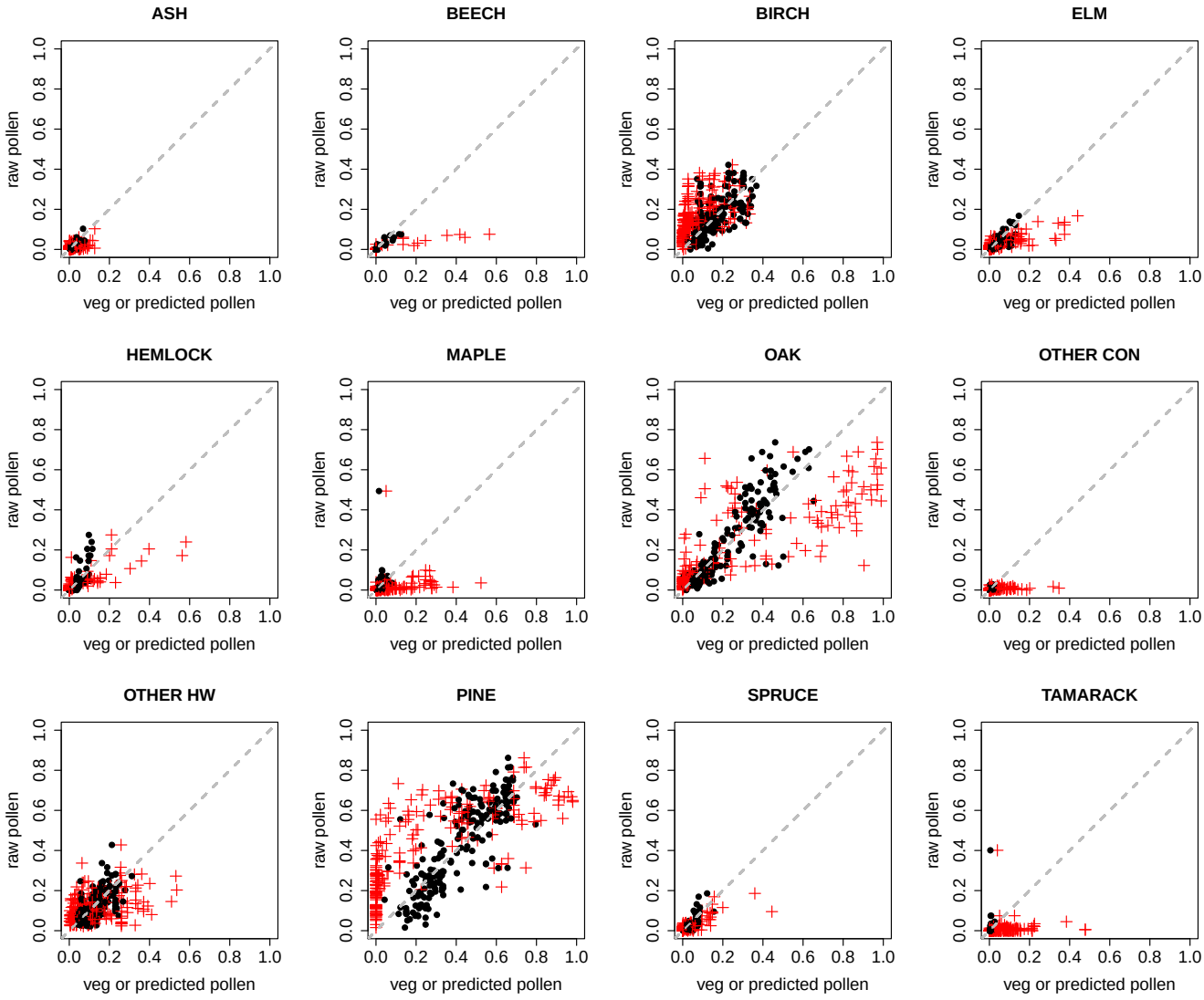

Figure 5: Pollen proportions plotted against local vegetation proportions (red crosses) or model-predicted pollen (black dots), by taxon.

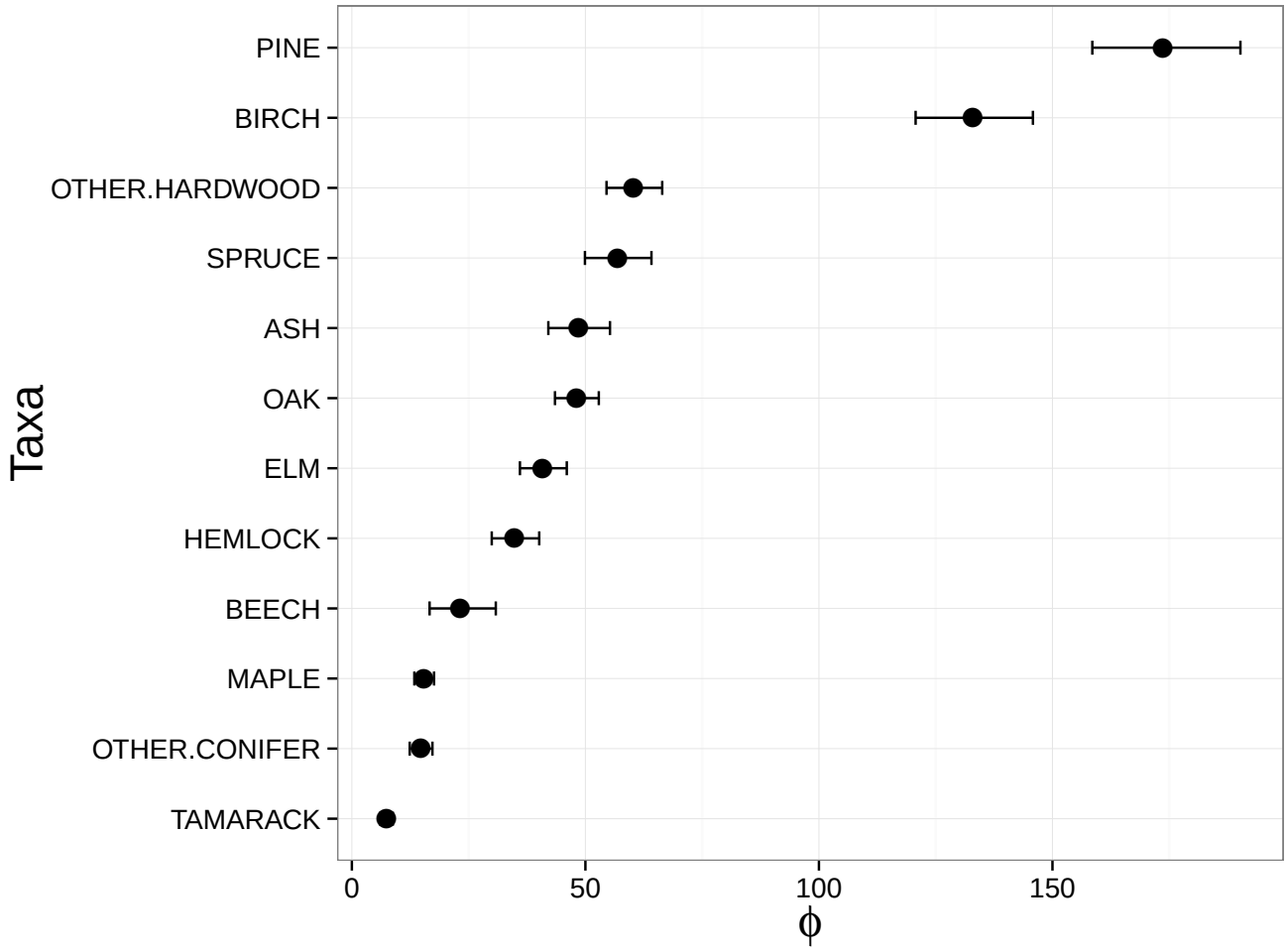

Figure 6: Mean values and 95% credible intervals for the estimated values of the differential production parameter ϕ.

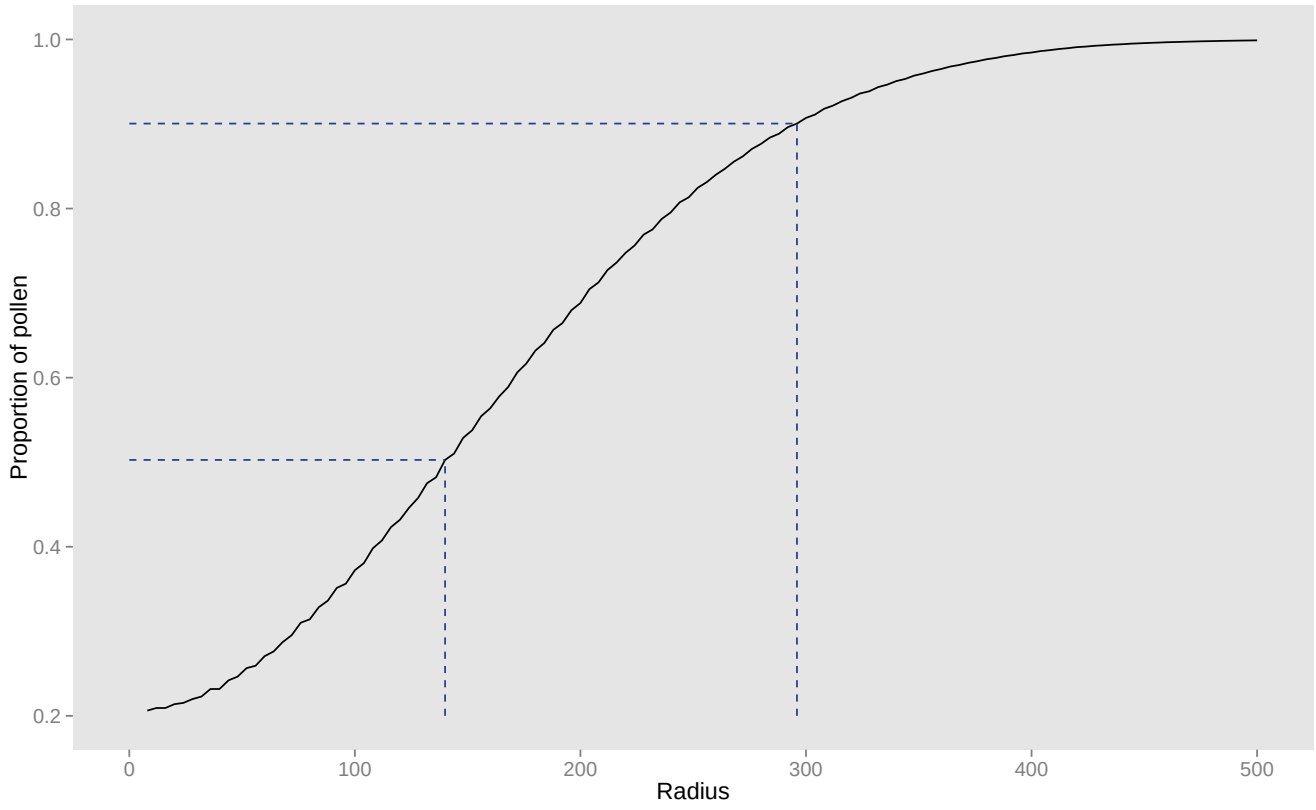

Figure 7: Here we consider pollen produced by a focal cell, and plot the proportion of deposited pollen as a function of the radius of a circle centered at the focal cell.

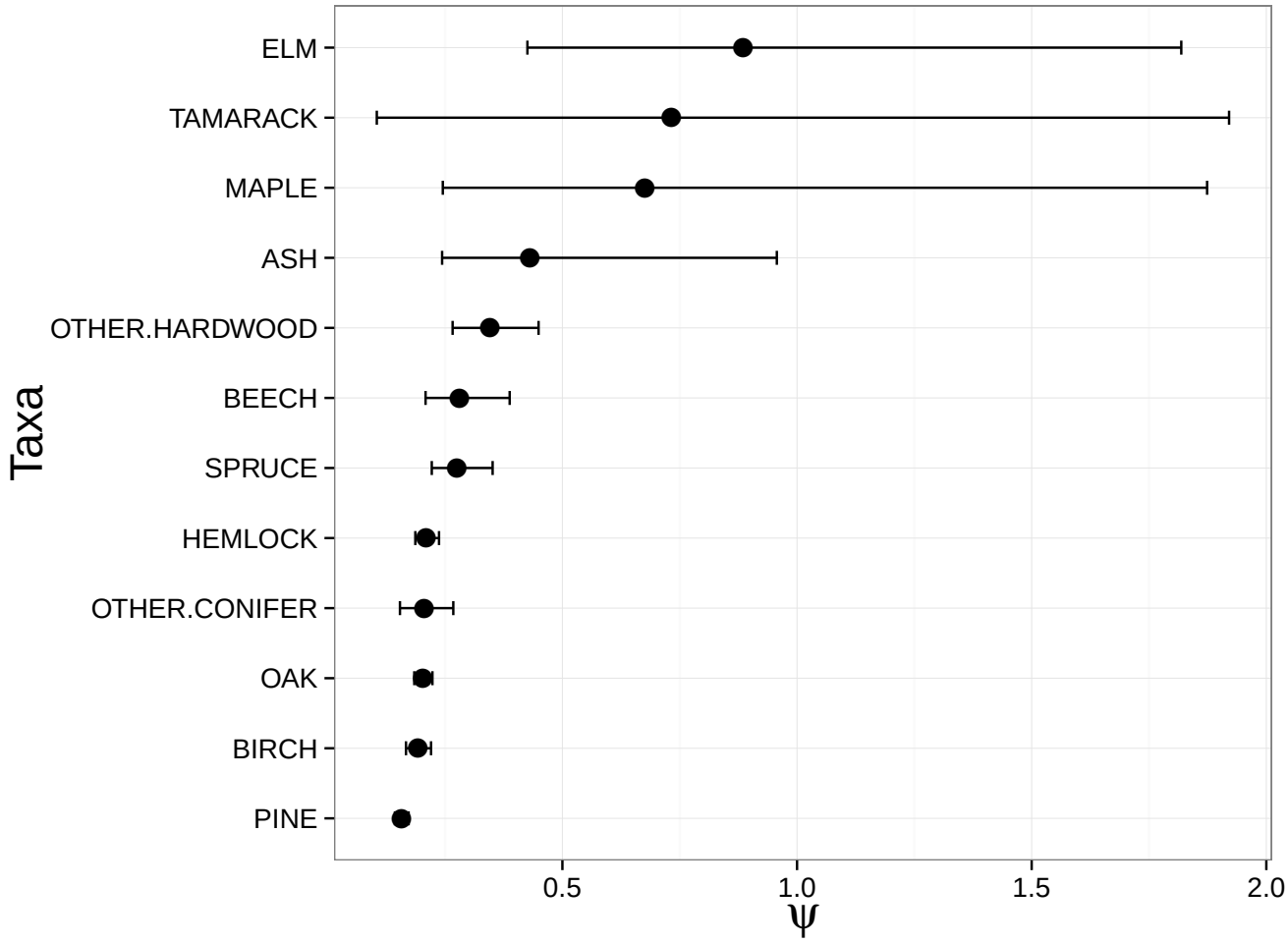

Figure 8: Mean values of 95% credible intervals for the estimated values of the dispersal kernel spread ψ for the case where we let ψ vary by taxon.

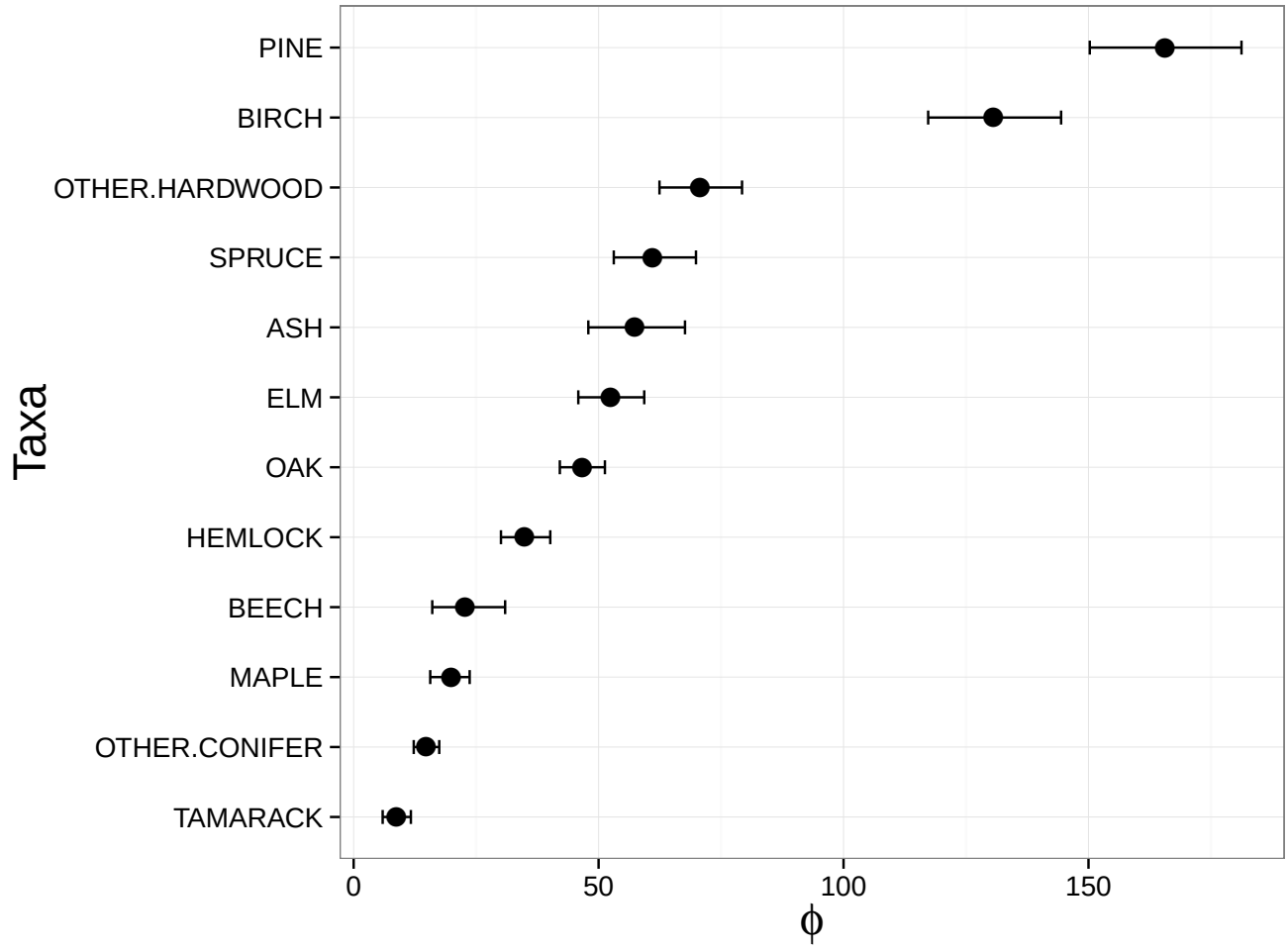

Figure 9: Mean values of 95% credible intervals for the estimated values of the differential production parameter ϕ for the case where ψ varied by taxon.

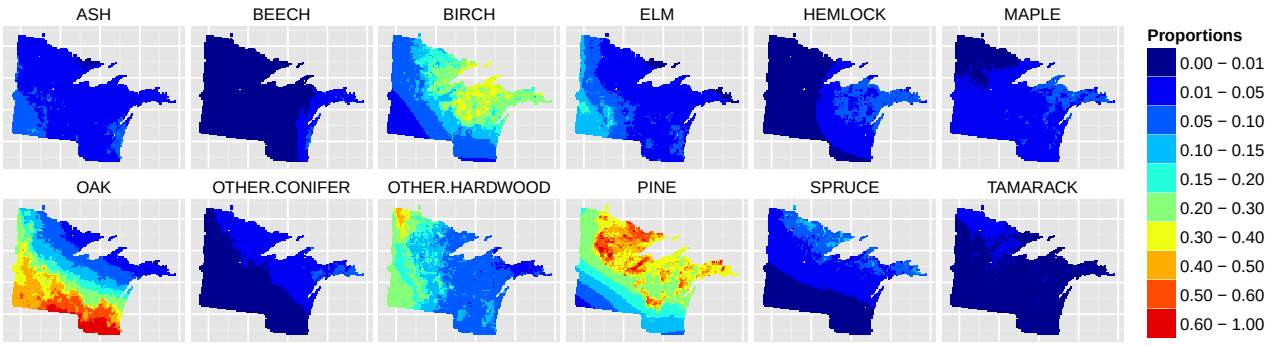

Figure 10: Heat maps of model-predicted pollen for each grid cell in the domain, by taxon.

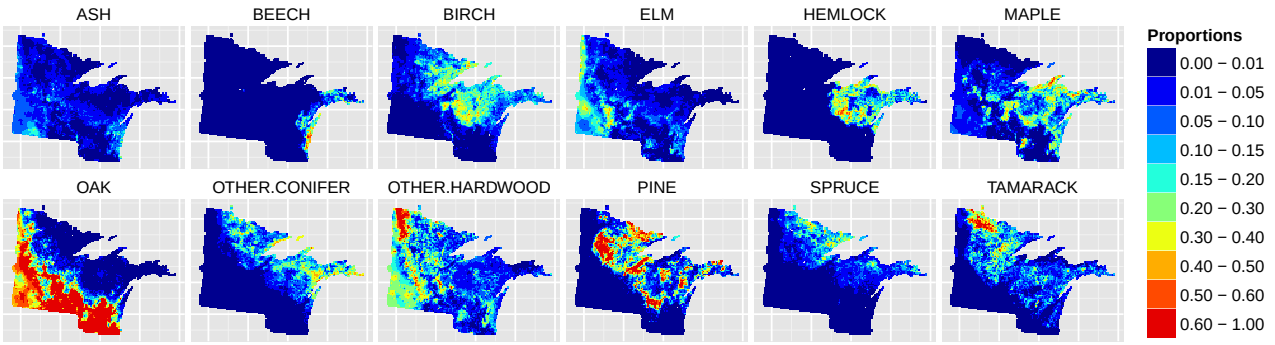

Figure 11: Heat maps of the PLS data, by taxon.

References

Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1351 – 1381, 2011.

Stephen T Jackson. Pollen source area and representation in small lakes of the northeastern united states. *Review of Palaeobotany and Palynology*, 63(1):53–76, 1990.

IC Prentice and TIII Webb. Pollen percentages, tree abundances and the fagerlind effect. *Journal of Quaternary Science*, 1(1):35–43, 1986.

Stan Development Team. Stan: A c++ library for probability and sampling, version 2.4, 2014. URL `http://mc-stan.org/`.