Spatiotemporal sampling

Alyssa Willson

2024-03-27

Purpose

The purpose of this document is to explore options for subsampling the STEPPS relative abundance data product to achieve relative independence in time and space. This is required to make inference about the drivers of relative abundance, especially in the particular case of understanding how the vegetation itself drives and reinforces the relative abundances of each taxon.

Data

The STEPPS data product consists of the relative abundance of 12 common and ecologically important tree taxa. The data product was developed by Andria Dawson et al. using a network of fossil pollen sites and observed fossil pollen-tree abundance relationships in the 1800s. To make a spatiotemporally continuous data product, the model smooths estimates of taxon-level relative abundance over both space and time. This represents our best reconstructions of the relative abundance of different taxa over the Upper Midwest region composed of Minnesota, Wisconsin, and Upper Michigan for the last 2,000 years of the pre-Industrial Holocene.

Proposed analysis

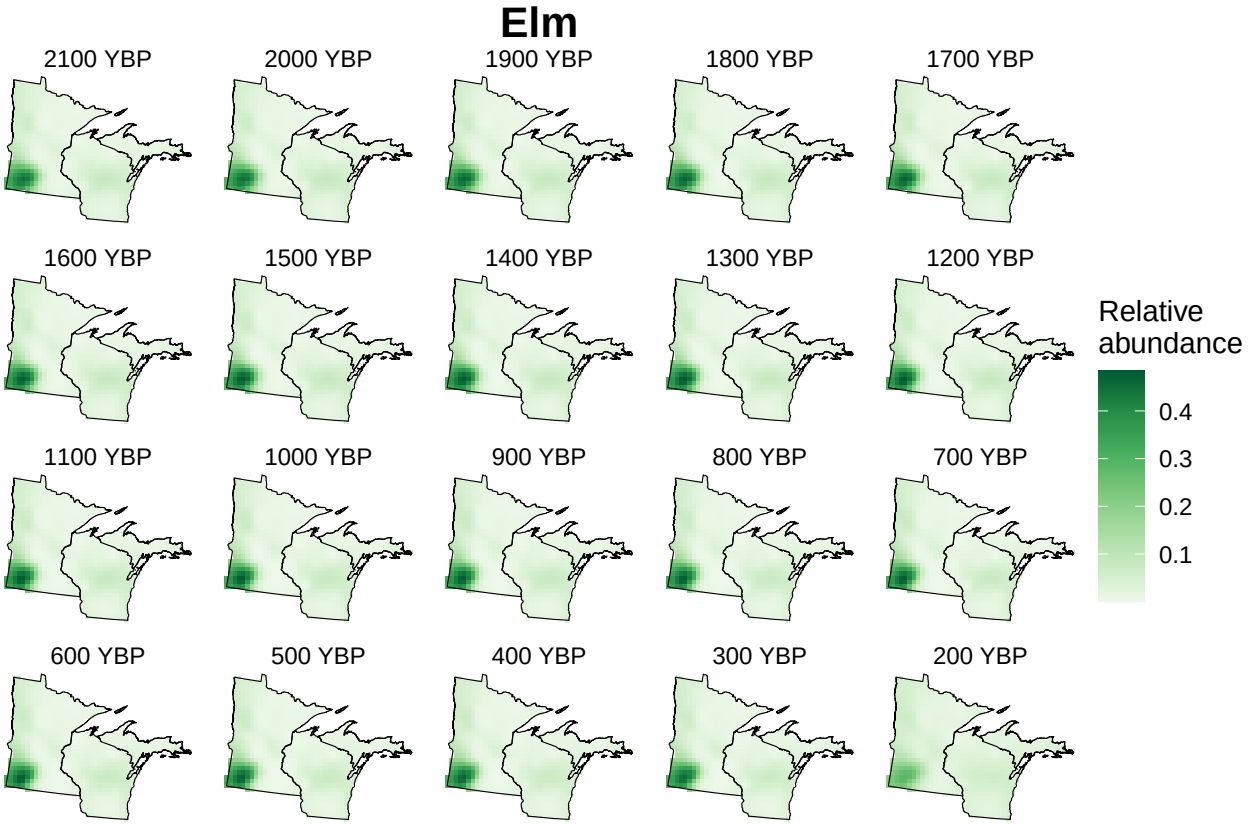
Here, we will use reconstructions of average annual temperature and precipitation, temperature and precipitation seasonality, and soil texture to predict the joint relative abundance of our 12 taxa. We will use GJAM, a species distribution model that allows us to leverage the covariance among response variables (relative abundances of each taxon) as well as covariates (climate and soil reconstructions) in a hierarchical structure to make predictions of relative abundance. The explicit quantification of response variable covariance is important because it provides an estimate of the degree to which taxa resist or favor associating with one another, after accounting for their joint dependence on the hypothesized dominant abiotic drivers of vegetation distributions. It is therefore important that the relative abundances are relatively independent in space and time, because we are interested in spatiotemporal processes of vegetation assembly (climate and soils also vary over space).

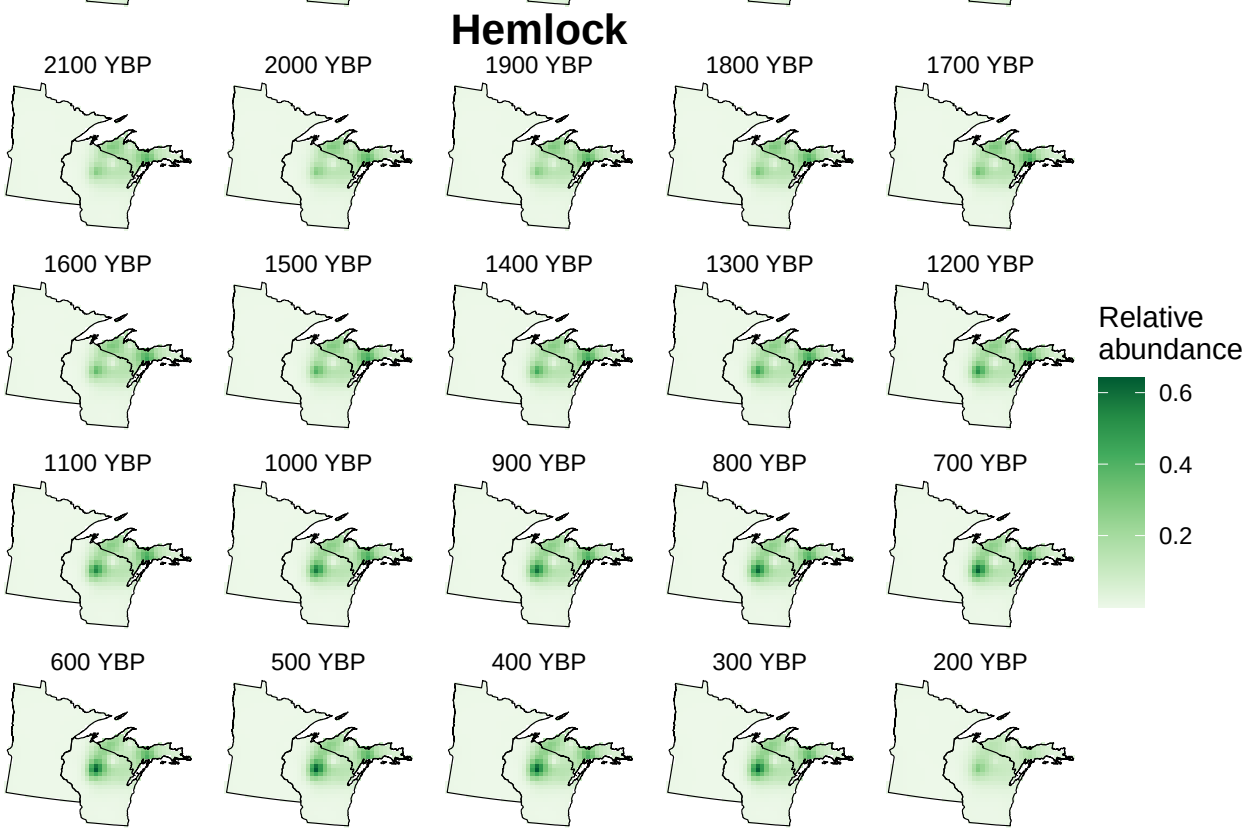
The first step in this analysis is to choose a subsampling scheme that allows us to retain important trends in relative abundance (e.g., higher relative abundance of pine in sand plains, higher relative abundance of hemlock in north central Wisconsin). This document iterates through several options of how to subsample in space and time.

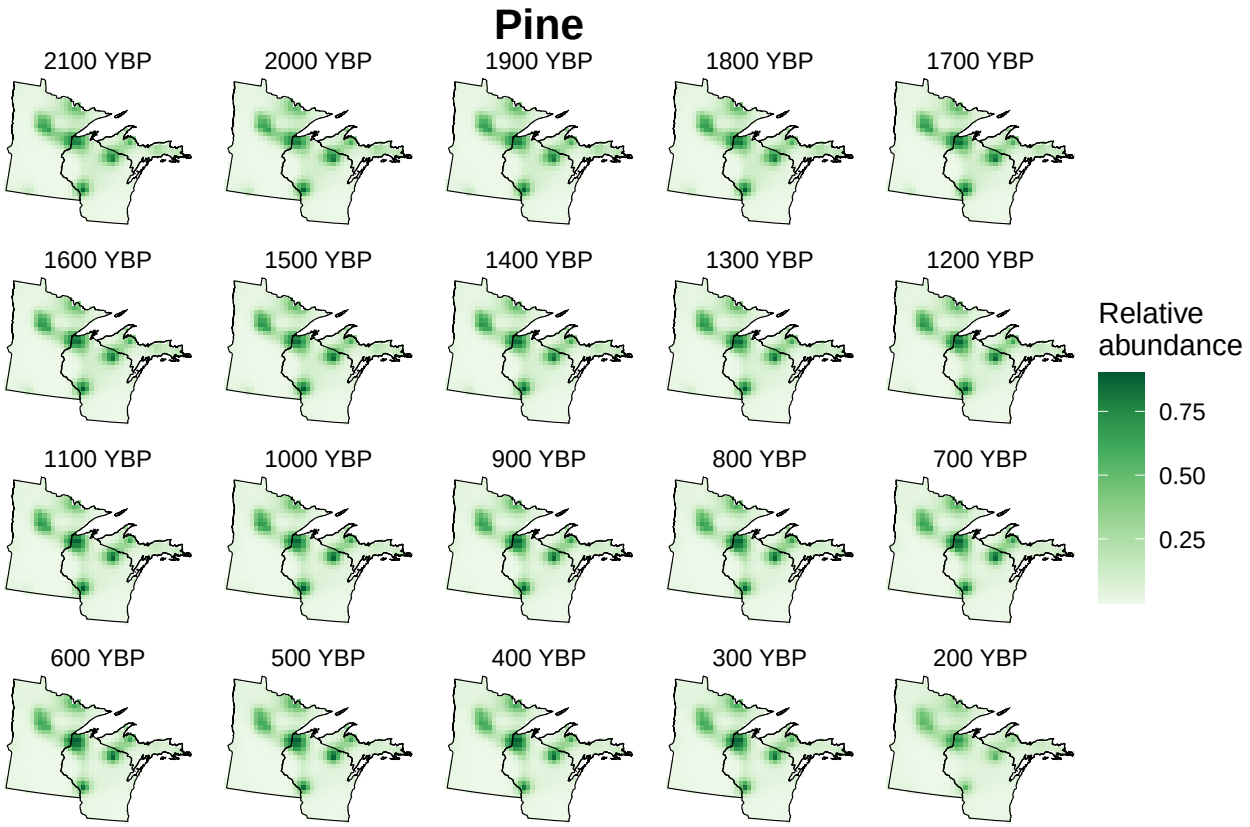
Understanding observed spatial structure

There are some spatial patterns that are very clear in the relative abundances and are likely at least partially an artifact of borrowing strength over space. For example, circles of high to low relative abundance over space are likely the result of high relative abundanace in the pollen record at a given location, with smoothing resulting in a relatively gradual decrease in relative abundance radiating outwards. Below are some examples

of this for elm, hemlock, and pine, which have particularly strong and non-monotonic trends in relative abundance.







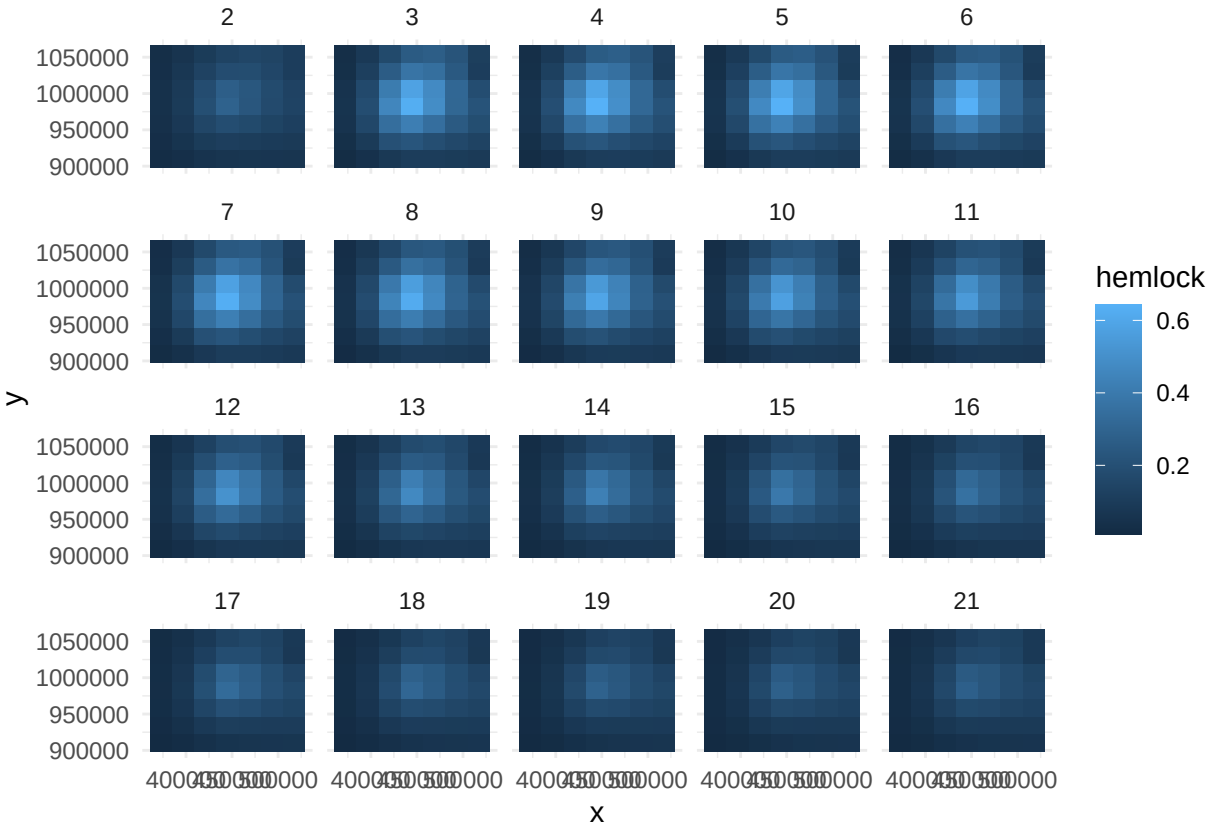
Quantifying spatial dependence

First, I want to see if we can quantify how far from these high abundance areas the decrease in abundance occurs. That is, how many cells over are still estimated to have higher than average relative abundance? This could give us an indication of how many cells we should skip between each cell we keep.

I'll start with hemlock because there is one clear example of a spike in relative abundance in north central Wisconsin. It seems like the trend in hemlock in this area decays on the order of about three cells in any direction.

```
# Subset for one time period
# We can do this because hemlock is always most abundant at this one location,
# the abundance just is greater towards the middle-end of the time period
hemlock_mat <- hemlock[,,1]

# Greatest relative abundance
max_abund <- max(hemlock_mat, na.rm = TRUE)
# Cell corresponding to greatest relative abundance
cell_max <- which(hemlock_mat == max_abund, arr.ind = TRUE)

# Make index of rows and columns +/-3 cells away from the cell
# with greatest relative abundance
row_surround <- (cell_max[1,1] - 3):(cell_max[1,1] + 3)
col_surround <- (cell_max[1,2] - 3):(cell_max[1,2] + 3)

# Look at only this smaller subset of area
sub <- hemlock[row_surround,col_surround,]
```

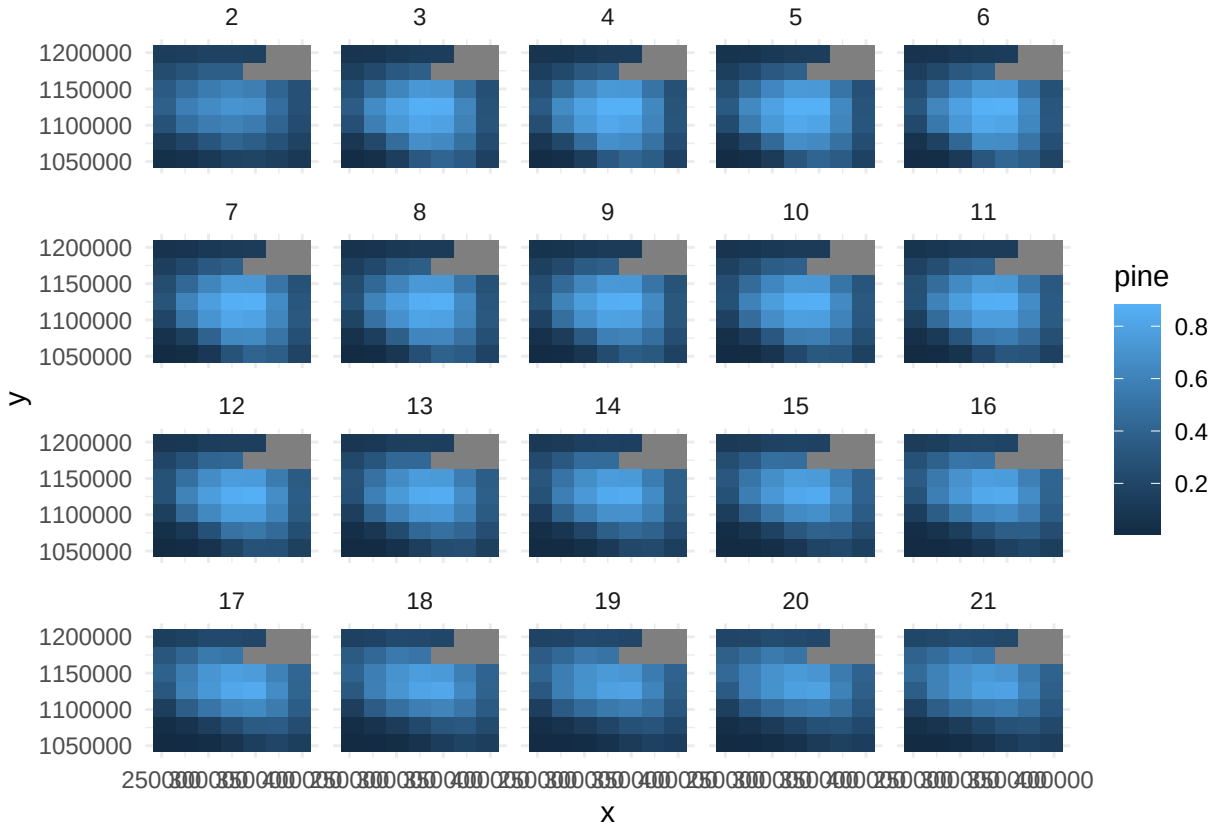
```
# Melt to dataframe
sub_melt <- melt_array(taxon_mat = sub, x = x[row_surround],
                       y = y[col_surround], time = time,
                       col_names = c('x', 'y', 'time', 'hemlock'))

# Plot
# Note that time is running  backward here
sub_melt |>
  ggplot2::ggplot() +
  ggplot2::geom_raster(ggplot2::aes(x = x, y = y, fill = hemlock)) +
  ggplot2::facet_wrap(~time) +
  ggplot2::theme_minimal()
```



Let's see if the spatial dependence for pine is the same because this is another taxon that has very strong peaks in relative abundance and then dissipating relative abundance radiating from those peaks. For this taxon, there are three really big peaks in abundance that I want to look at.

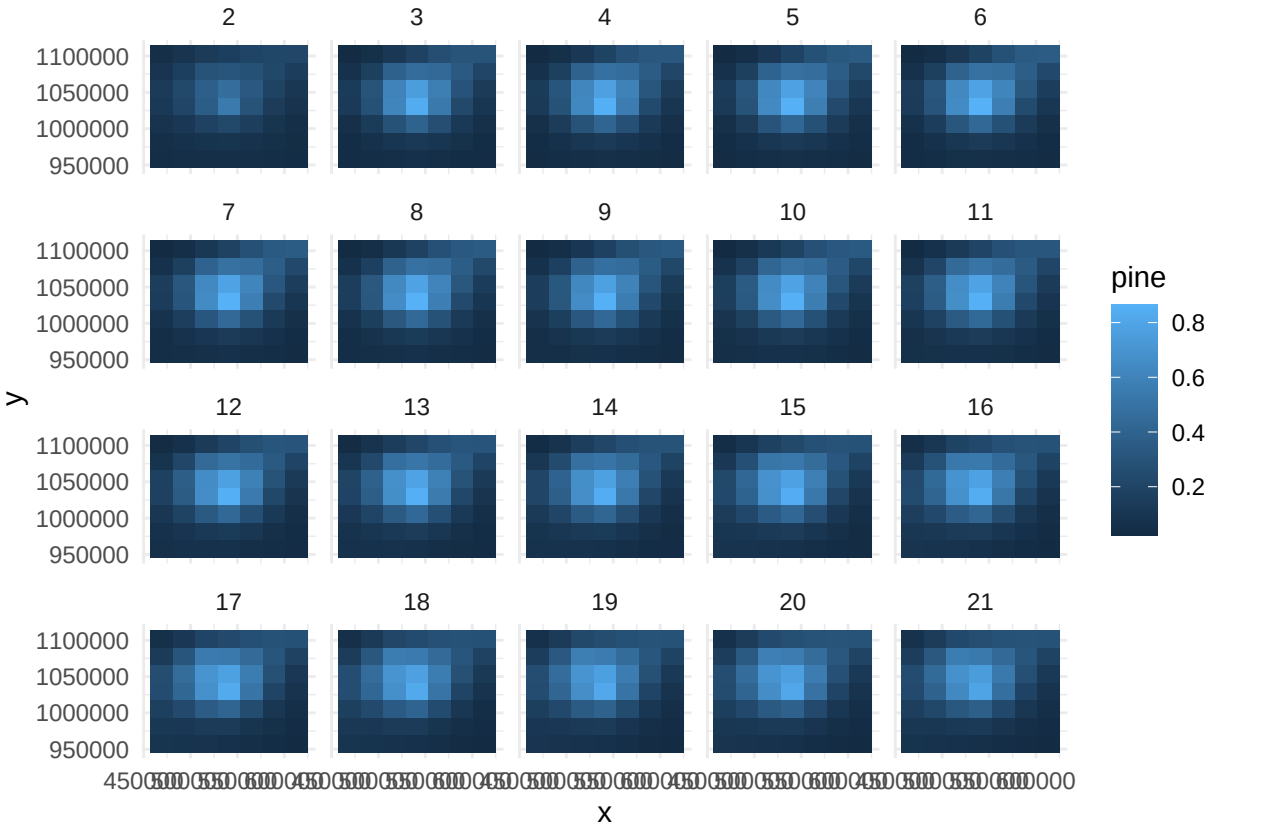
```
# Pine at one time period over space
pine_mat <- pine[,,1]

# Maximum abundance at this one point in time
max_abund <- max(pine_mat, na.rm = TRUE)
# Cell corresponding to maximum abundance
cell_max <- which(pine_mat == max_abund, arr.ind = TRUE)

# Rows surrounding cell with maximum abundance
row_surround1 <- (cell_max[1,1] - 3):(cell_max[1,1] + 3)
col_surround1 <- (cell_max[1,2] - 3):(cell_max[1,2] + 3)
```

```
# Subset for just this area
sub1 <- pine[row_surround1, col_surround1,]

# Melt to dataframe
sub_melt <- melt_array(taxon_mat = sub1, x = x[row_surround1],
                       y = y[col_surround1], time = time,
                       col_names = c('x', 'y', 'time', 'pine'))

# Plot
sub_melt |>
  ggplot2::ggplot() +
  ggplot2::geom_raster(ggplot2::aes(x = x, y = y, fill = pine)) +
  ggplot2::facet_wrap(~time) +
  ggplot2::theme_minimal()
```




```
# Remove the values corresponding to this peak
# so we can look at the next largest abundance
# outside the area we already plotted
pine[row_surround1,col_surround1,] <- NA

# Resubset for one time period
pine_mat <- pine[,,1]

# Find the cell corresponding to the next highest
# relative abundance
max_abund <- max(pine_mat, na.rm = TRUE)
cell_max <- which(pine_mat == max_abund, arr.ind = TRUE)
```

```
# Find surrounding cells
row_surround2 <- (cell_max[1,1] - 3):(cell_max[1,1] + 3)
col_surround2 <- (cell_max[1,2] - 3):(cell_max[1,2] + 3)

# Subset again for this different area
sub2 <- pine[row_surround2,col_surround2,]

# Melt to data frame
sub_melt <- melt_array(taxon_mat = sub2, x = x[row_surround2],
                       y = y[col_surround2], time = time,
                       col_names = c('x', 'y', 'time', 'pine'))

# Plot
sub_melt |>
  ggplot2::ggplot() +
  ggplot2::geom_raster(ggplot2::aes(x = x, y = y, fill = pine)) +
  ggplot2::facet_wrap(~time) +
  ggplot2::theme_minimal()
```



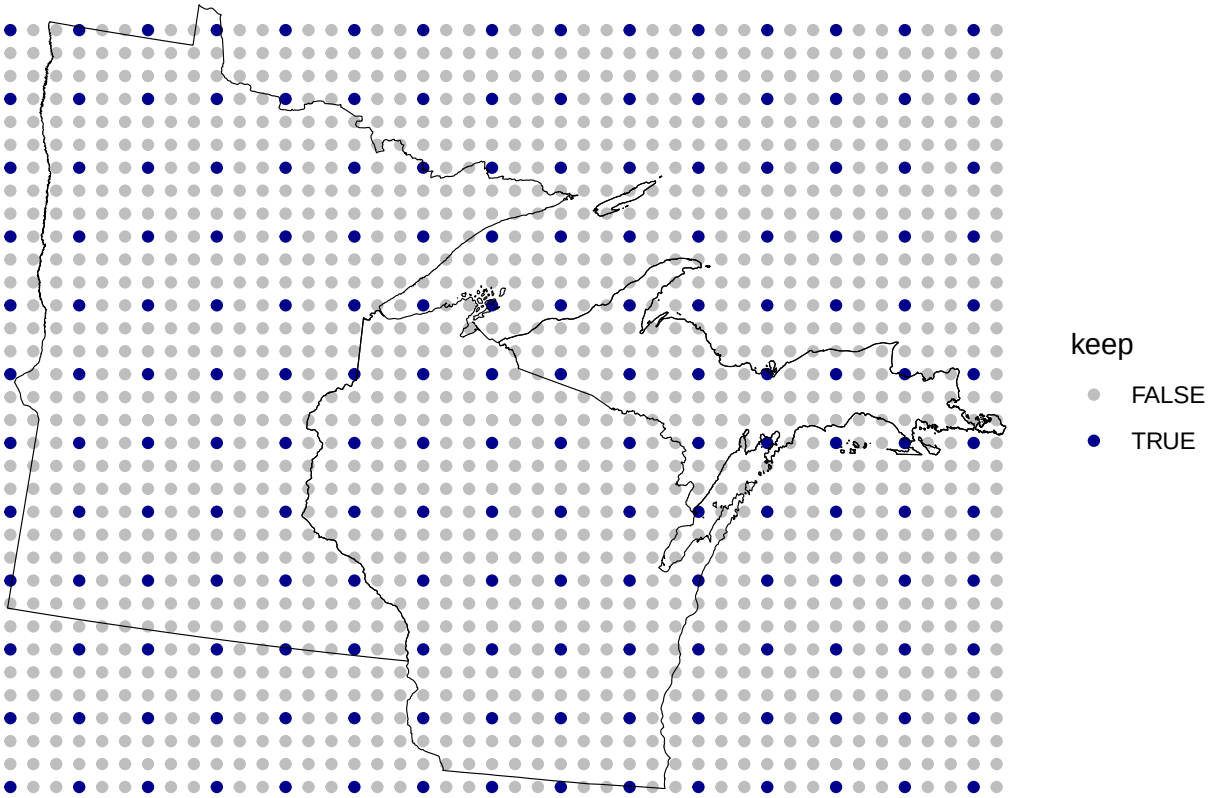
All of these locations seem to indicate that if we sample every third cell, we are missing a lot of the spatial dependence. We can now start trying to sample.

Sampling in space

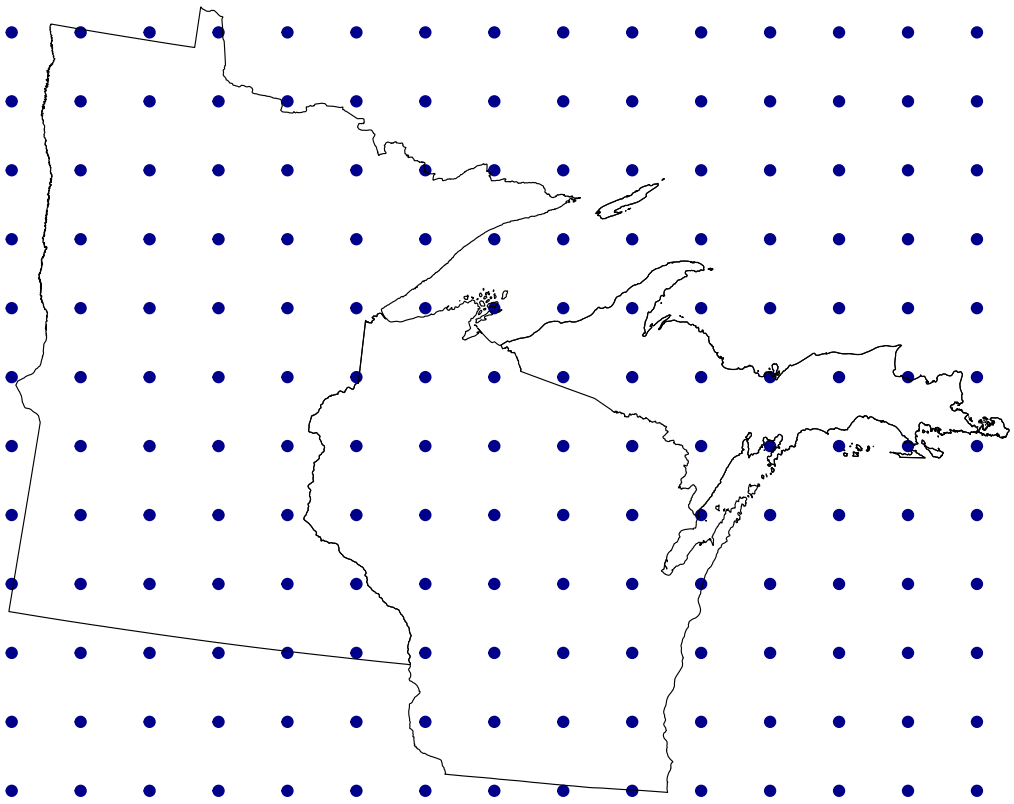
The easiest way I can think of is to create a regular grid sampling from our latitudes and longitudes to get every third. We can then start our samples at different locations to get different variations.

Start: x = 1, y = 1

```
# Take every 3rd x and y coordinate starting at one
x_ind <- seq(from = 1, to = length(x), by = 3)
y_ind <- seq(from = 1, to = length(y), by = 3)

# Empty matrix
locs <- matrix(, nrow = length(x), ncol = length(y))

# Set cells we are keeping to TRUE
locs[x_ind, y_ind] <- TRUE

# Add x and y coordinates as dimension names
dimnames(locs) <- list(x,y)
# Melt to dataframe
locs_melt <- reshape2::melt(locs)
# Add column names
colnames(locs_melt) <- c('x', 'y', 'keep')
# Format
locs_melt <- dplyr::mutate(locs_melt,
                           keep = dplyr::if_else(is.na(keep), FALSE, keep))

# Plot locations we're going to keep
# We want most of the points to fall within our spatial domain and capture
# known patterns of relative abundance
locs_melt |>
  ggplot2::ggplot() +
  ggplot2::geom_point(ggplot2::aes(x = x, y = y, color = keep)) +
  ggplot2::geom_sf(data = states, color = 'black', fill = NA) +
  ggplot2::theme_void() +
  ggplot2::scale_color_manual(values = c('gray', 'darkblue'))
```




```
locs_melt |>
  dplyr::filter(keep == TRUE) |>
  ggplot2::ggplot() +
  ggplot2::geom_point(ggplot2::aes(x = x, y = y), color = 'darkblue') +
  ggplot2::geom_sf(data = states, color = 'black', fill = NA) +
  ggplot2::theme_void()
```

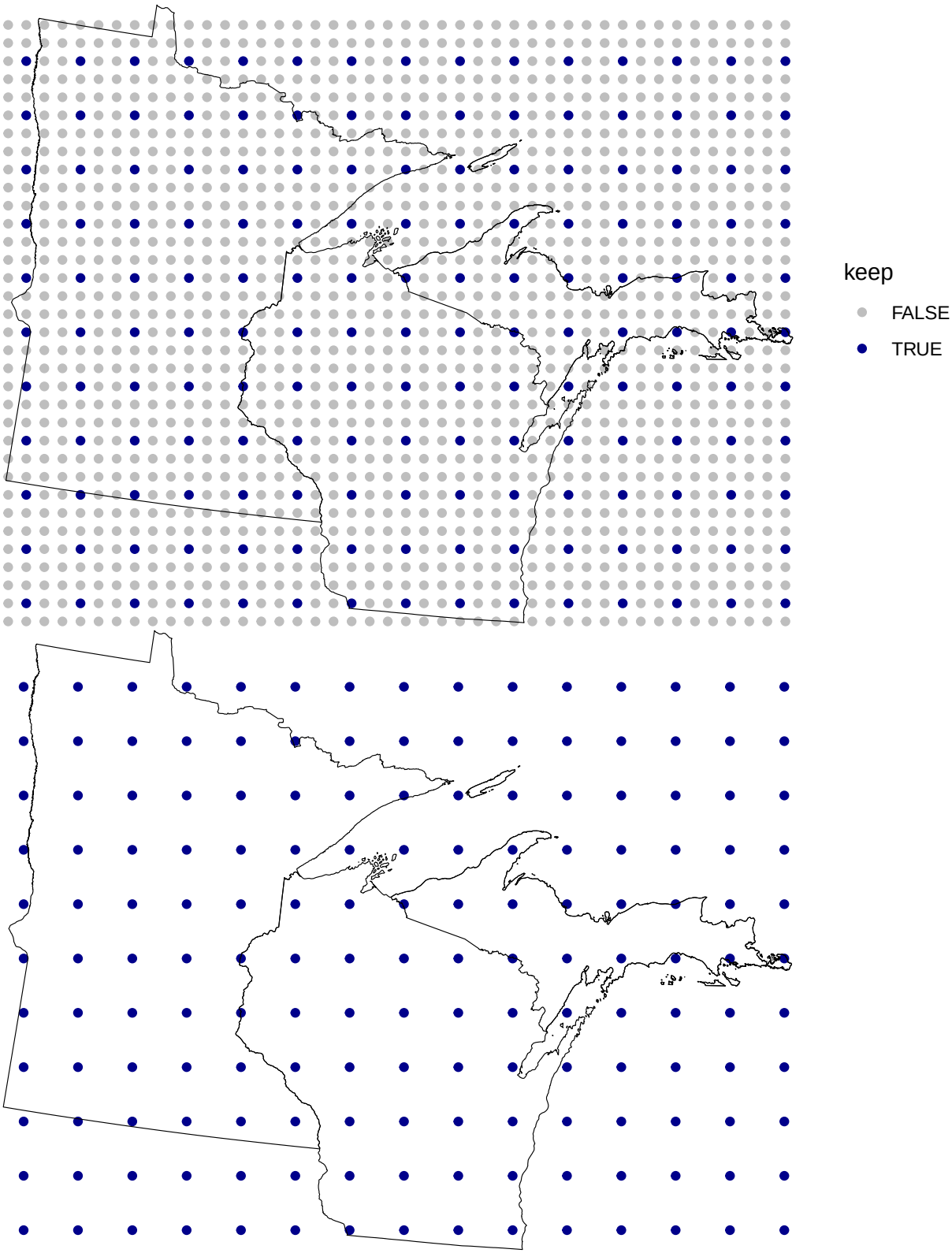


```
# Check how many points that we are keeping fall within our domain
# We would also like to make sure that we keep as many points as possible
test_melt <- dplyr::filter(ash_melt, time == 2)

n_points <- locs_melt |>
  dplyr::filter(keep == TRUE) |>
  dplyr::left_join(y = test_melt, by = c('x', 'y')) |>
  dplyr::filter(!is.na(ash)) |>
  dplyr::summarize(n = dplyr::n())
n_points
```

```
##    n
## 1 73
```

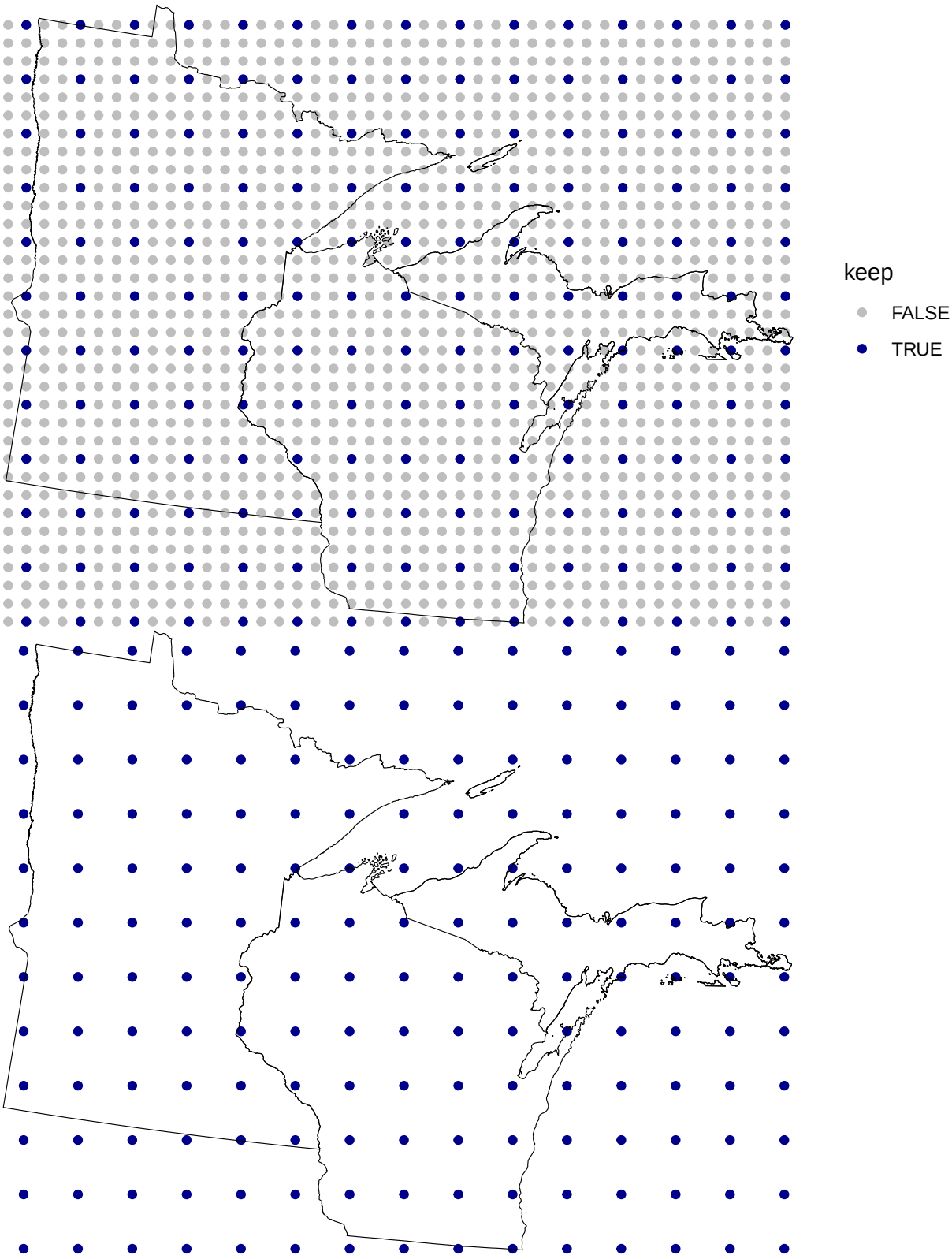
Here are some repeats of this for different starting values. They always sample every third point.

Start: x = 2, y = 2



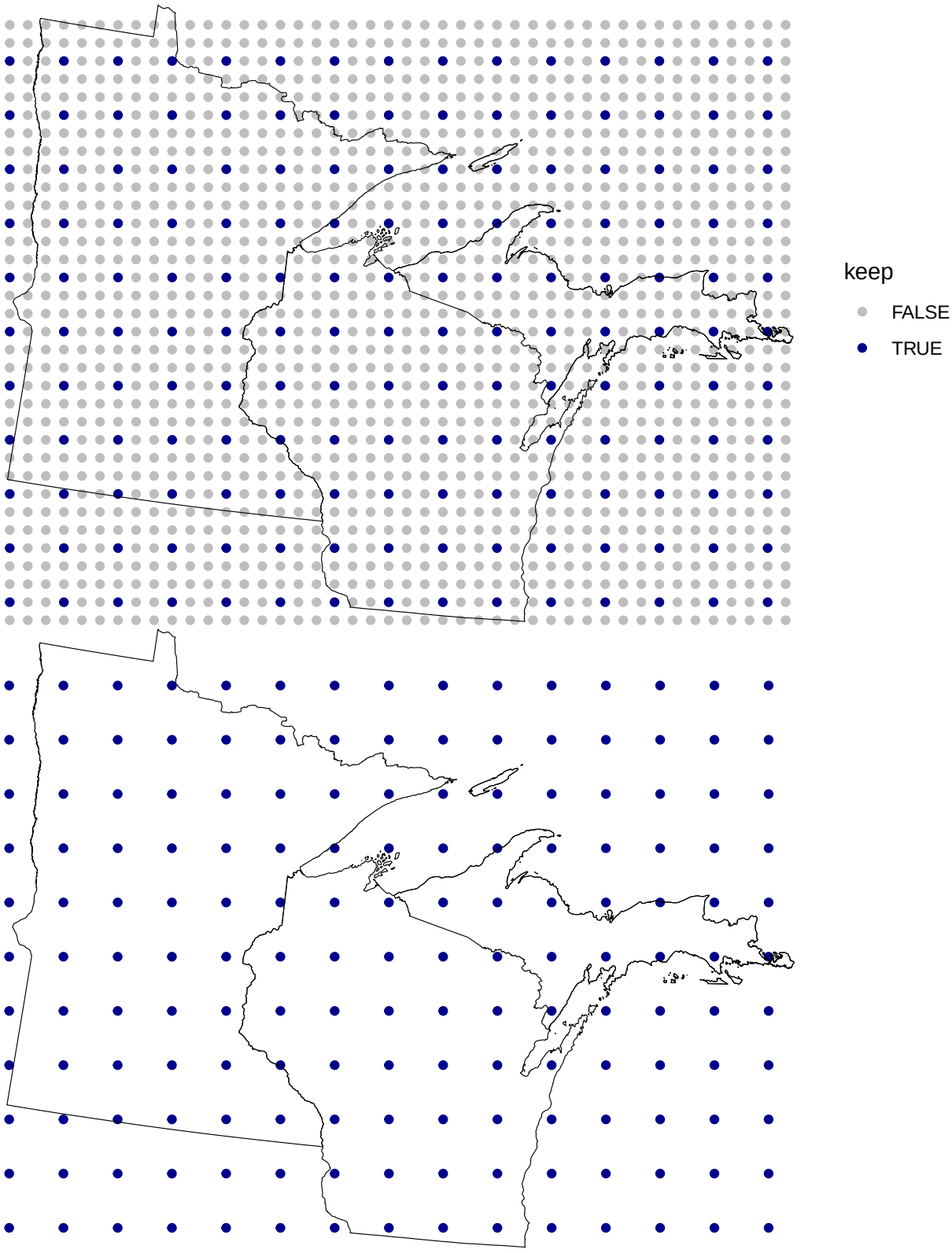

```
##    n
## 1 77
```

Start: x = 2, y = 1



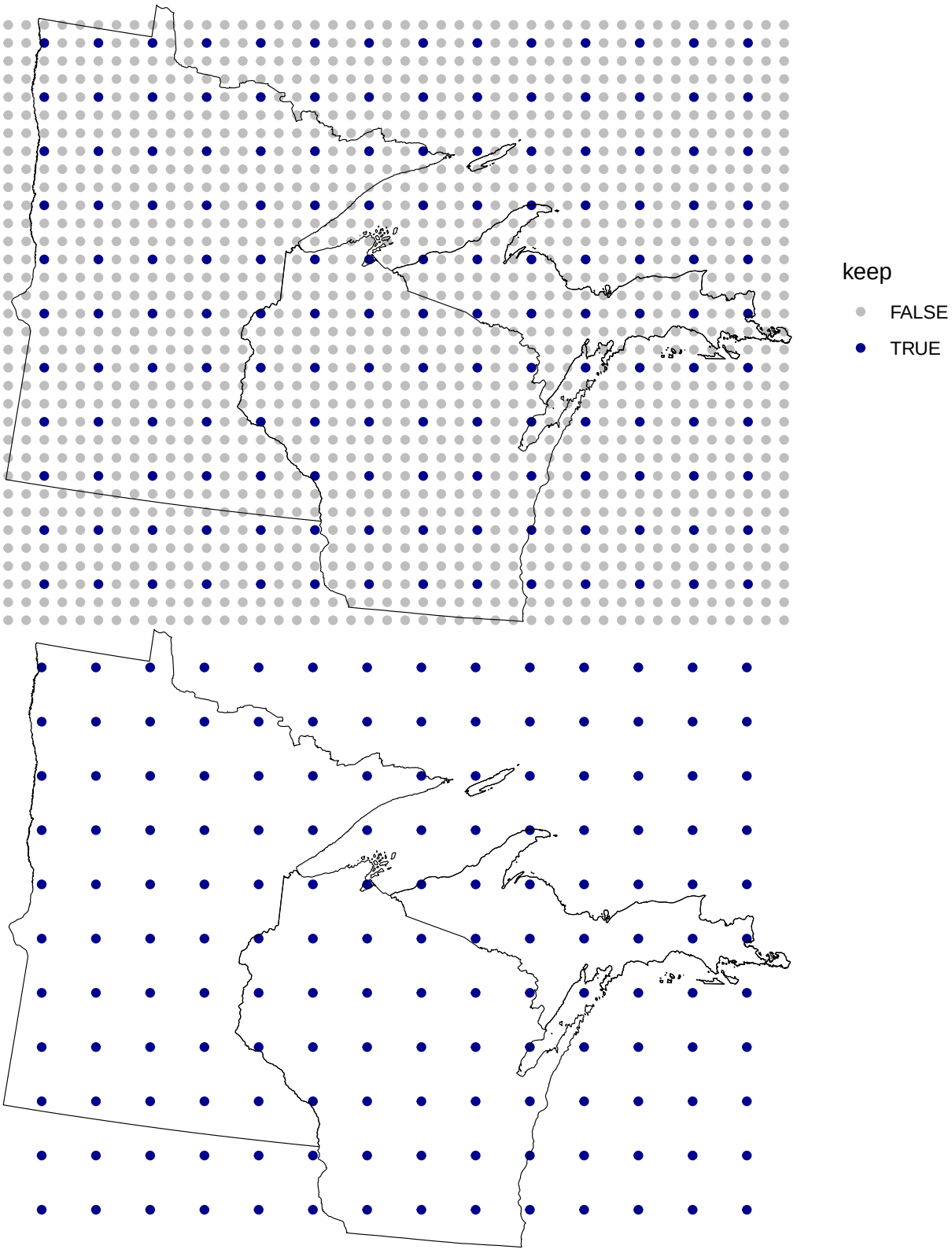

```
##    n
## 1 73
```

Start: x = 1, y = 2



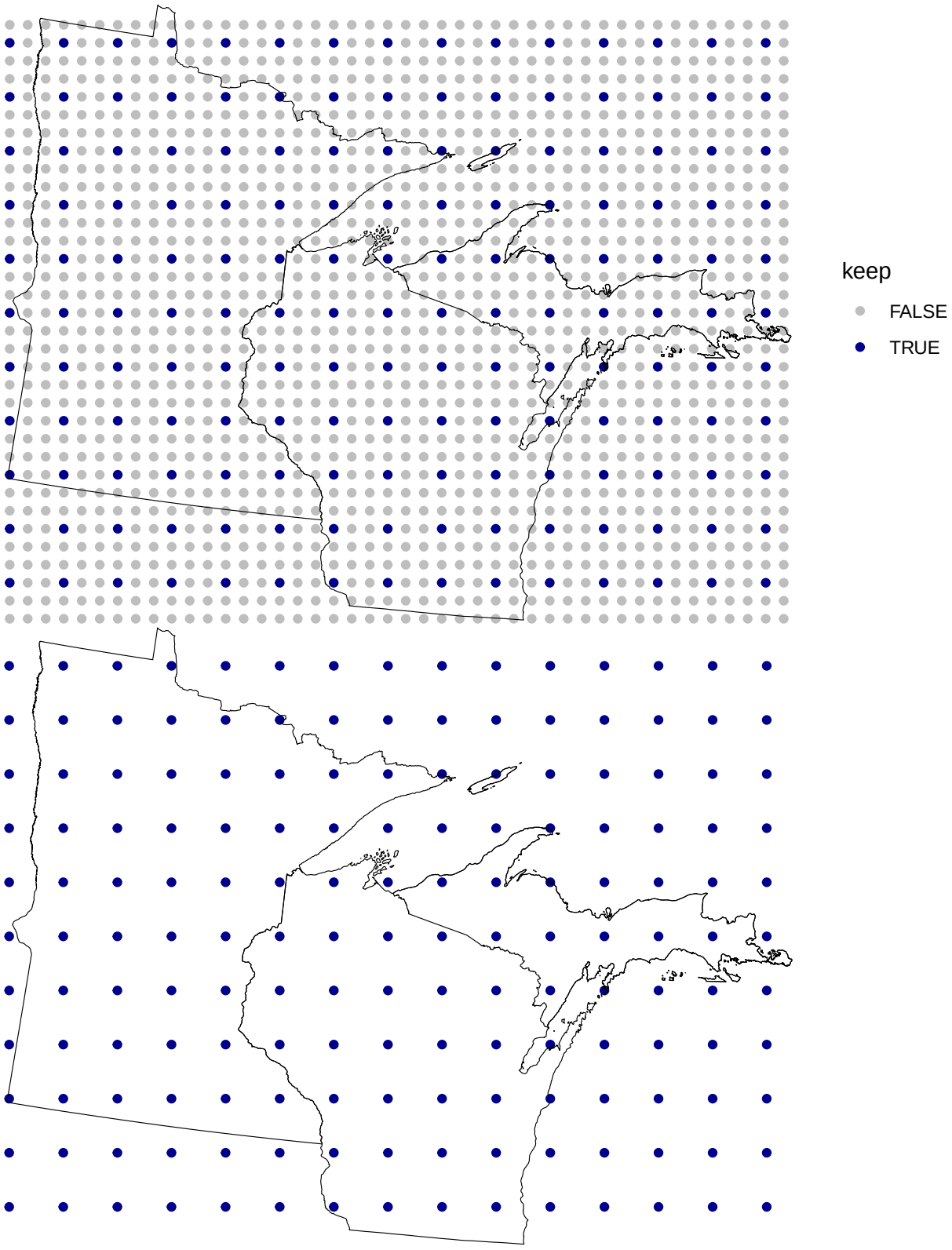

```
##    n
## 1 80
```

Start: x = 3, y = 3



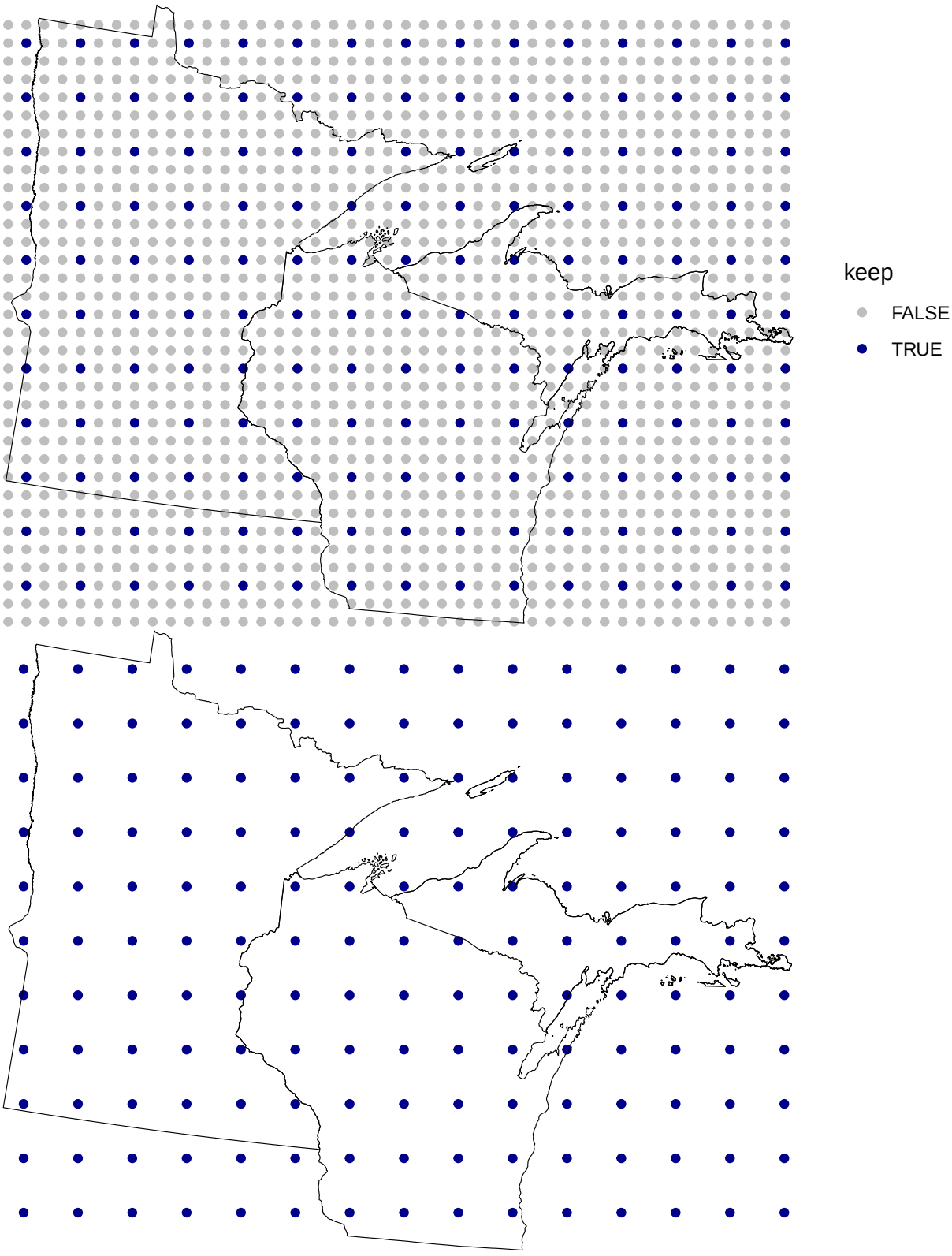

```
##    n
## 1 81
```

Start: x = 1, y = 3



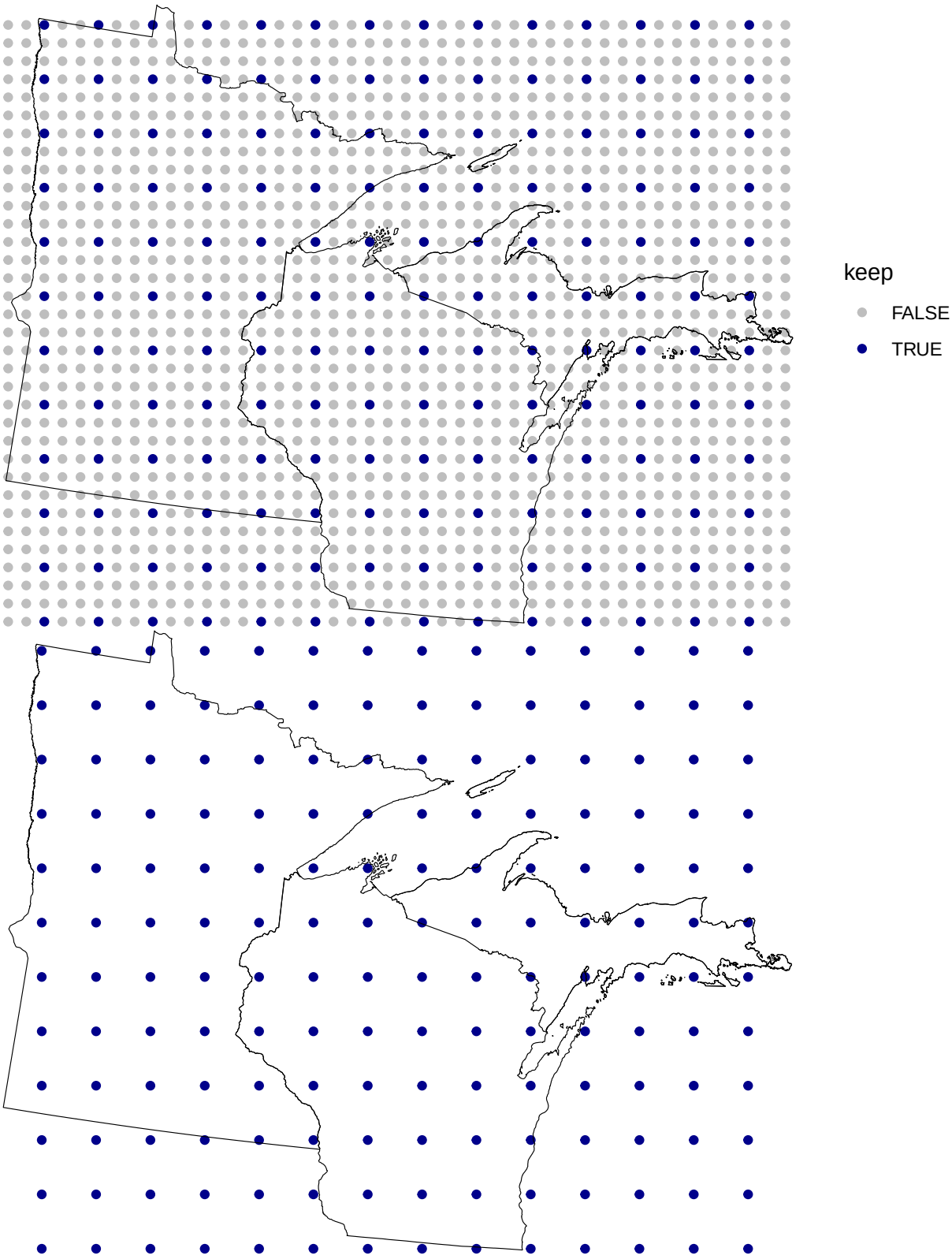

```
##    n
## 1 80
```

Start: x = 2, y = 3



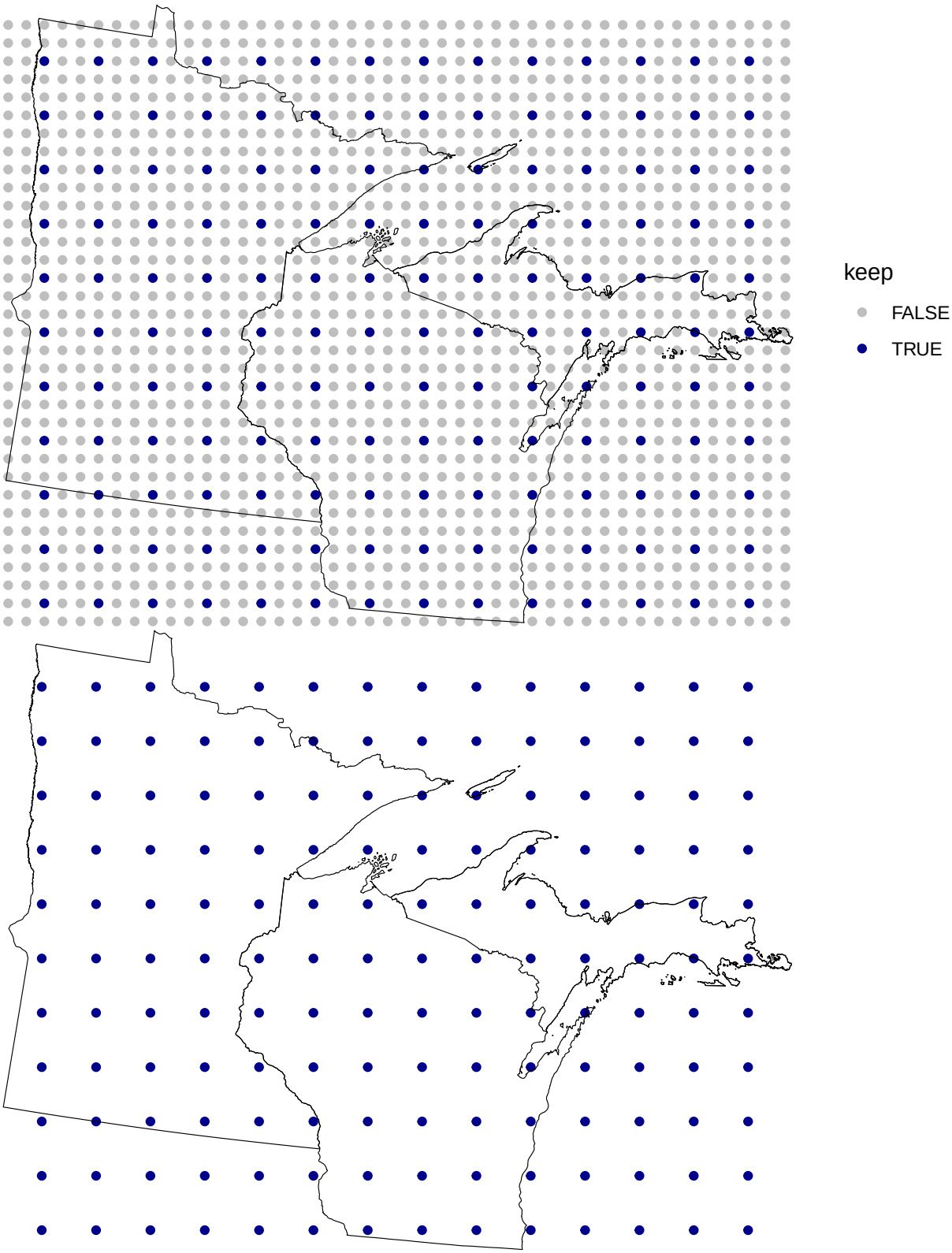

```
##    n
## 1 80
```

Start: x = 3, y = 1




```
##    n
## 1 79
```

Start: x = 3, y = 2

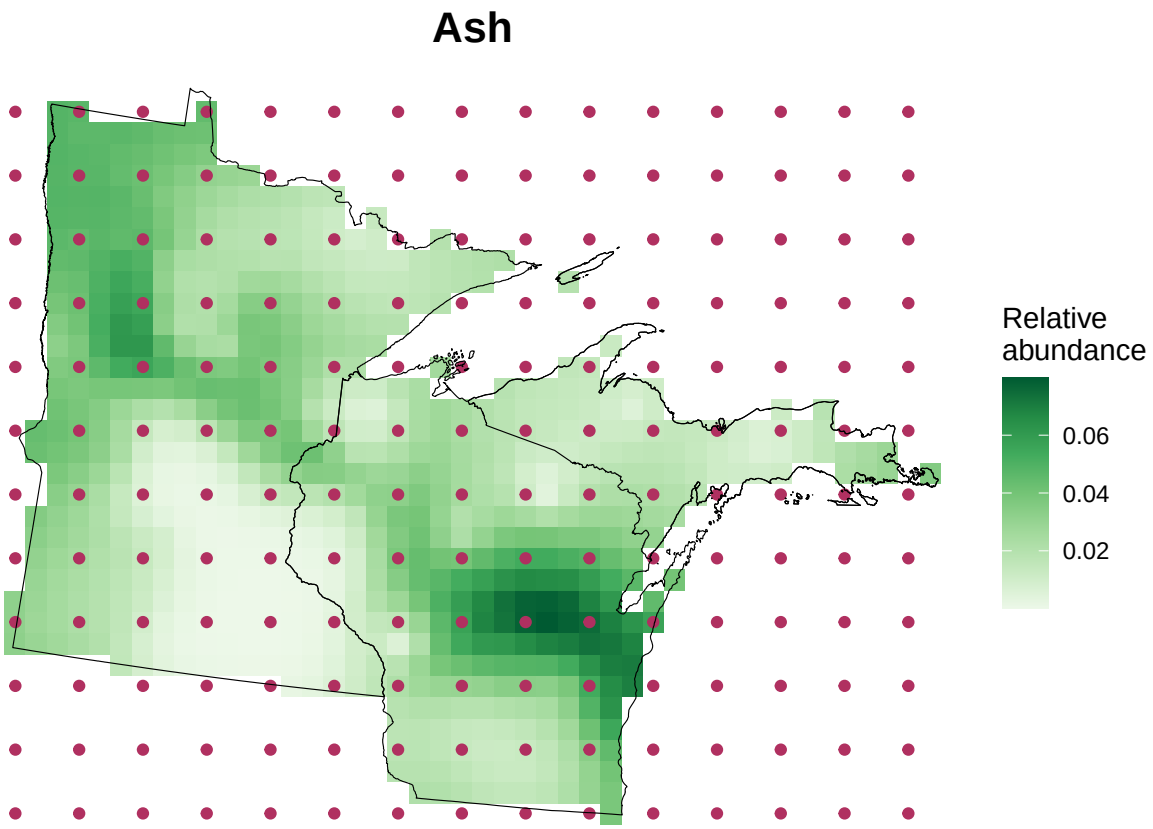



```
##    n
## 1 81
```

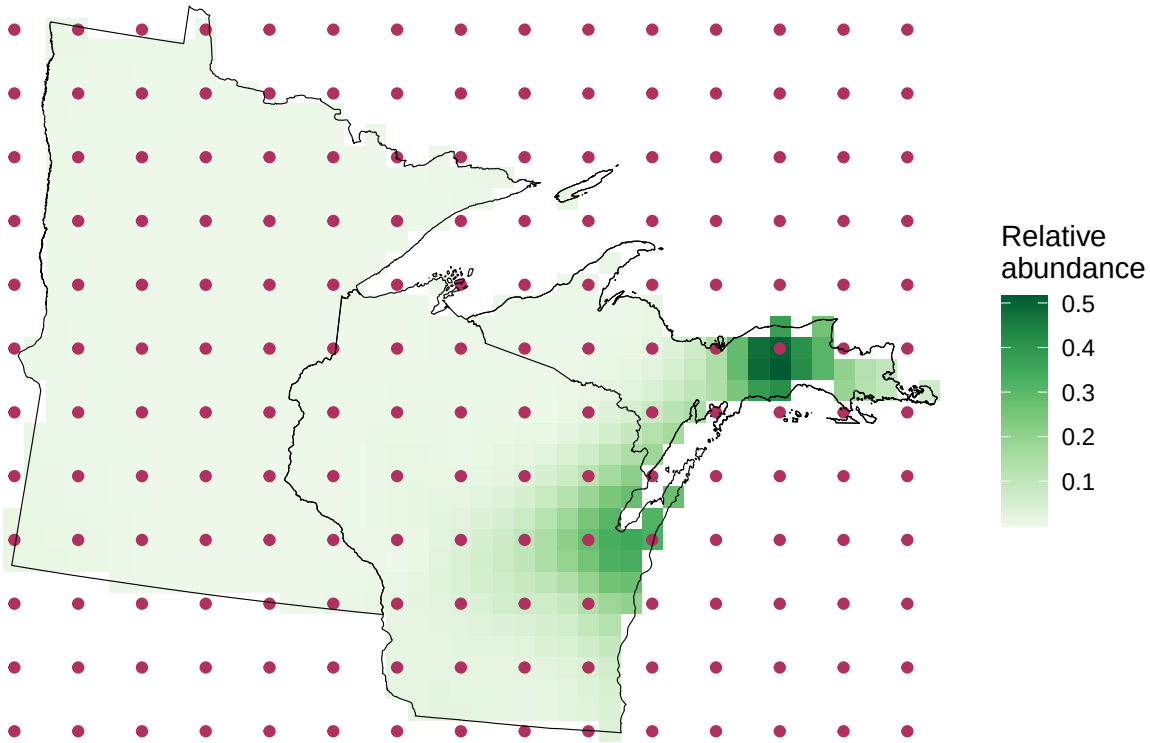
Check samples against relative abundance

A major goal of the subsampling is to intentionally try to keep cells with strong patterns in relative abundance. For example, we'd really like to keep where hemlock relative abundance is very high in central Wisconsin as well as where pine relative abundances are very high in the few cells identified above. Let's just make a ton of figures to look at whether the point distributions made above correspond to these patterns in relative abundance.

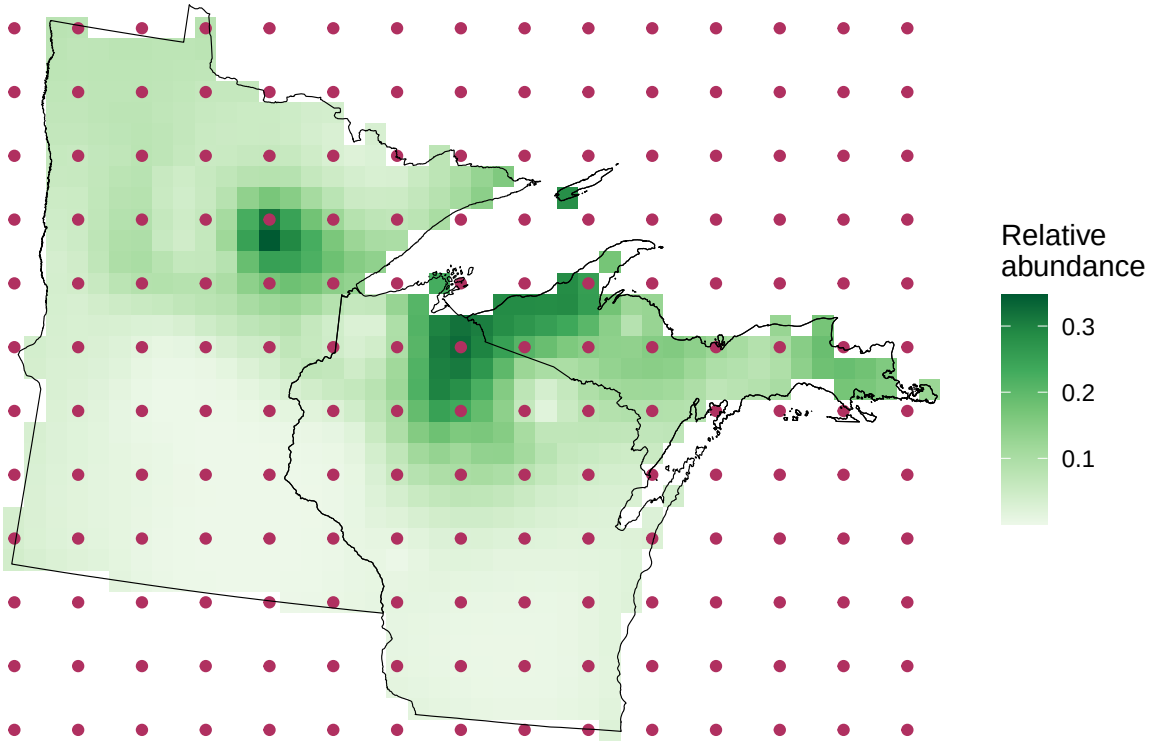
x = 1, y = 1

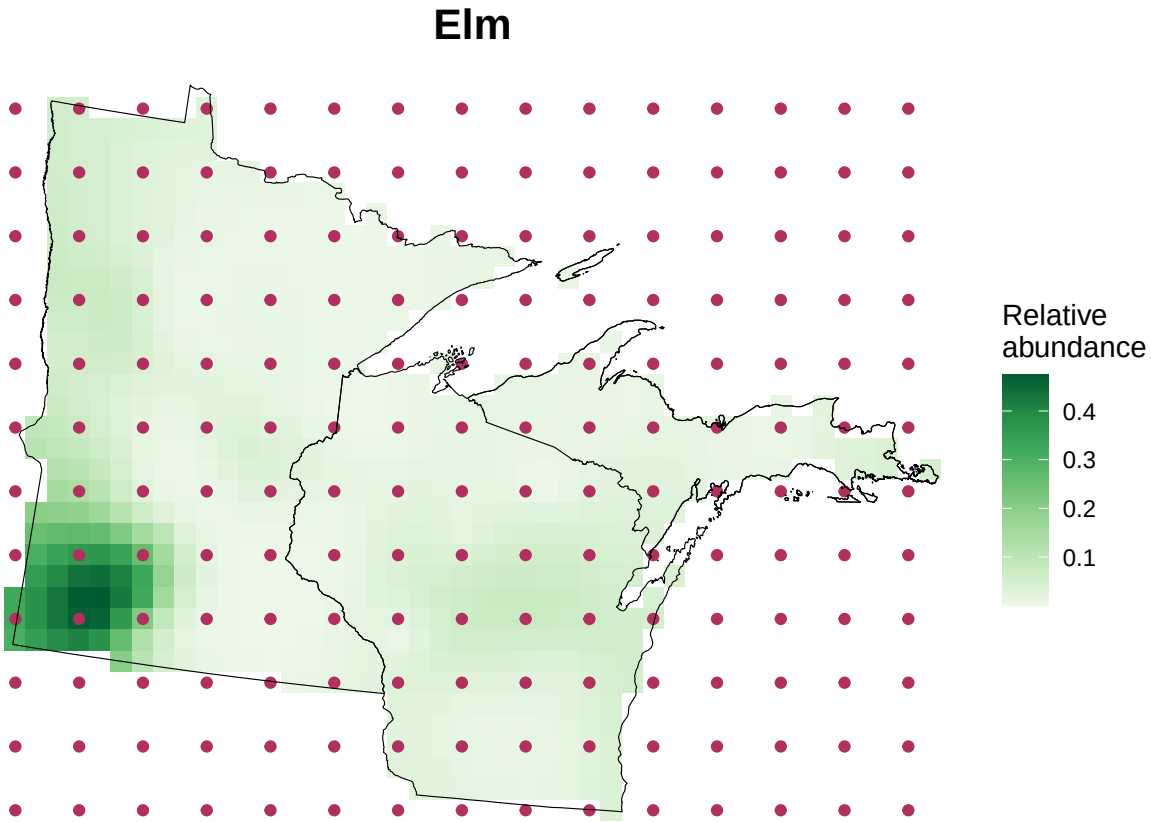


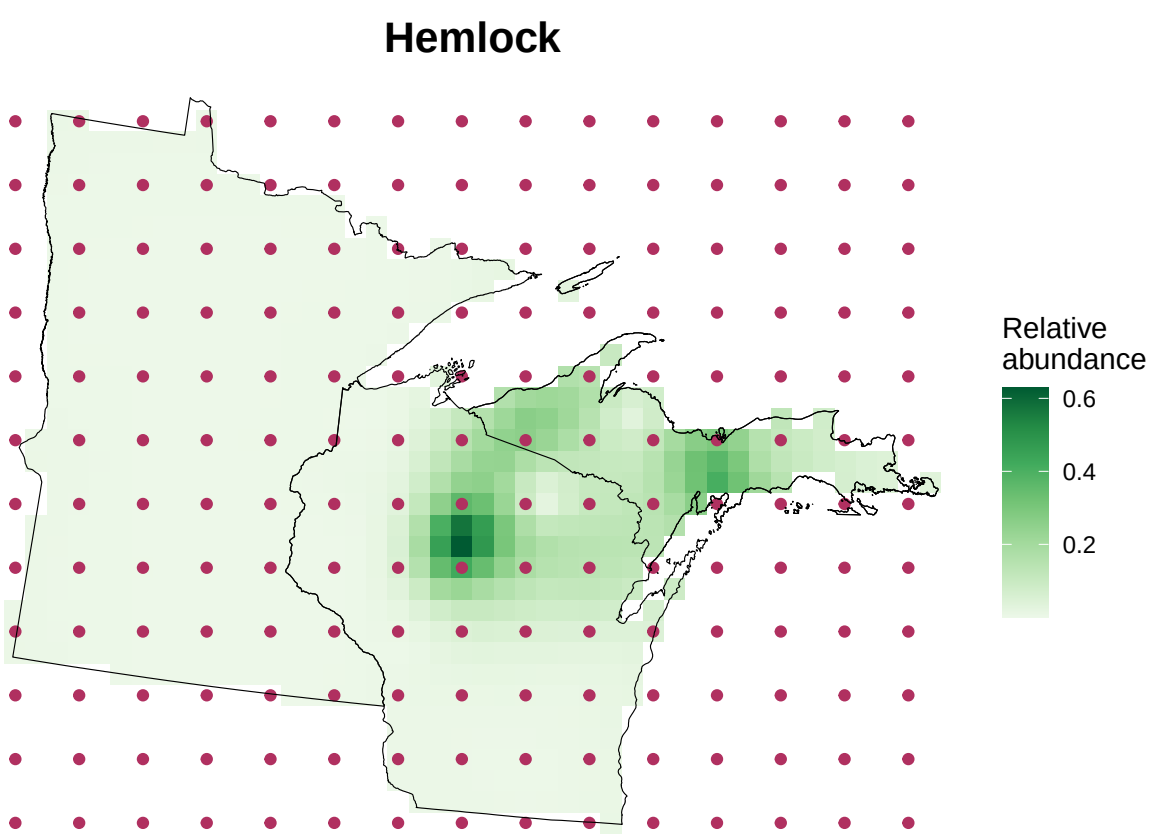
Beech


Relative
abundance
0.5
0.4
0.3
0.2
0.1

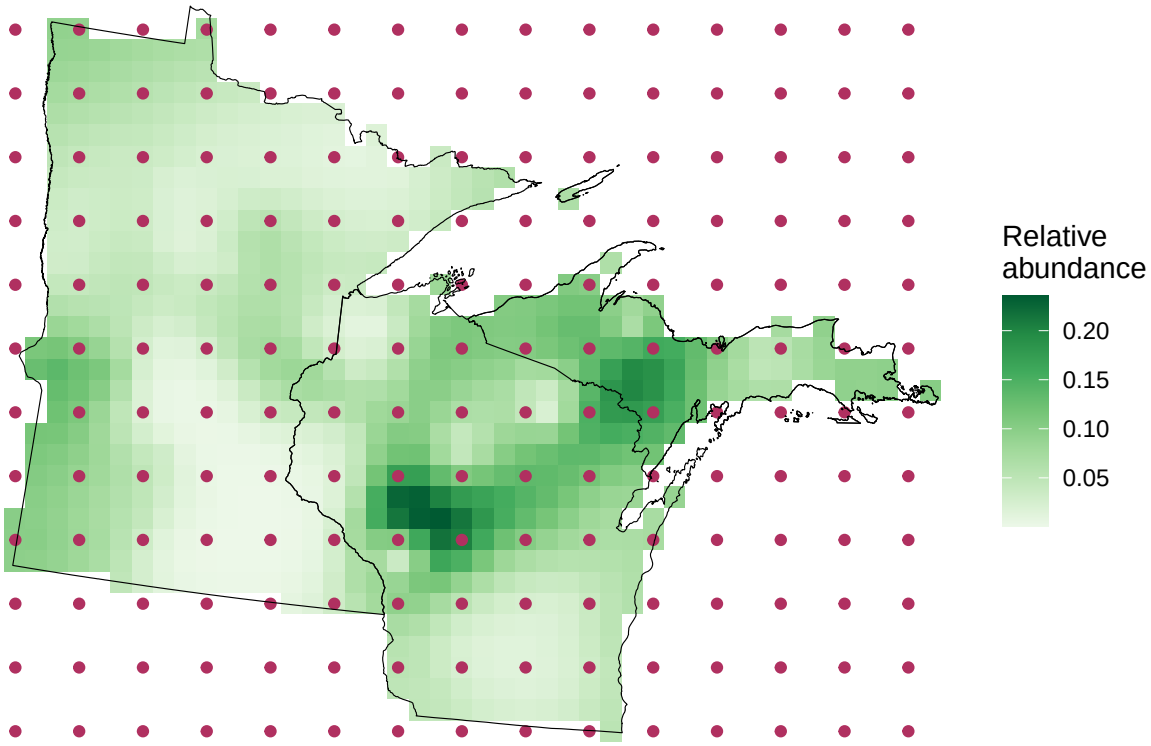
Birch


Relative
abundance
0.3
0.2
0.1

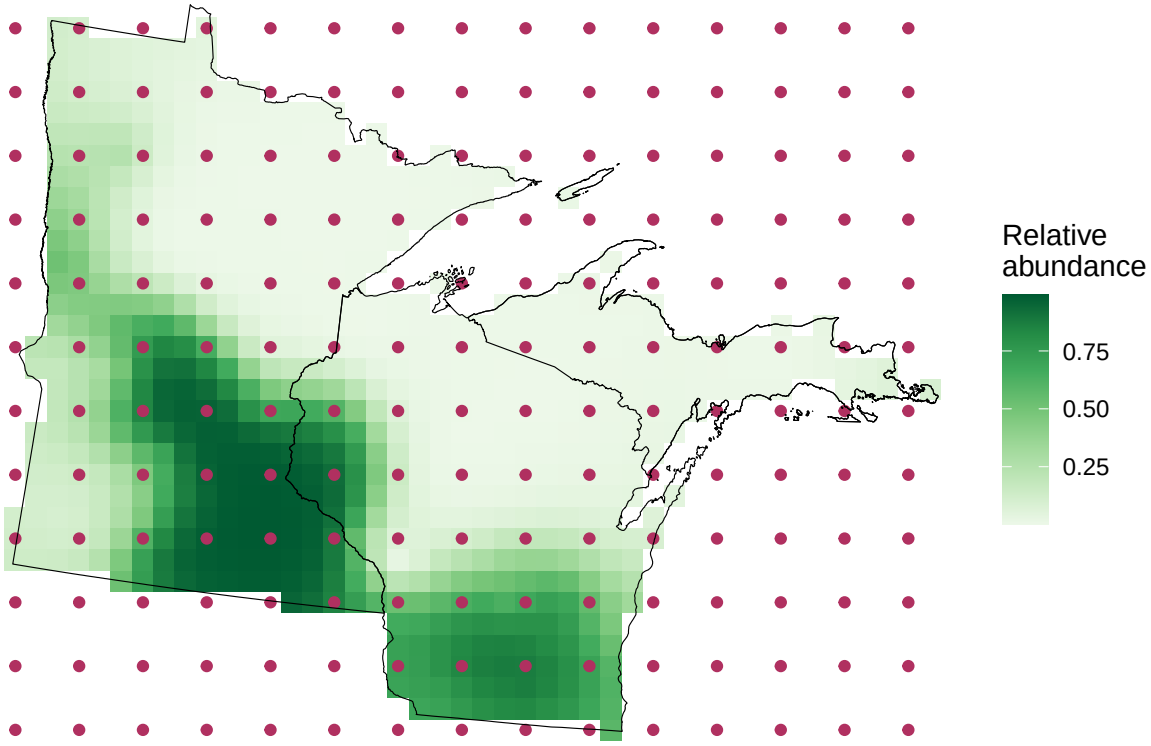

Elm
Relative abundance
0.4
0.3
0.2
0.1


Hemlock
Relative abundance
0.6
0.4
0.2

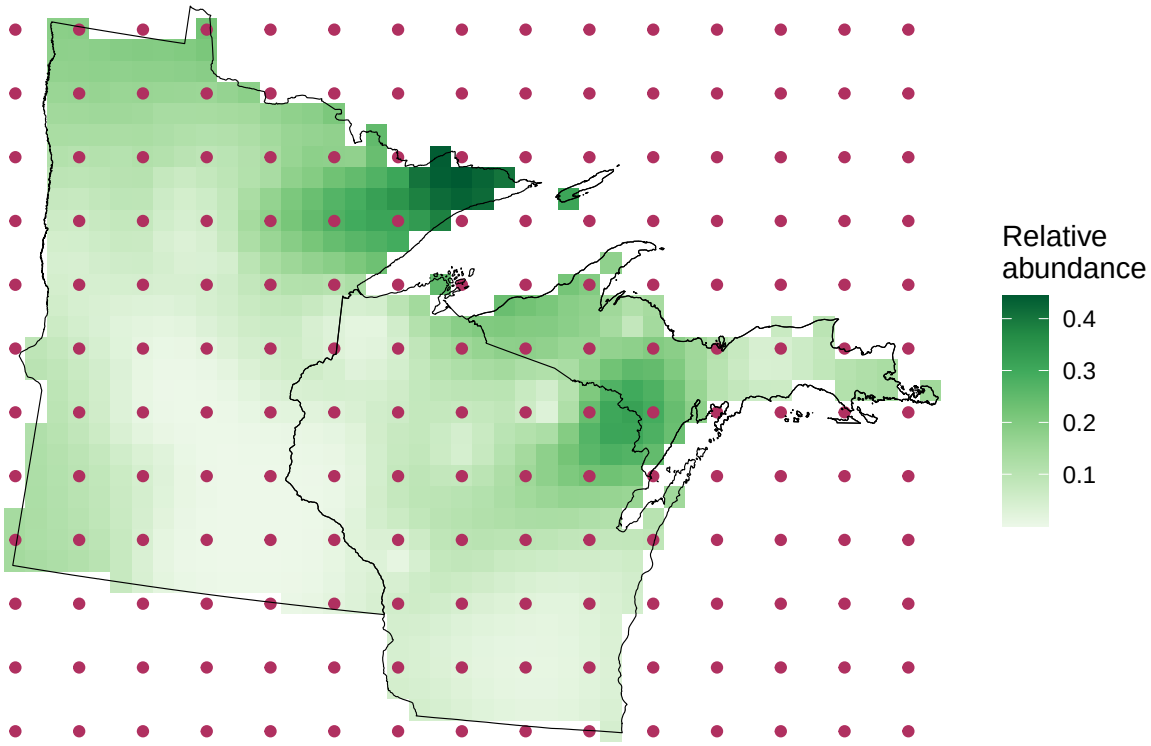
Maple


Relative
abundance
0.20
0.15
0.10
0.05

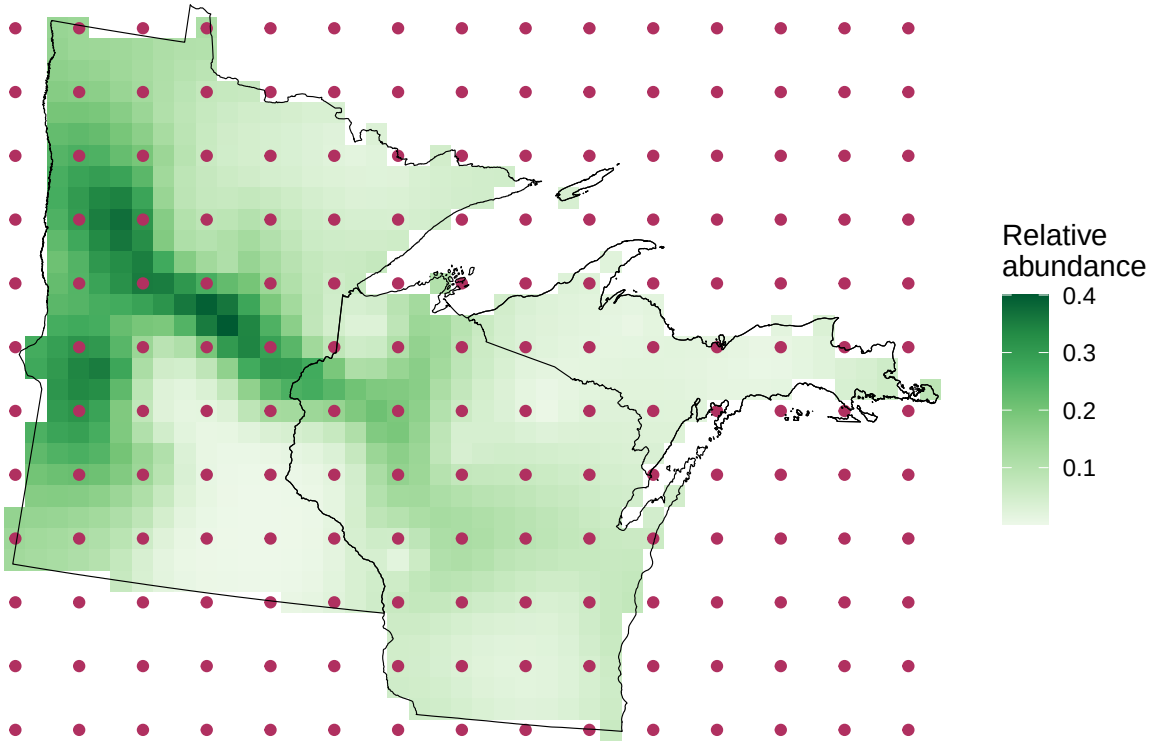
Oak

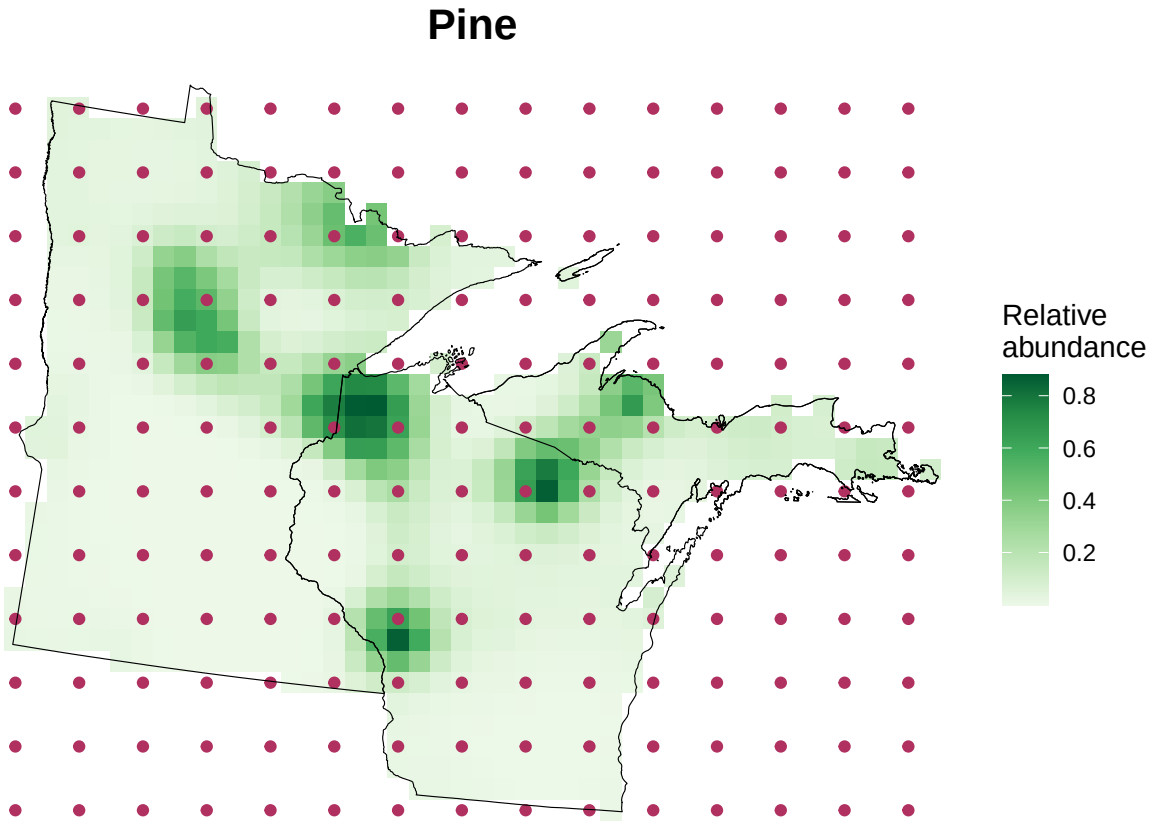

Relative
abundance
0.75
0.50
0.25

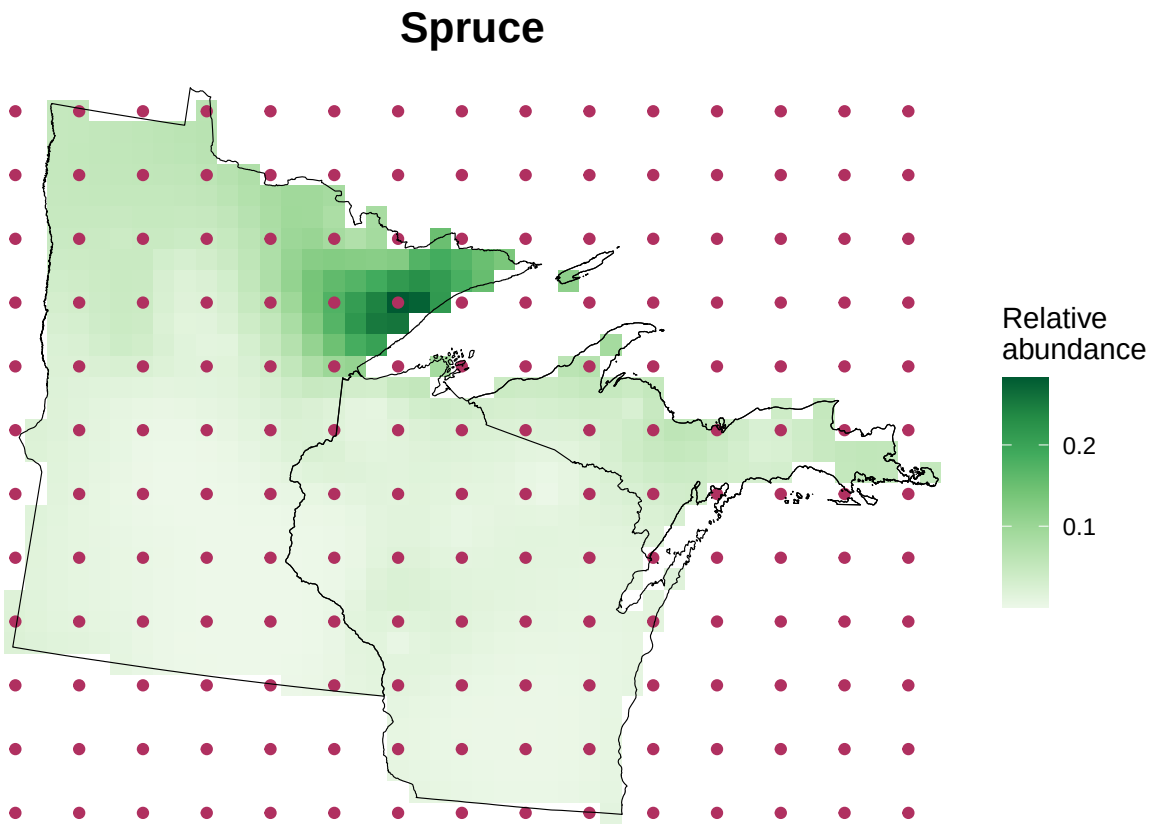
Other conifer



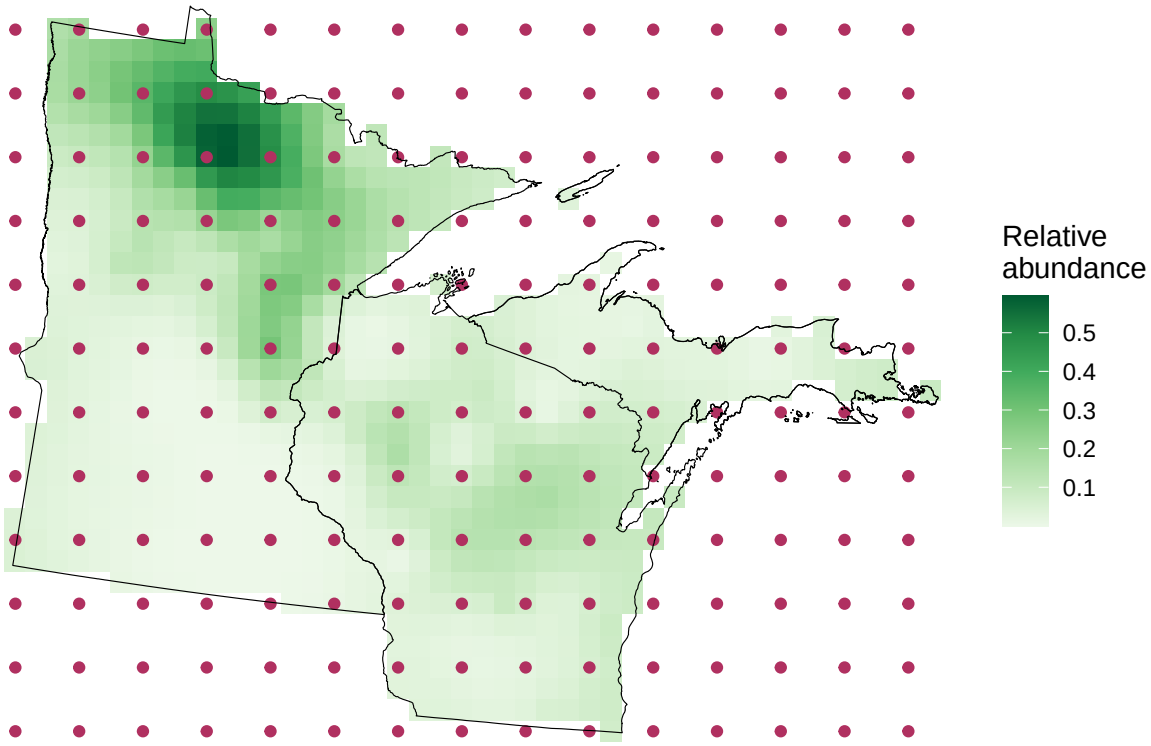
Other hardwood




Pine
Relative abundance
0.8
0.6
0.4
0.2

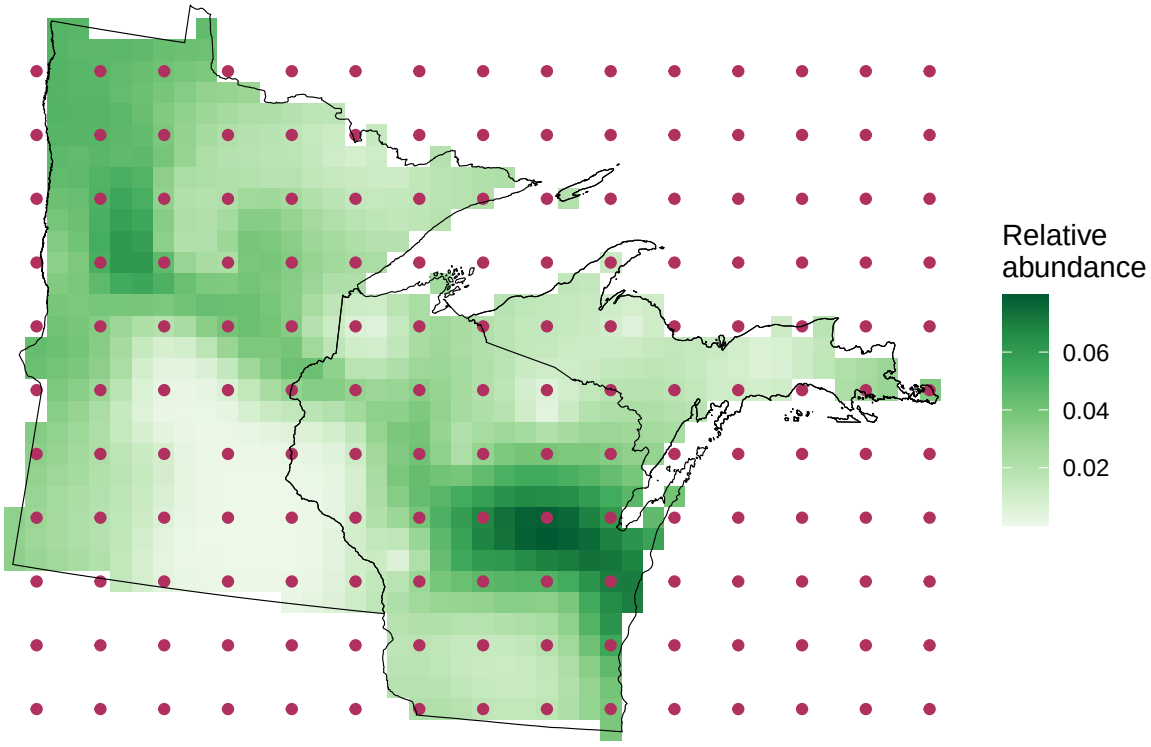

Spruce
Relative abundance
0.2
0.1

Tamarack

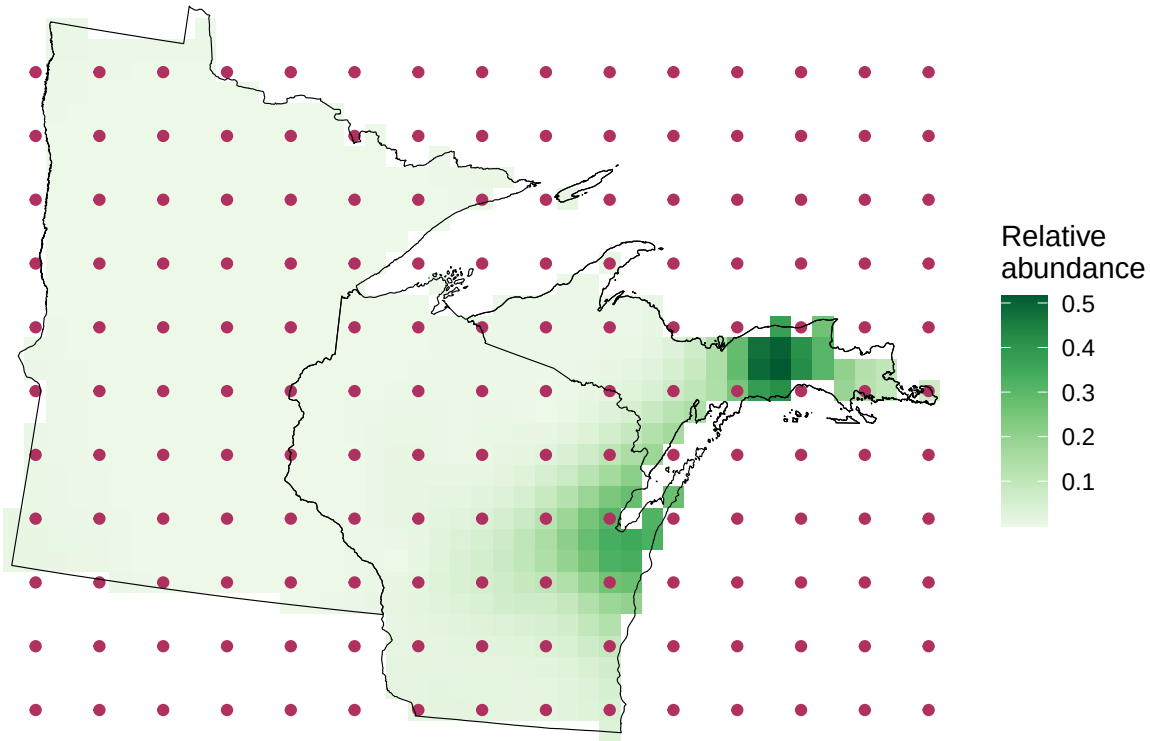


$\mathbf{x = 2, y = 2}$

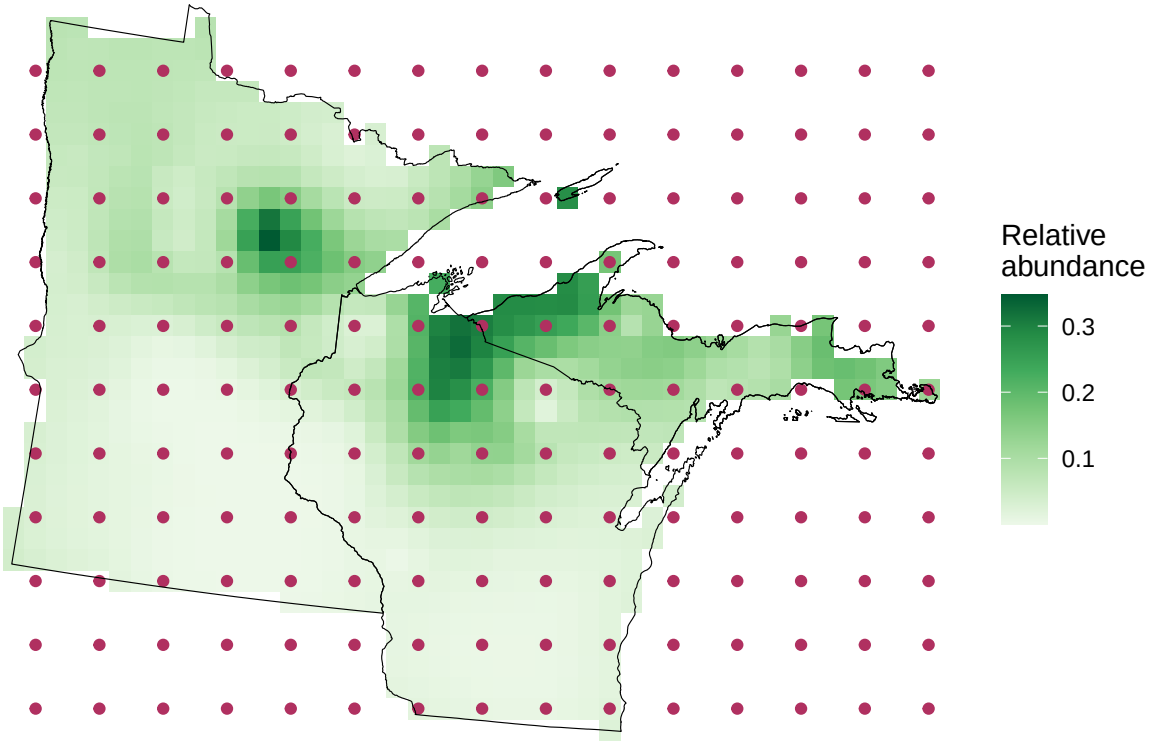
Ash



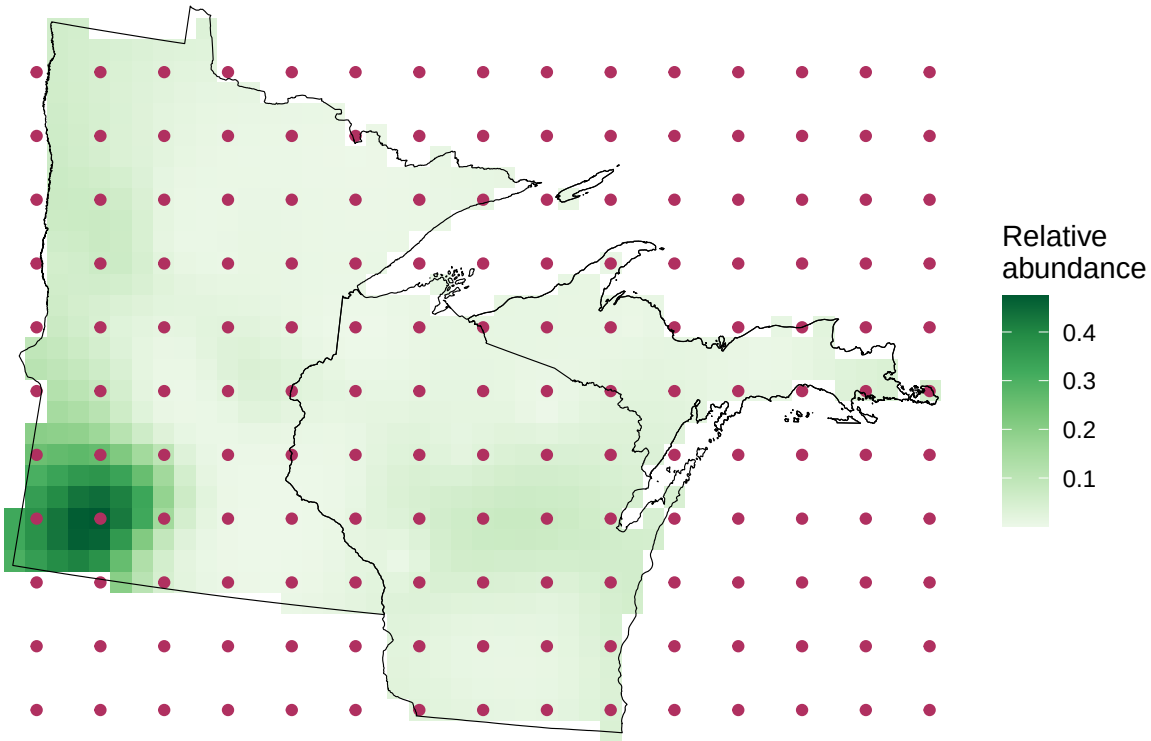
Beech


Relative
abundance
0.5
0.4
0.3
0.2
0.1

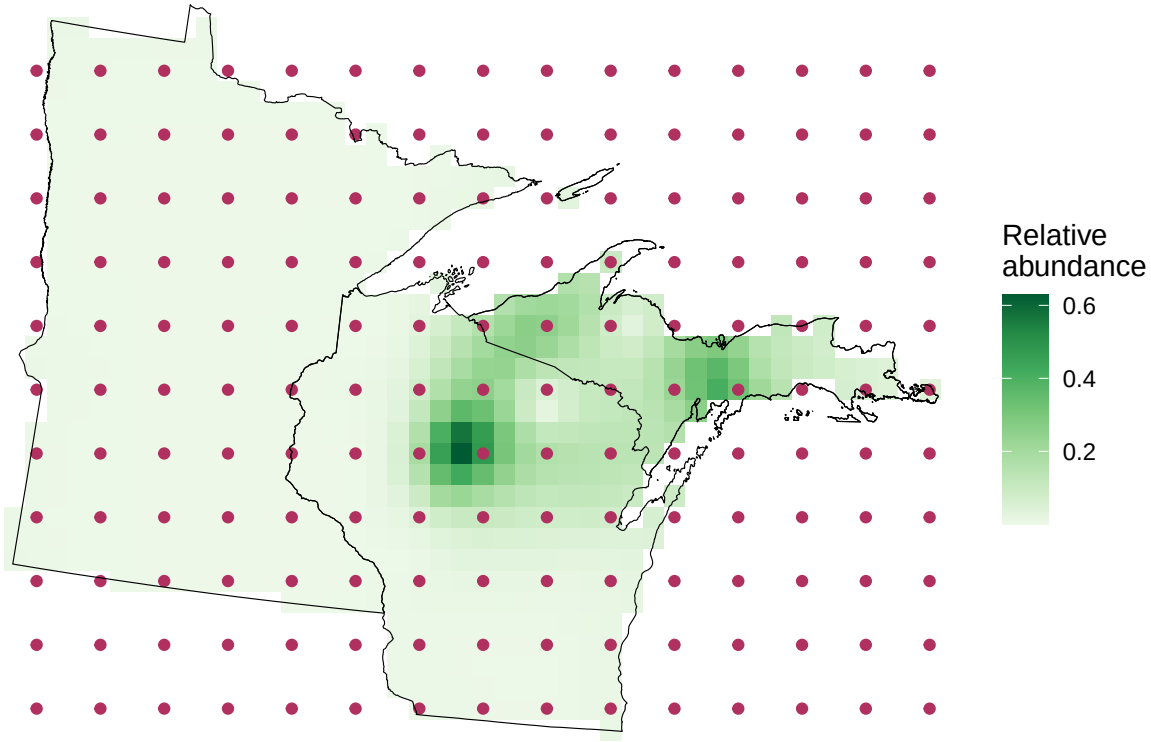
Birch


Relative
abundance
0.3
0.2
0.1

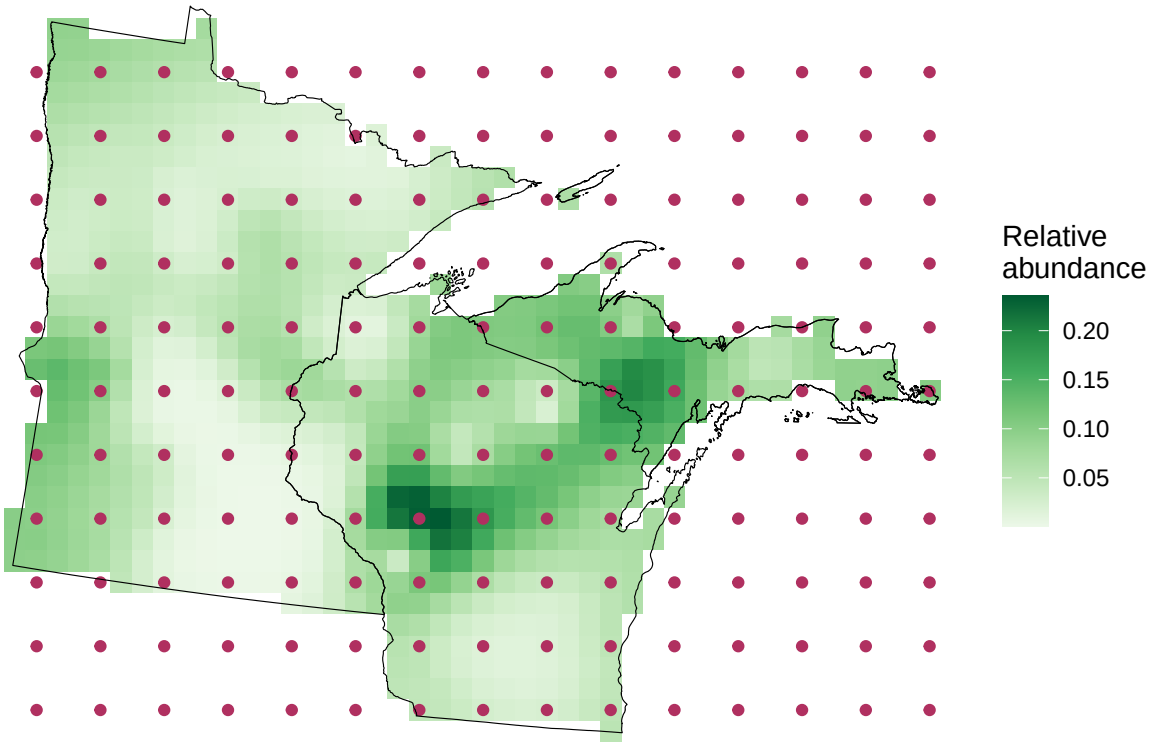
Elm



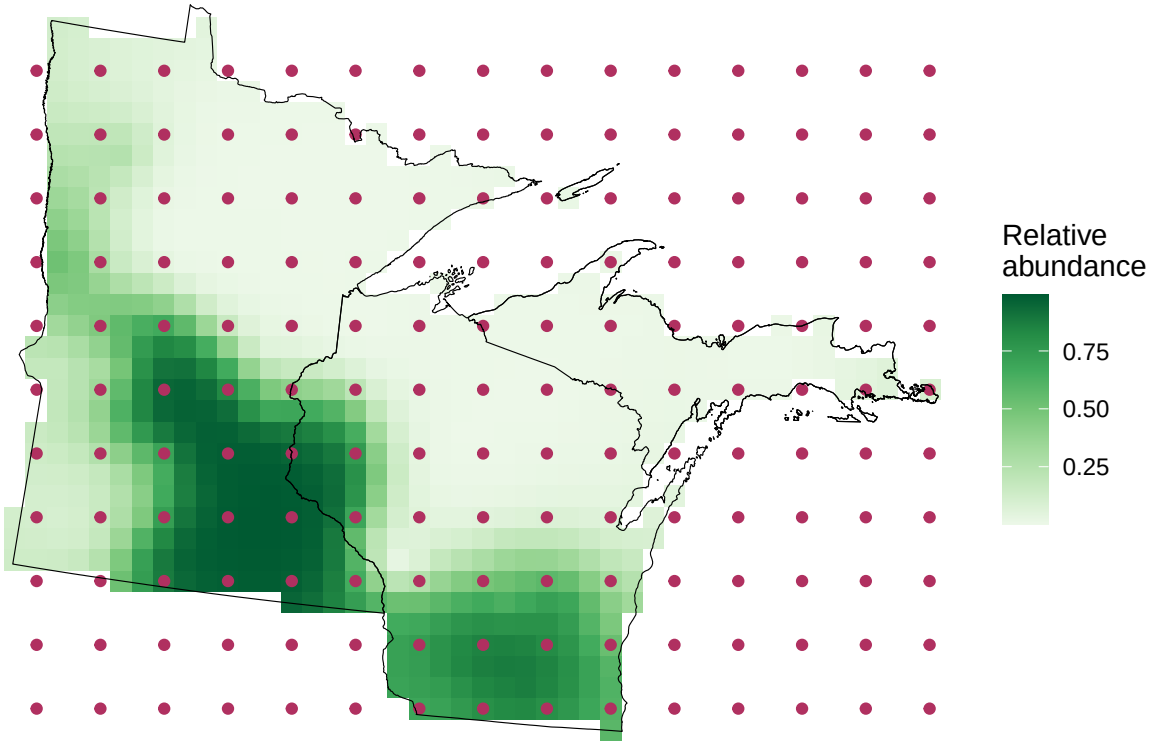
Hemlock



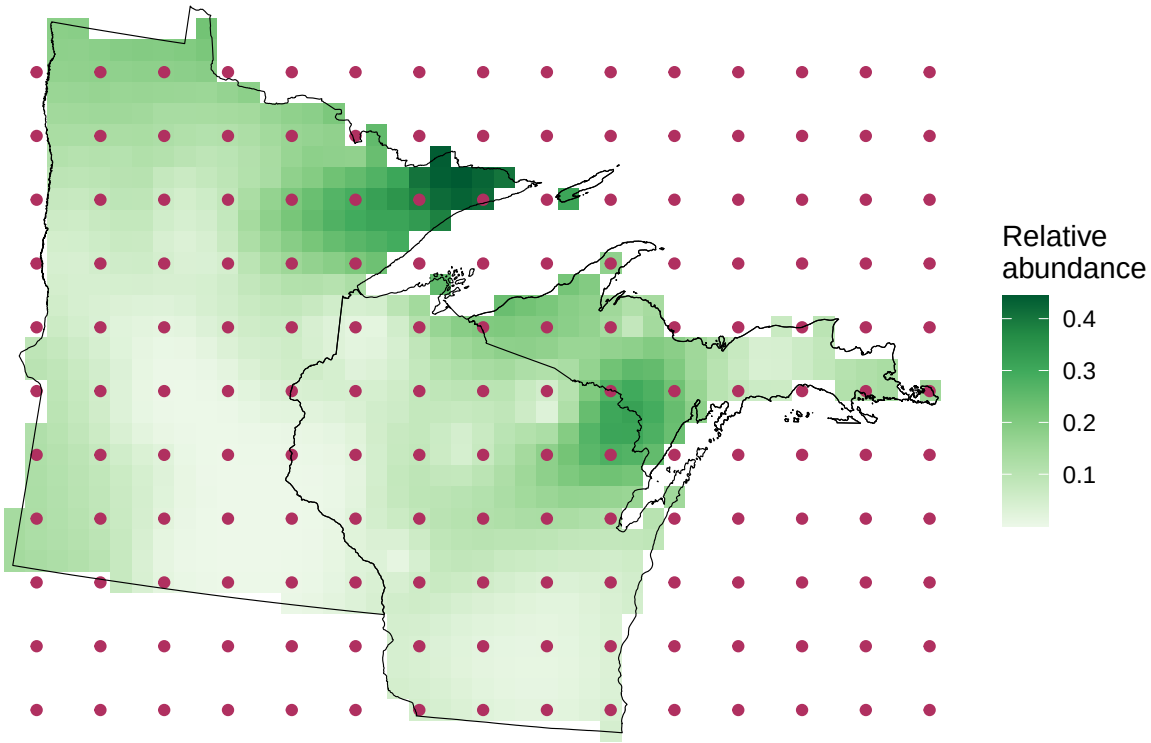
Maple


Relative
abundance
0.20
0.15
0.10
0.05

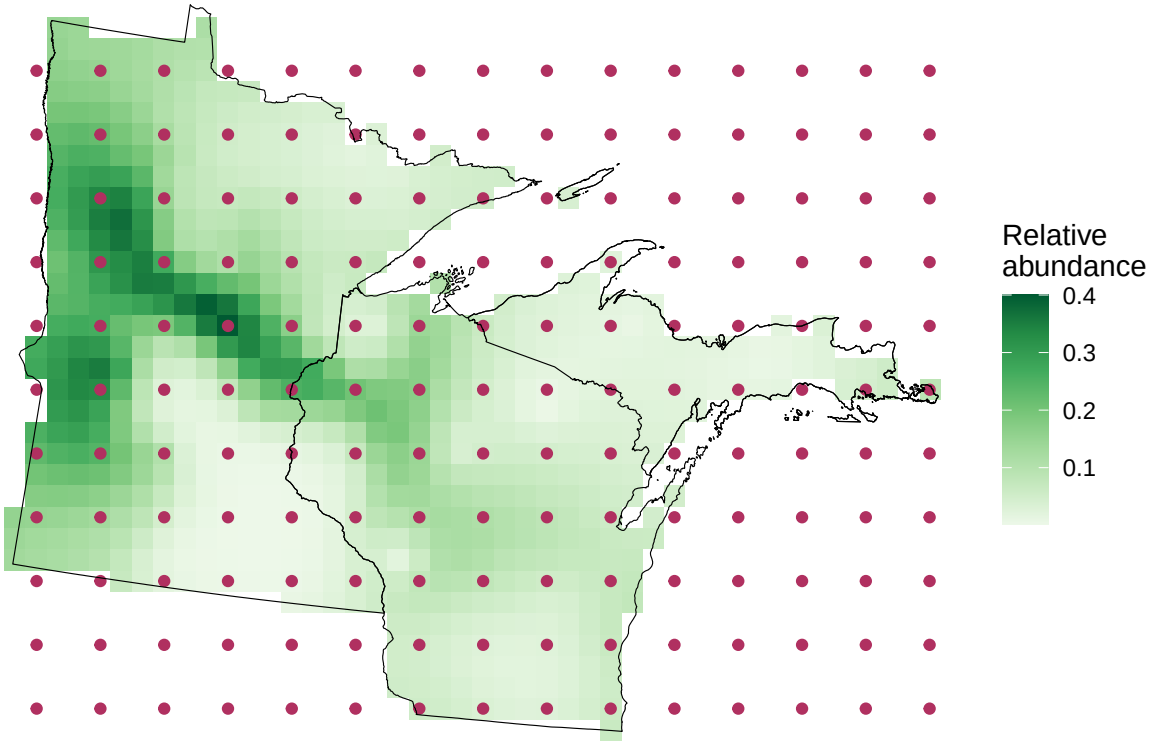
Oak


Relative
abundance
0.75
0.50
0.25

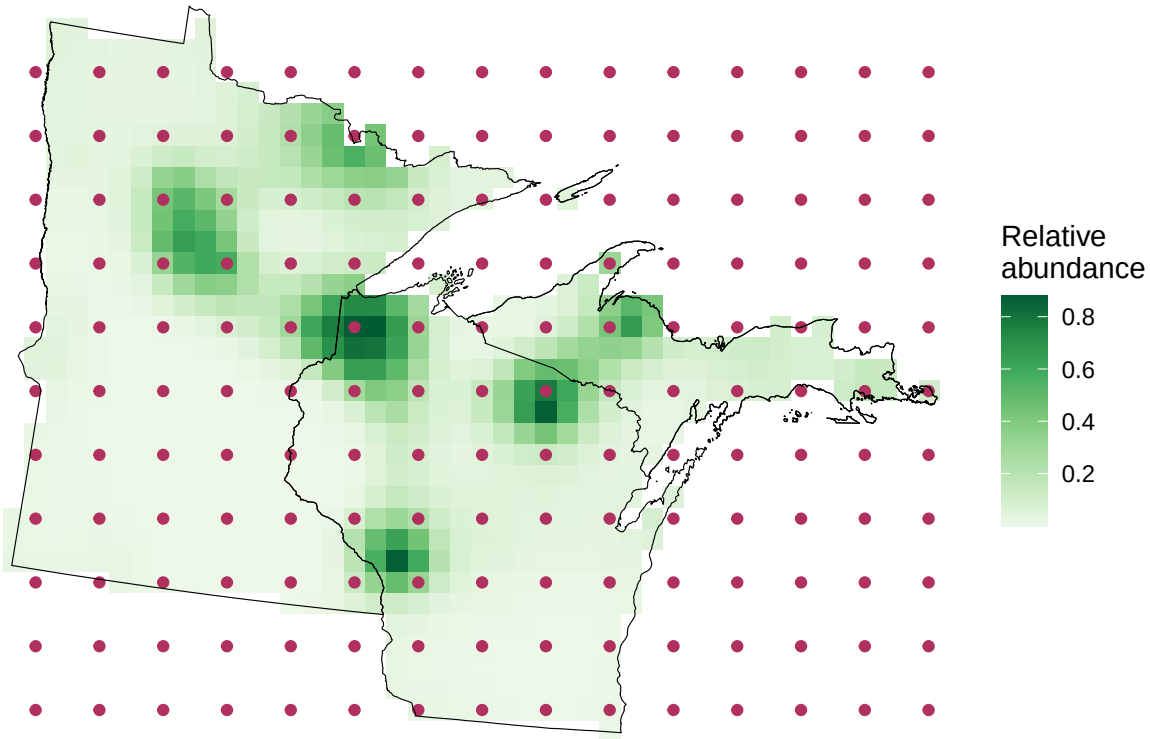
Other conifer



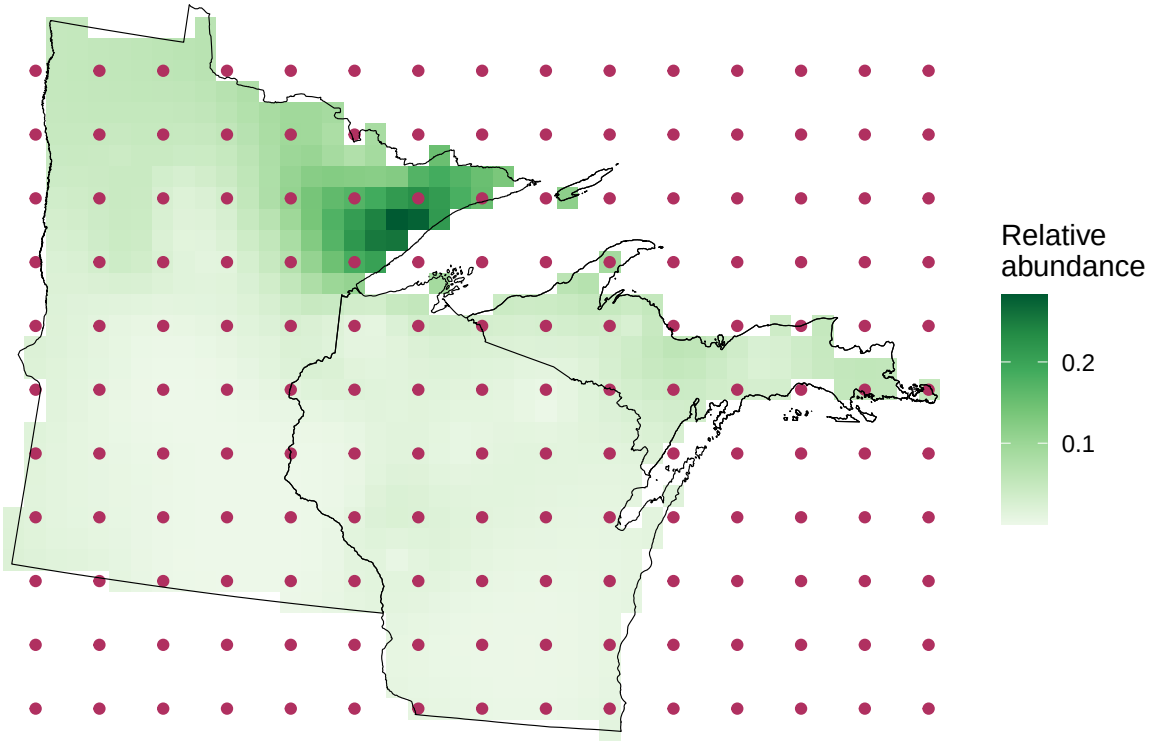
Other hardwood



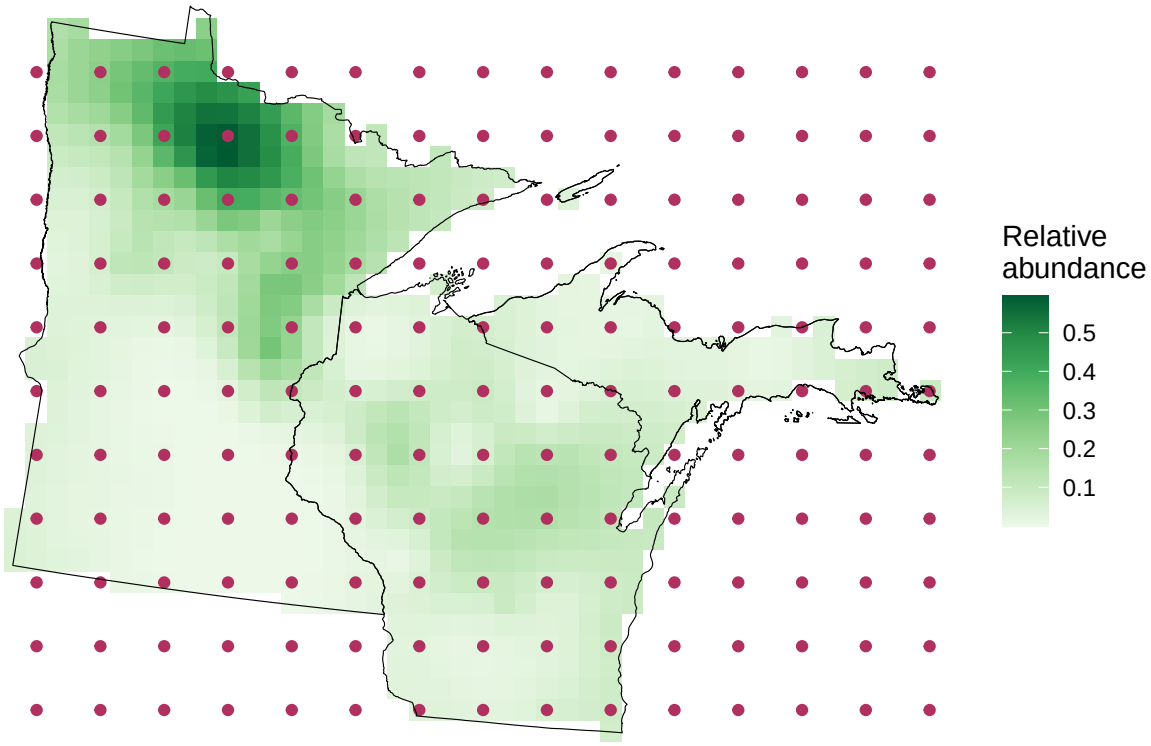
Pine



Spruce

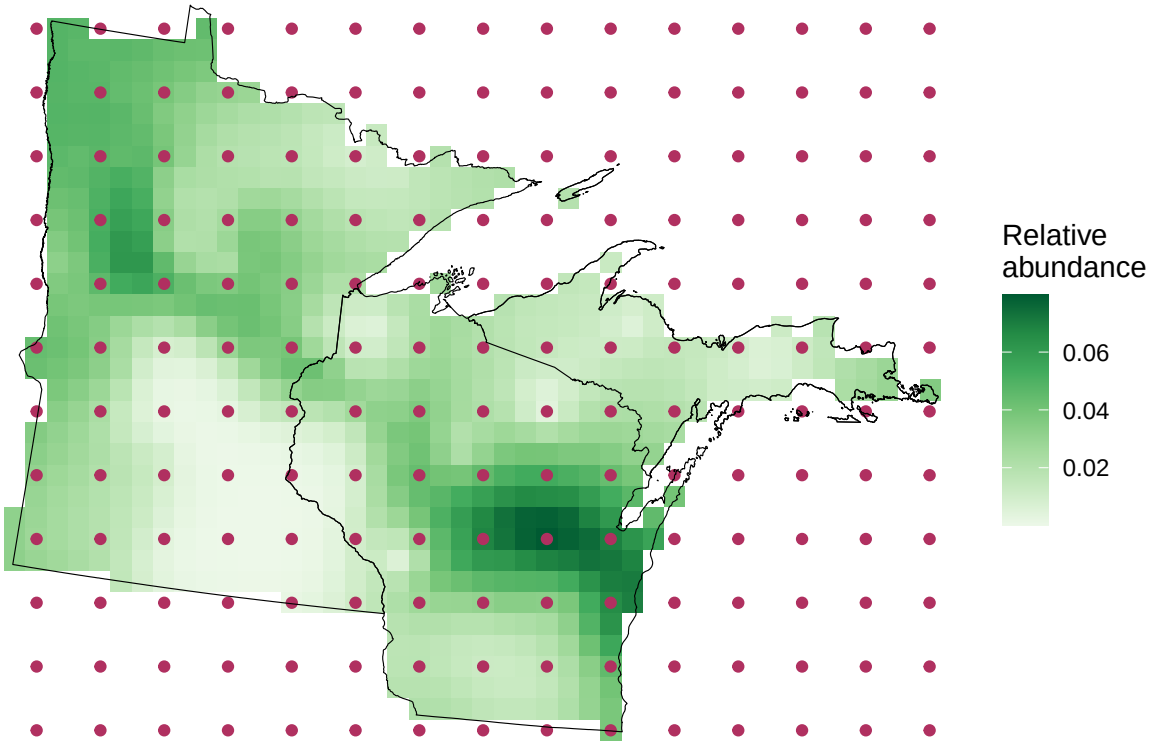


Tamarack

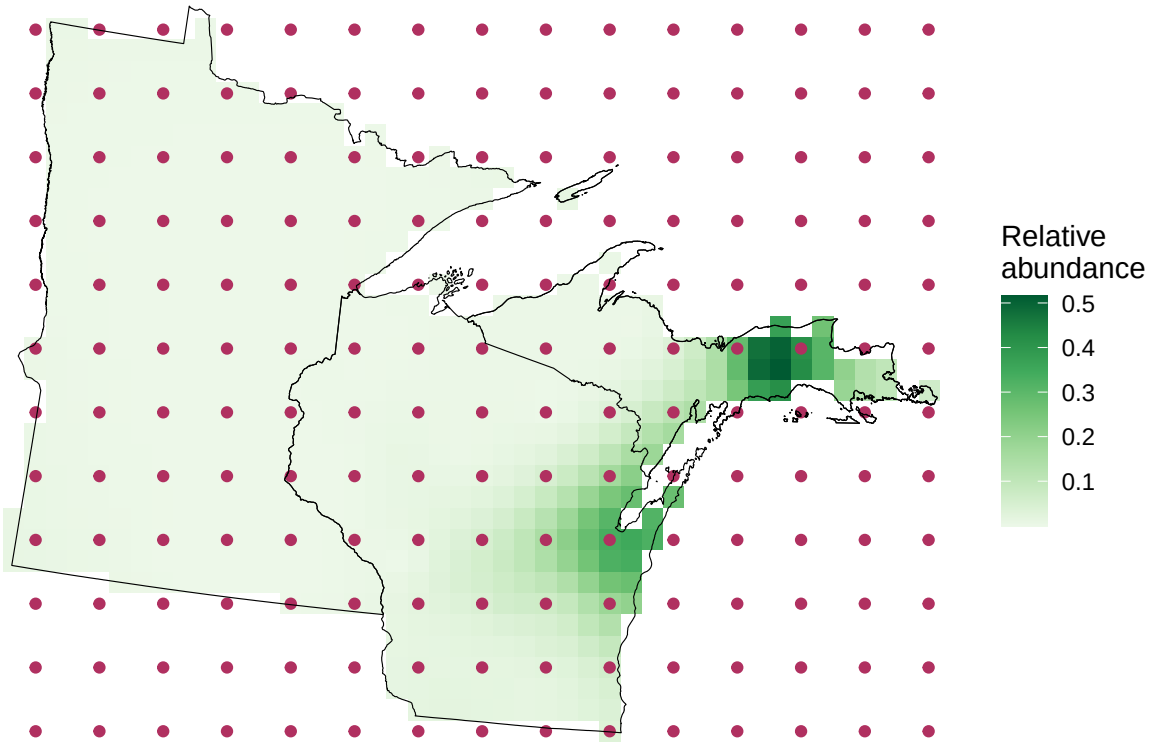


$\mathbf{x = 2, y = 1}$

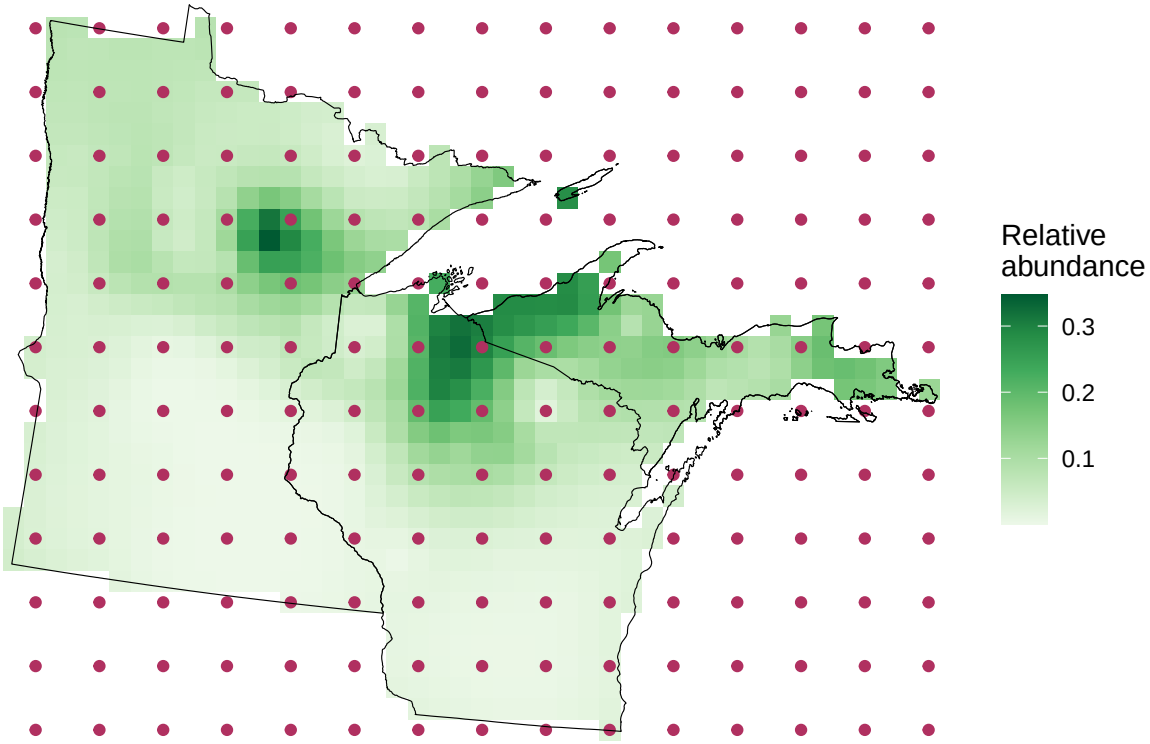
Ash

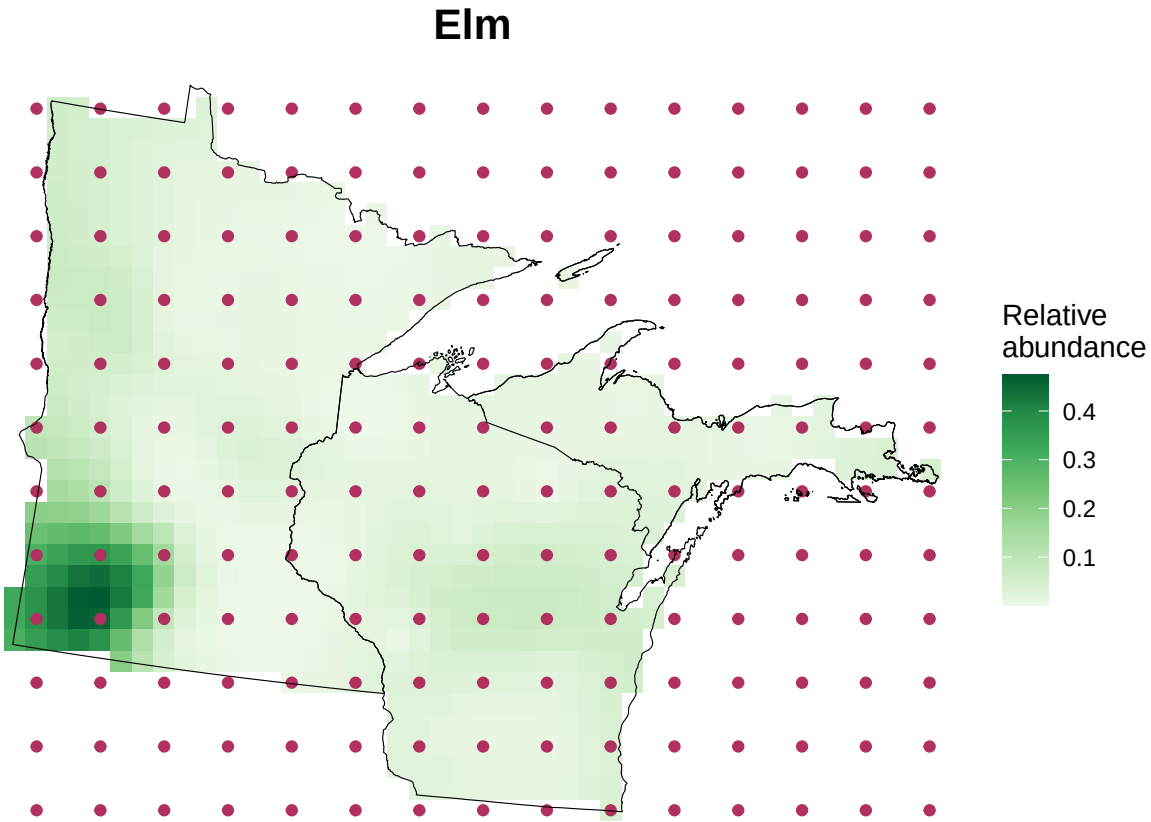


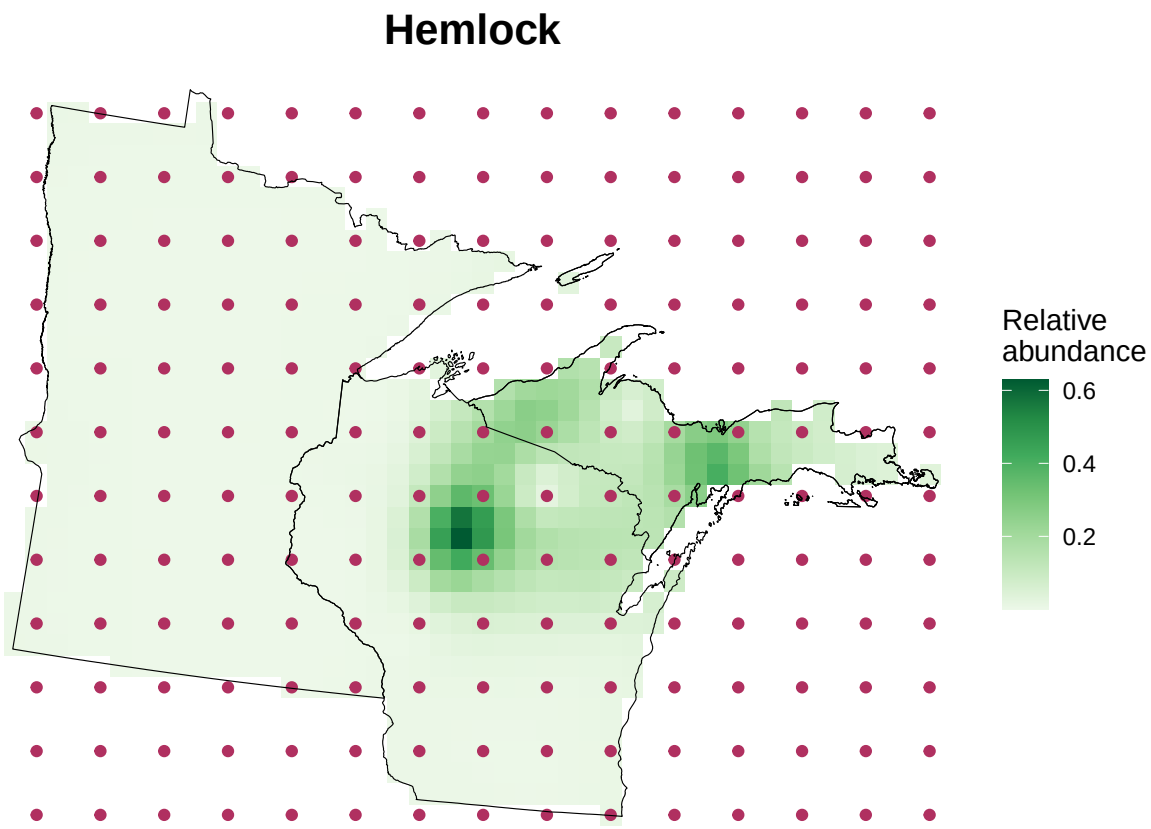
Beech


Relative
abundance
0.5
0.4
0.3
0.2
0.1

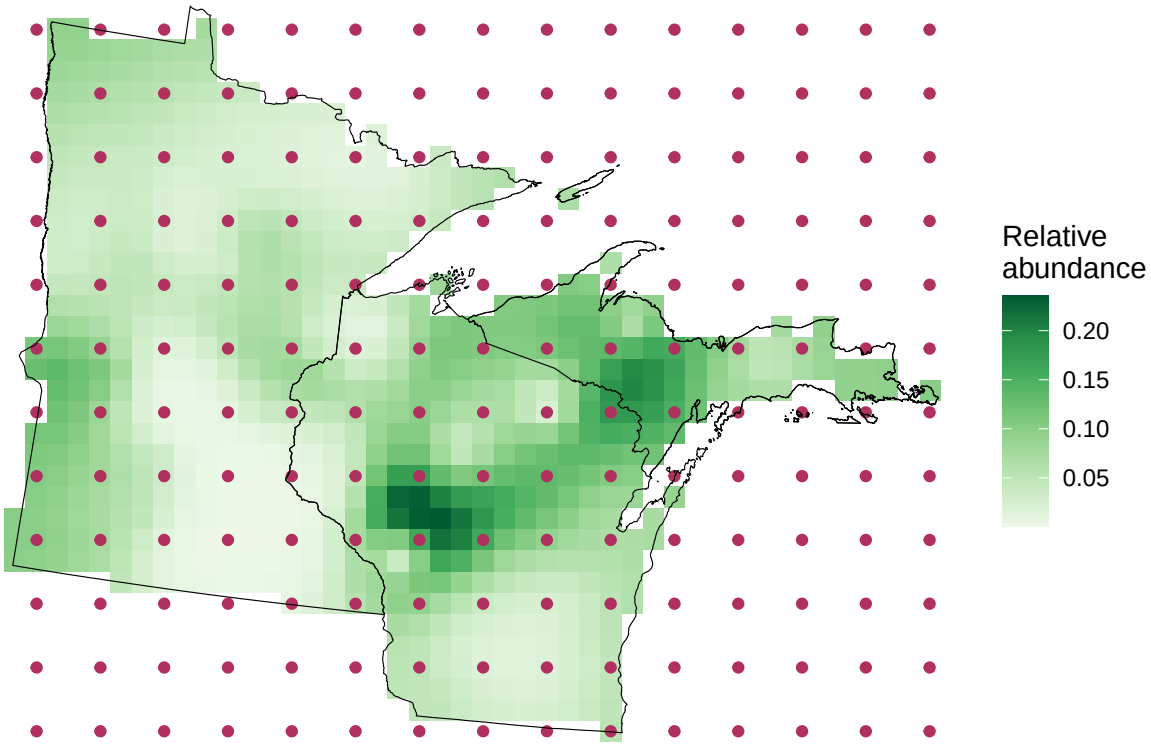
Birch


Relative
abundance
0.3
0.2
0.1

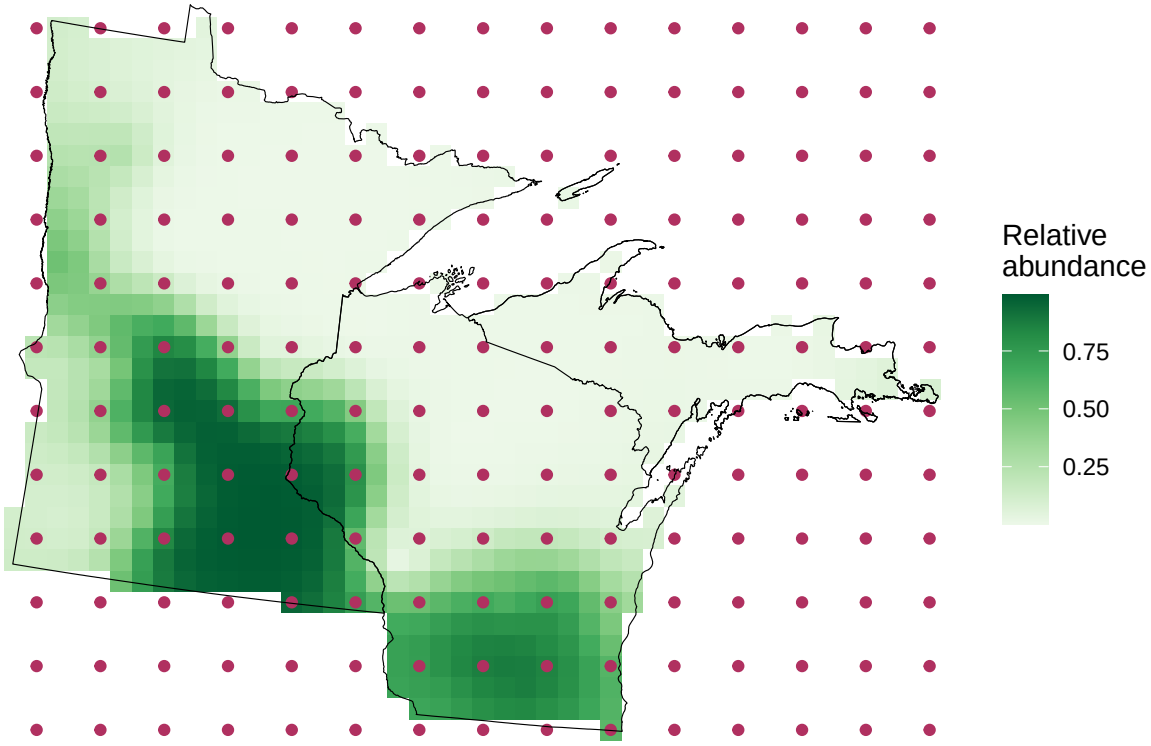

Elm
Relative abundance
0.4
0.3
0.2
0.1

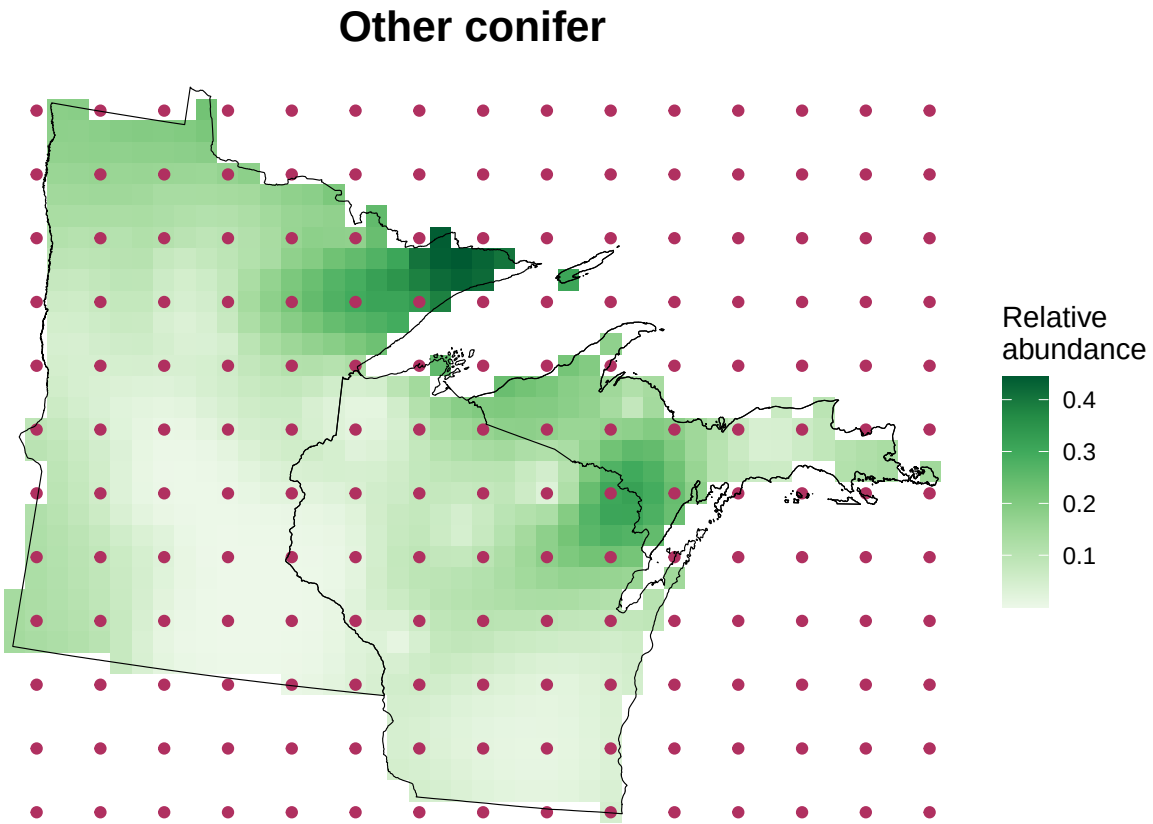

Hemlock
Relative abundance
0.6
0.4
0.2

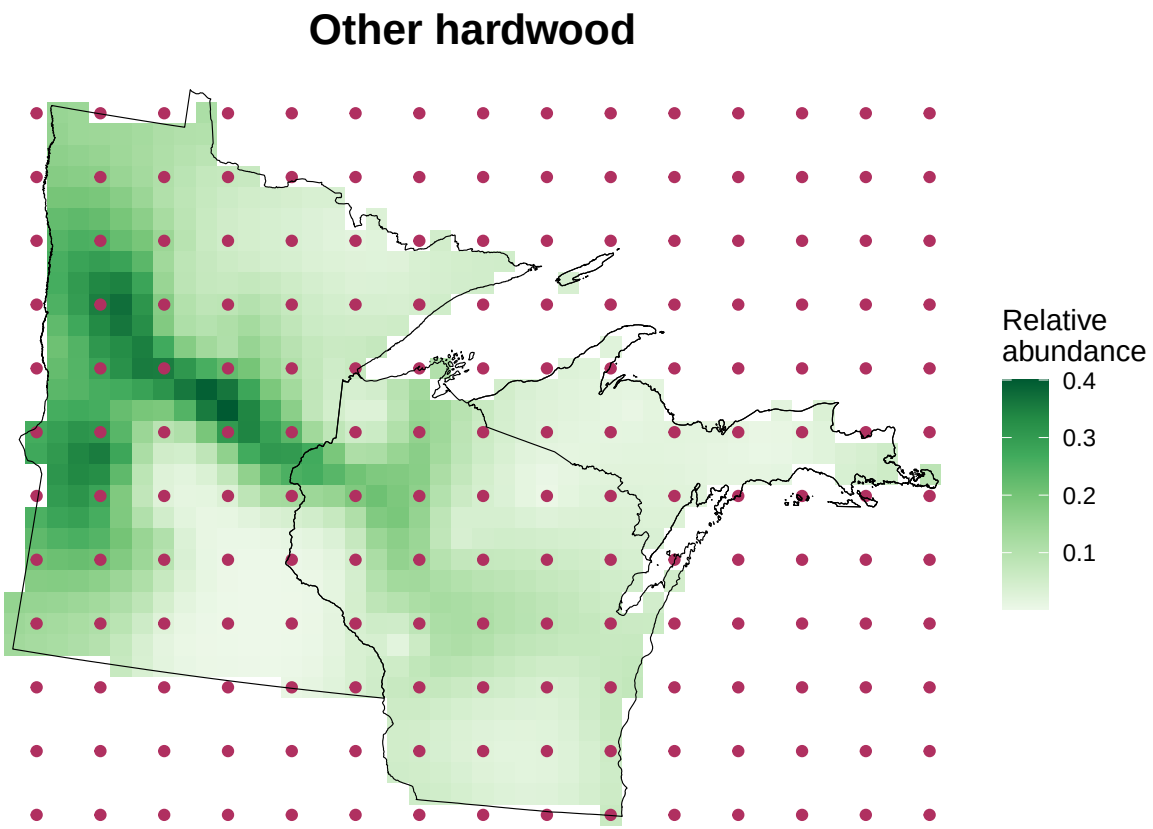
Maple

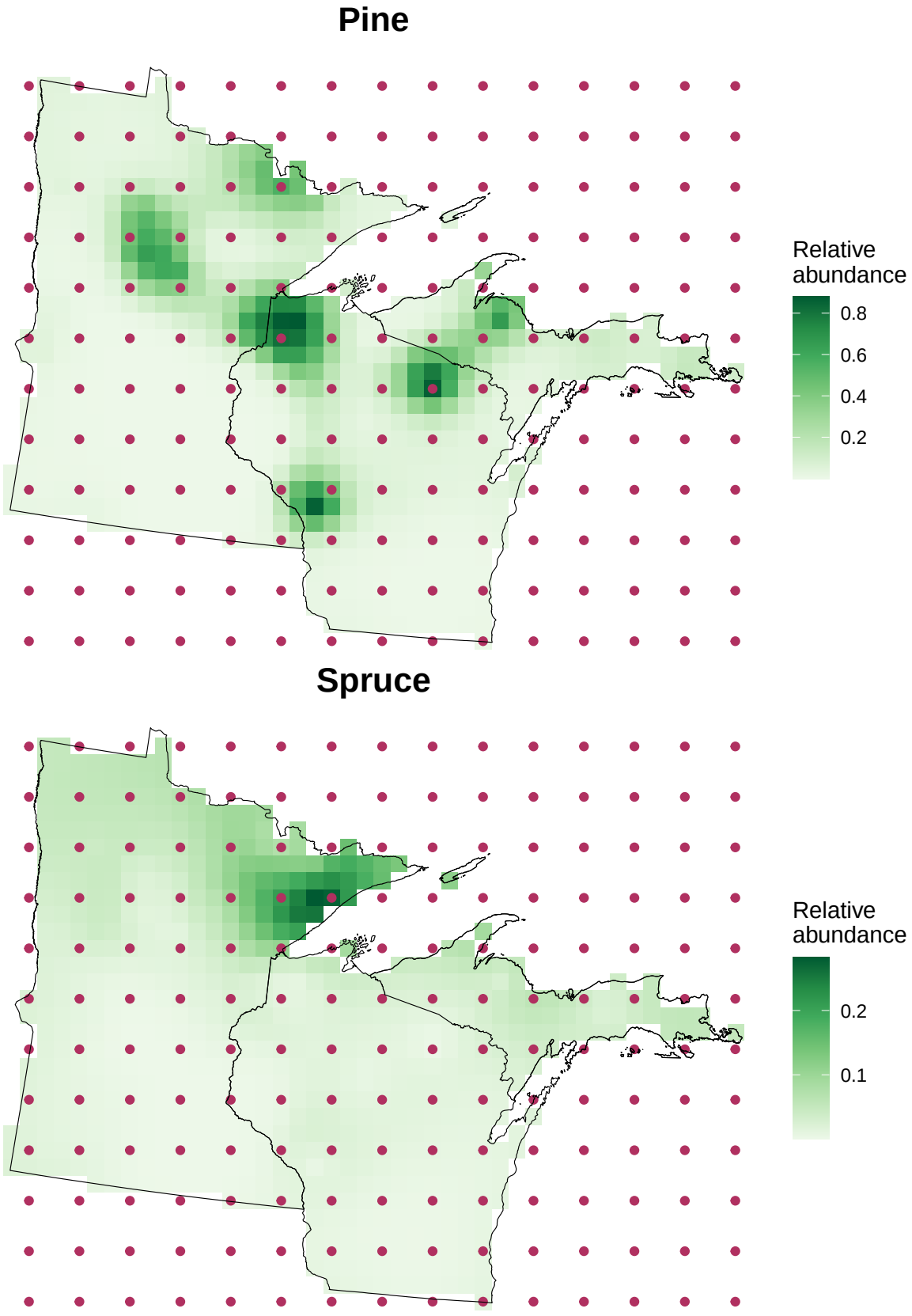

Relative
abundance
0.20
0.15
0.10
0.05

Oak

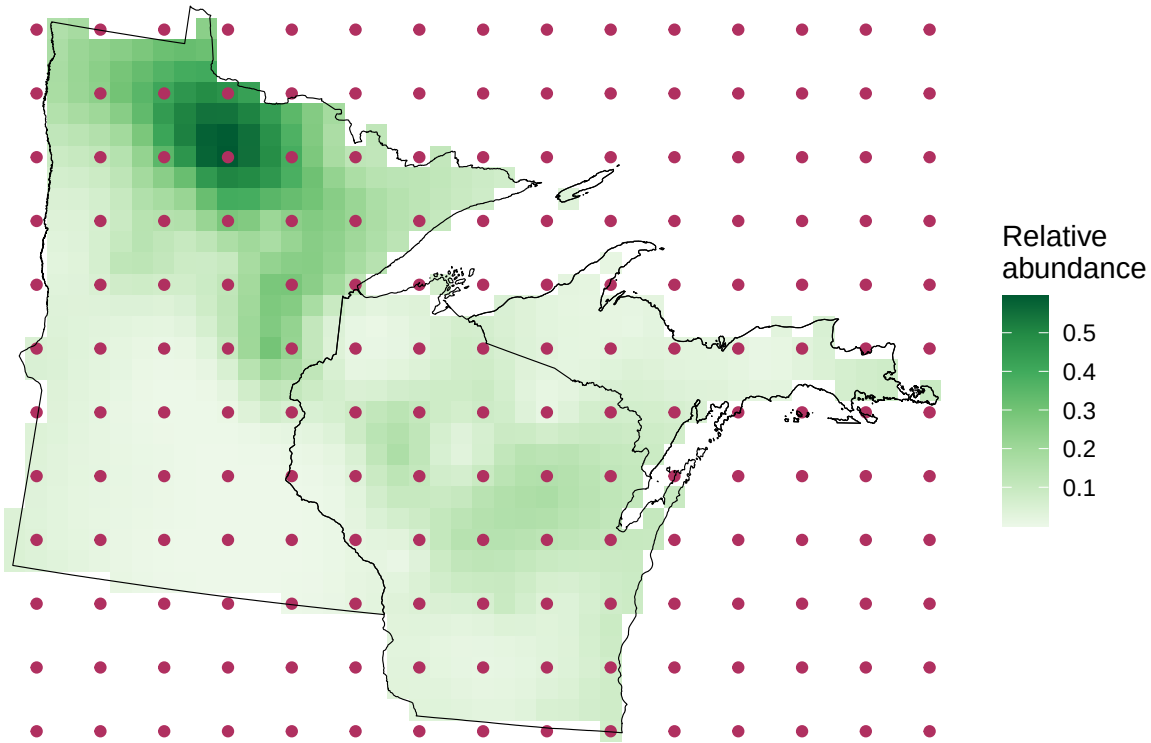

Relative
abundance
0.75
0.50
0.25


Other conifer
Relative abundance
0.4
0.3
0.2
0.1


Other hardwood
Relative abundance
0.4
0.3
0.2
0.1

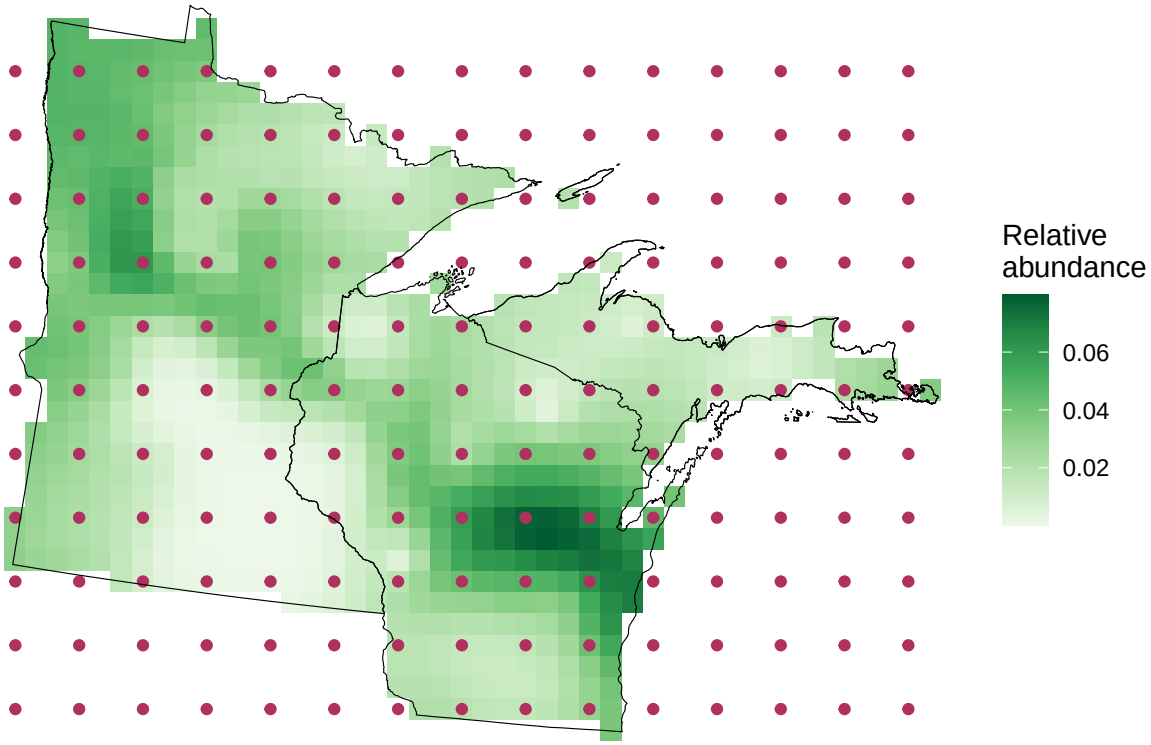

Pine
Relative abundance
0.8
0.6
0.4
0.2
Spruce
Relative abundance
0.2
0.1

Tamarack

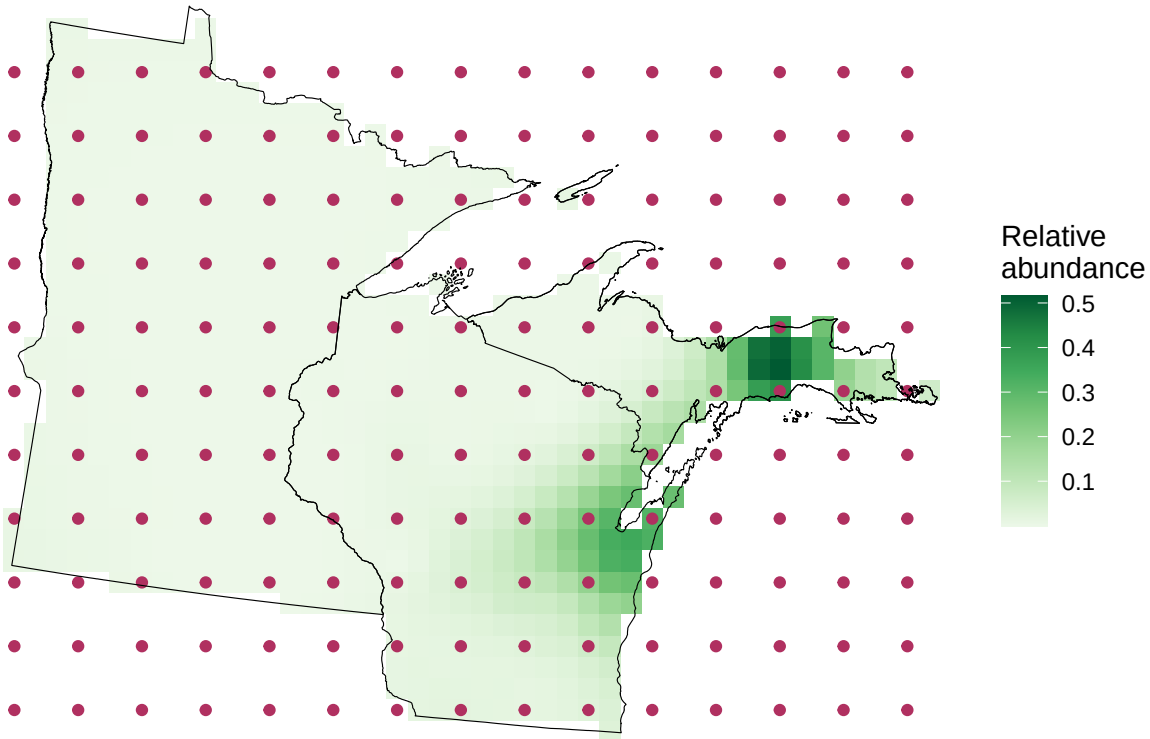


$\mathbf{x = 1, y = 2}$

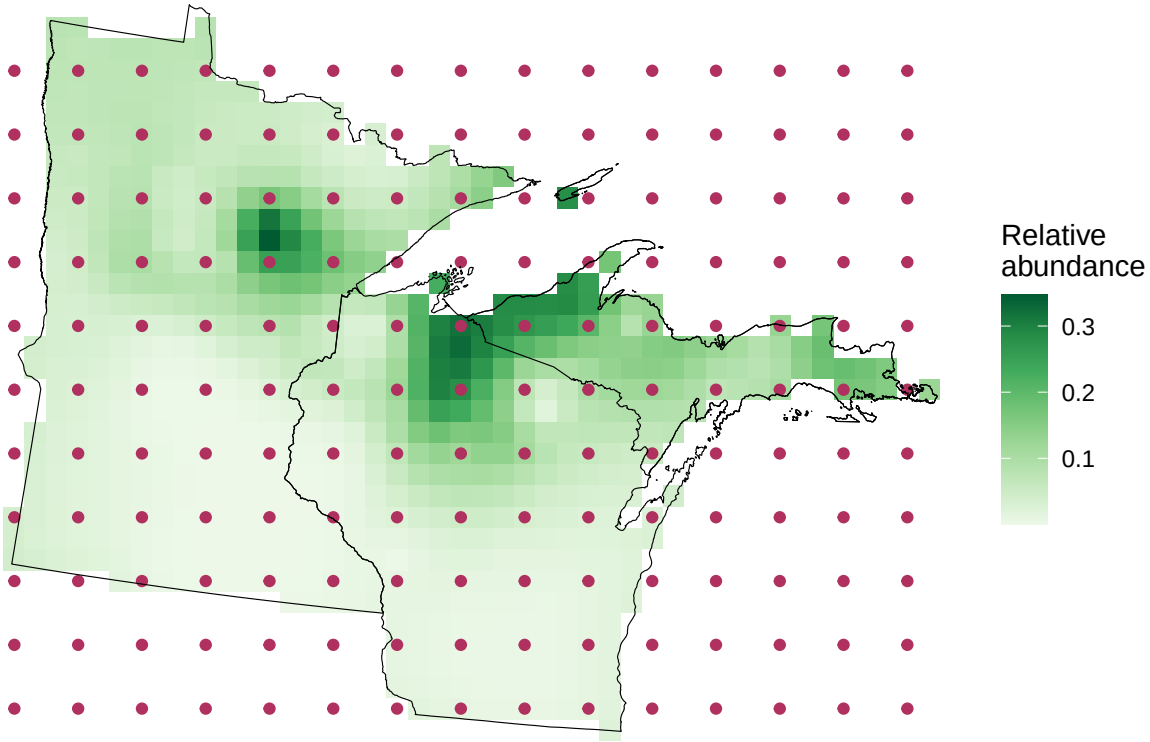
Ash



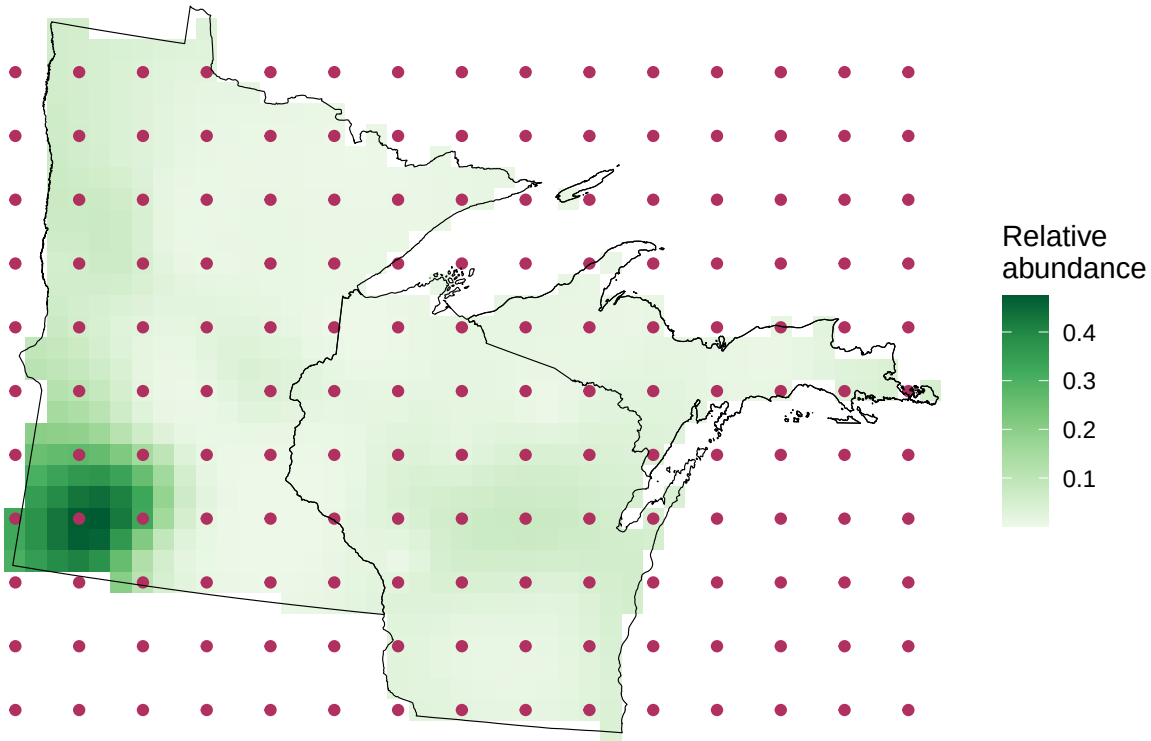
Beech


Relative
abundance
0.5
0.4
0.3
0.2
0.1

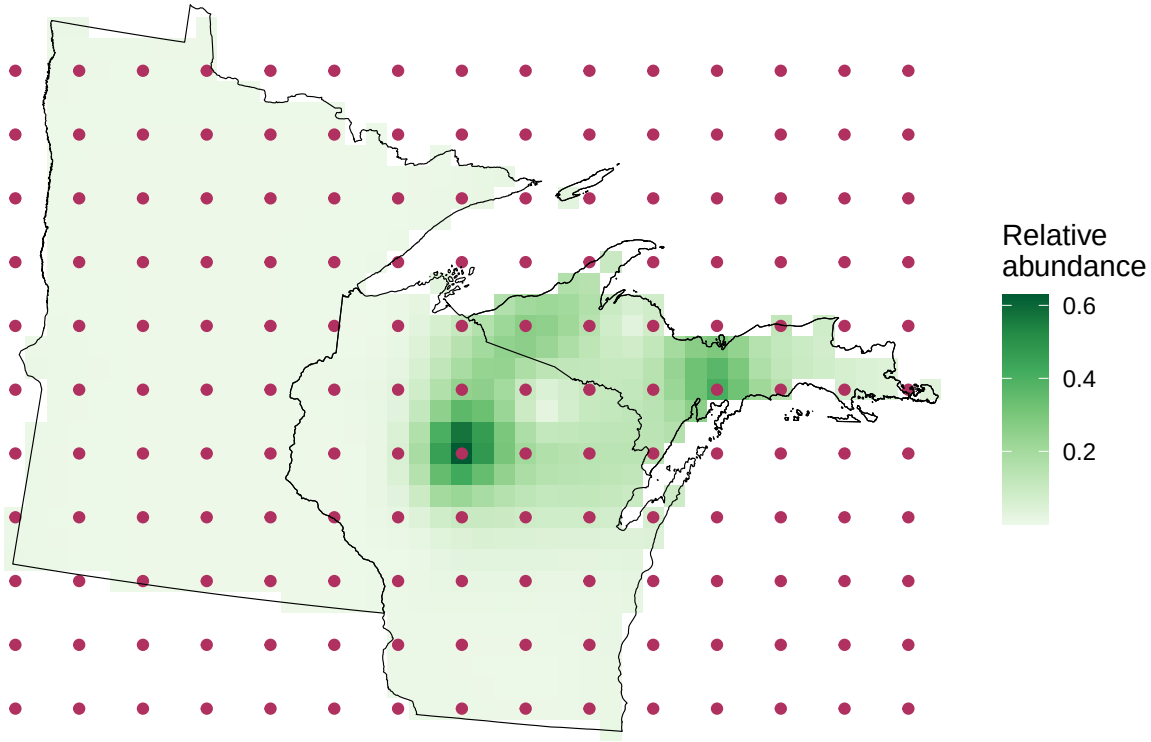
Birch


Relative
abundance
0.3
0.2
0.1

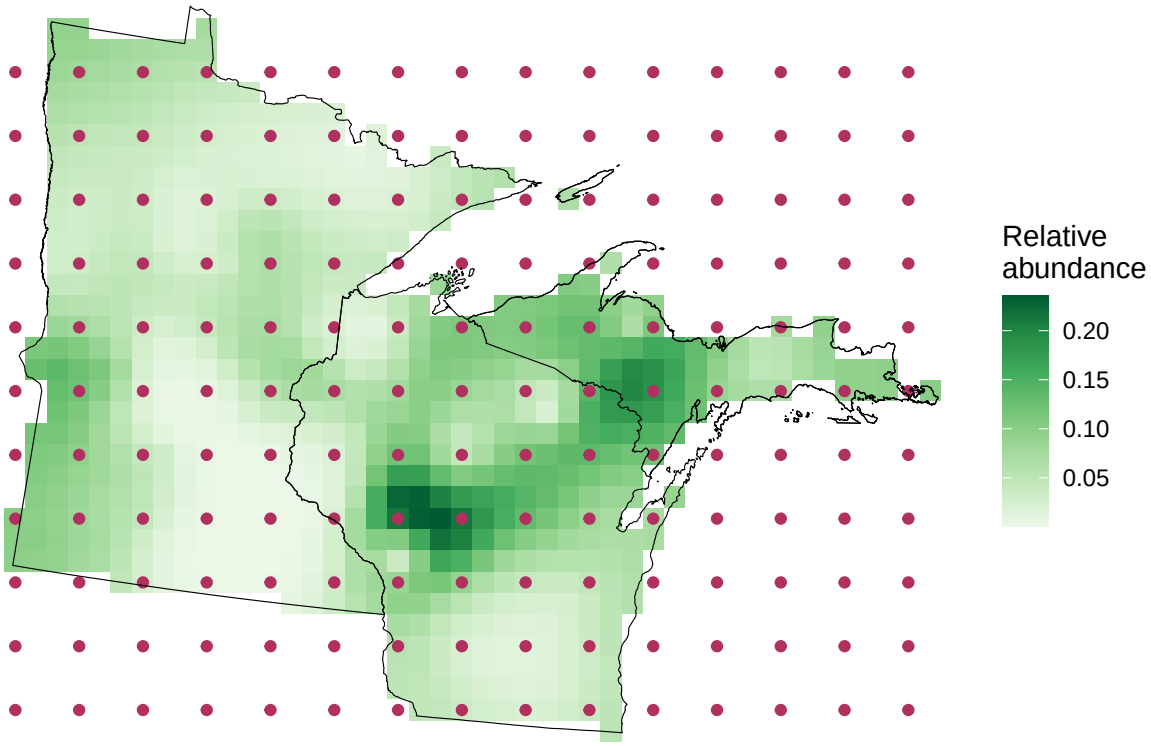
Elm



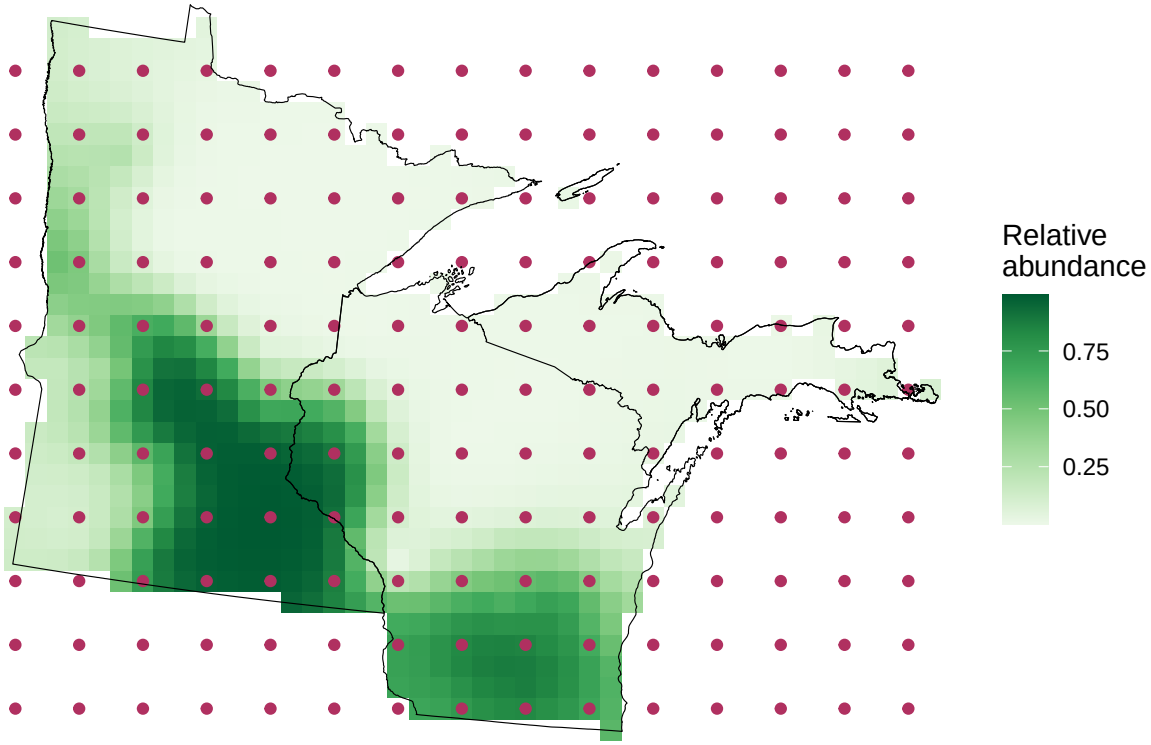
Hemlock



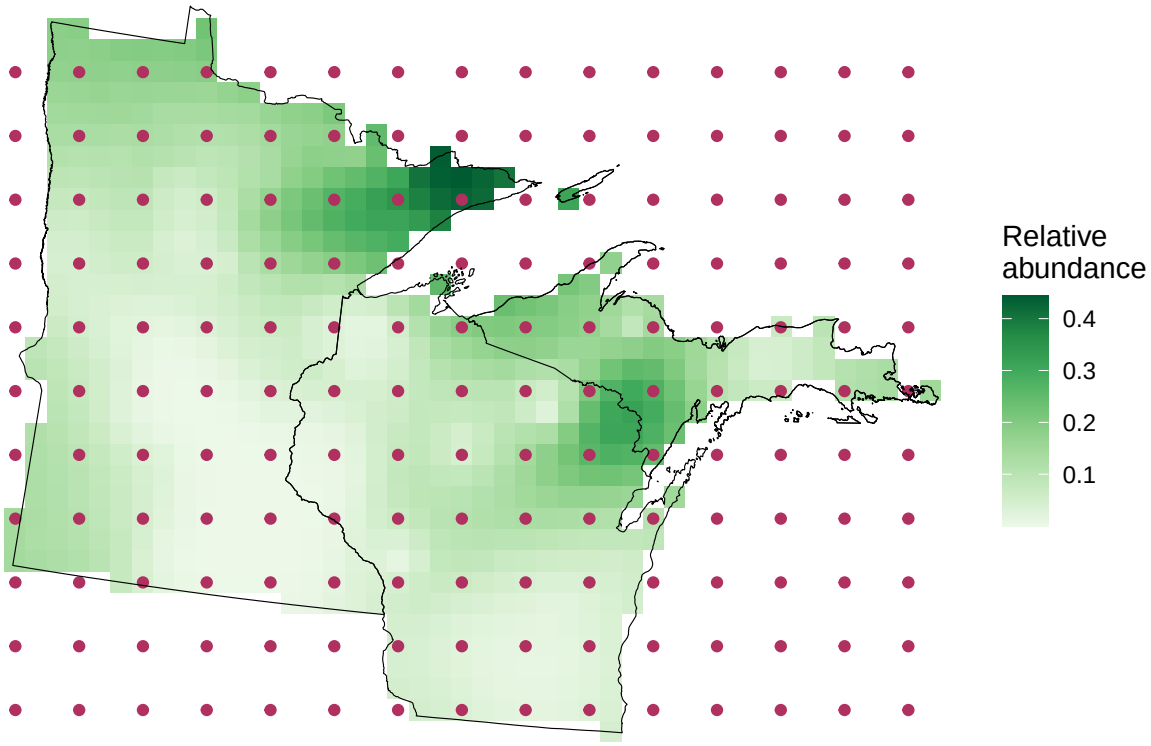
Maple


Relative
abundance
0.20
0.15
0.10
0.05

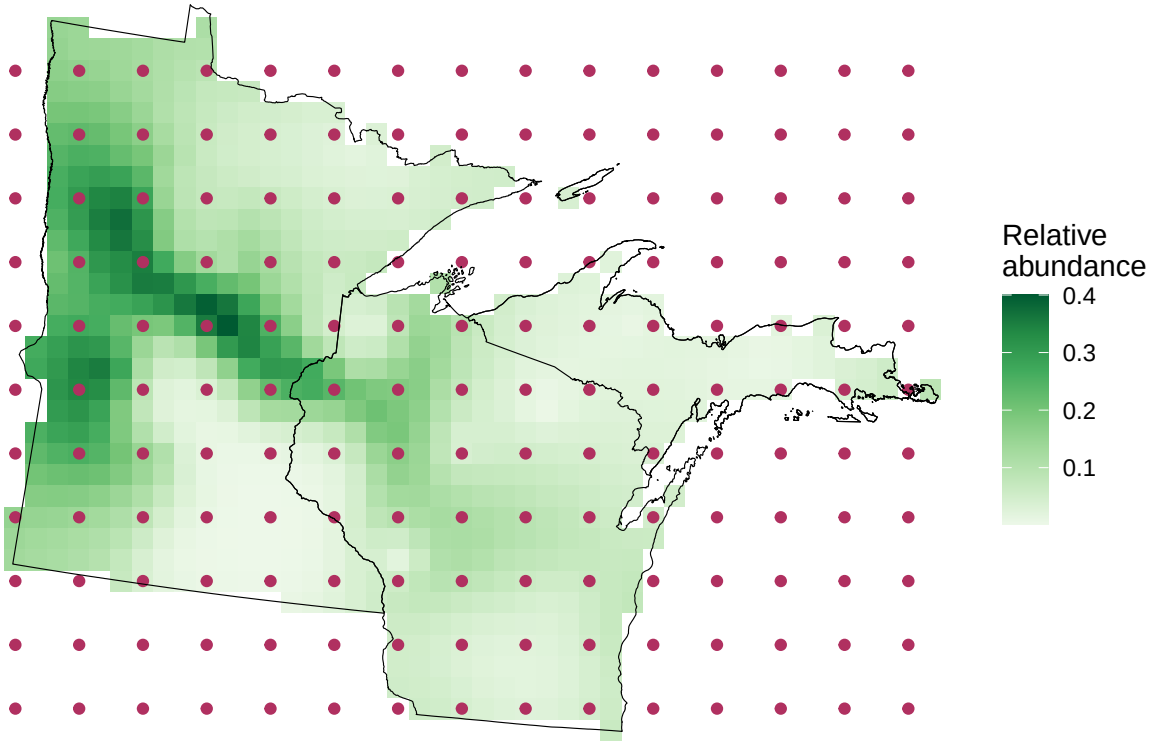
Oak


Relative
abundance
0.75
0.50
0.25

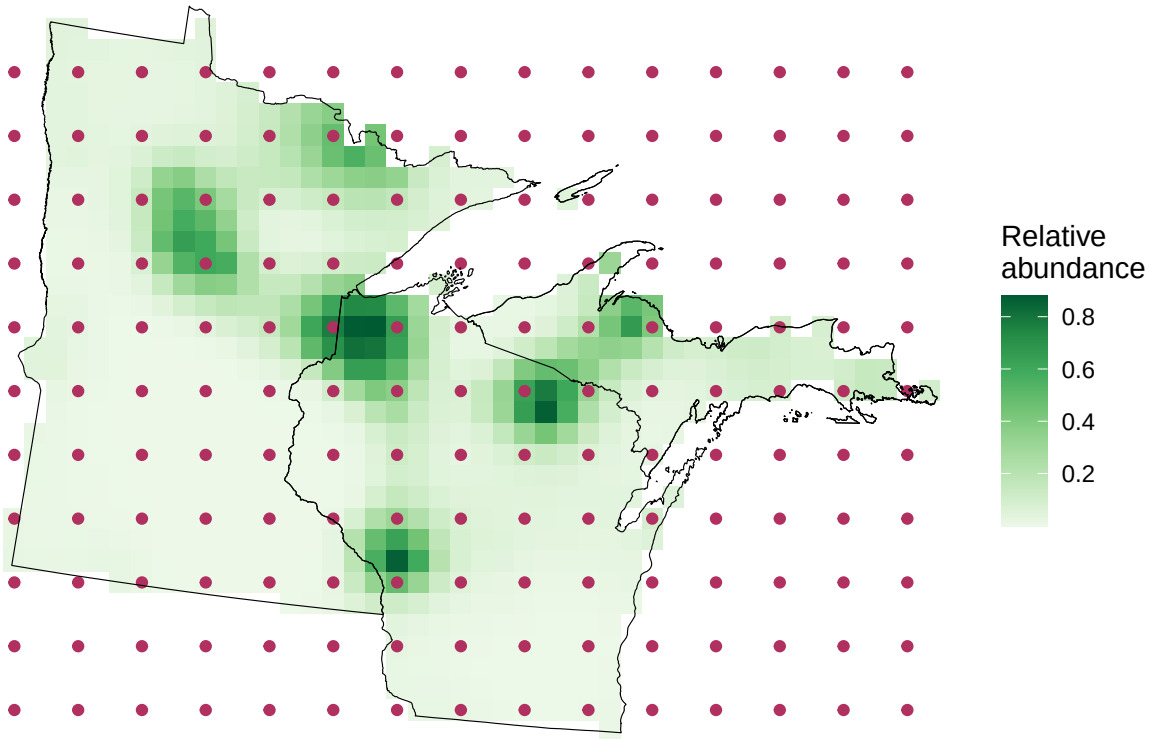
Other conifer



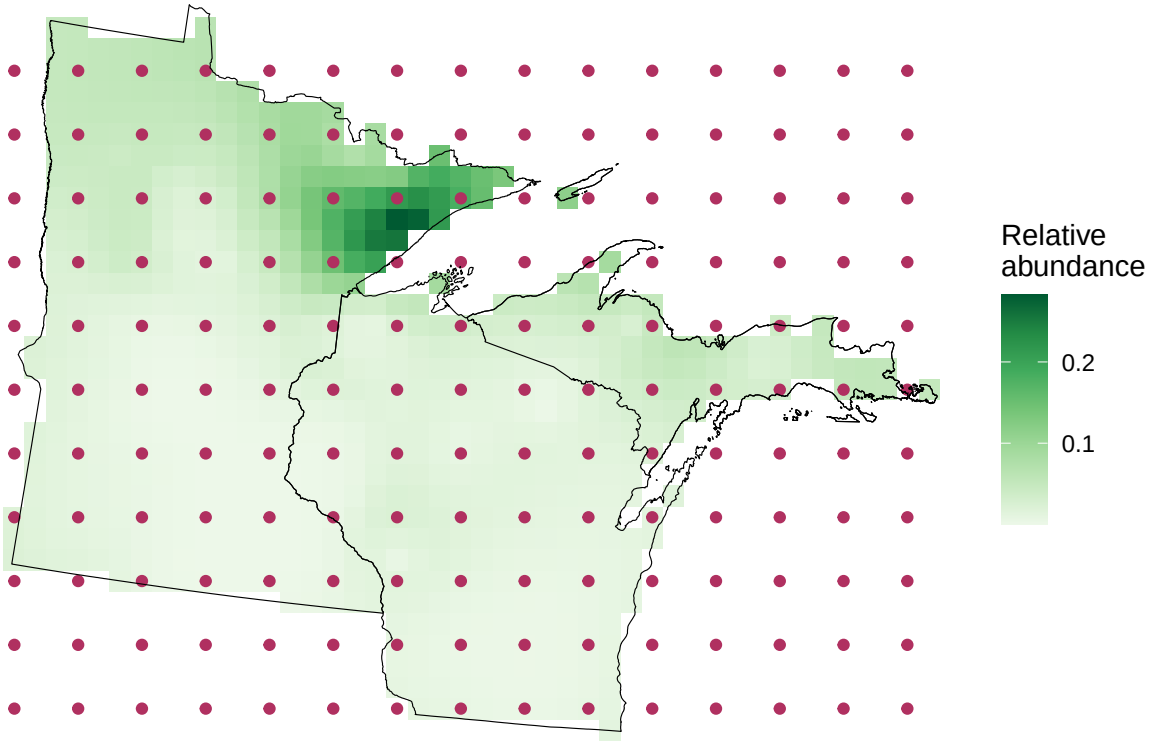
Other hardwood



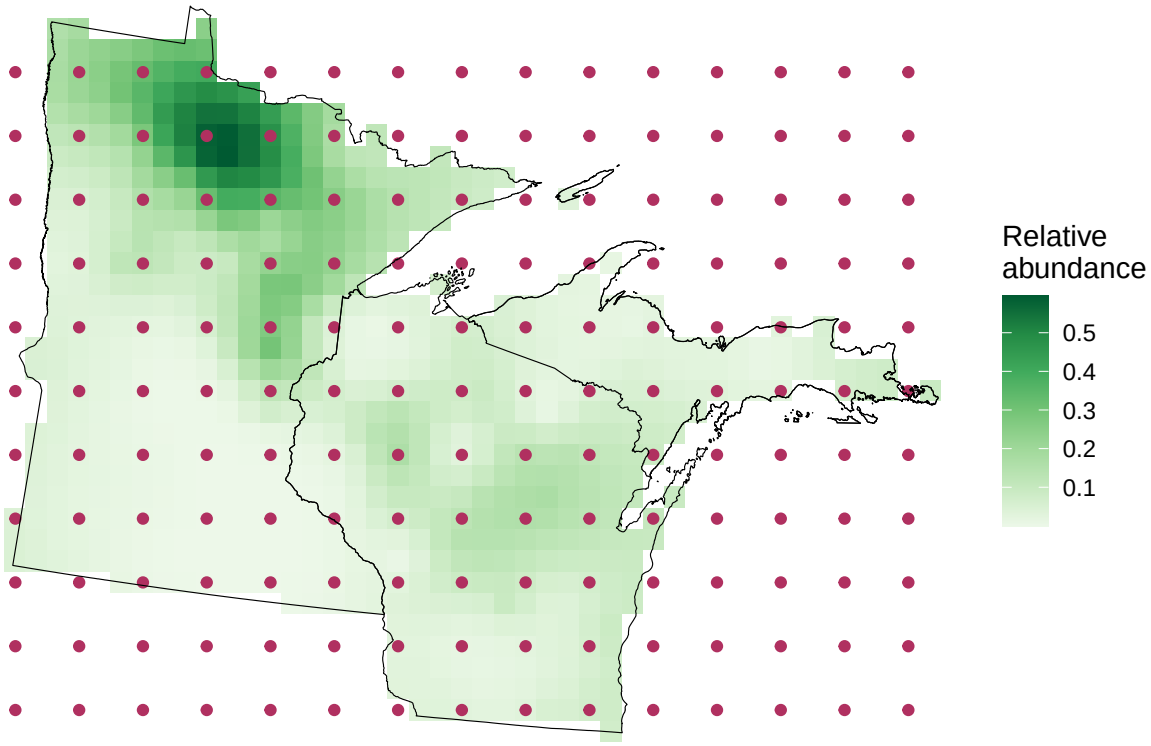
Pine



Spruce

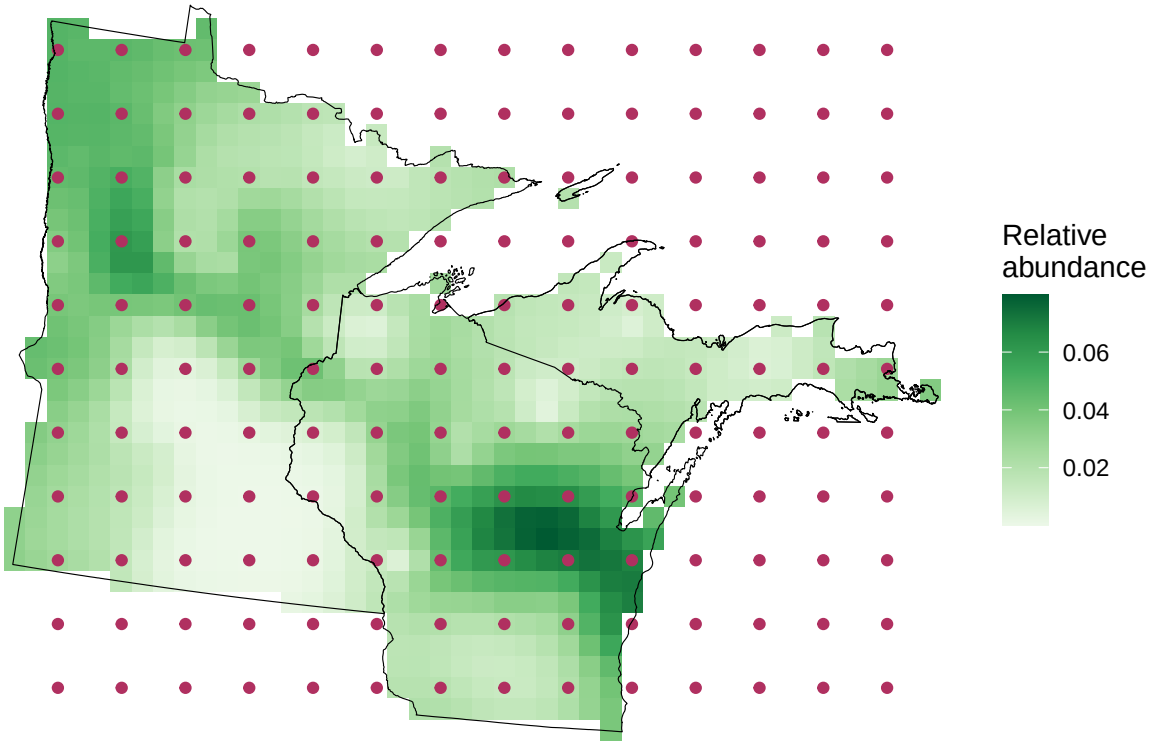


Tamarack

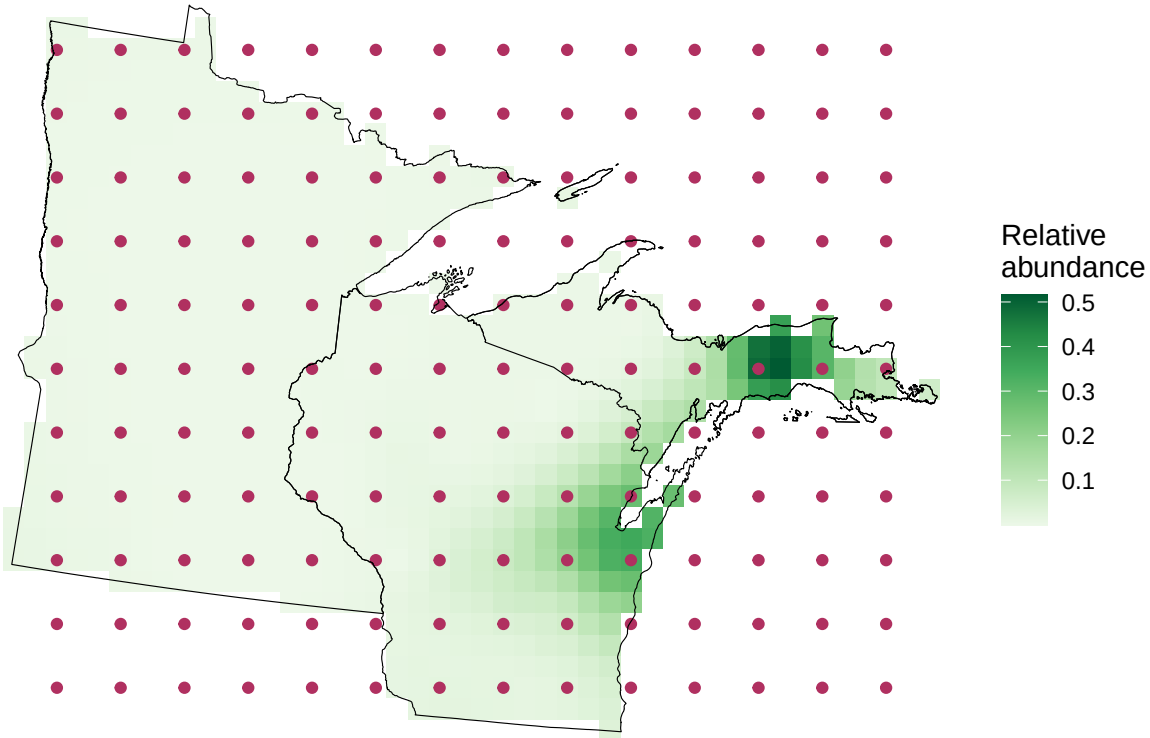


$\mathbf{x = 3, y = 3}$

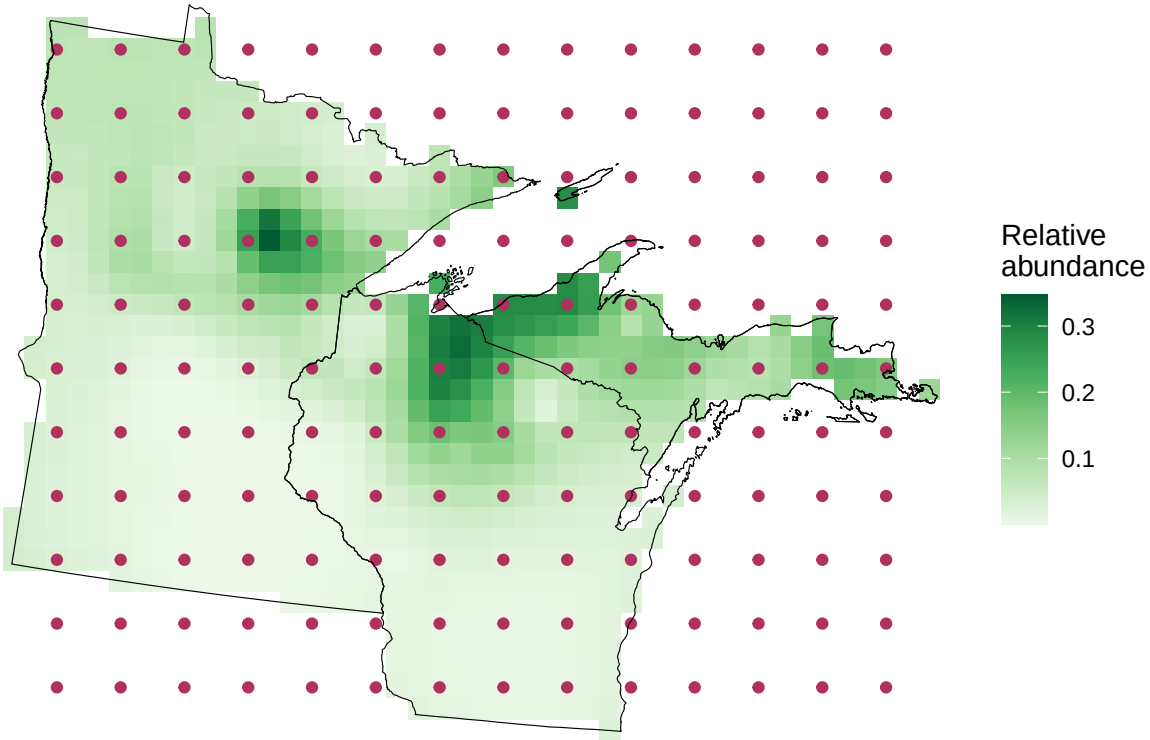
Ash



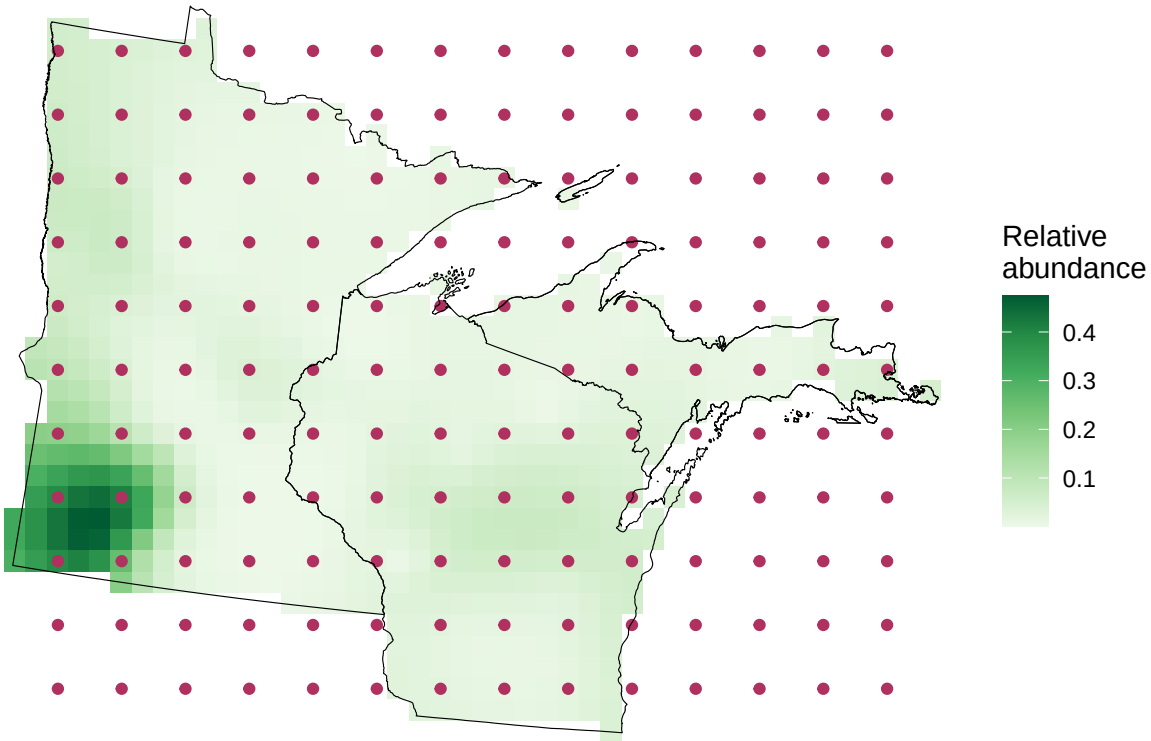
Beech


Relative
abundance
0.5
0.4
0.3
0.2
0.1

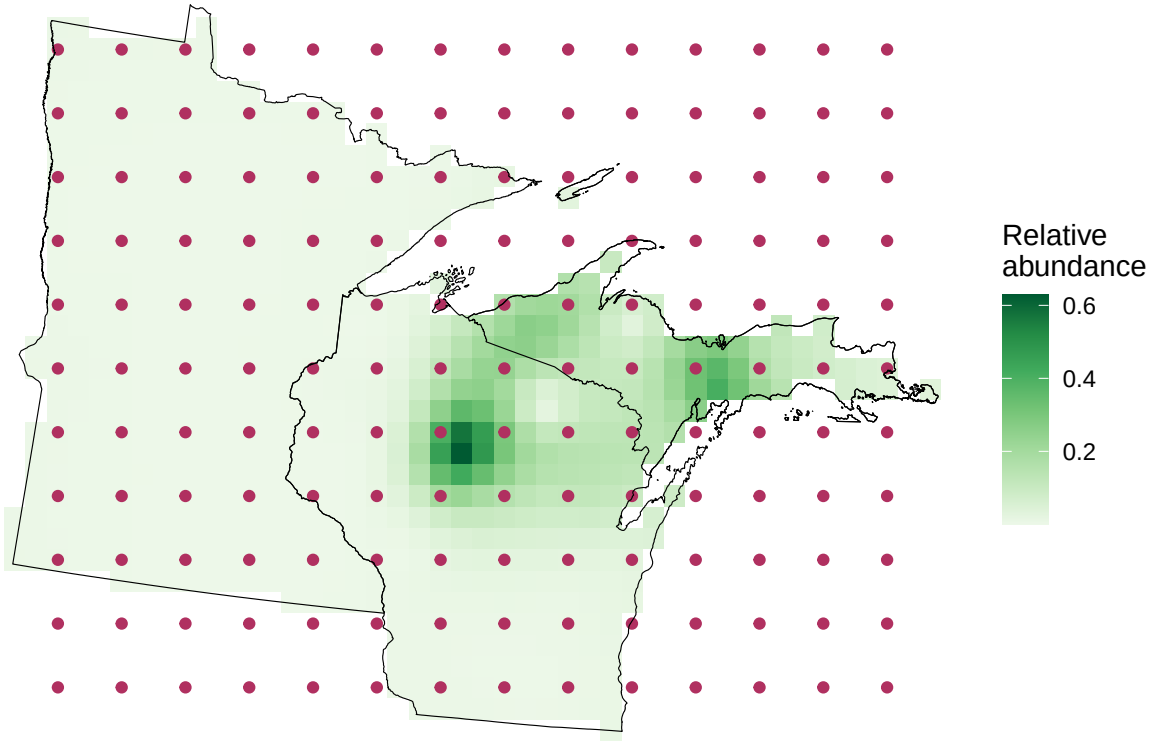
Birch


Relative
abundance
0.3
0.2
0.1

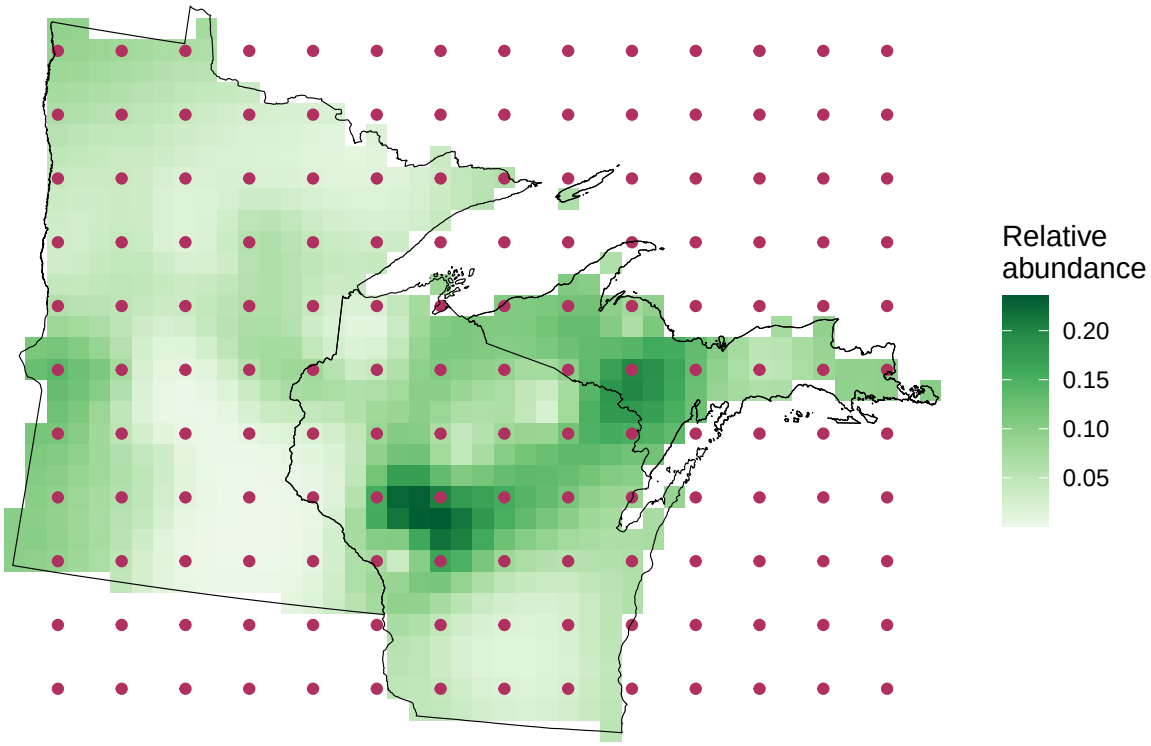
Elm



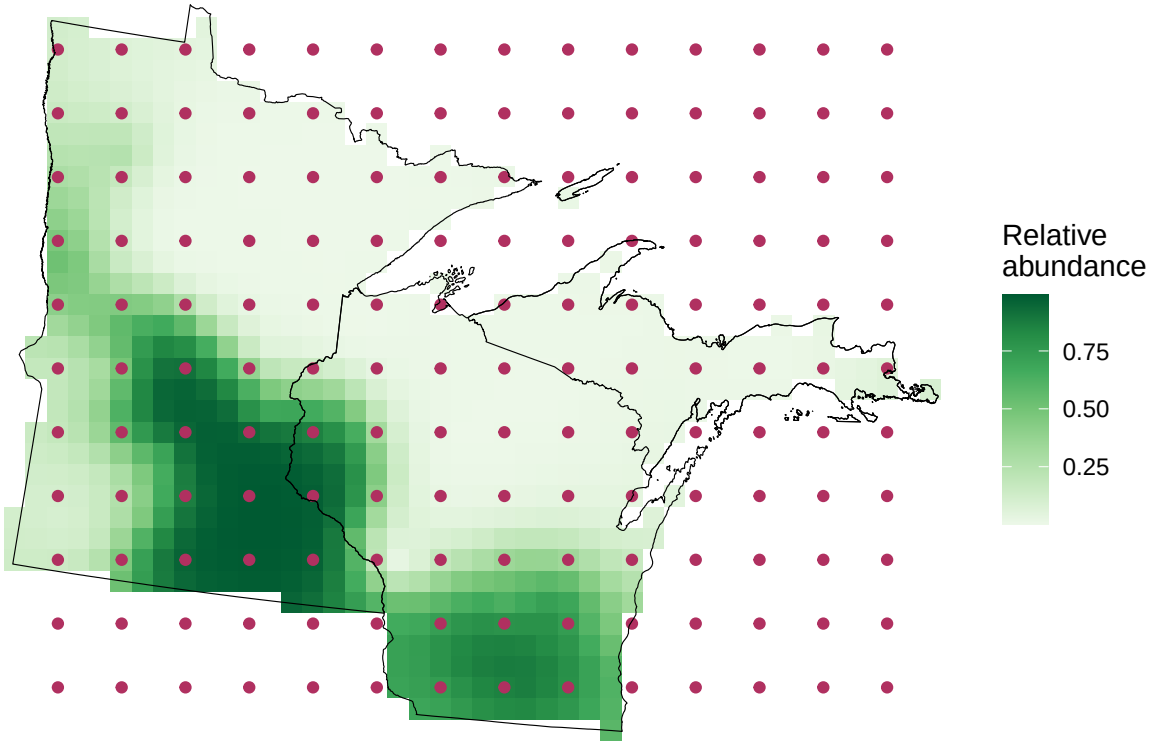
Hemlock



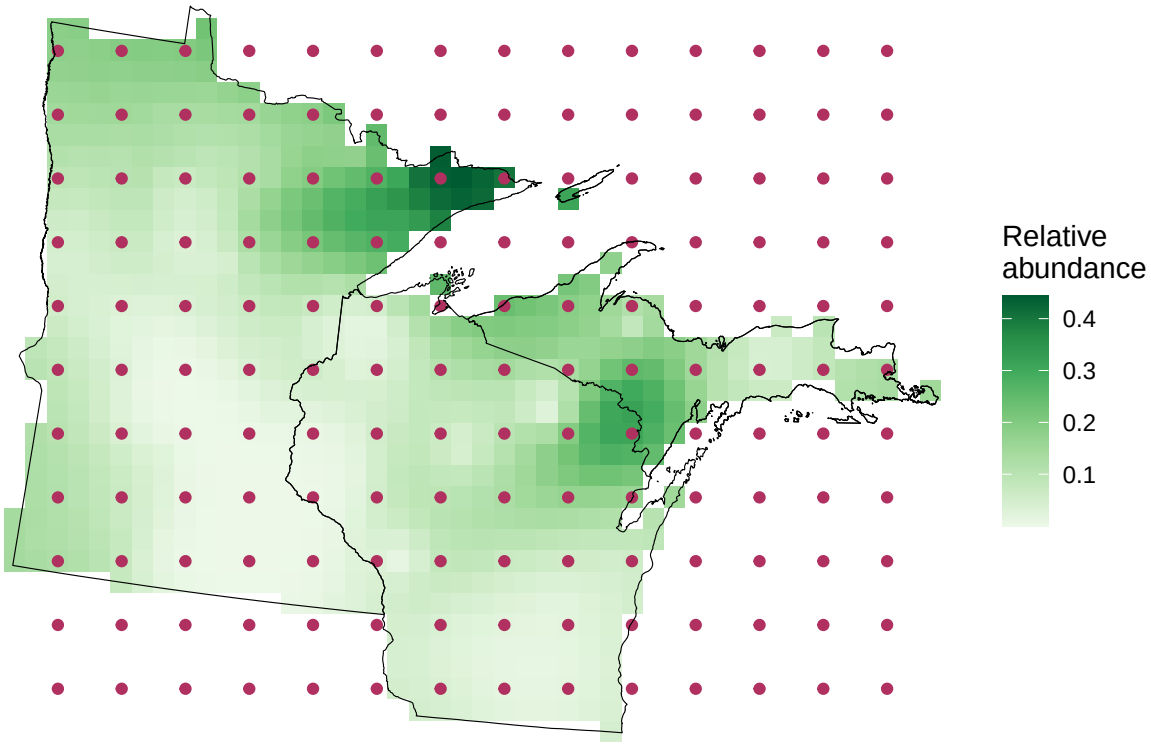
Maple


Relative
abundance
0.20
0.15
0.10
0.05

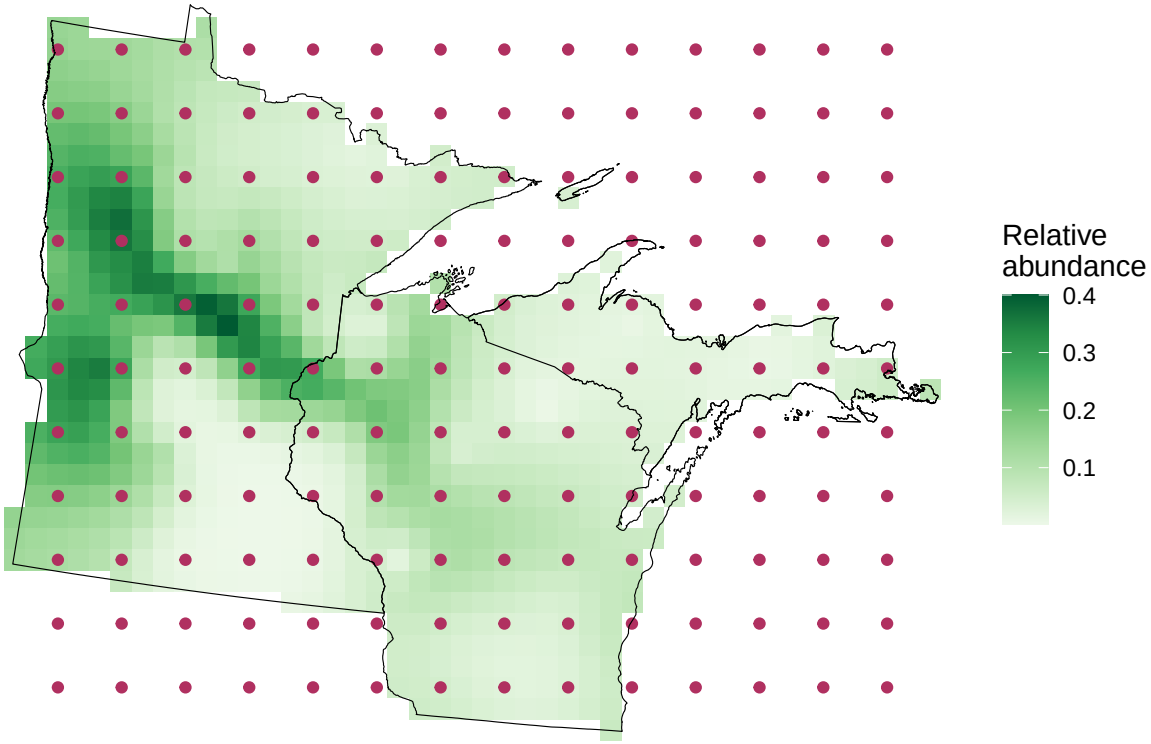
Oak


Relative
abundance
0.75
0.50
0.25

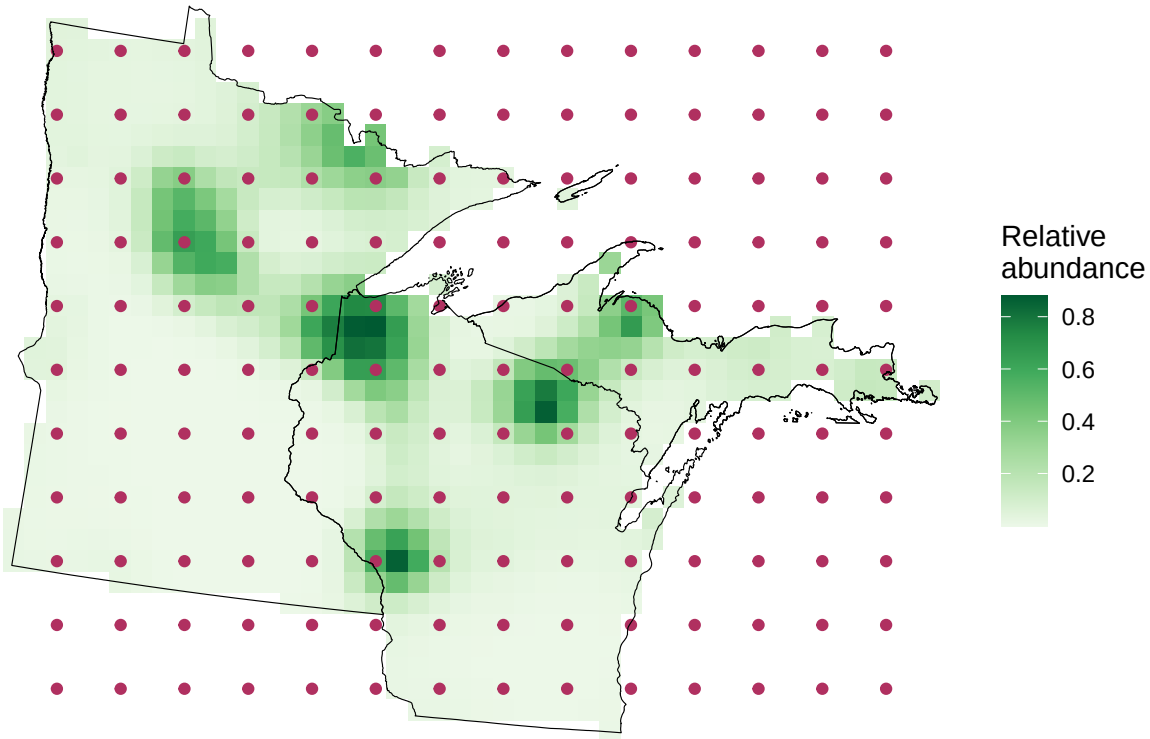
Other conifer



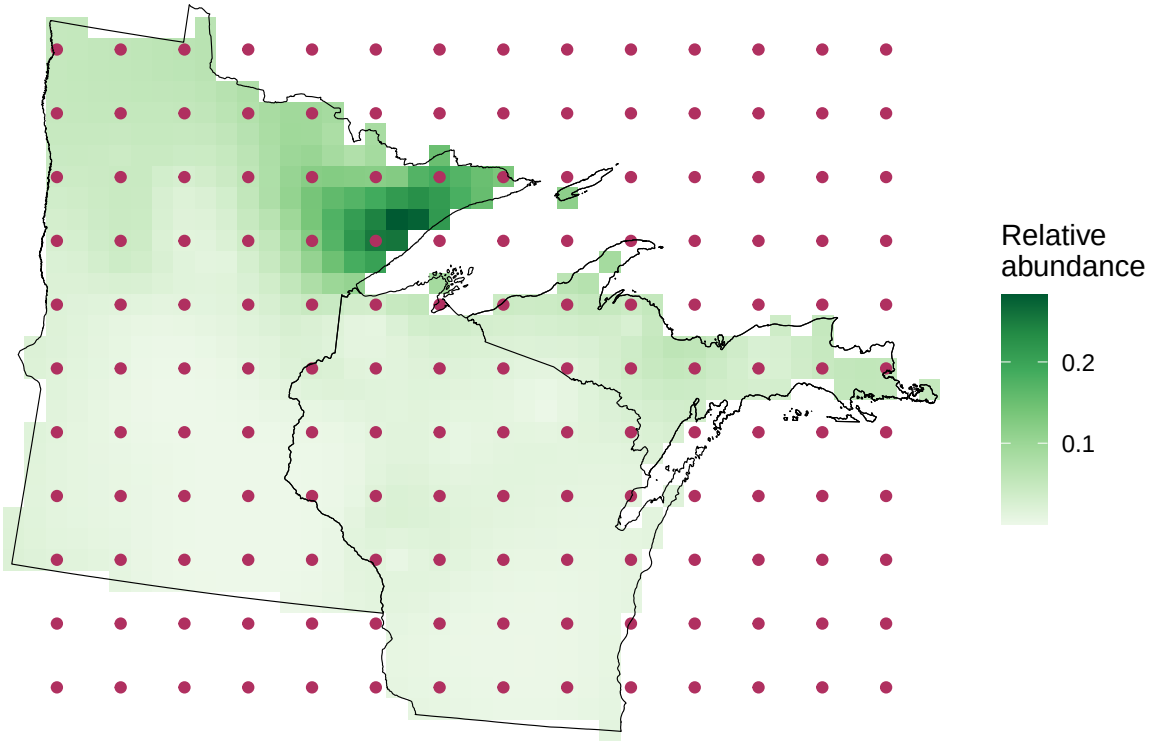
Other hardwood



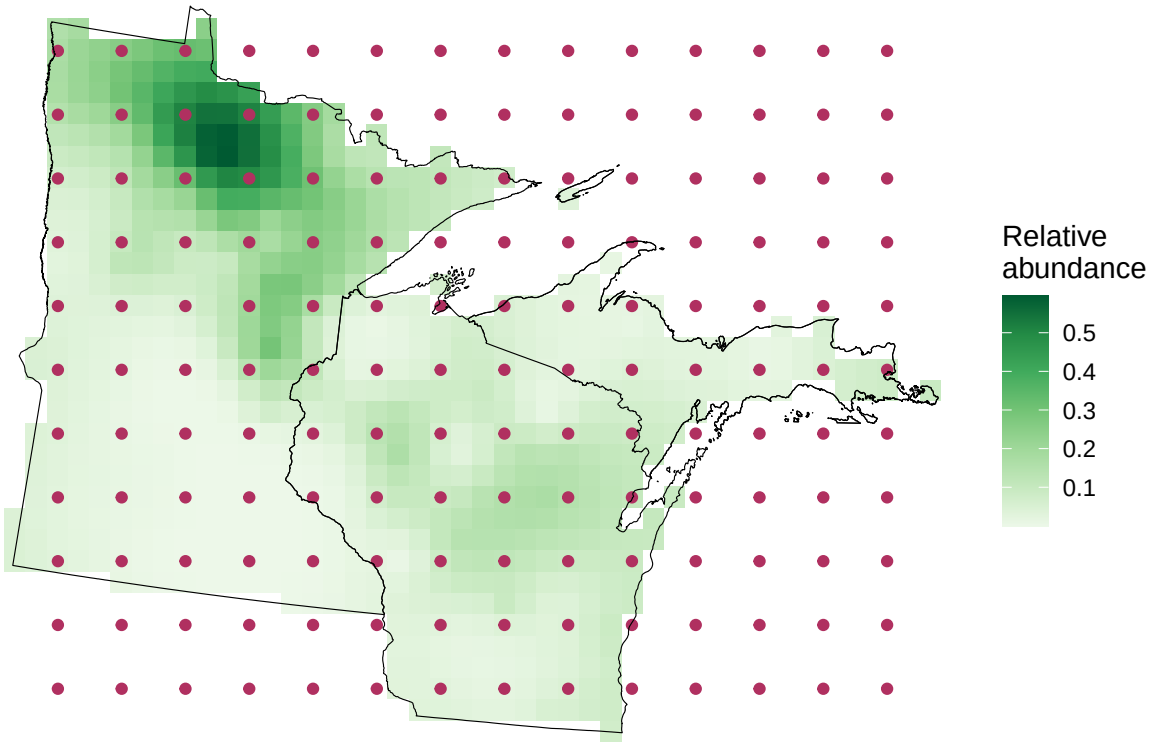
Pine



Spruce

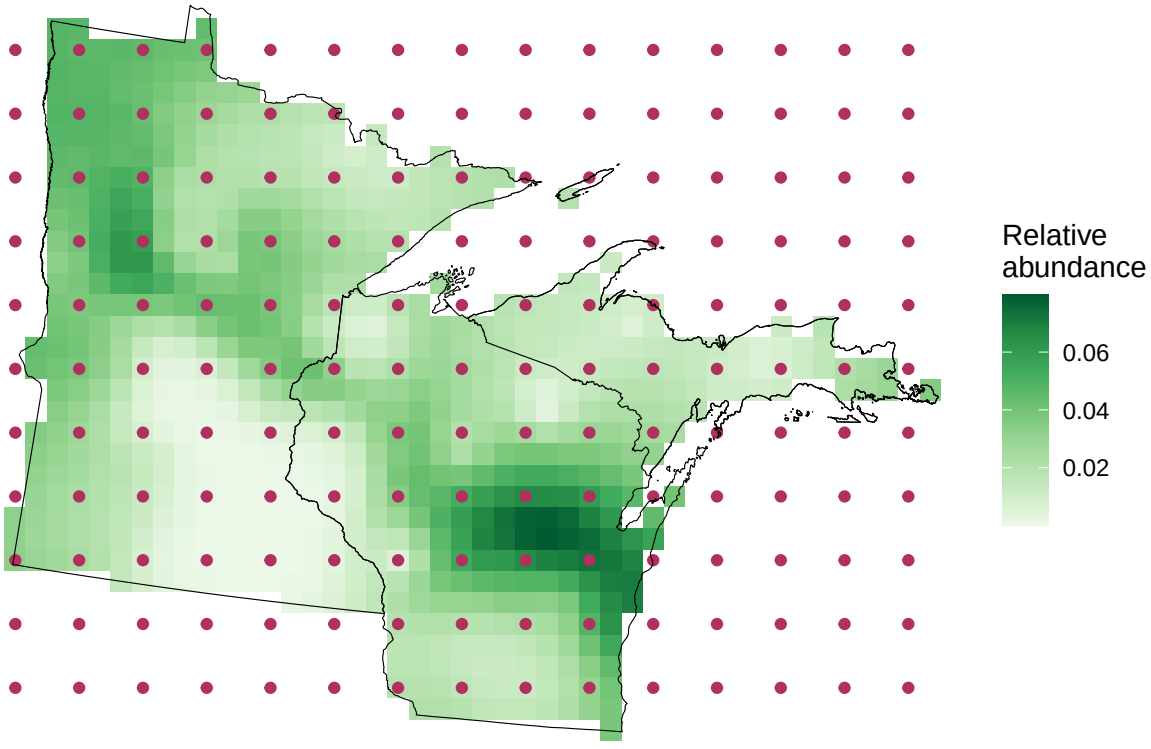


Tamarack

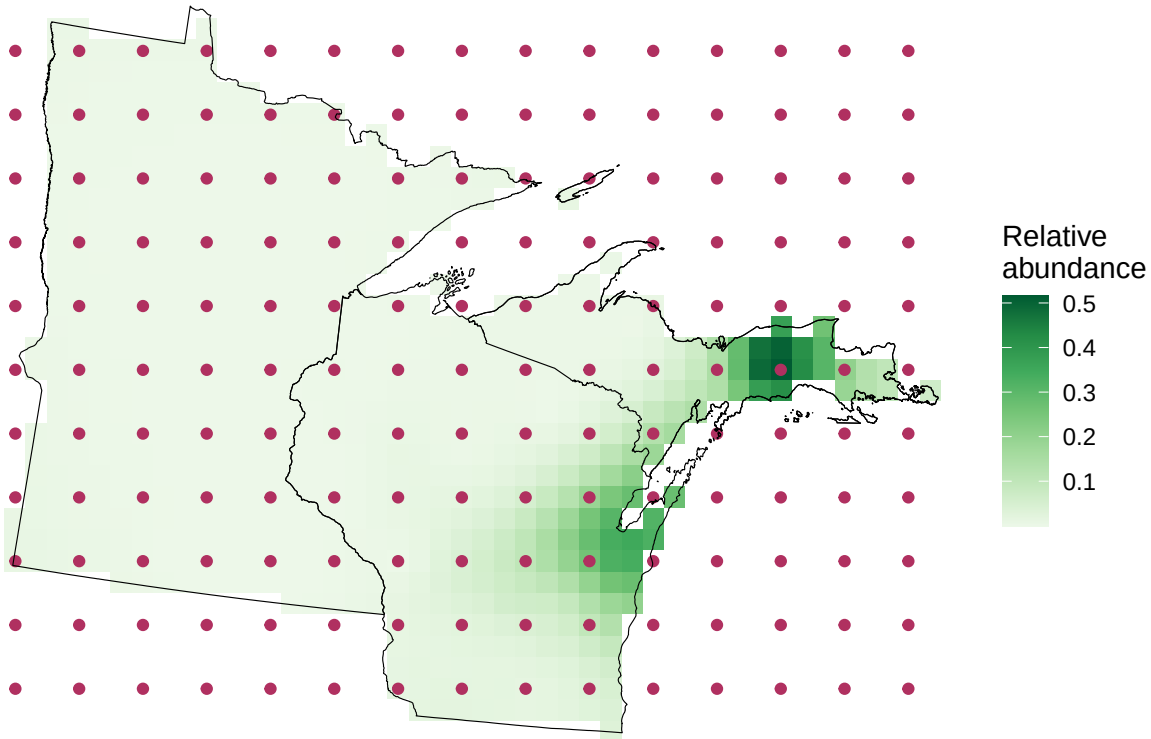


$\mathbf{x = 1, y = 3}$

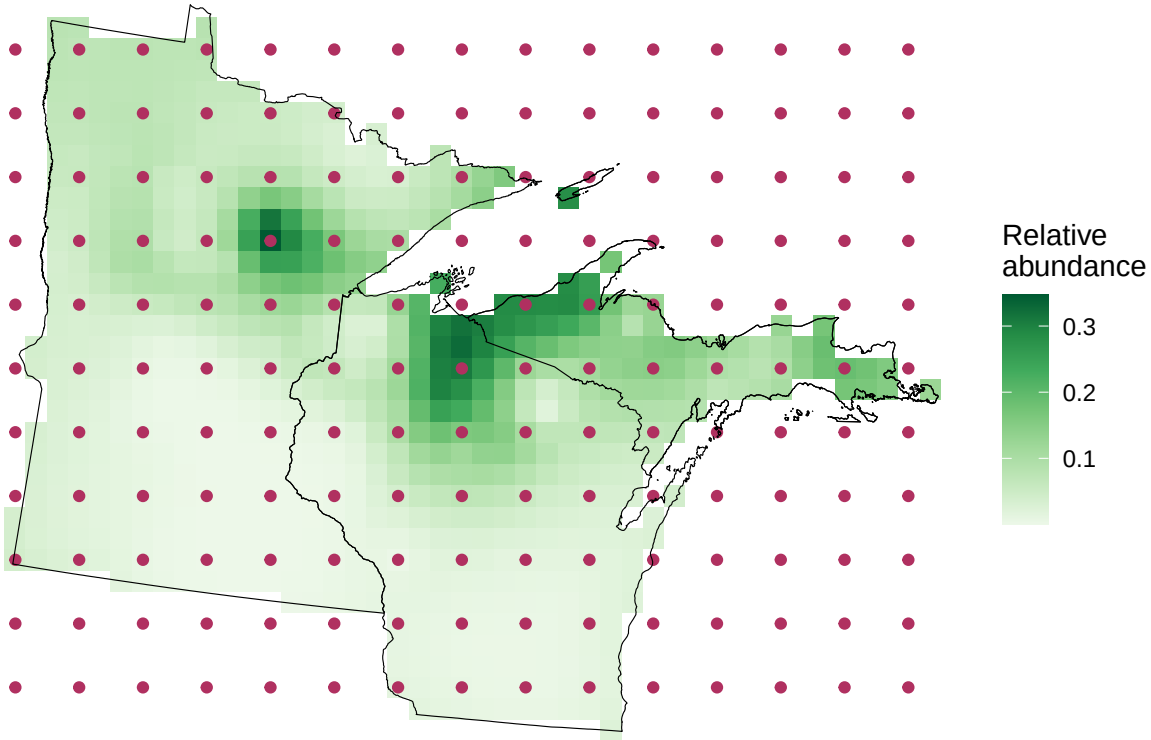
Ash



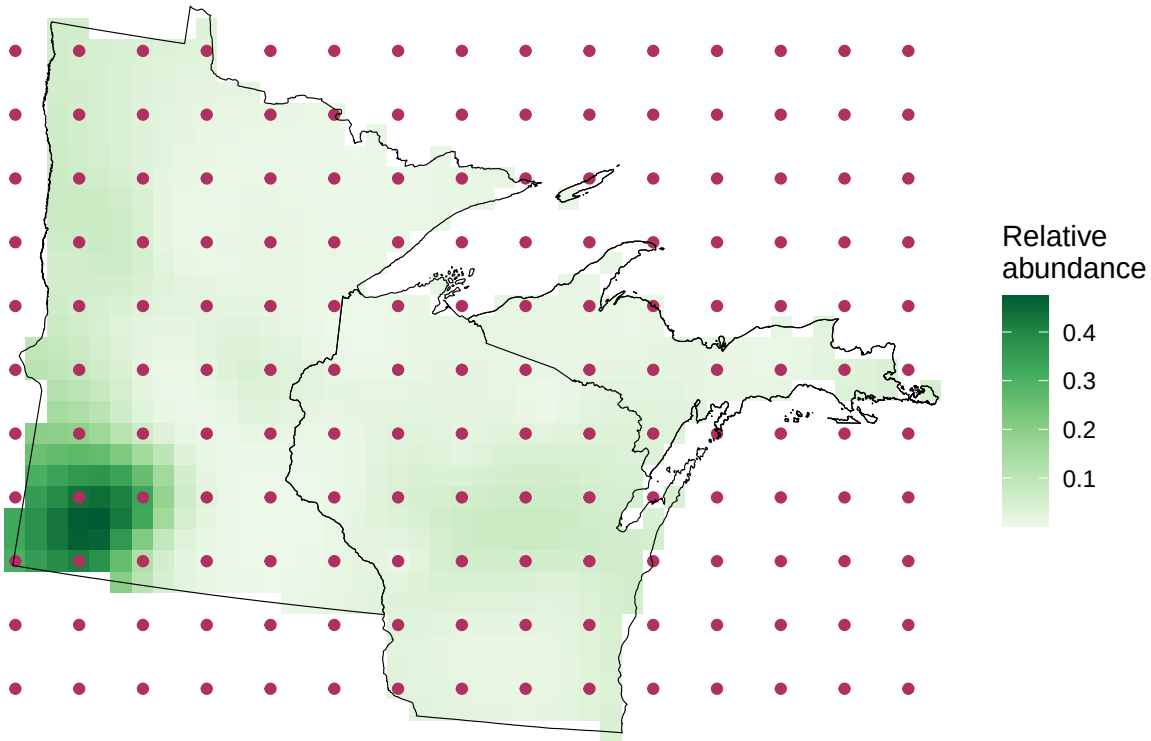
Beech


Relative
abundance
0.5
0.4
0.3
0.2
0.1

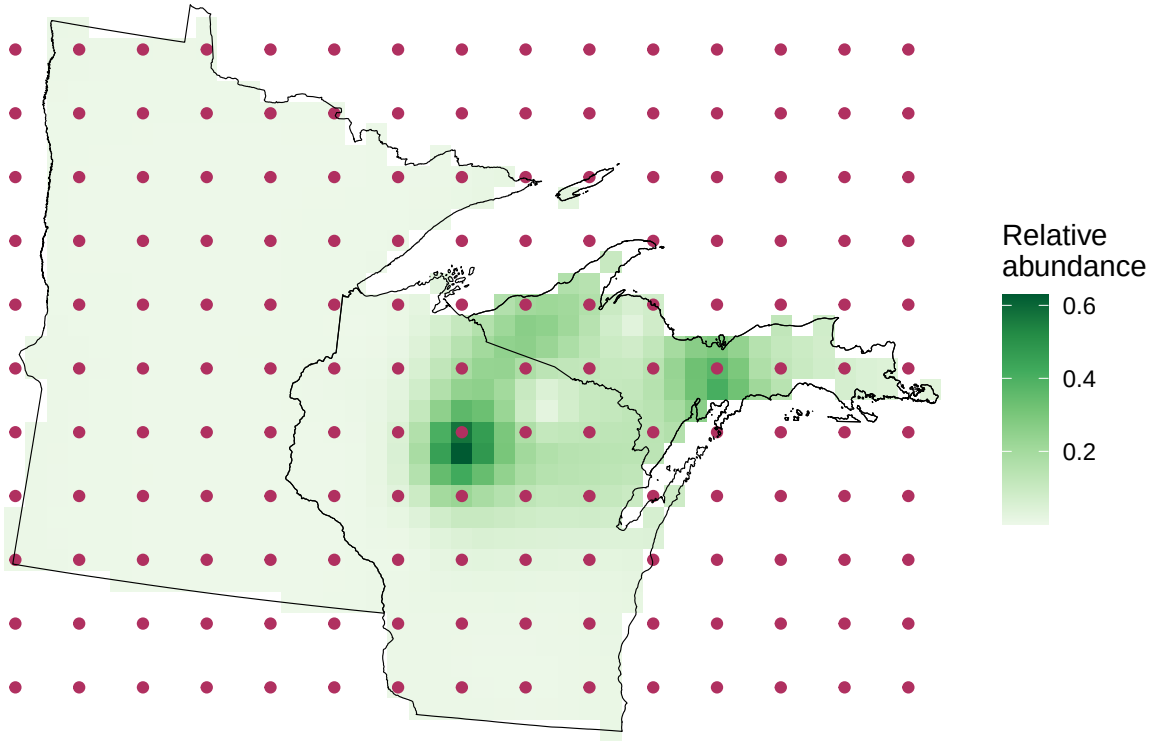
Birch


Relative
abundance
0.3
0.2
0.1

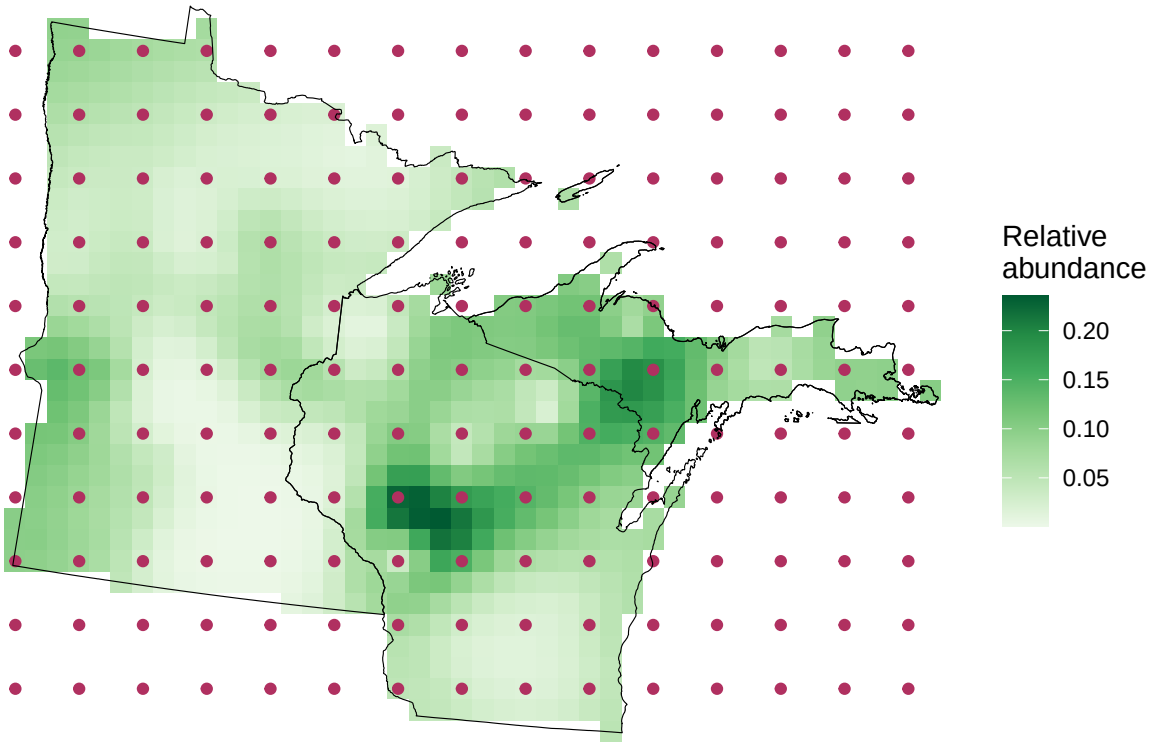
Elm



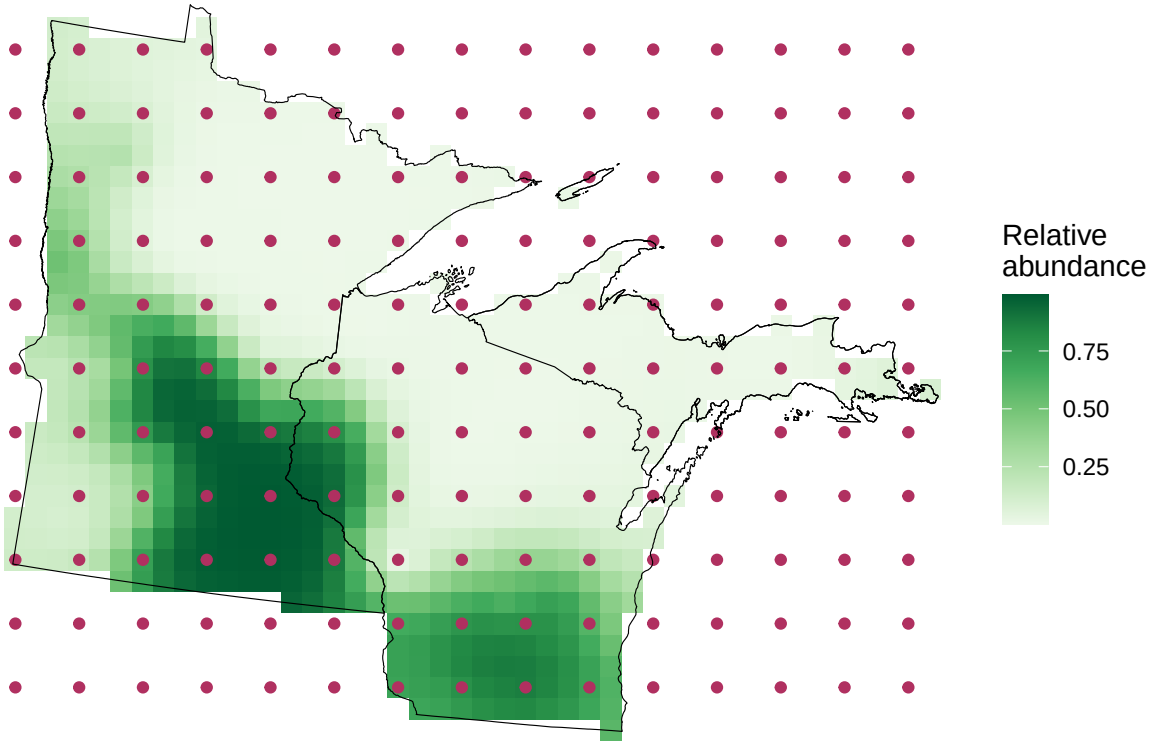
Hemlock



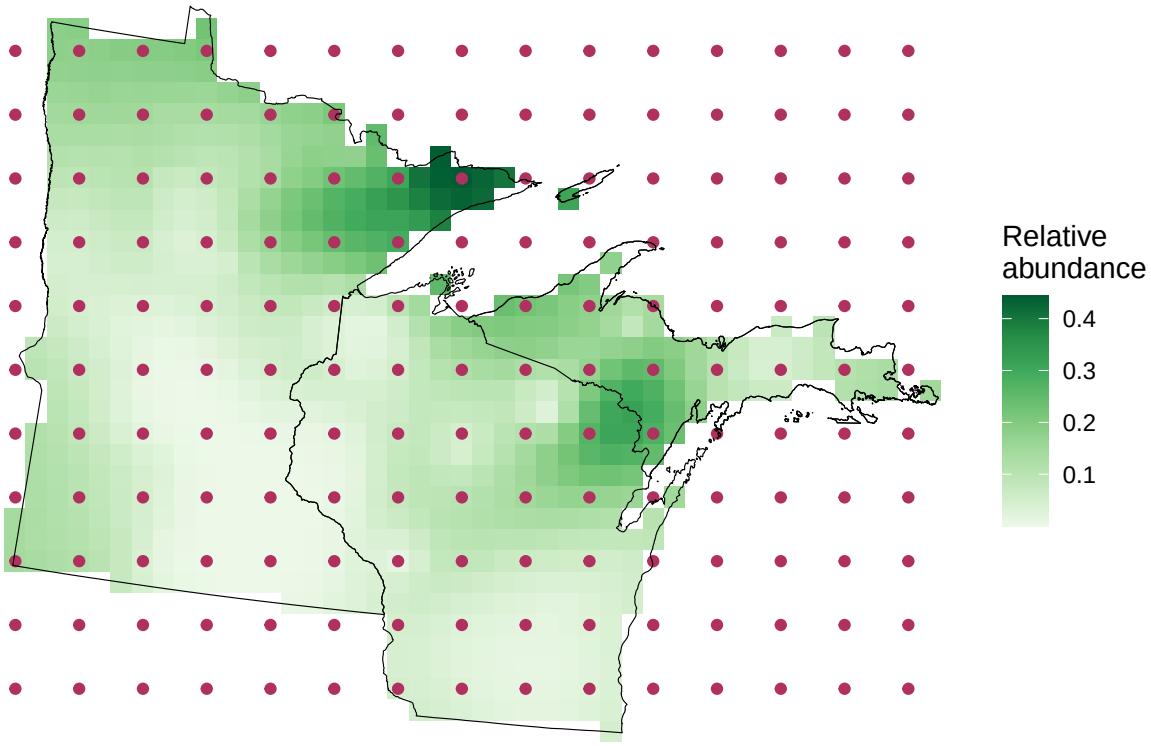
Maple


Relative
abundance
0.20
0.15
0.10
0.05

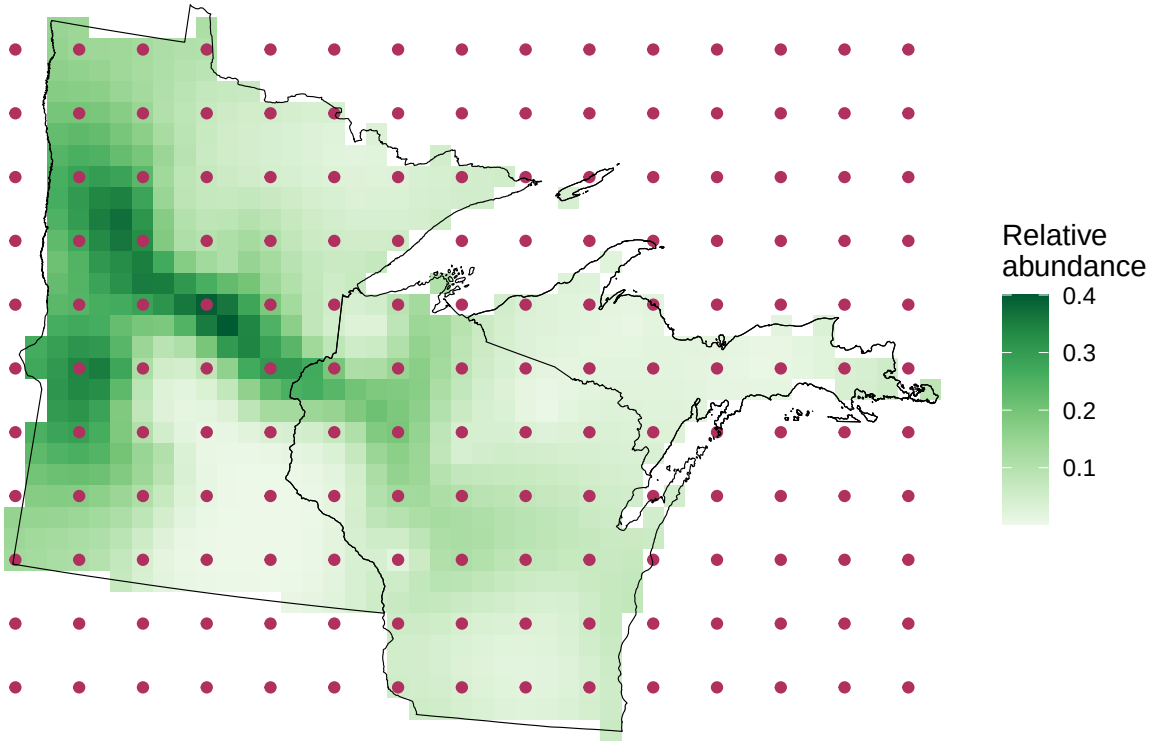
Oak


Relative
abundance
0.75
0.50
0.25

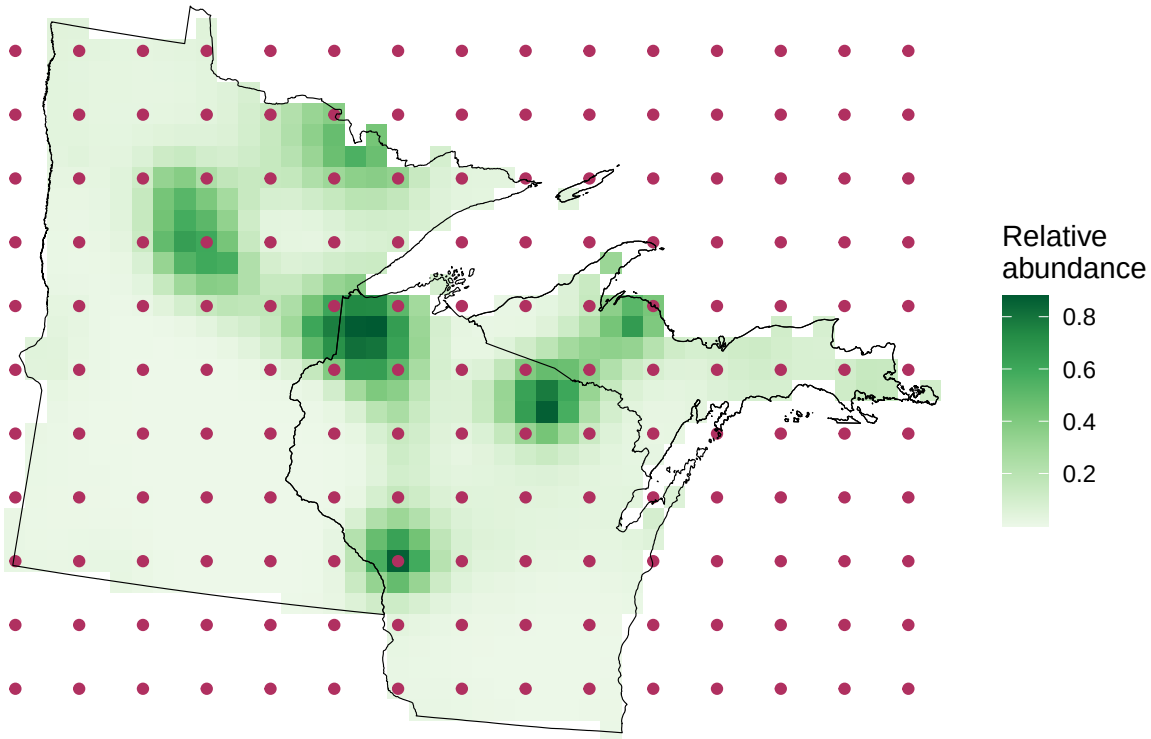
Other conifer



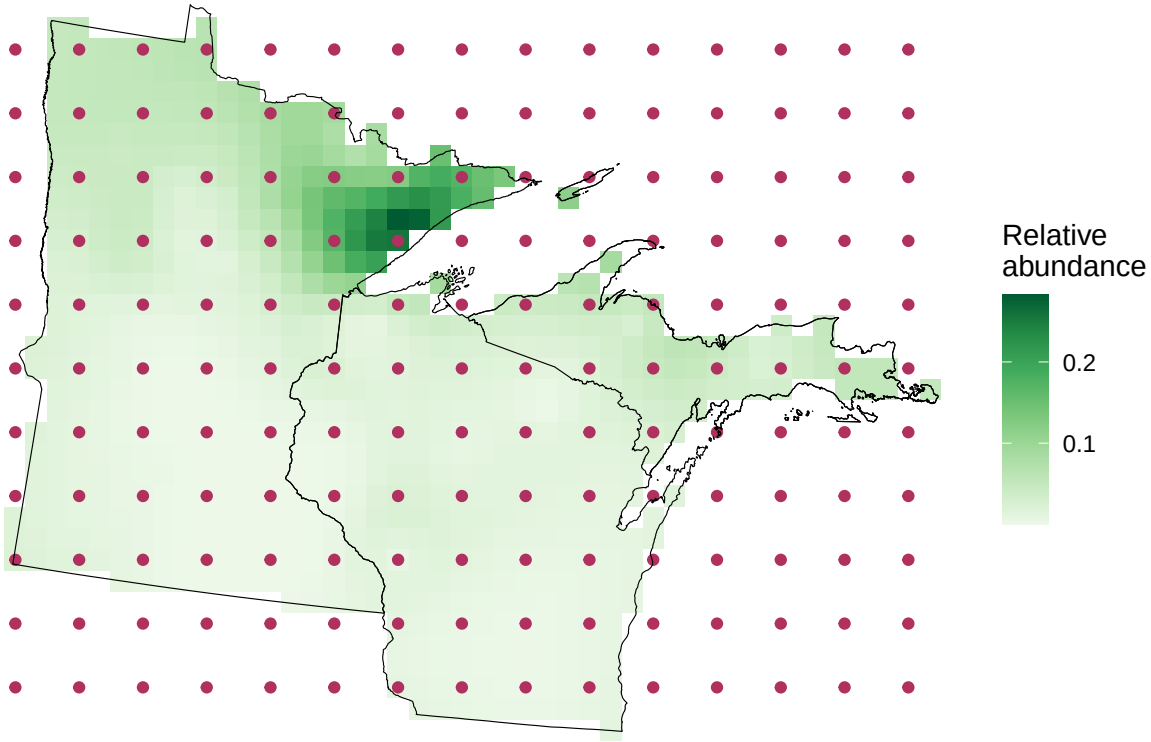
Other hardwood



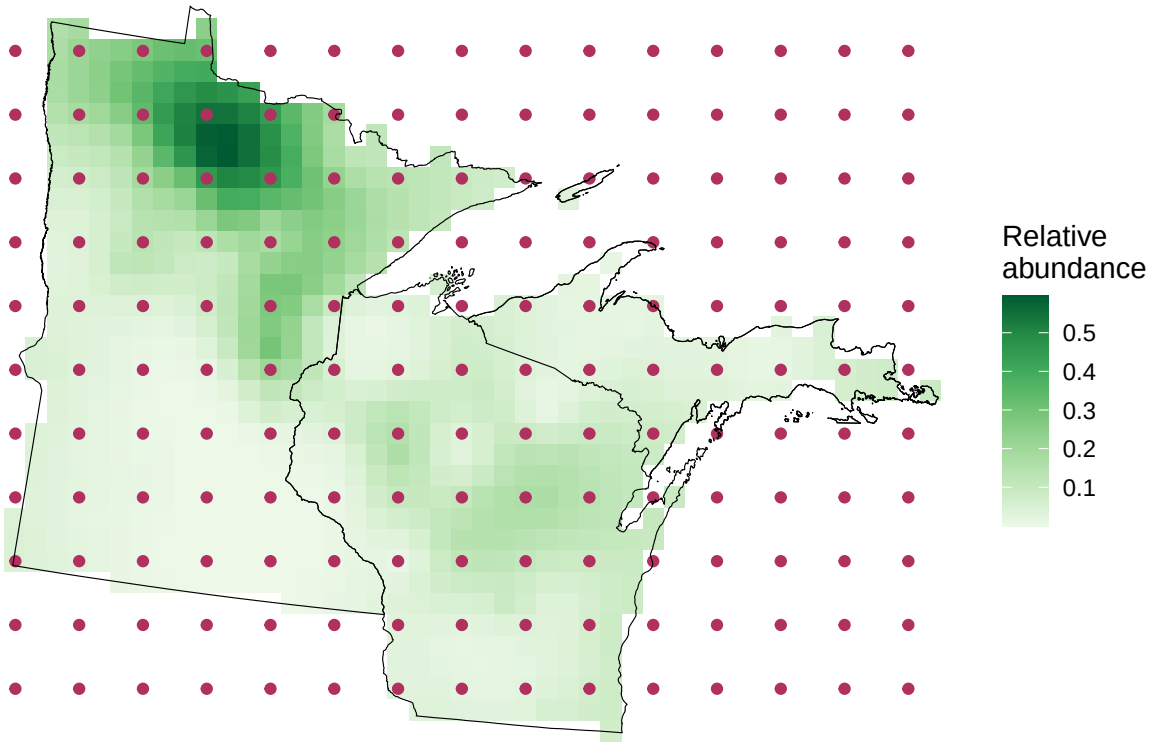
Pine



Spruce

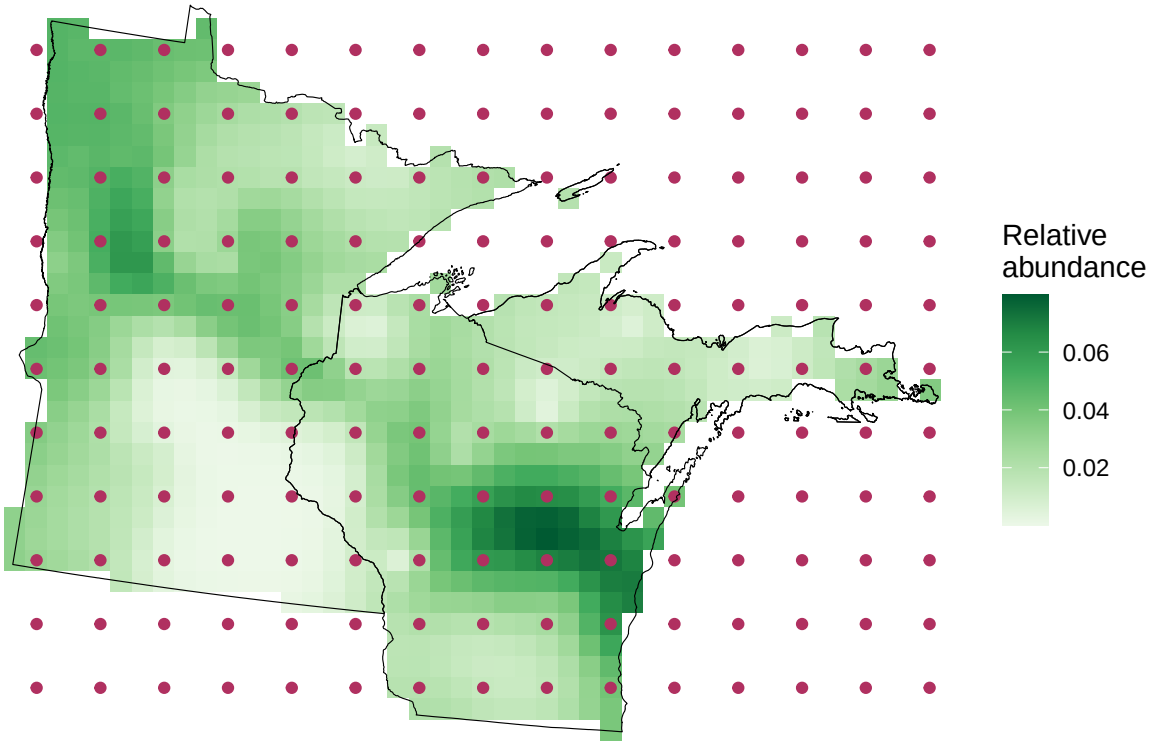


Tamarack

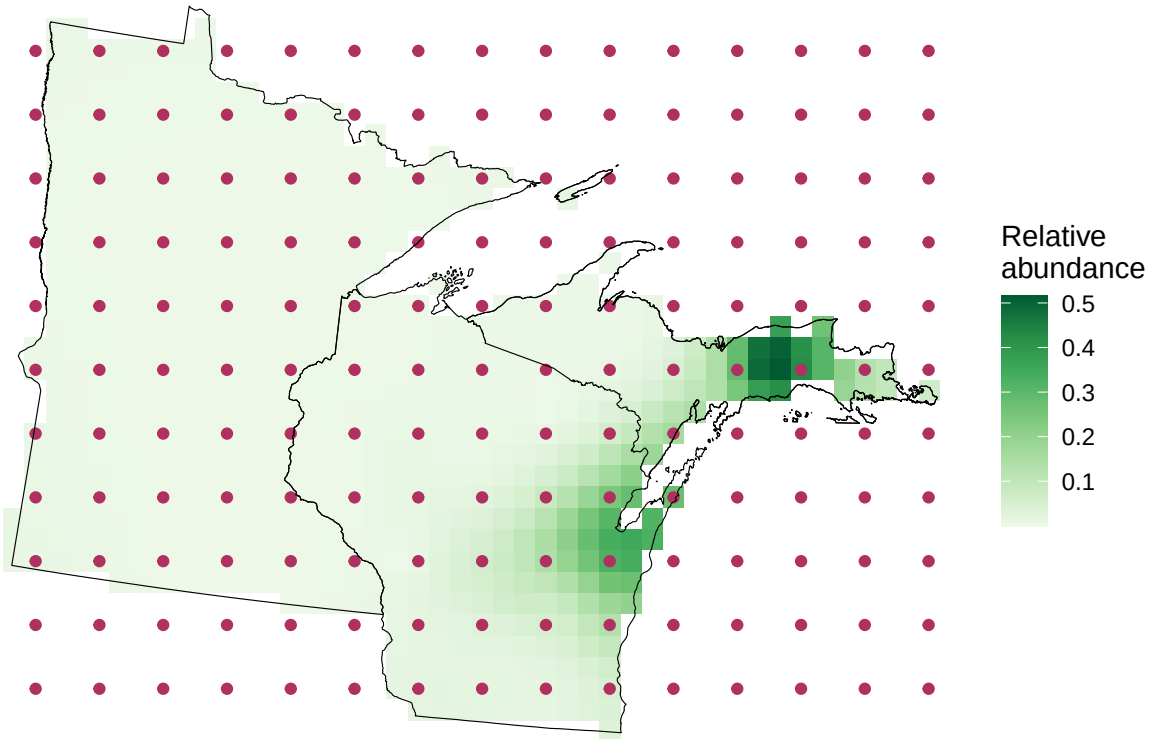


$\mathbf{x = 2, y = 3}$

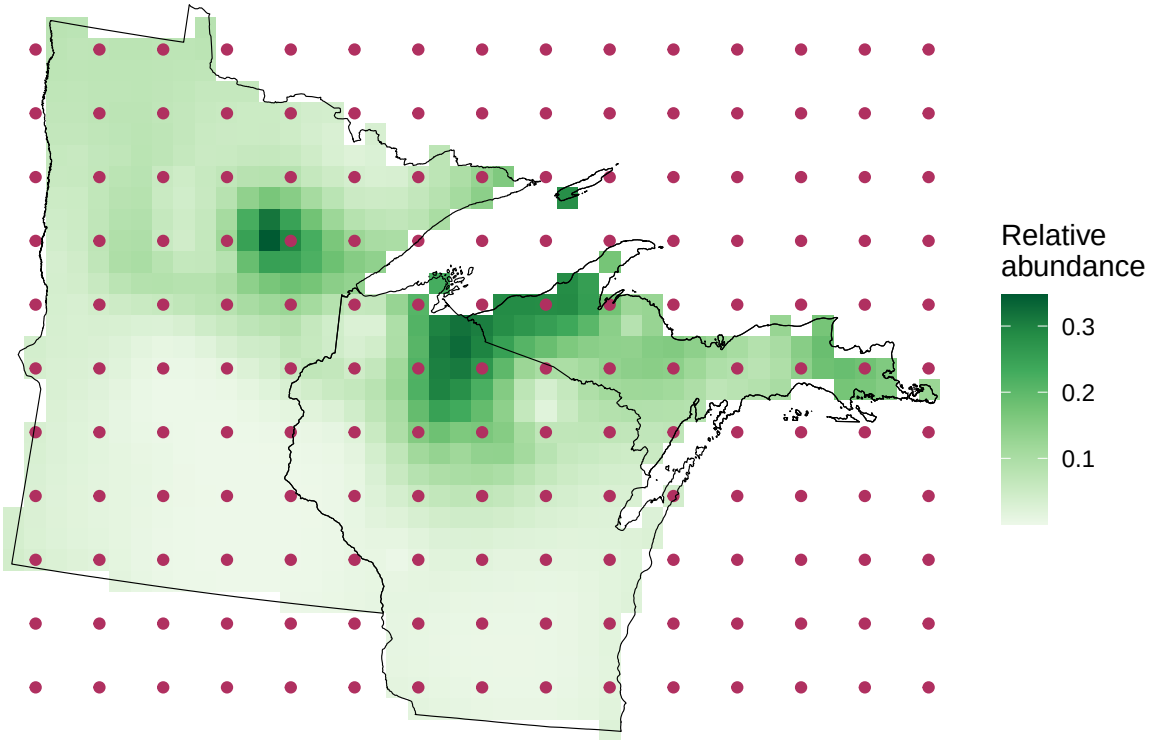
Ash



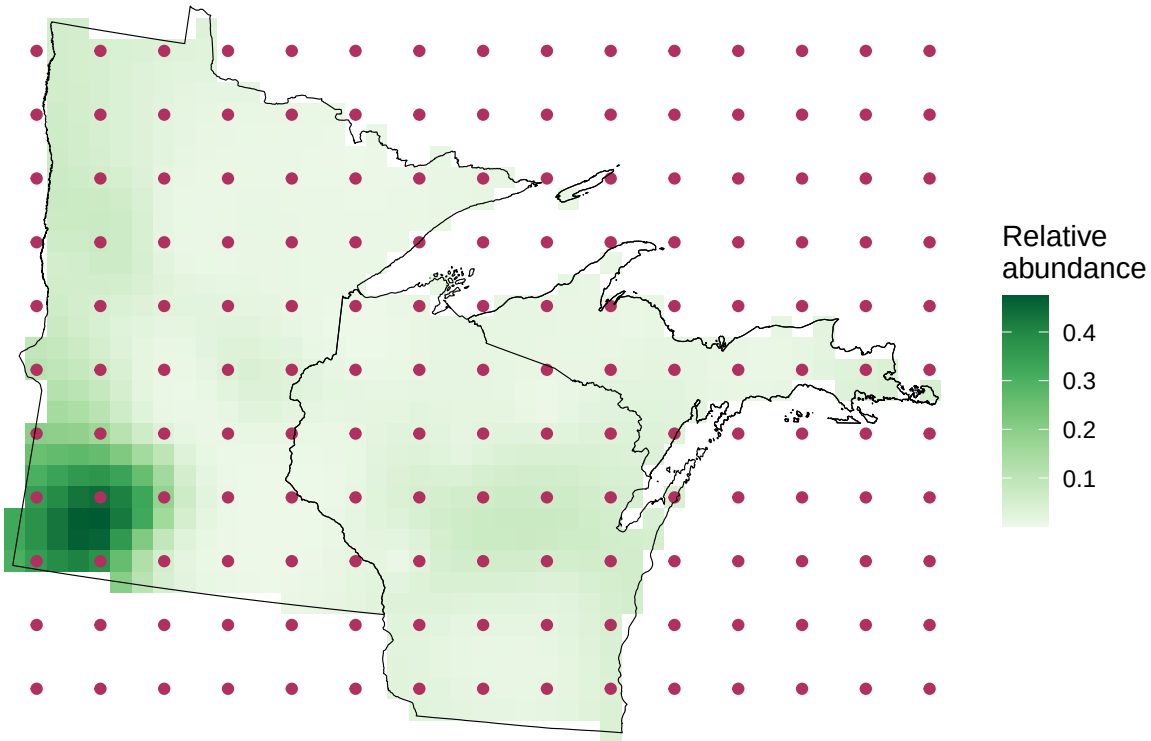
Beech


Relative
abundance
0.5
0.4
0.3
0.2
0.1

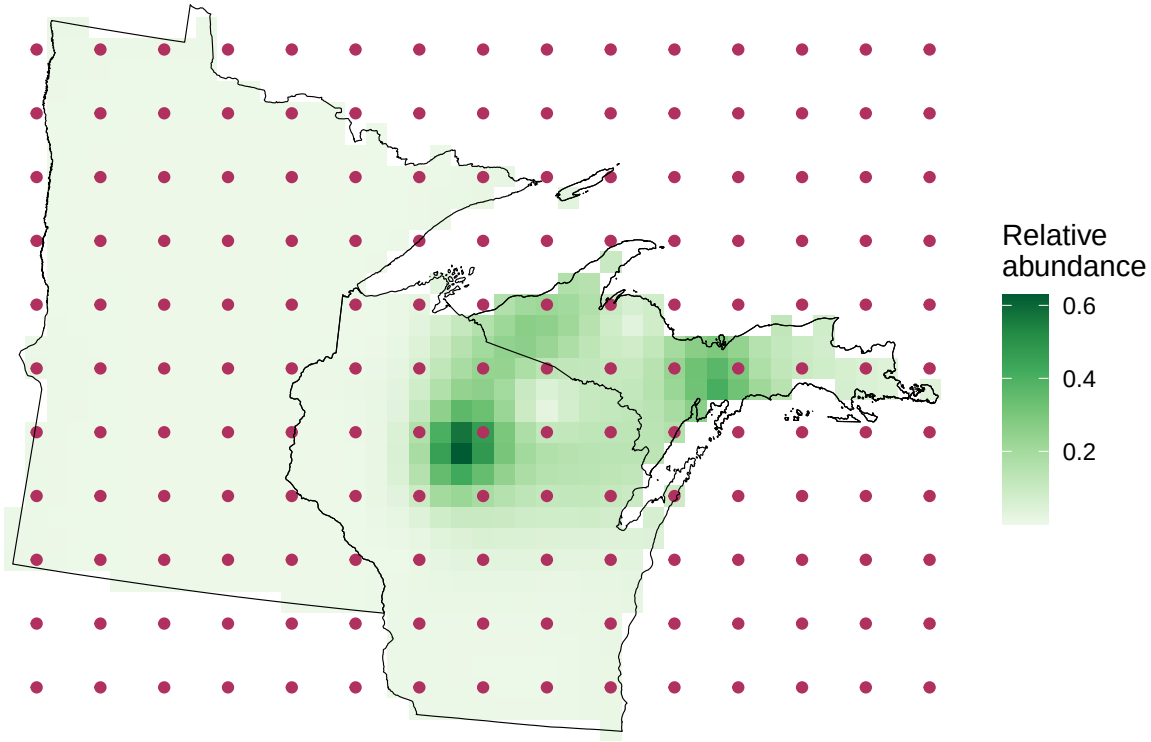
Birch


Relative
abundance
0.3
0.2
0.1

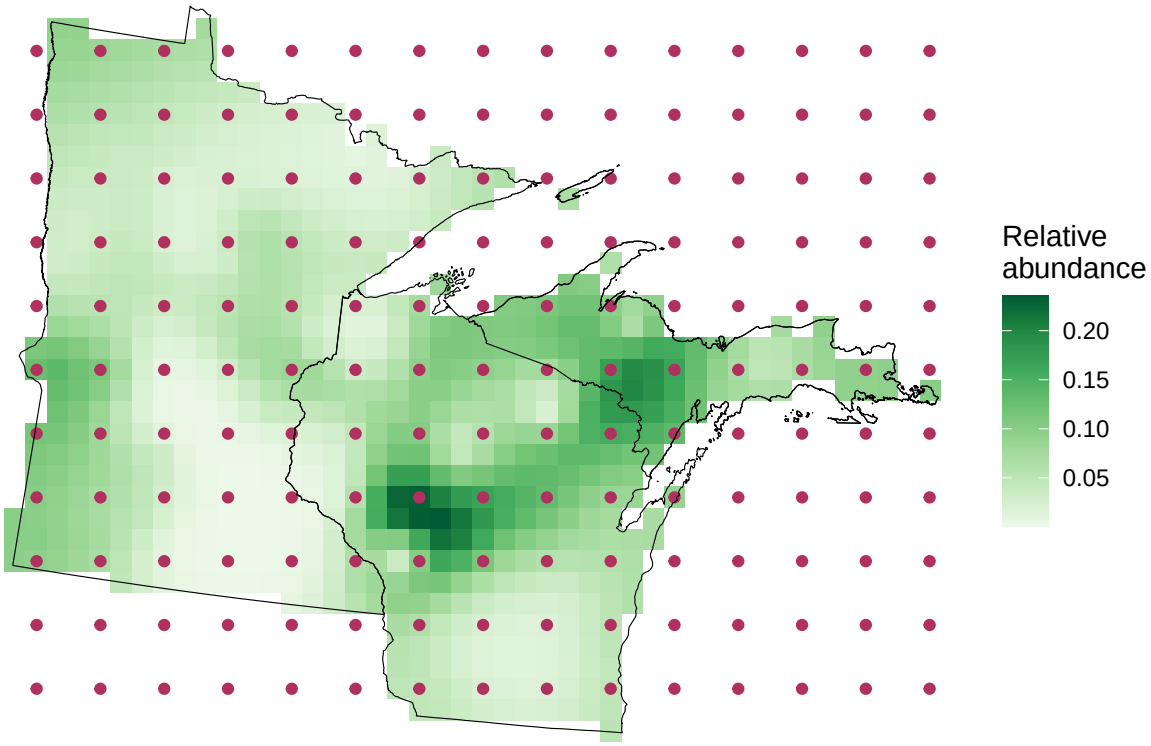
Elm



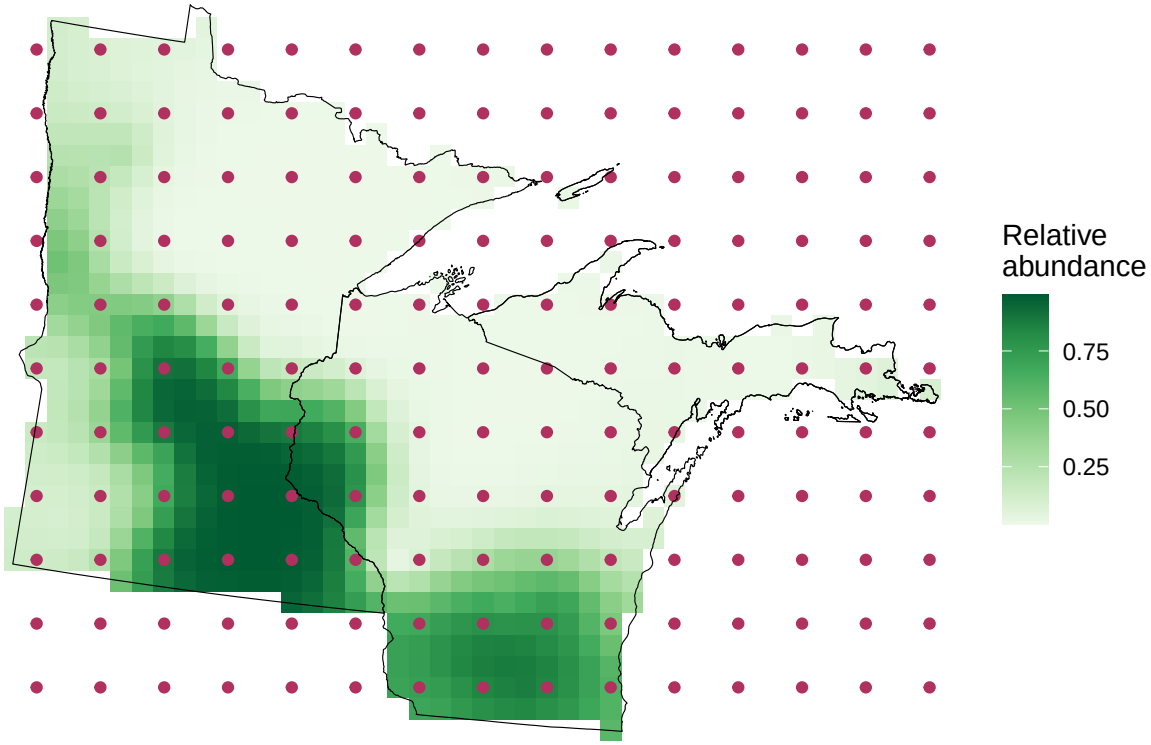
Hemlock



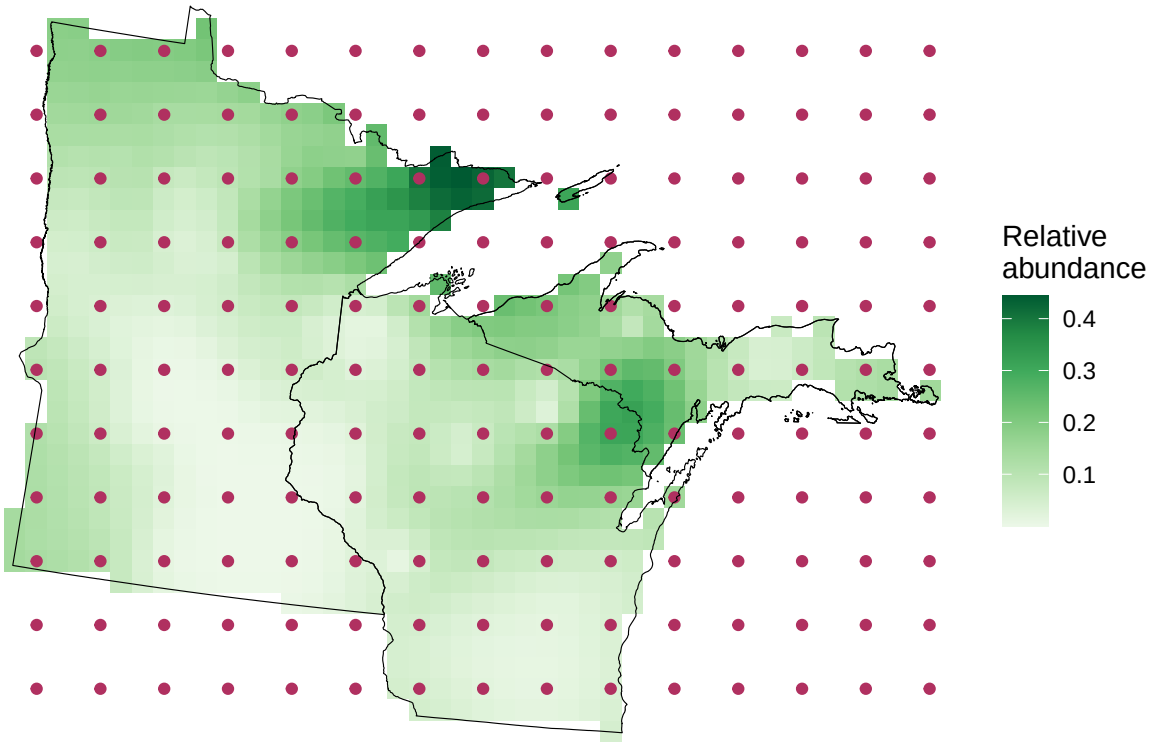
Maple


Relative
abundance
0.20
0.15
0.10
0.05

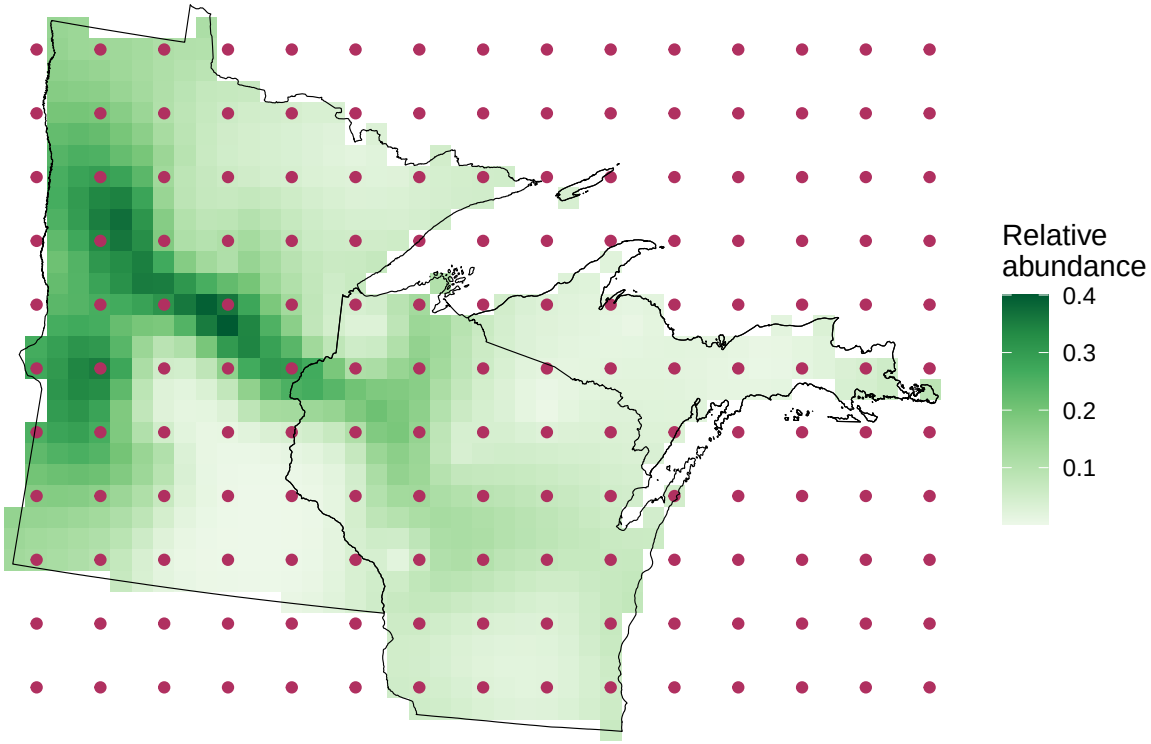
Oak


Relative
abundance
0.75
0.50
0.25

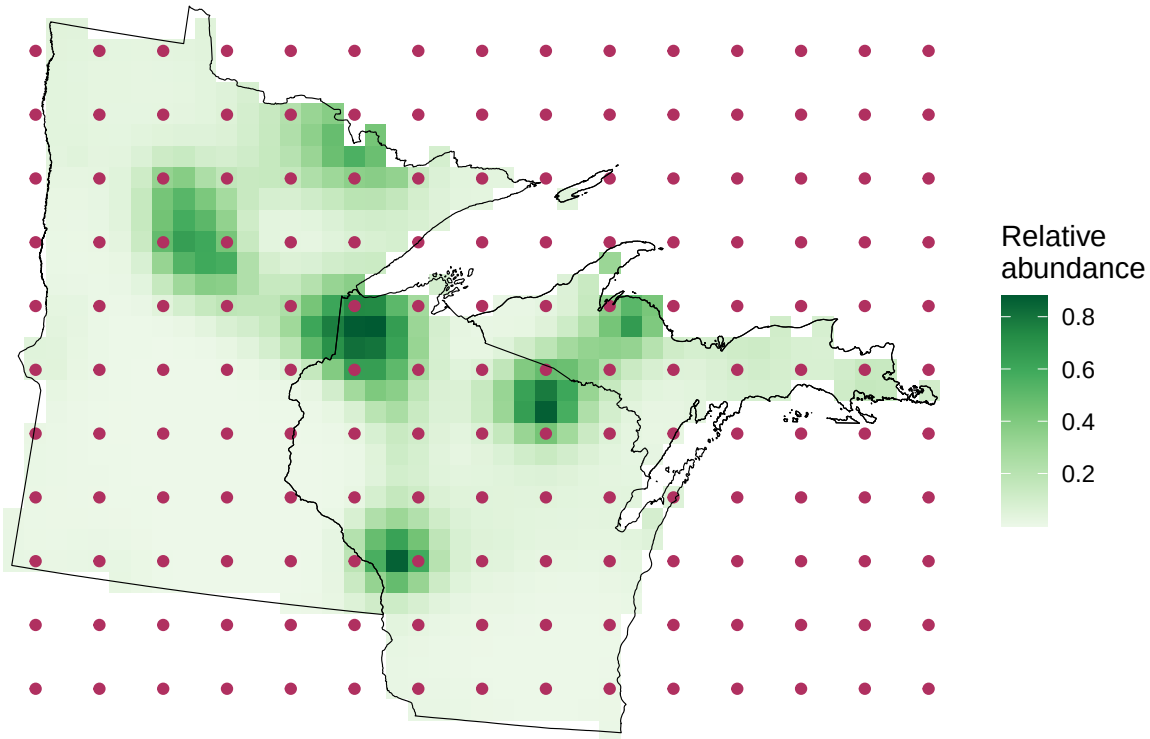
Other conifer



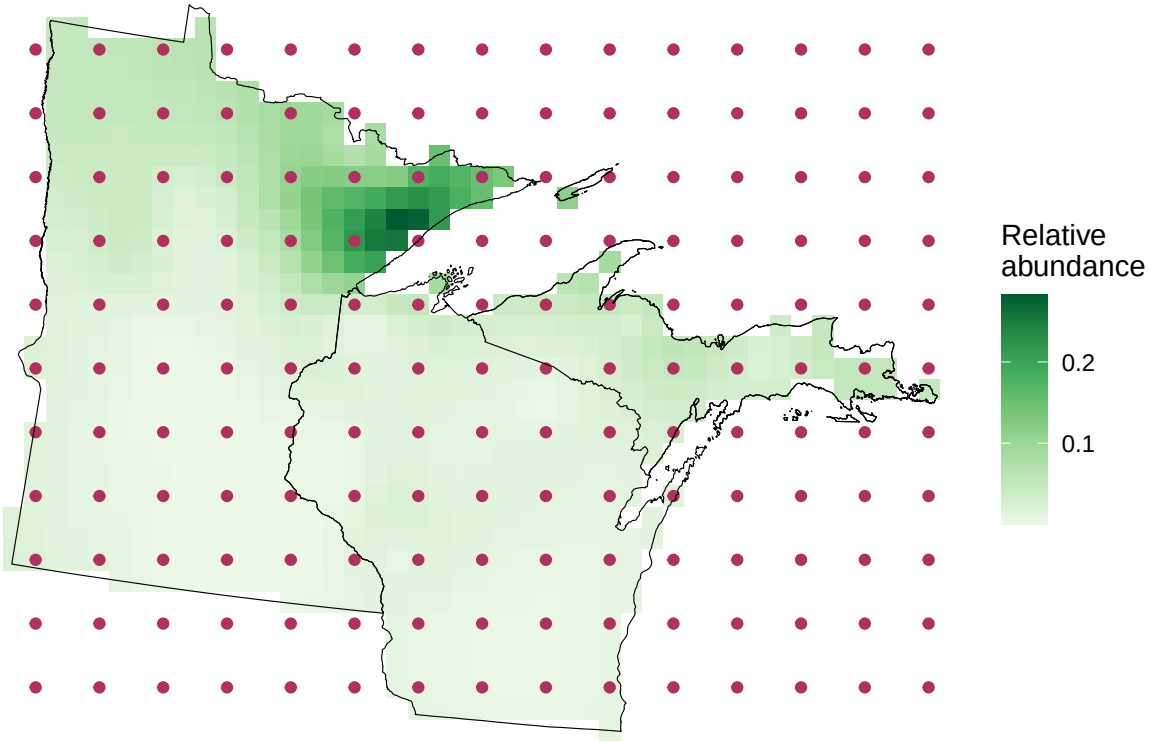
Other hardwood



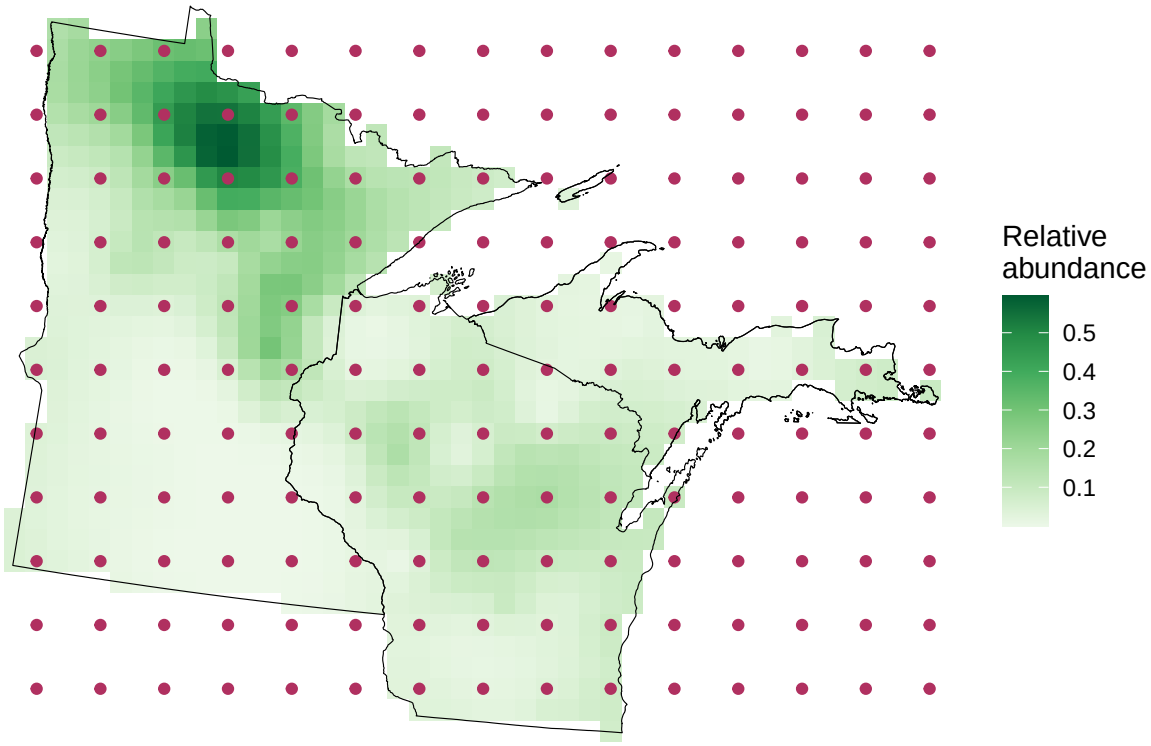
Pine



Spruce

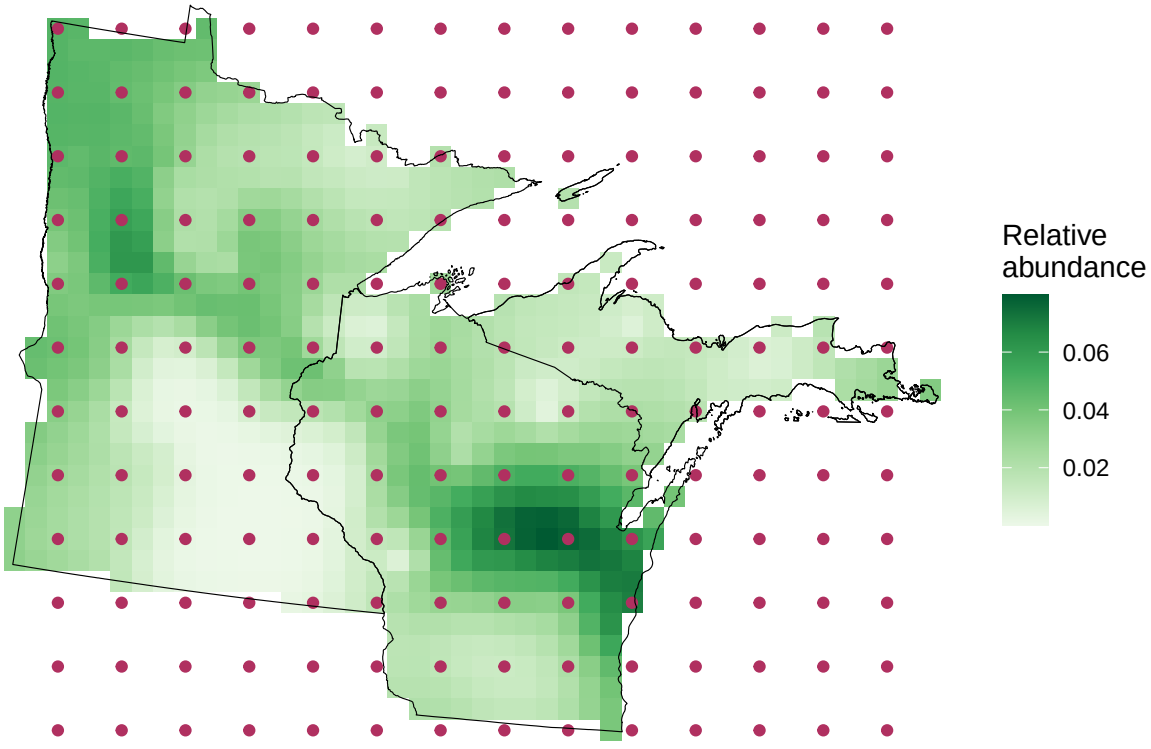


Tamarack

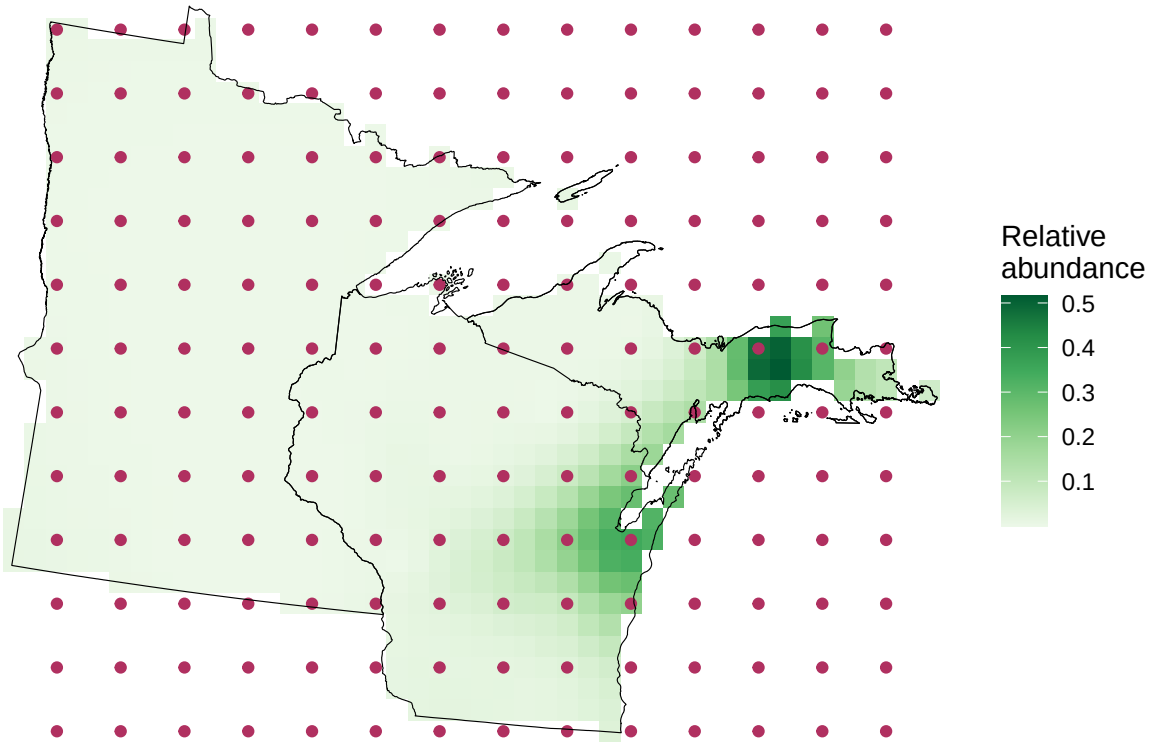


$\mathbf{x = 3, y = 1}$

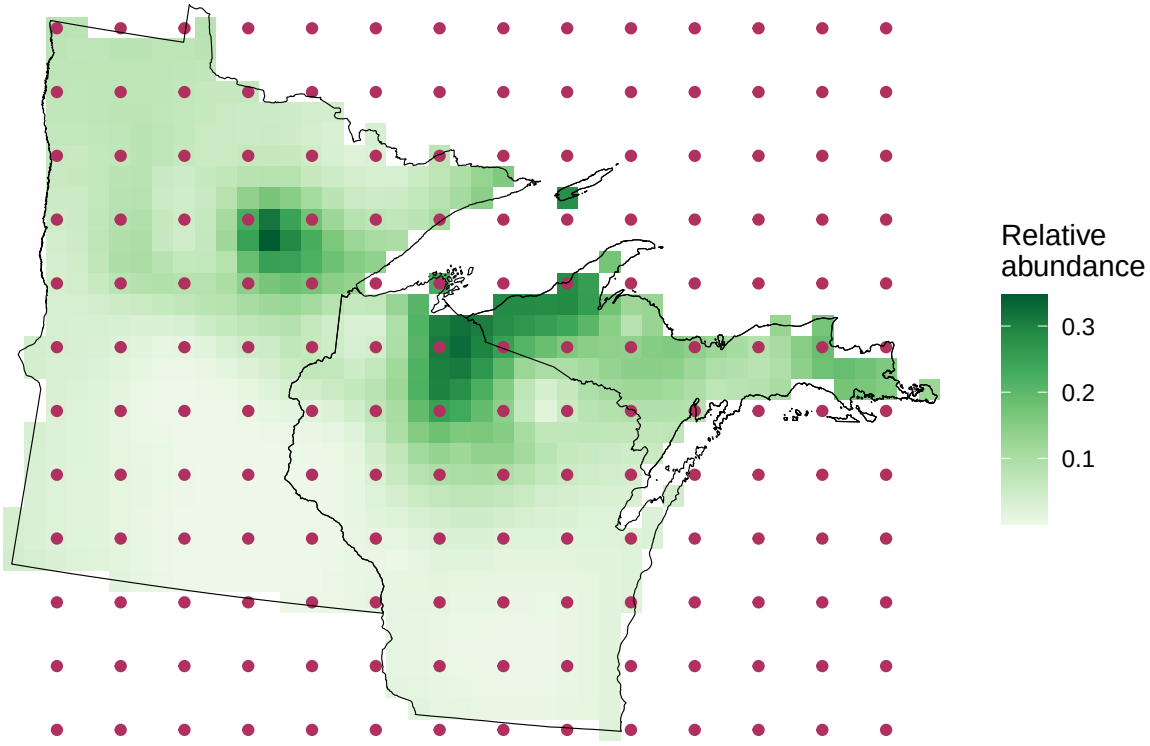
Ash

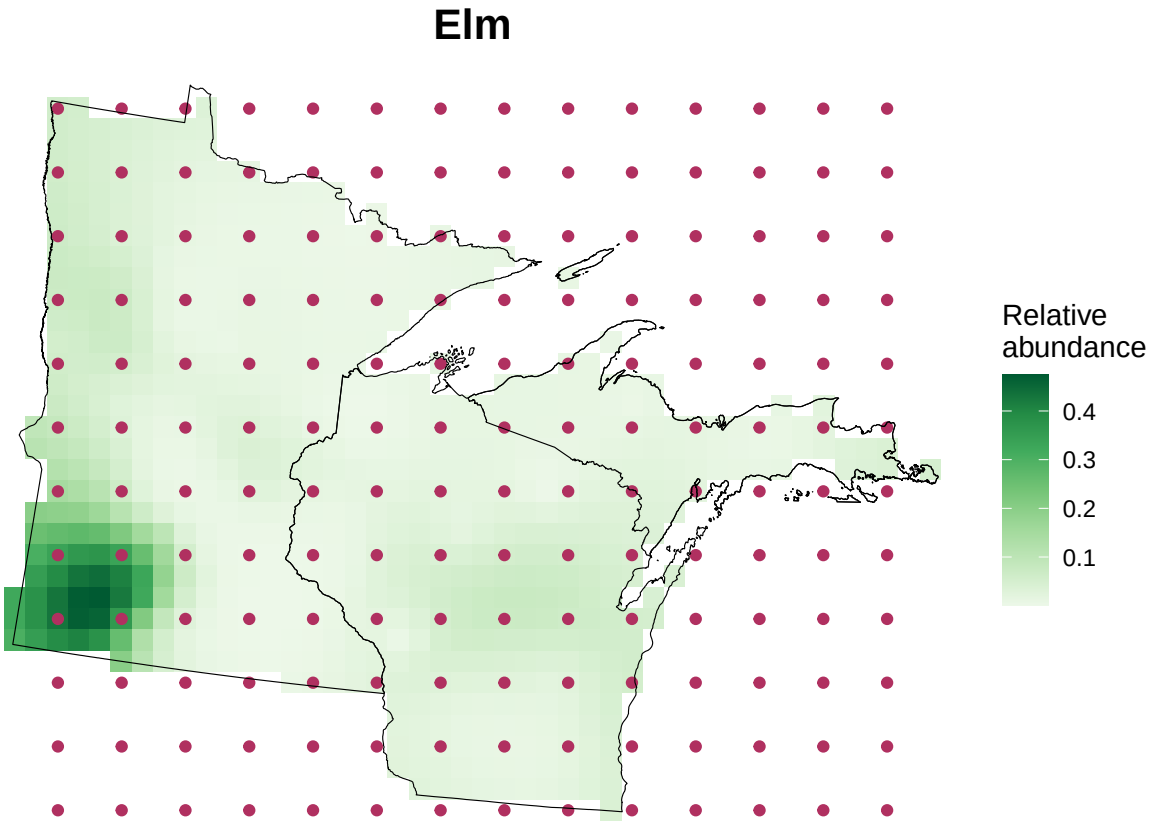


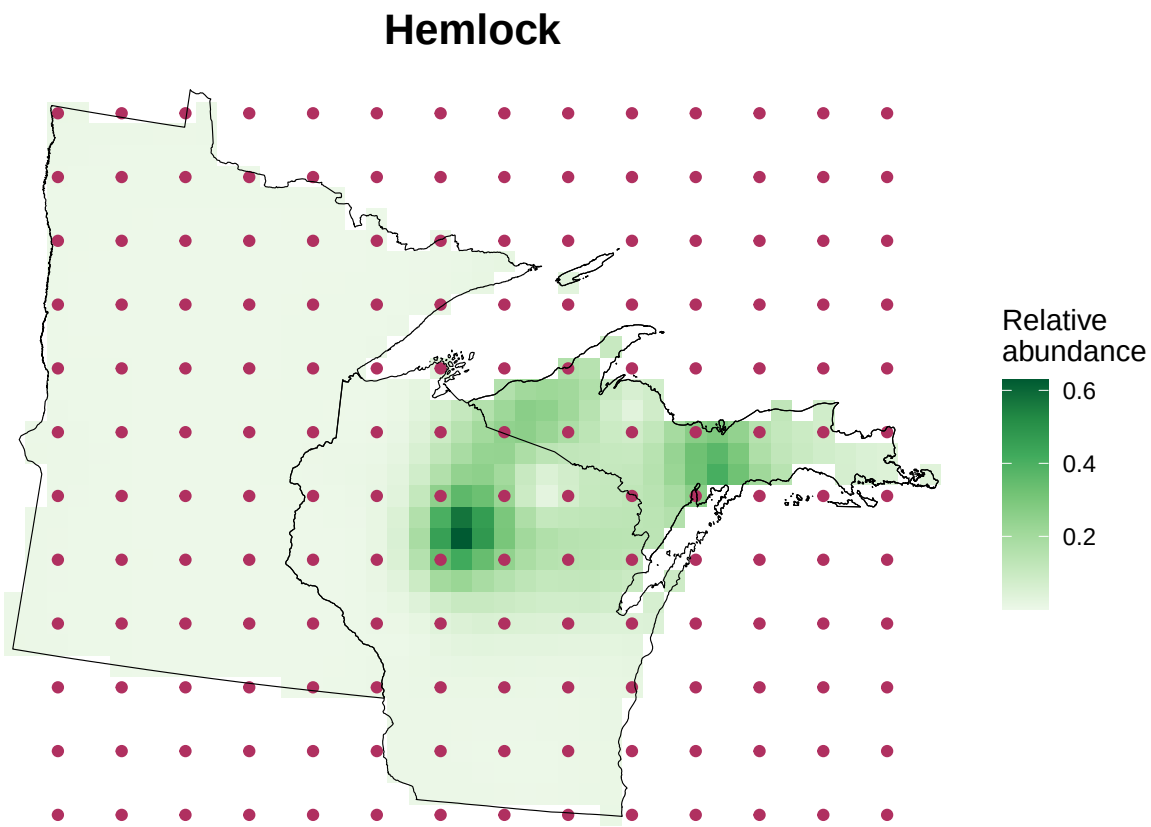
Beech


Relative
abundance
0.5
0.4
0.3
0.2
0.1

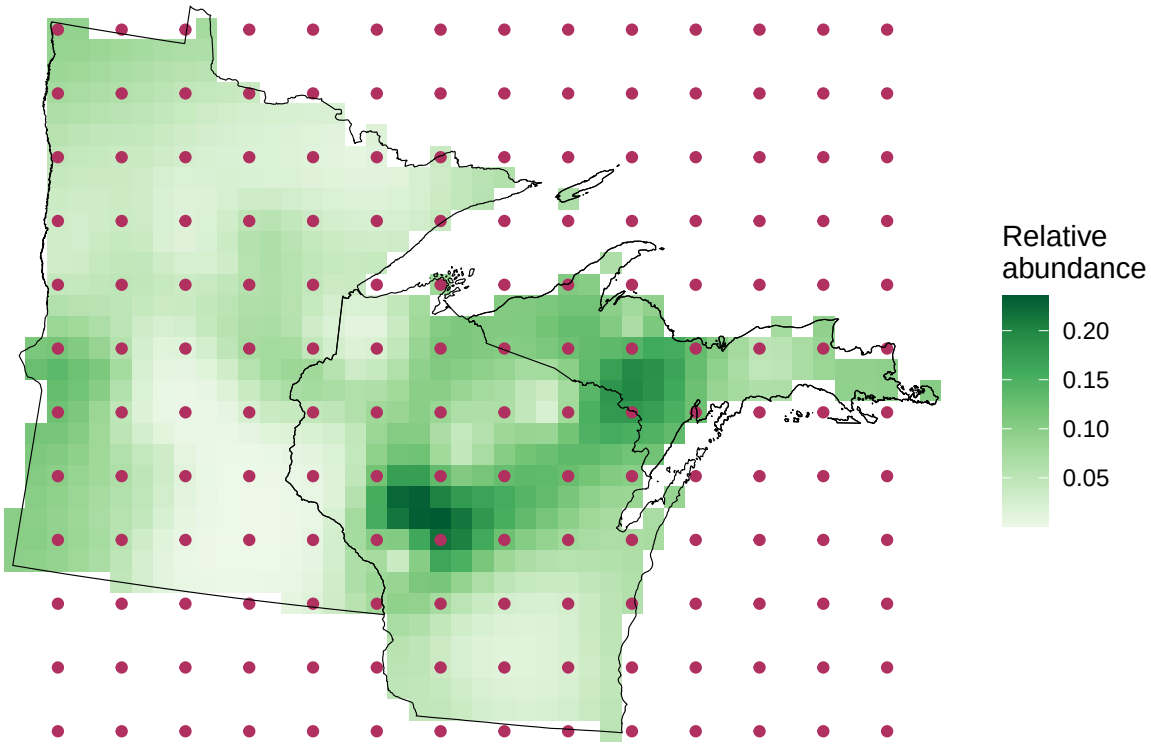
Birch


Relative
abundance
0.3
0.2
0.1

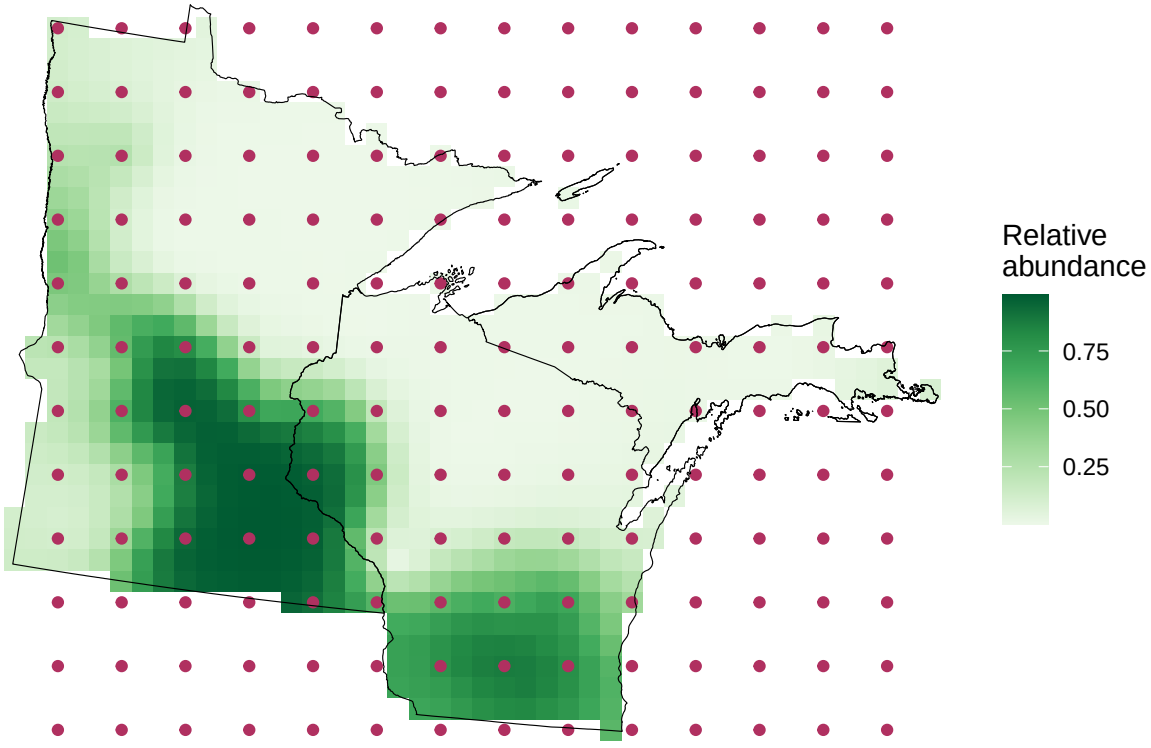

Elm
Relative abundance
0.4
0.3
0.2
0.1


Hemlock
Relative abundance
0.6
0.4
0.2

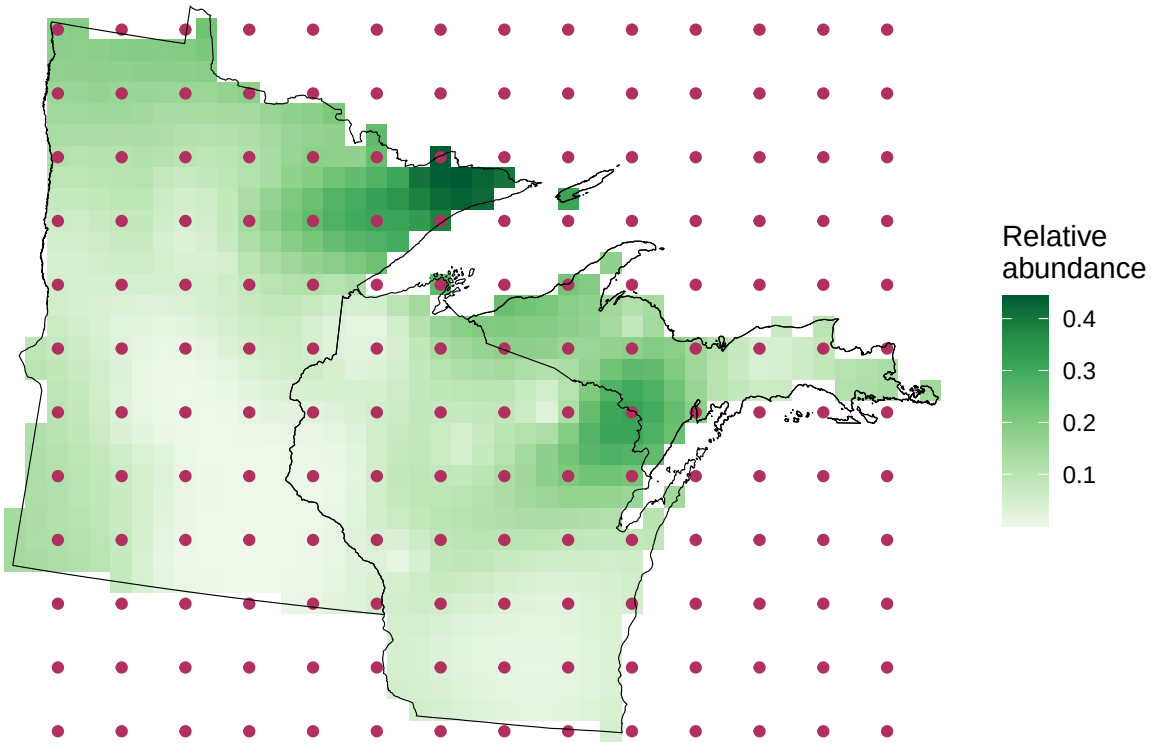
Maple


Relative
abundance
0.20
0.15
0.10
0.05

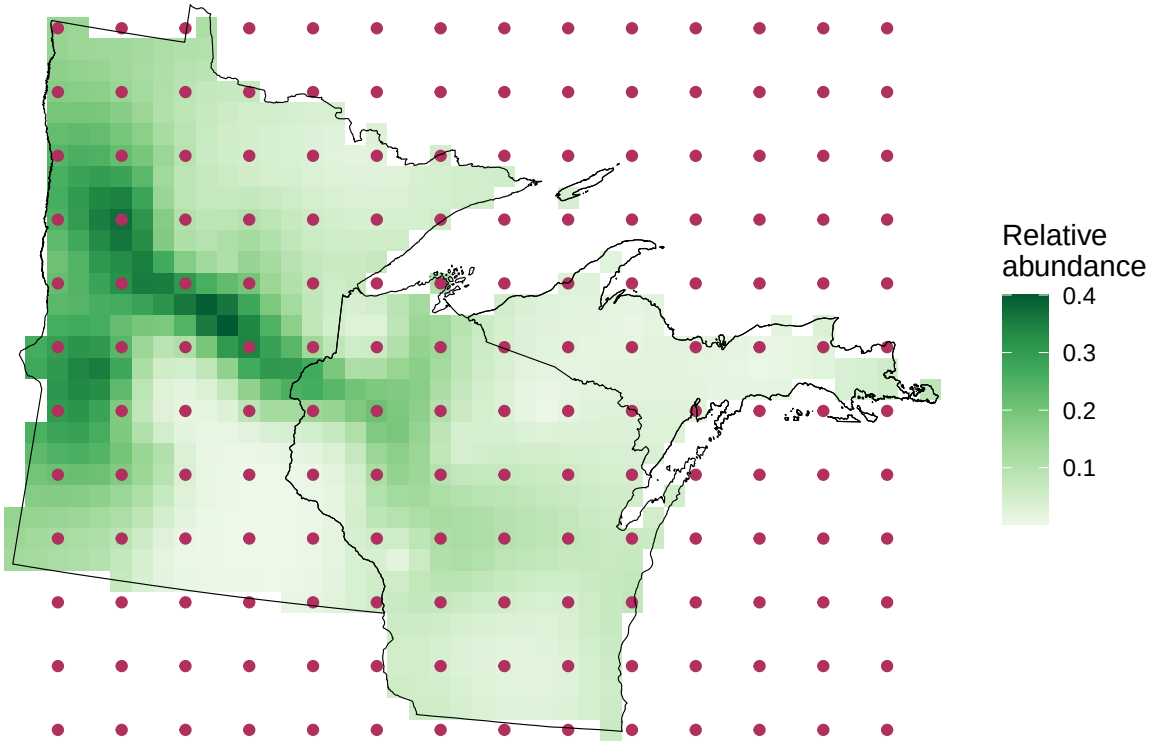
Oak

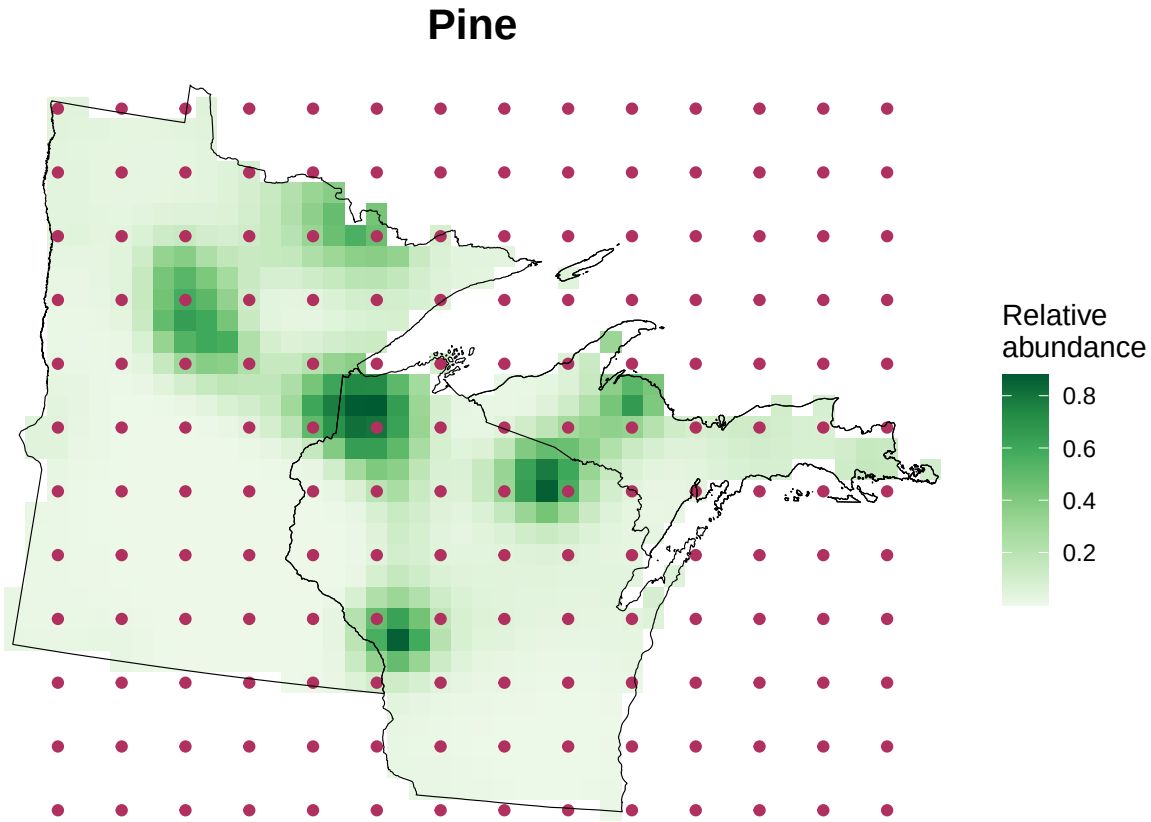

Relative
abundance
0.75
0.50
0.25

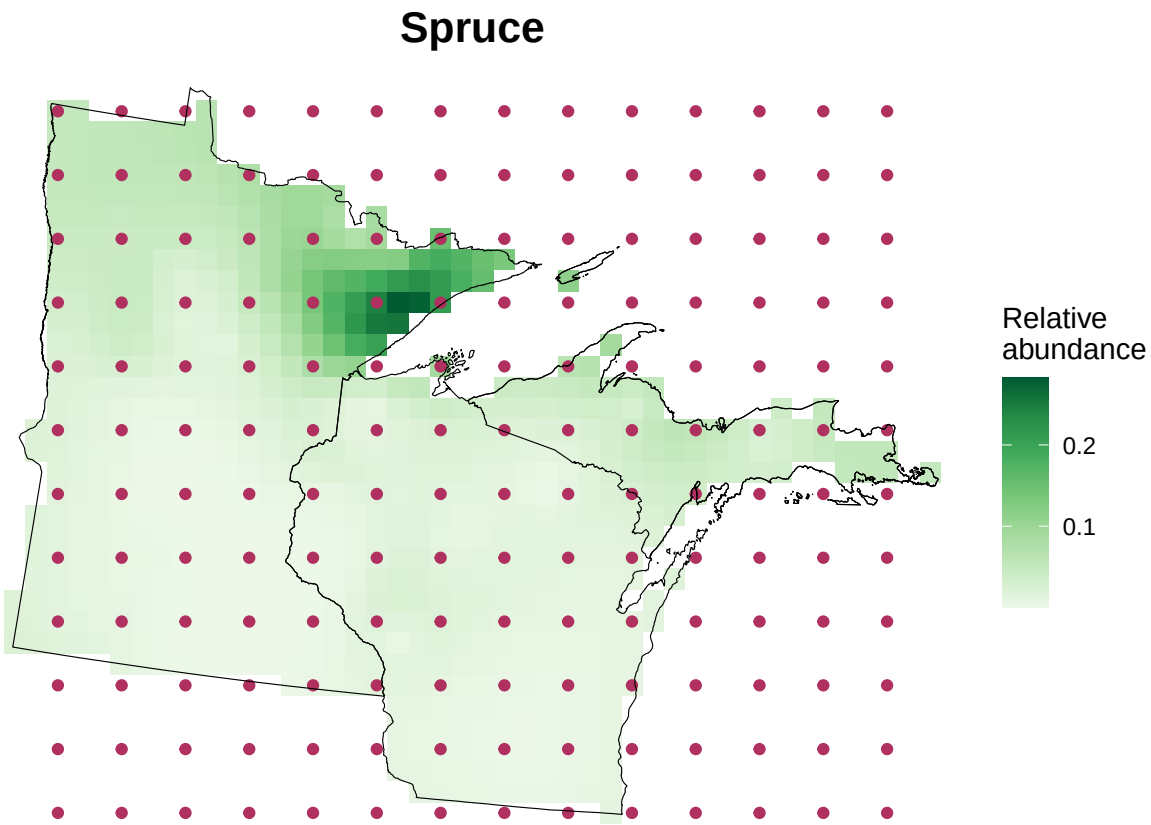
Other conifer



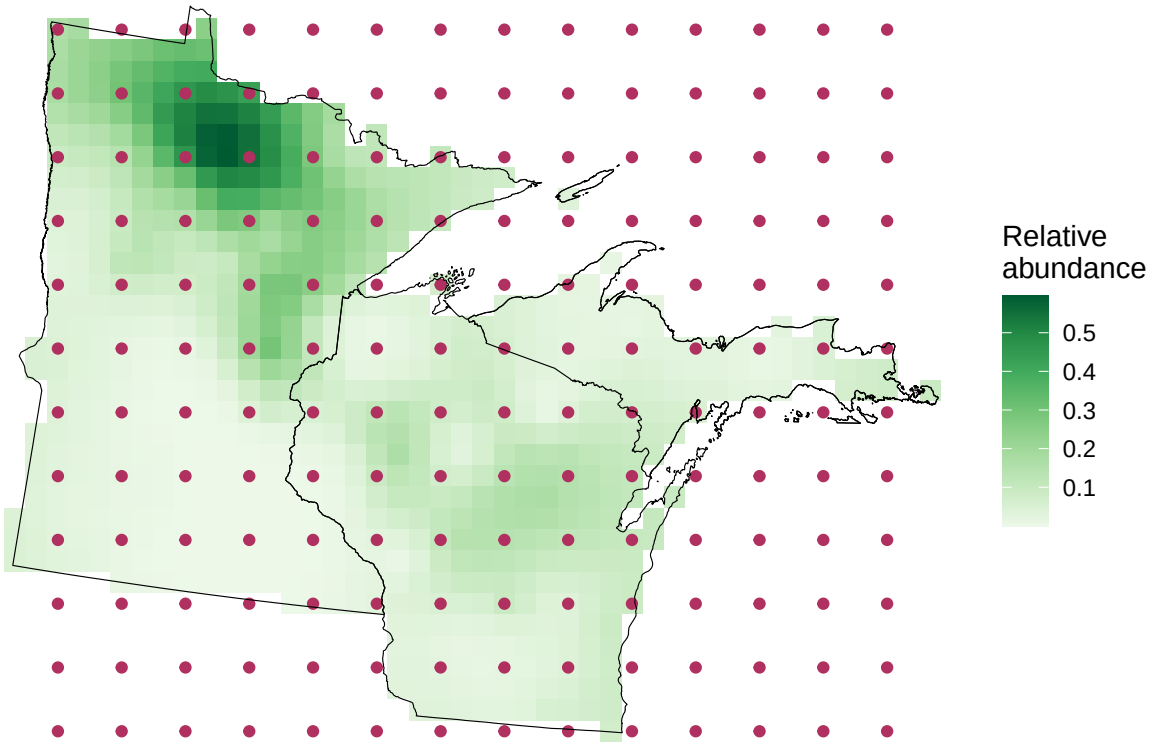
Other hardwood




Pine
Relative abundance
0.8
0.6
0.4
0.2

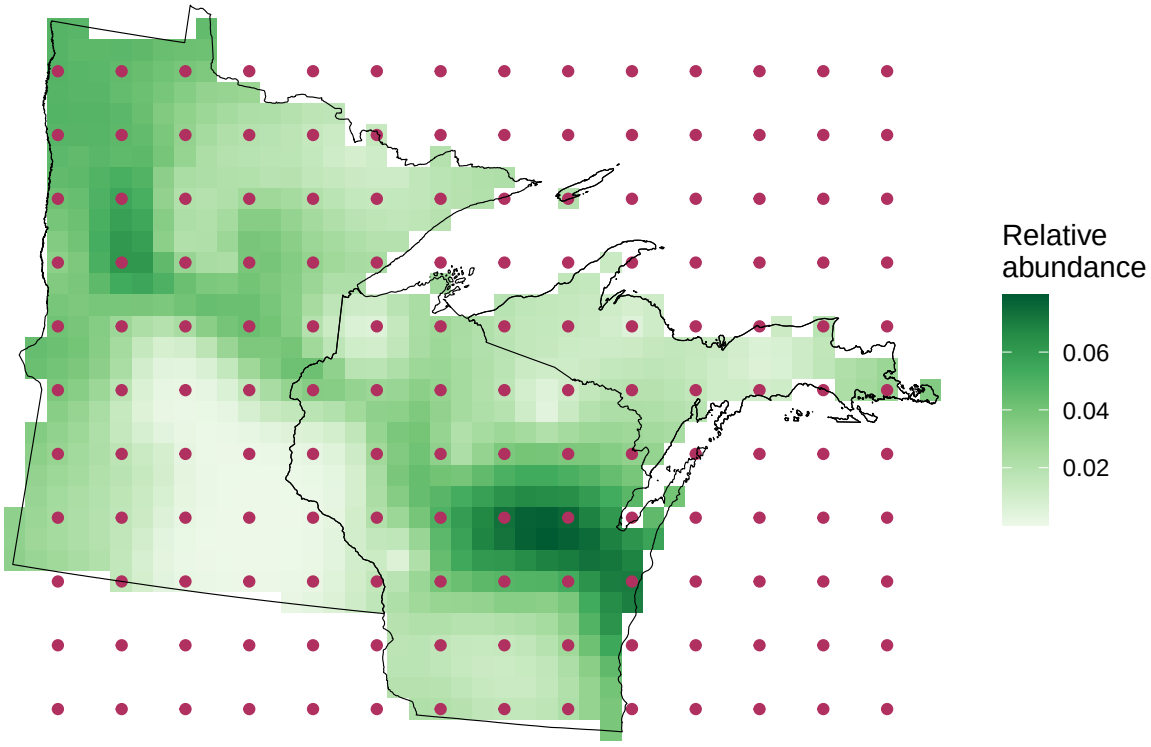

Spruce
Relative abundance
0.2
0.1

Tamarack

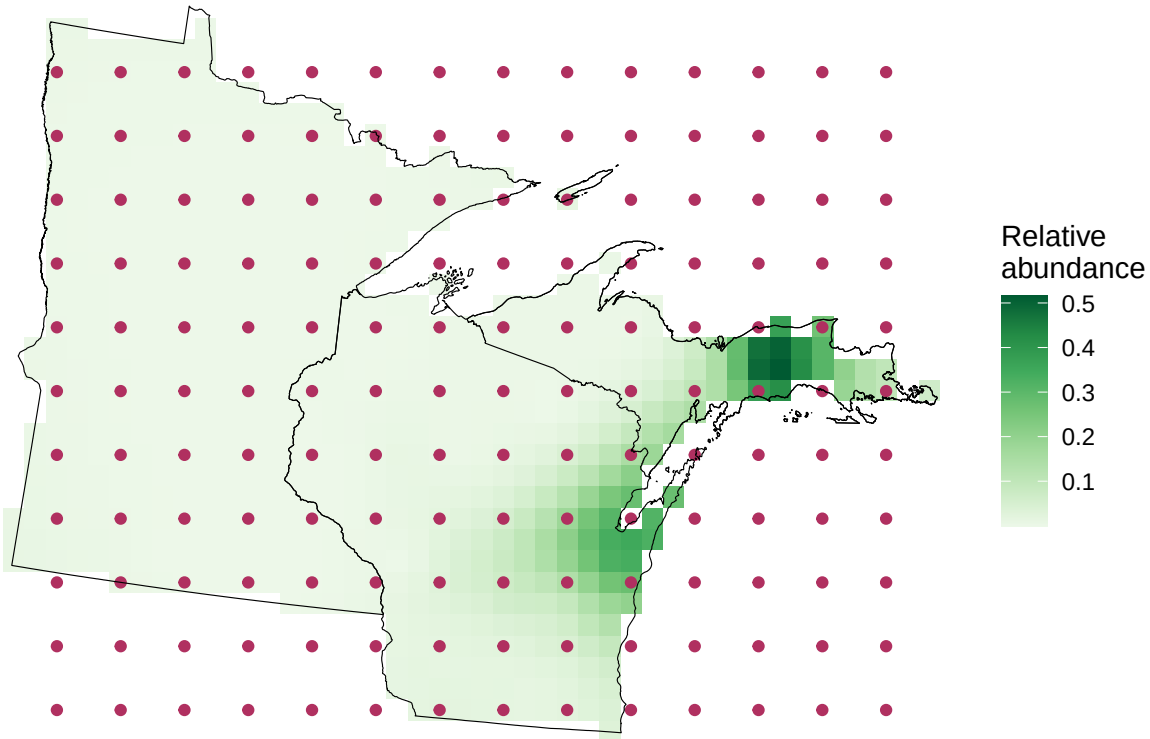


$\mathbf{x = 3, y = 2}$

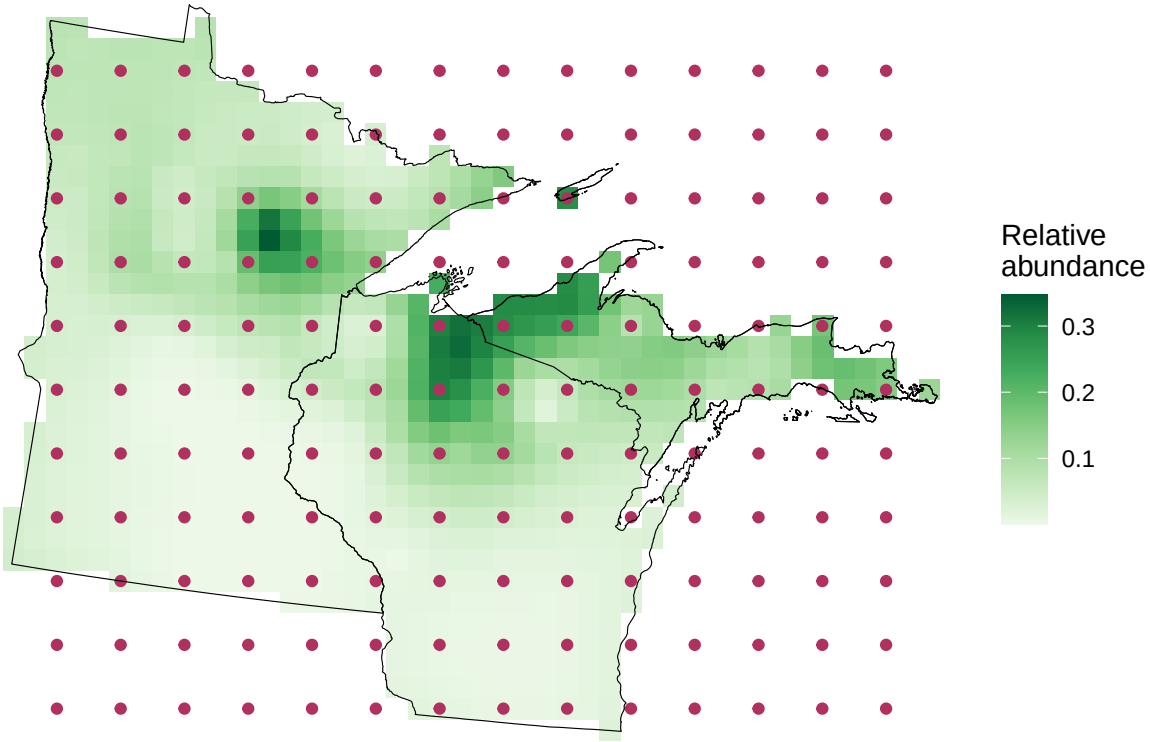
Ash



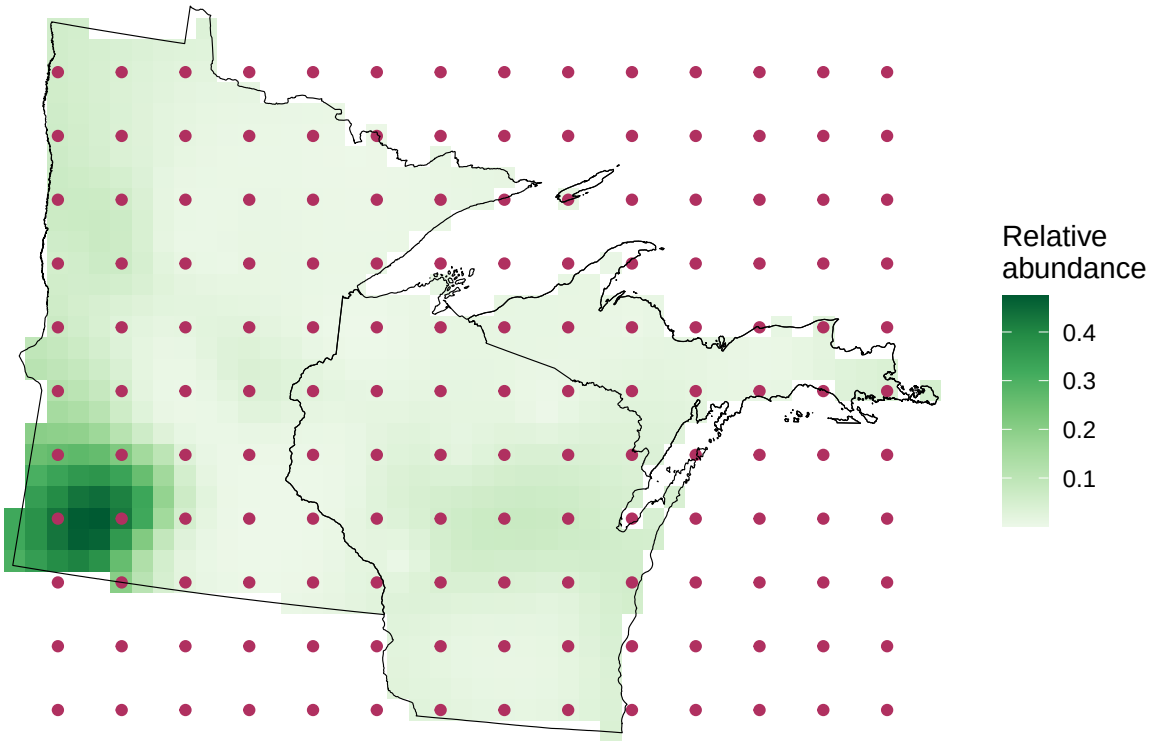
Beech


Relative
abundance
0.5
0.4
0.3
0.2
0.1

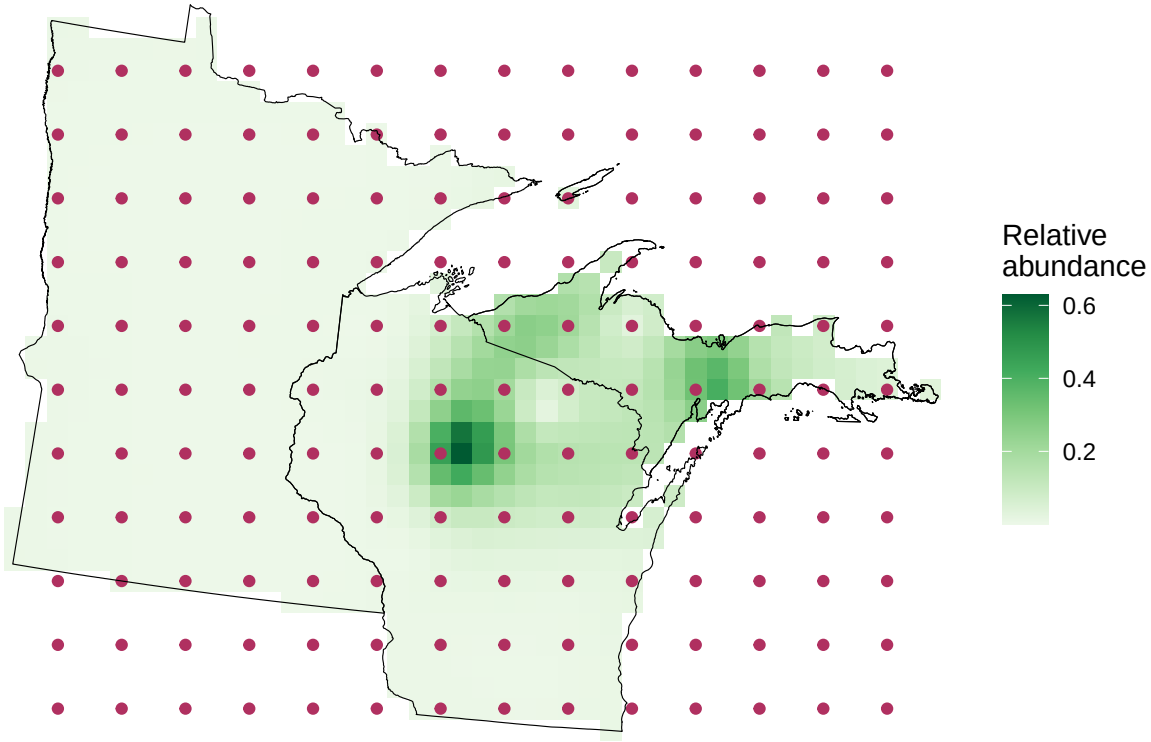
Birch


Relative
abundance
0.3
0.2
0.1

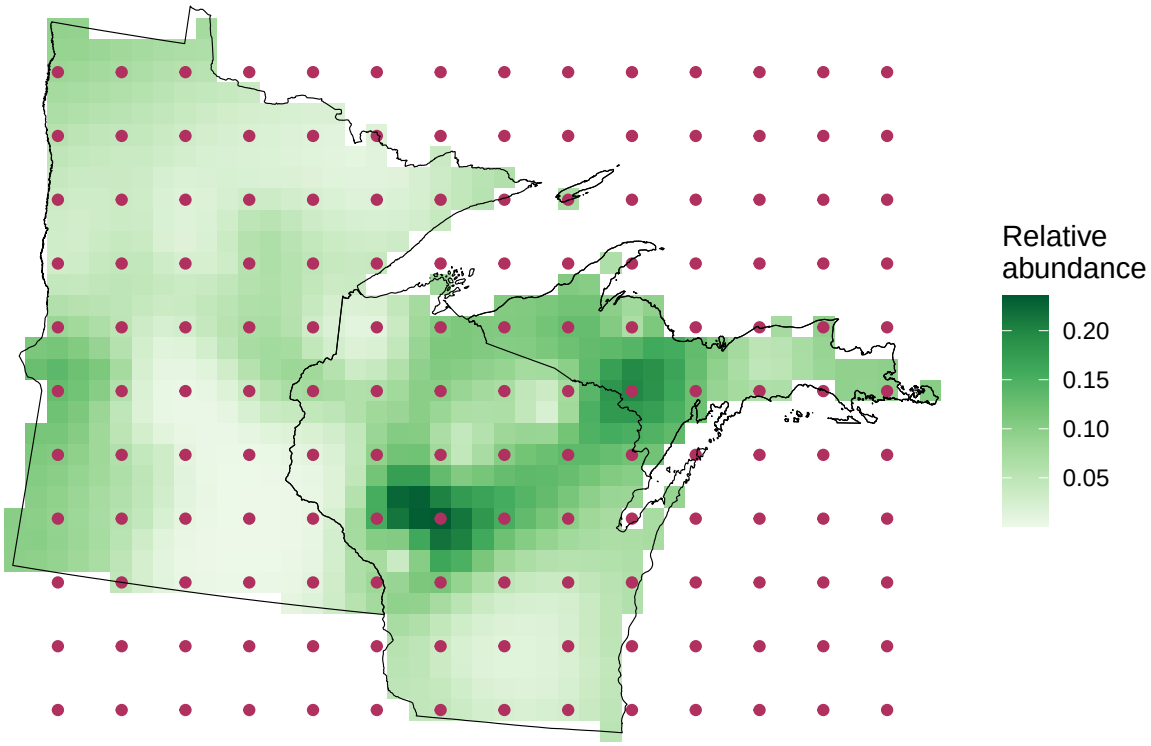
Elm



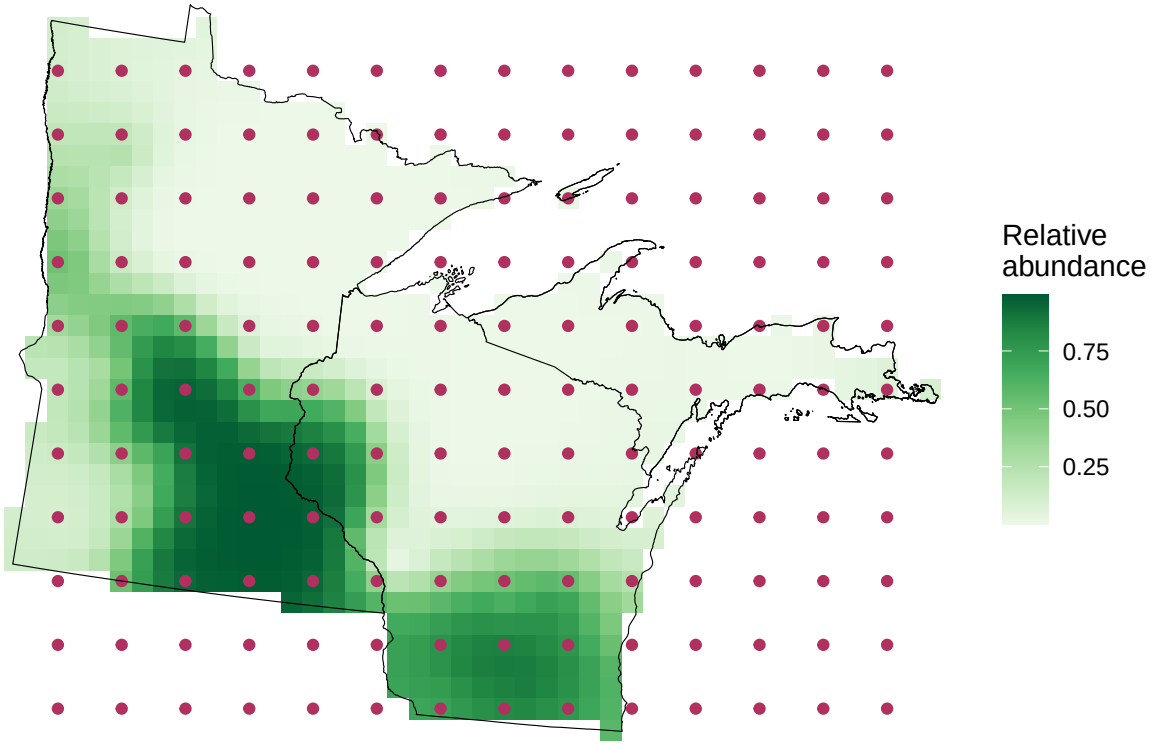
Hemlock



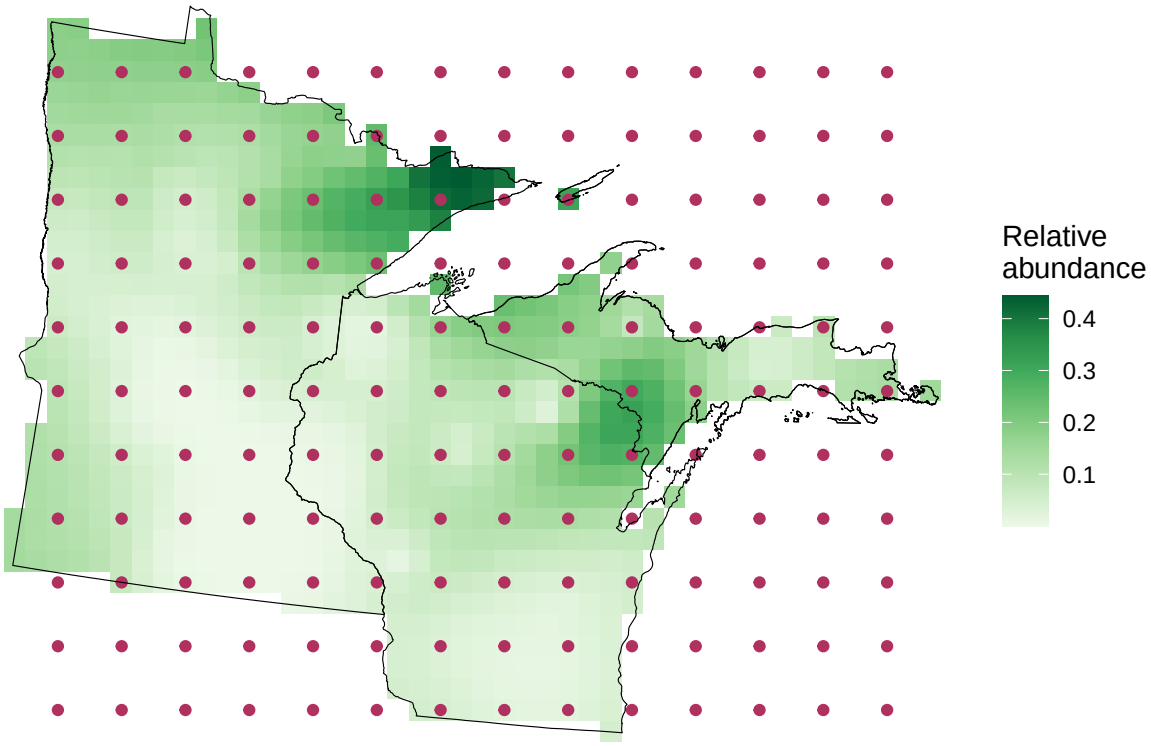
Maple


Relative
abundance
0.20
0.15
0.10
0.05

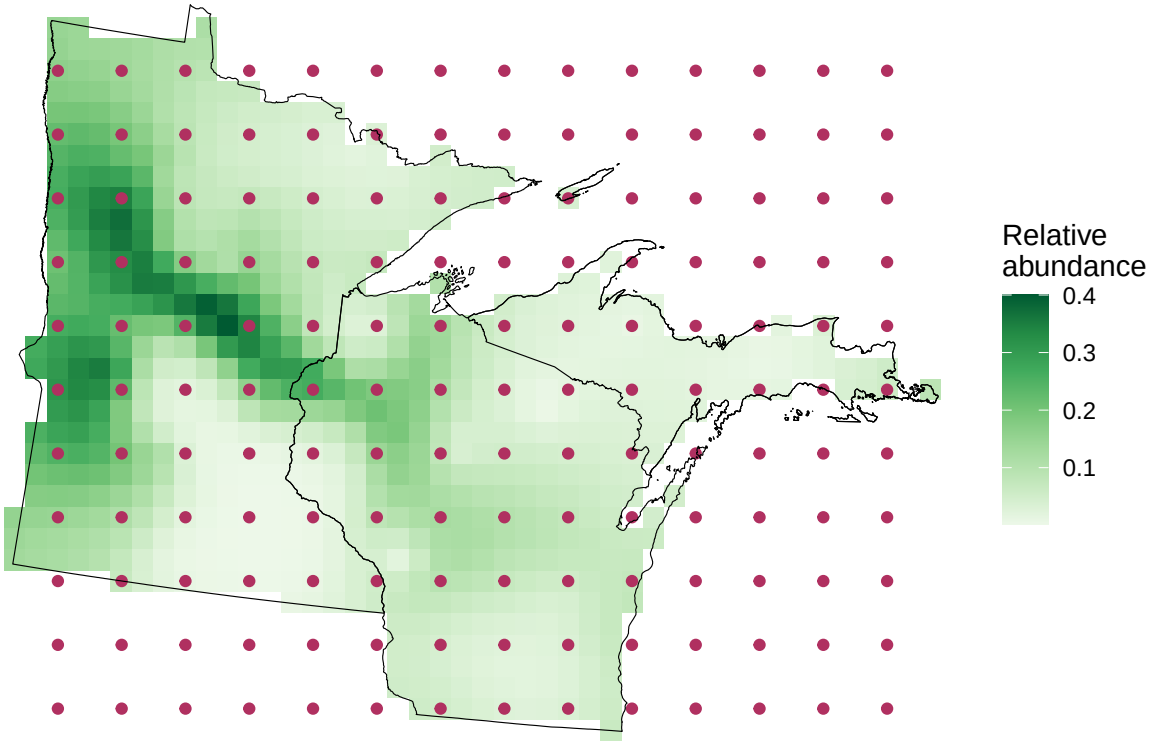
Oak


Relative
abundance
0.75
0.50
0.25

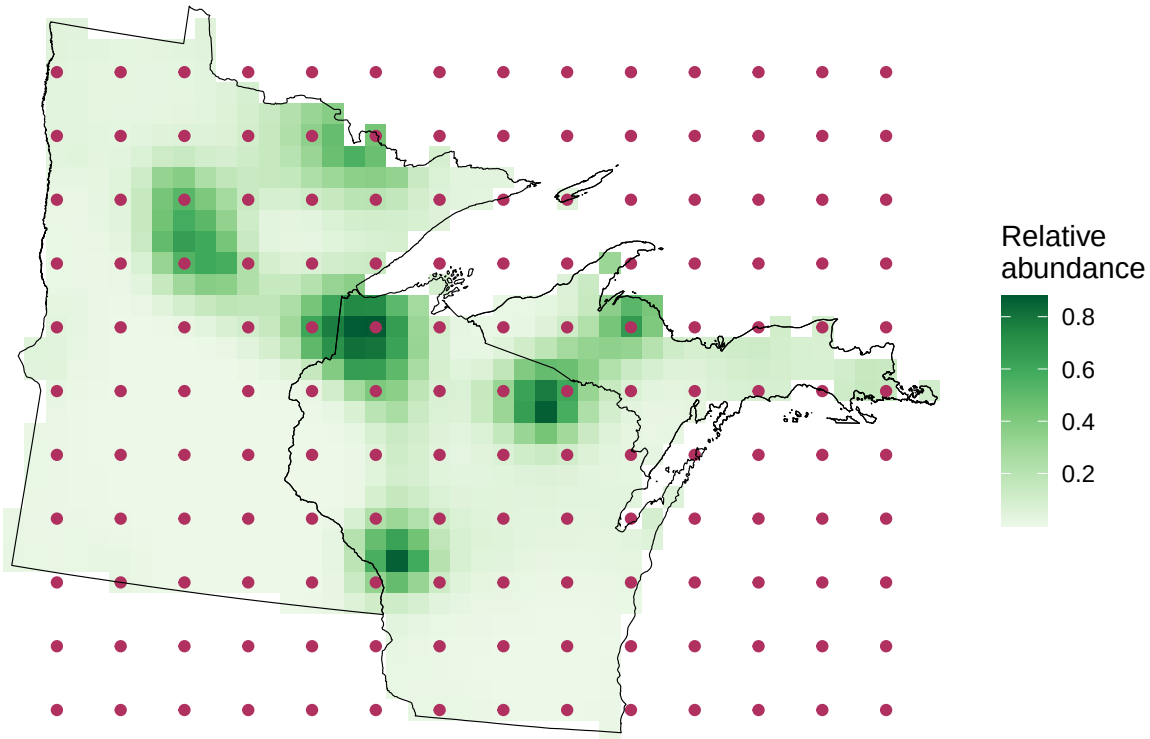
Other conifer



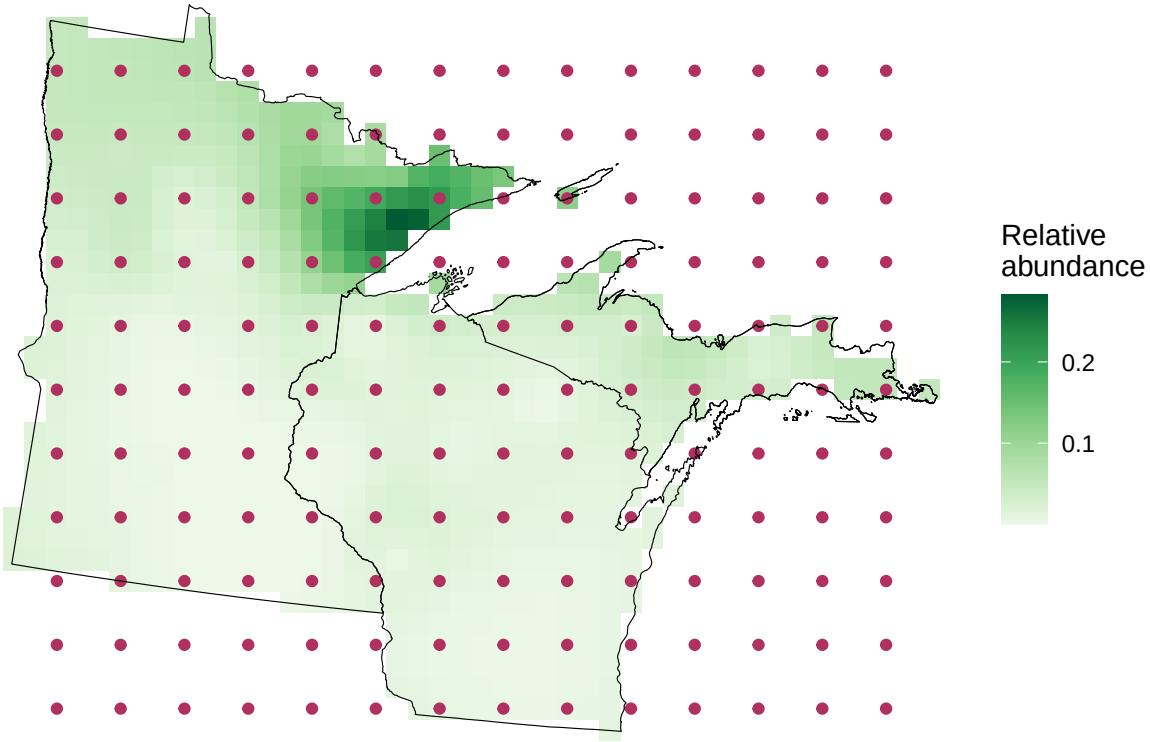
Other hardwood

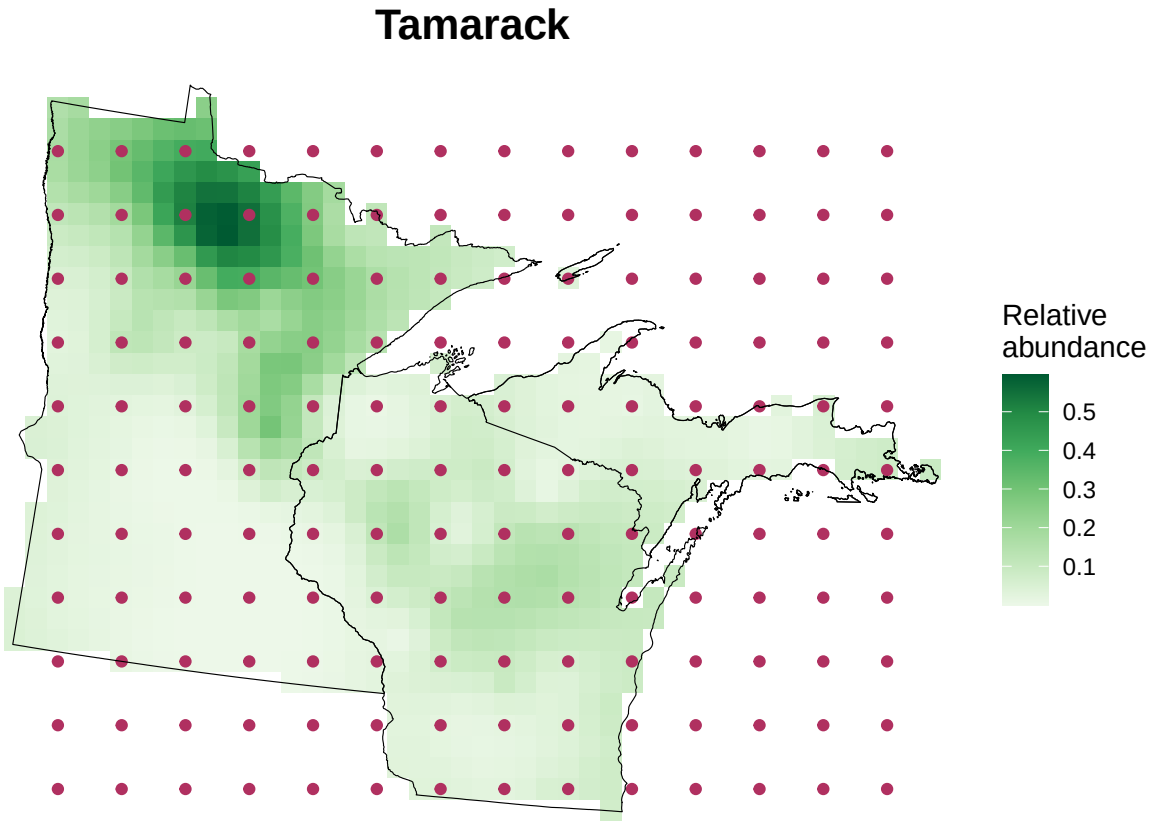


Pine



Spruce





Summary

None of the options are perfect. I have ranked each configuration in terms of number of grid cells retained in our spatial domain and the degree to which the variation in relative abundance of each taxon is captured, both separately and in a composite score. The scoring is completely subjective and based on visual assessment of each plot. Here is a summary of the different statistics for each configuration:

Sampling scheme	Number of cells	Ash	Beech	Birch	Elm	Hemlock	Maple	Oak	Other Conifer	Other Hardwood	Pine	Spruce	Tamarack	Composite	Percent
x=1, y=1	73	8	7	8	10	3	2	10	1	4	2	10	10	75	63
x=2, y=2	77	10	1	2	10	6	9	10	7	9	8	6	10	88	73
x=2, y=1	73	8	4	2	9	1	3	10	2	5	8	10	10	72	60
x=1, y=2	80	9	2	4	10	10	9	10	9	8	6	5	9	91	76
x=3, y=3	81	6	7	7	7	2	4	10	10	7	2	4	4	70	58
x=1, y=3	80	5	8	9	9	8	8	10	10	10	5	8	6	96	80
x=2, y=3	80	6	3	2	7	6	9	10	7	9	6	6	7	78	65
x=3, y=1	79	10	9	7	7	2	3	10	5	5	6	8	7	65	54
x=3, y=2	81	8	3	6	10	7	7	10	9	7	6	4	6	83	69

Not all taxa contribute equally to the forest (or savanna) community in this region. For example, ash only ever represents 8% of the stems in any grid cell, while oak relative abundance can approach 80%. To take this into account, I simply multipled each taxon's score by the maximum relative abundance according to the legend of the taxon relative abundance figures. Then, I added the scores together to create the composite again.

Sampling scheme	Number of cells	Ash	Beech	Birch	Elm	Hemlock	Maple	Oak	Other Conifer	Other Hardwood	Pine	Spruce	Tamarack	Composite	Percent
x=1, y=1	73	0.48	3.5	2.4	4	1.8	0.4	7.5	0.4	1.6	1.6	2	5	30.68	60
x=2, y=2	77	0.6	0.5	0.6	4	3.6	1.8	7.5	2.8	3.6	6.4	1.2	5	37.6	74
x=2, y=1	73	0.48	2	0.6	3.6	0.6	0.6	7.5	0.8	2	6.4	2	5	31.58	62
x=1, y=2	80	0.54	1	1.2	4	6	1.8	7.5	3.6	3.2	4.8	1	4.5	39.14	77
x=3, y=3	81	0.36	3.5	2.1	2.8	1.2	0.8	7.5	4	2.8	1.6	0.8	2	29.46	58
x=1, y=3	80	0.3	4	2.7	3.6	4.8	1.6	7.5	4	4	4	1.6	3	41.1	80
x=2, y=3	80	0.36	1.5	0.6	2.8	3.6	1.8	7.5	2.8	3.6	4.8	1.2	3.5	34.06	67
x=3, y=1	79	0.6	4.5	2.1	2.8	1.2	0.6	7.5	2	2	4.8	1.6	3.5	33.2	65
x=3, y=2	81	0.48	1.5	1.8	4	4.2	1.4	7.5	3.6	2.8	4.8	0.8	3	41.88	82

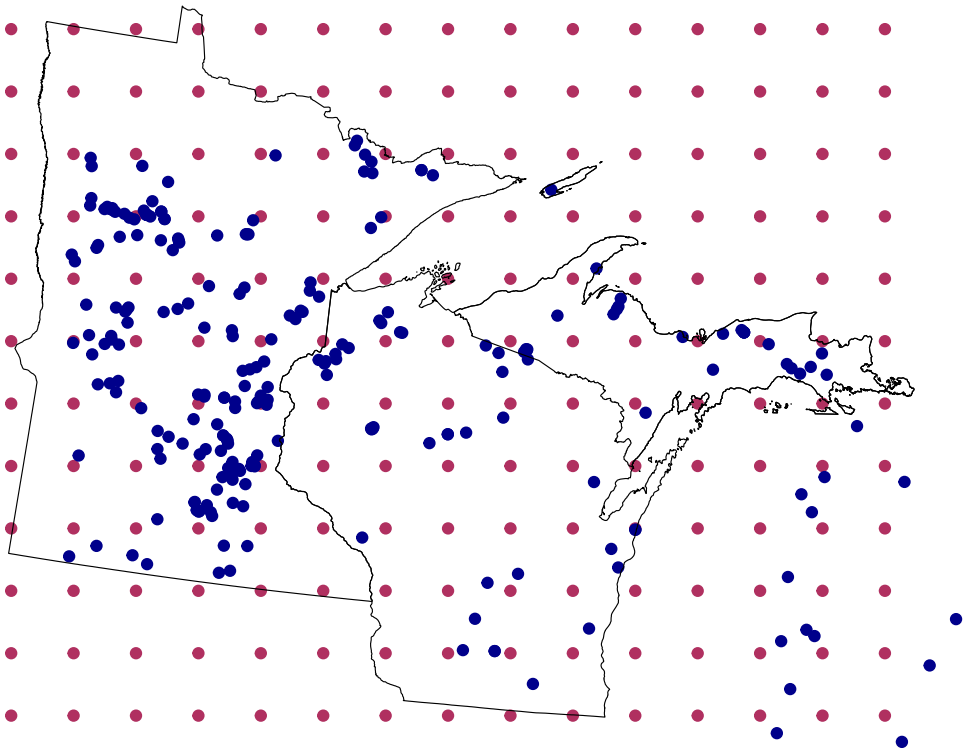
The results are fairly consistent between these two experiments. So far, it seems like x = 1, y = 3 is probably the best, with x = 1, y = 2 next and maybe x = 3, y = 2 doing fairly well.

Pollen sampling locations

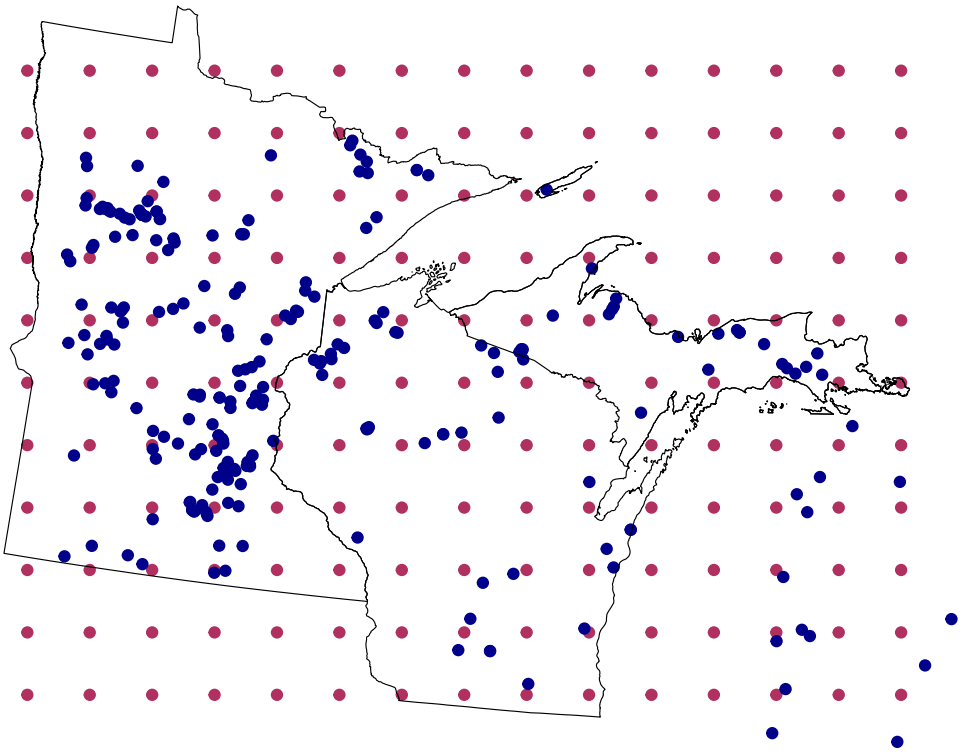
Next, I'd like to see how close our sampling schemes can get to the locations of the pollen sampling sites from which these relative abundances are estimated.

```
# Read in locations of plots from STEPPS github repo
ts <- readRDS('../data/raw/pollen_ts.RDS')

# Extract just unique sites
sites <- ts |>
  dplyr::select(id, sitename, lat, long) |>
  dplyr::distinct()

# Change coordinates
sites <- sf::st_as_sf(sites, coords = c('long', 'lat'),
                      crs = 'EPSG:4326')
sites <- sf::st_transform(sites, crs = 'EPSG:3175')
sites <- sfheaders::sf_to_df(sites, fill = TRUE)
sites <- dplyr::select(sites, -sfg_id, -point_id)

# Take every 3rd x and y coordinate starting at one
x_ind <- seq(from = 1, to = length(x), by = 3)
y_ind <- seq(from = 1, to = length(y), by = 3)

# Empty matrix
locs <- matrix(, nrow = length(x), ncol = length(y))

# Set cells we are keeping to TRUE
locs[x_ind, y_ind] <- TRUE

# Add x and y coordinates as dimension names
dimnames(locs) <- list(x,y)
# Melt to dataframe
locs_melt <- reshape2::melt(locs)
# Add column names
colnames(locs_melt) <- c('x', 'y', 'keep')
# Format
locs_melt <- dplyr::mutate(locs_melt,
                           keep = dplyr::if_else(is.na(keep), FALSE, keep))

locs_melt <- locs_melt |> dplyr::filter(keep == TRUE)

sites |>
  ggplot2::ggplot() +
  ggplot2::geom_point(data = locs_melt, ggplot2::aes(x = x, y = y), color = 'maroon') +
  ggplot2::geom_point(ggplot2::aes(x = x, y = y), color = 'darkblue') +
  ggplot2::geom_sf(data = states, color = 'black', fill = NA) +
  ggplot2::theme_void() +
  ggplot2::ggtitle('x = 1, y = 1') +
  ggplot2::theme(plot.title = ggplot2::element_text(size = 16, face = 'bold', hjust = 0.5))
```

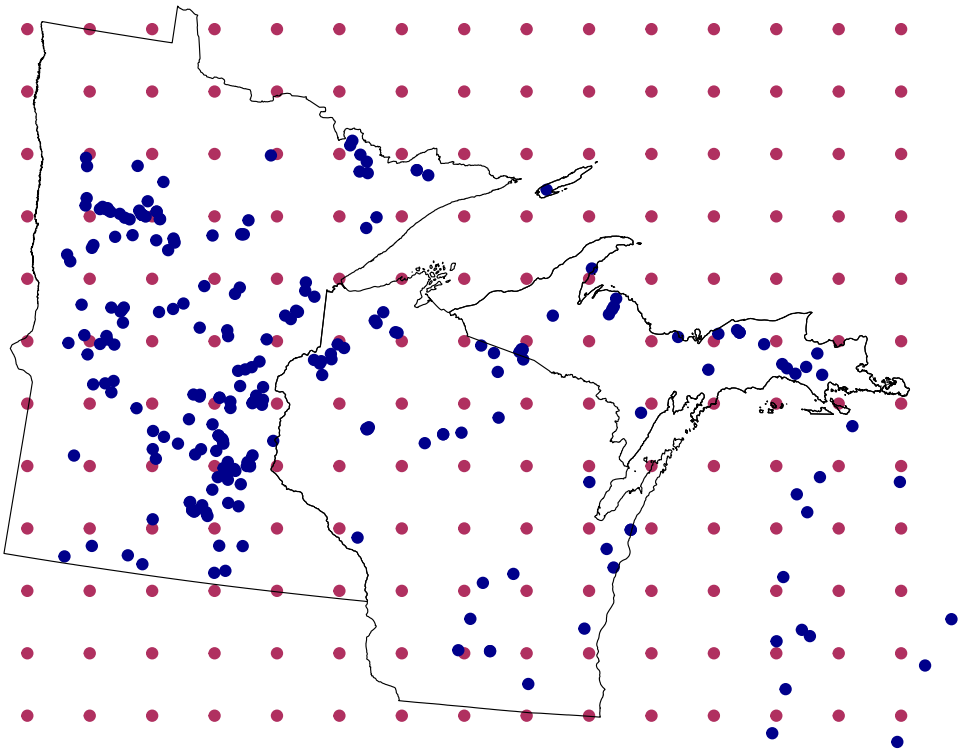
x = 1, y = 1



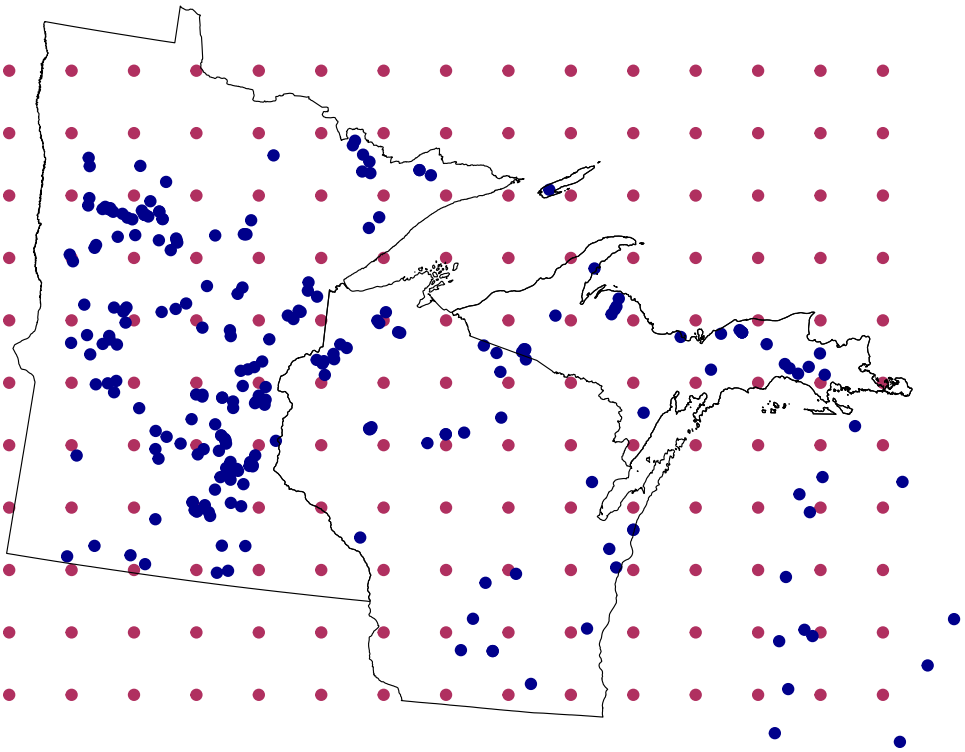
x = 2, y = 2



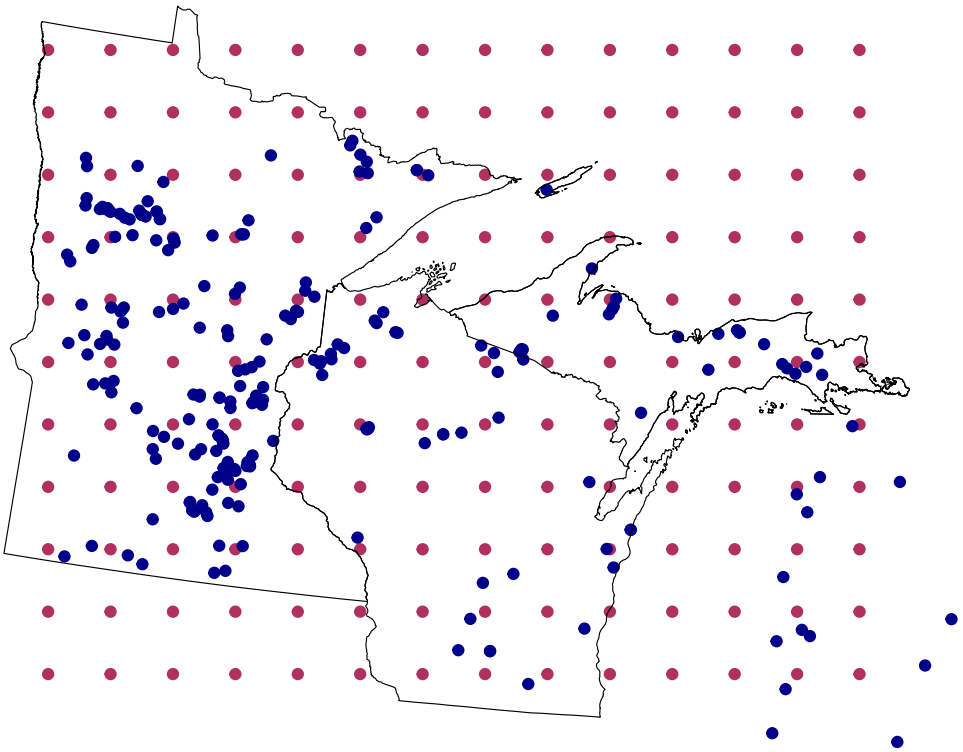
x = 2, y = 1



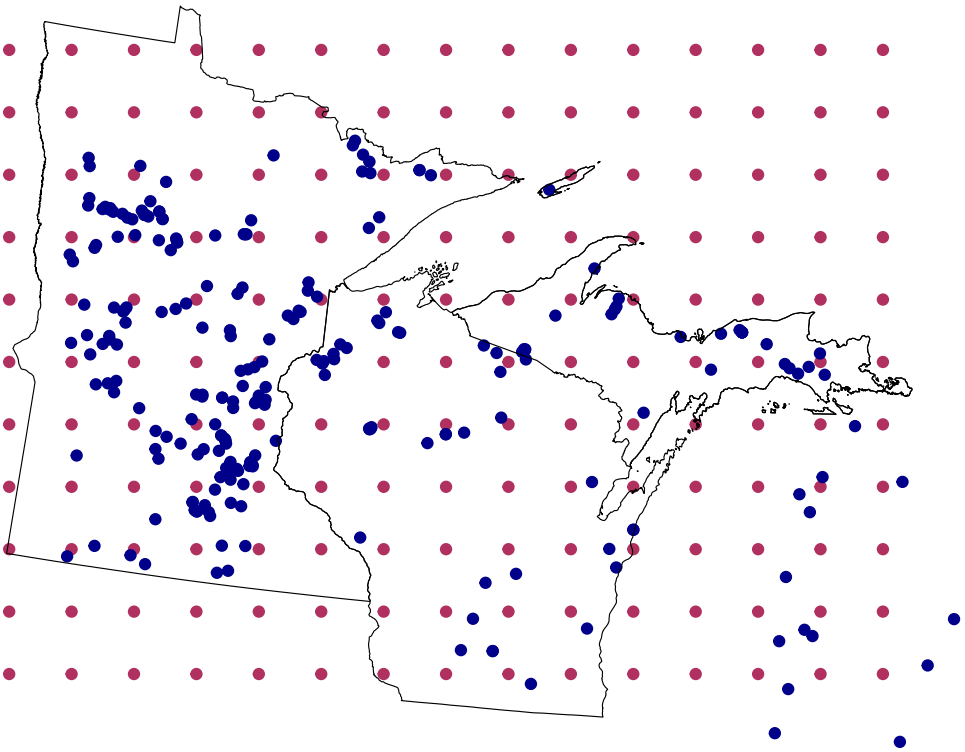
x = 1, y = 2



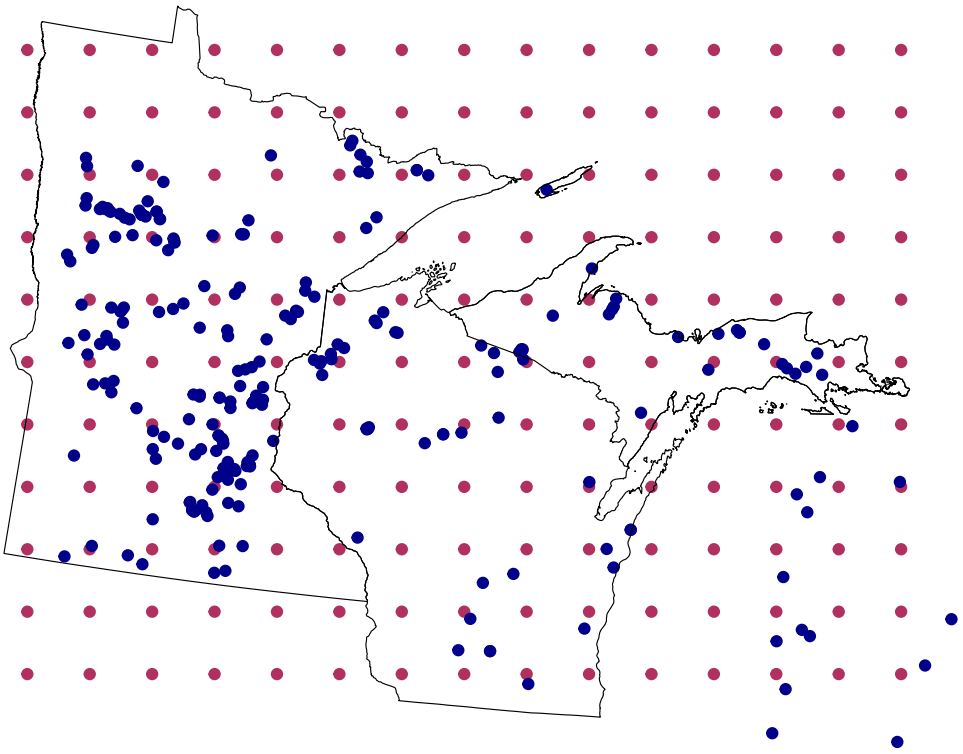
x = 3, y = 3



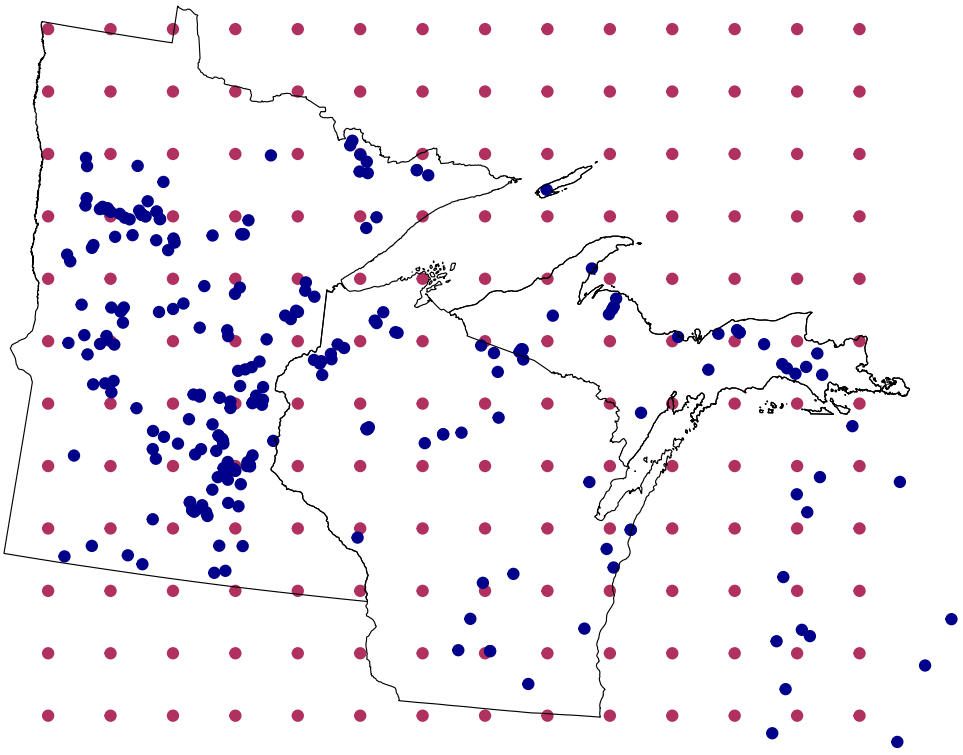
x = 1, y = 3

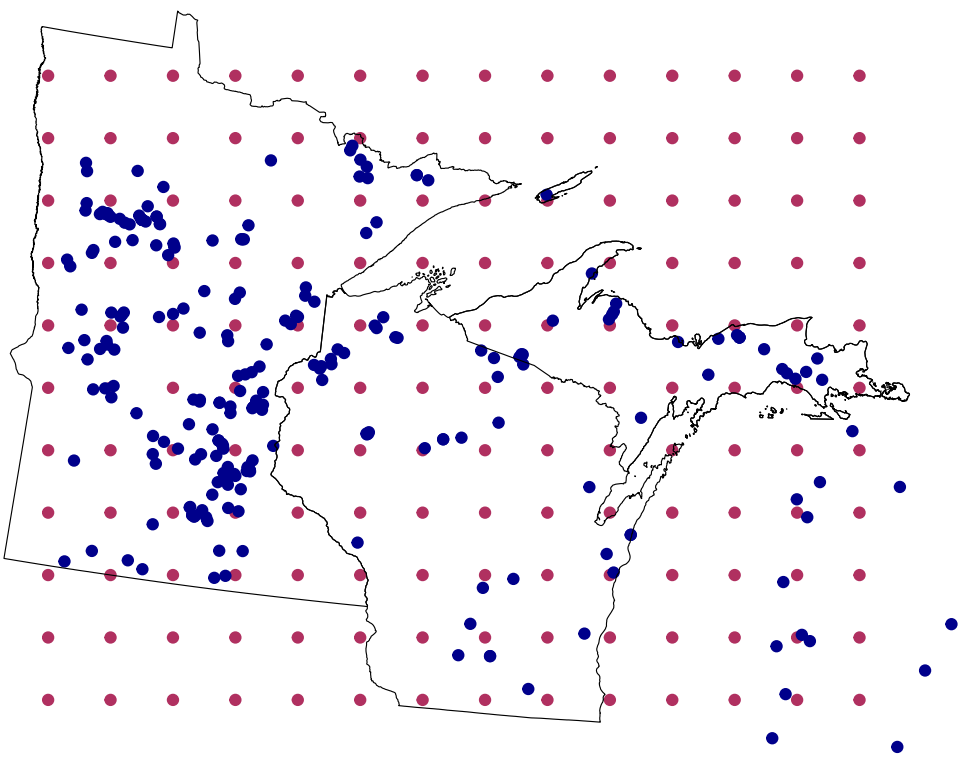


x = 2, y = 3



x = 3, y = 1




x = 3, y = 2