ECOLOGY WORKSHOP

Species Accumulation Curves

Gilles Colling
Division of BioInvasions, Global Change & Macroecology
University of Vienna

January 26–27, 2026

CONTENTS

Day 1

Theory

Biodiversity & Species Accumulation Curves

Day 1: Theory

Biodiversity & Species Accumulation Curves

Ecology Workshop

2026-01-25

What is Biodiversity?

Definition

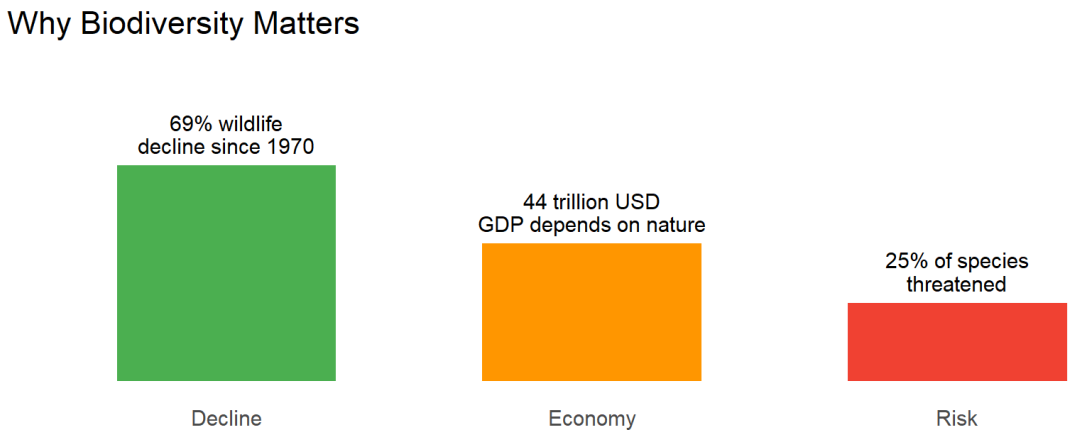
Biodiversity = the variety of life in an area

Compare two forests:

Oak, Oak, Oak, Oak, Oak	Oak, Beech, Maple, Pine, Birch
5 trees, **1 species**	5 trees, **5 species**

Which forest is more biodiverse? Forest B - even though both have the same number of trees.

Why Does Biodiversity Matter?



- **Ecosystem services**: Clean air, water filtration, pollination
- **Medicine**: 70% of cancer drugs come from natural sources
- **Food security**: Genetic diversity protects crops from disease

The Sampling Problem

The Counting Problem

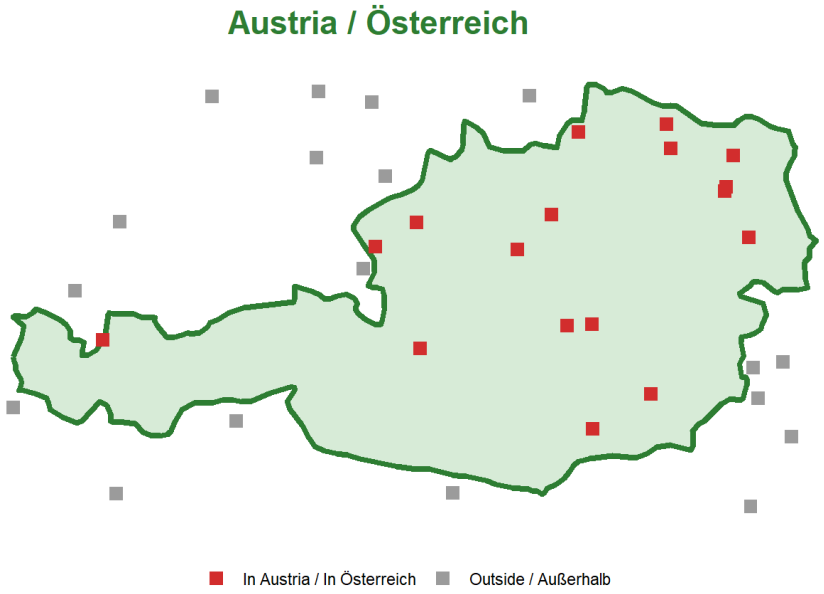
How many plant species are in Austria?

Can we count them all? Consider:

- Austria = 83,879 km^2
- Some species are tiny (mosses, lichens)
- Some only appear in certain seasons
- Some are extremely rare

The answer: Sampling!

We study small areas carefully and use statistics to estimate the total.



Key insight: We study small plots carefully, then use patterns to understand the whole.

Discussion Question

Imagine you want to know how many different types of candy exist in a candy store.

You have **5 minutes** and can look in **10 jars** (out of 100 total jars).

Which strategy is better?

A. Look at the 10 jars closest to the door

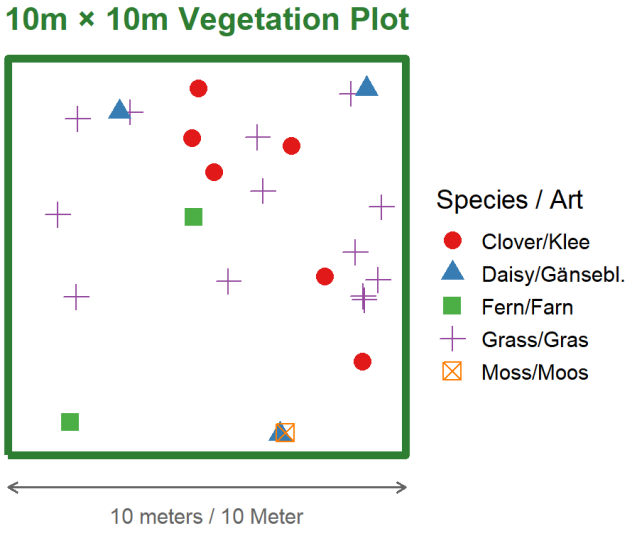
B. Look at 10 jars spread throughout the store

Why does this matter for ecology?

Vegetation Plots

A **vegetation plot** is a small area (typically 1-100 m^2) where ecologists record every plant species.

Location (coordinates)	47.0707° N, 15.4395° E
Date	2023-06-15
All species present	Festuca rubra, Trifolium repens, ...
Cover (%) per species	25%, 10%, ...



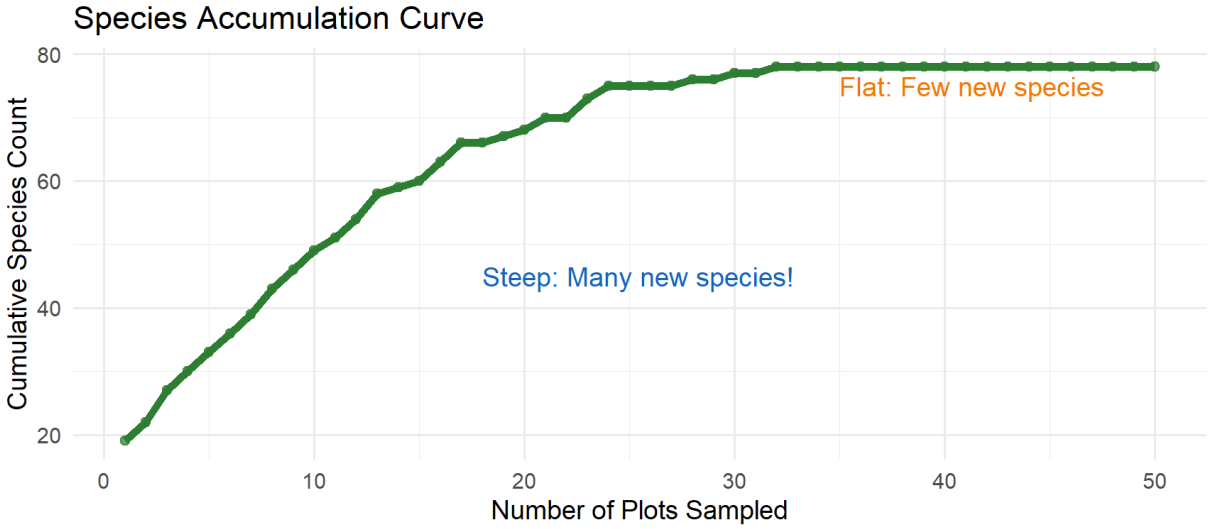
Species Accumulation Curves

The Core Idea

As we sample more plots, we find more species - but at a **decreasing rate**.

Why?

- First plots: Mostly **common** species (easy to find)
- Later plots: Mostly **rare** species (hard to find)
- Eventually: Very few new species per plot



Why Does the Curve Flatten?

Think of it like a **card game**:

1st card	Definitely new!
10th card	Probably still new
30th card	Maybe seen it before
50th card	Very likely a repeat

The more species you've already found, the harder it is to find new ones.

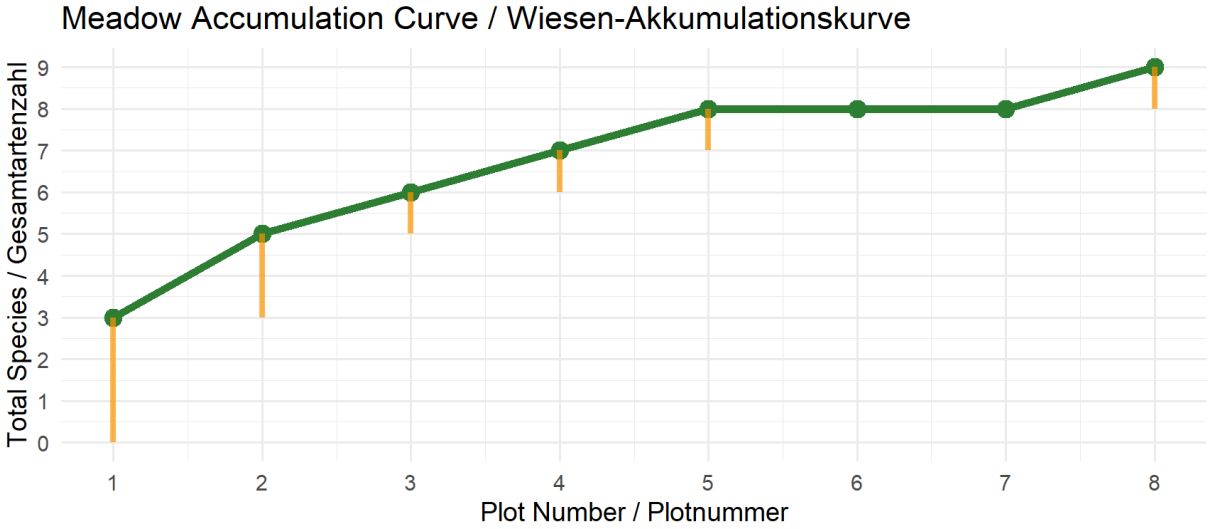
Let's Build One Together

Imagine we're sampling a meadow. Each "plot" reveals some species:

1	Daisy, Clover, Grass	3	3
2	Buttercup, Dandelion, Grass	2	5
3	Daisy, Thistle, Clover	1	6
4	Grass, Plantain, Daisy	1	7
5	Clover, Buttercup, Yarrow	1	8
6	Daisy, Grass, Clover	0	8
7	Dandelion, Plantain, Grass	0	8
8	Clover, Daisy, Sorrel	1	9

Notice: By plot 6-7, we're not finding anything new!

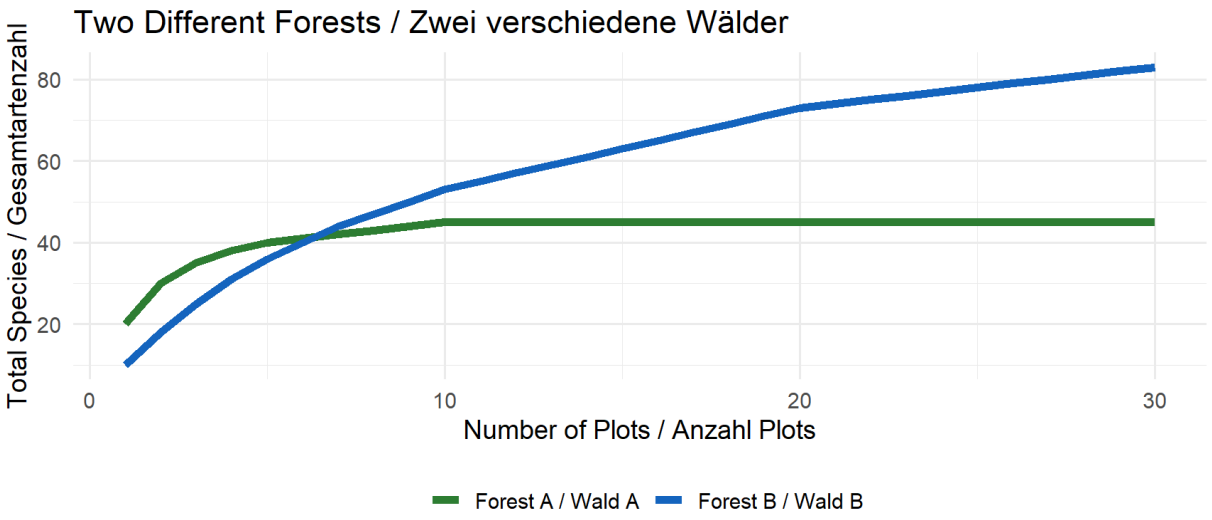
Visualizing Our Example



The **orange segments** show how many new species each plot adds. Notice how they get shorter!

Discussion Question

Look at these two curves:



Questions:

1. Which forest probably has more total species?
2. Which forest has species spread more evenly?
3. In which forest would you need to sample MORE plots to find everything?

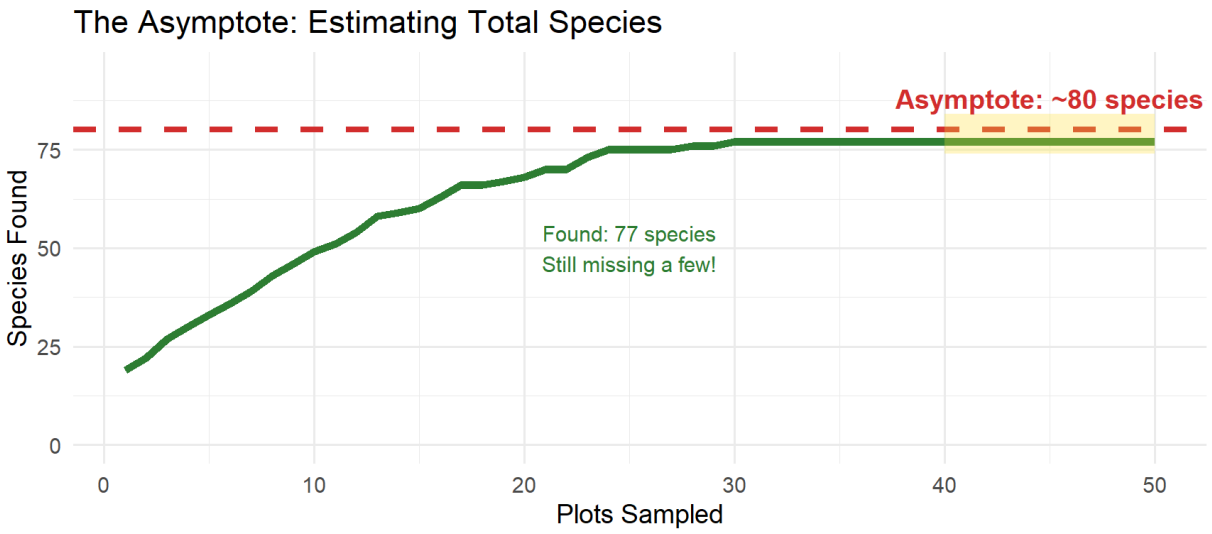
The Asymptote

Predicting the Unknowable

The **asymptote** is the line the curve approaches but never quite reaches.

It represents: **the estimated total number of species**

Even without sampling everything, we can estimate how much we're missing!



The European Vegetation Archive

EVA Database

EVA (European Vegetation Archive) is the largest vegetation database in Europe:

- **1.6+ million** plots
- **82** countries
- Data from **1900 to today**

For Austria, we have access to ~52,000 plots with location data.

Native vs. Alien Species

Our dataset distinguishes between:

Naturally occurring	Introduced by humans
Part of original ecosystem	Arrived after 1492
Co-evolved with other species	May outcompete natives

Research Question for Day 2: Do alien species accumulate differently than native species?

Why Not Excel?

Excel is Great For…

You probably know Excel well. It's perfect for:

- Quick calculations

- Simple tables and charts
- Small datasets (hundreds of rows)

Could we build accumulation curves in Excel? Yes! For a small example:

Excel: Manual formulas for each row

A Plot	B Species	C New	D Total
1	Rose, Daisy...	3	=C2
2	Clover, Grass...	2	=D2+C3
3	Rose, Clover...	1	=D3+C4
...	...	...	...

This works fine for 5 plots. But what about **52,000 plots**?

But Real Research Needs More

Our Austrian dataset has **52,000 plots** and **850,000 species observations**.

Load 850,000 rows	Crashes or freezes	2 seconds
Repeat analysis 100 times	Click 100 times	One loop
Share exact method	"I clicked here, then there..."	Share the code
Find a bug	Where did I click wrong?	Read line 47
Update with new data	Start over	Re-run script

Reproducibility: The Key Difference

In science, others must be able to **reproduce** your results.

With Excel: > "I sorted column B, then filtered for values > 10, then made a chart, then... I think I also removed some outliers?"

With R:

```
data %>%
  filter(value > 10) %>%
  ggplot(aes(x, y)) +
  geom_line()
```

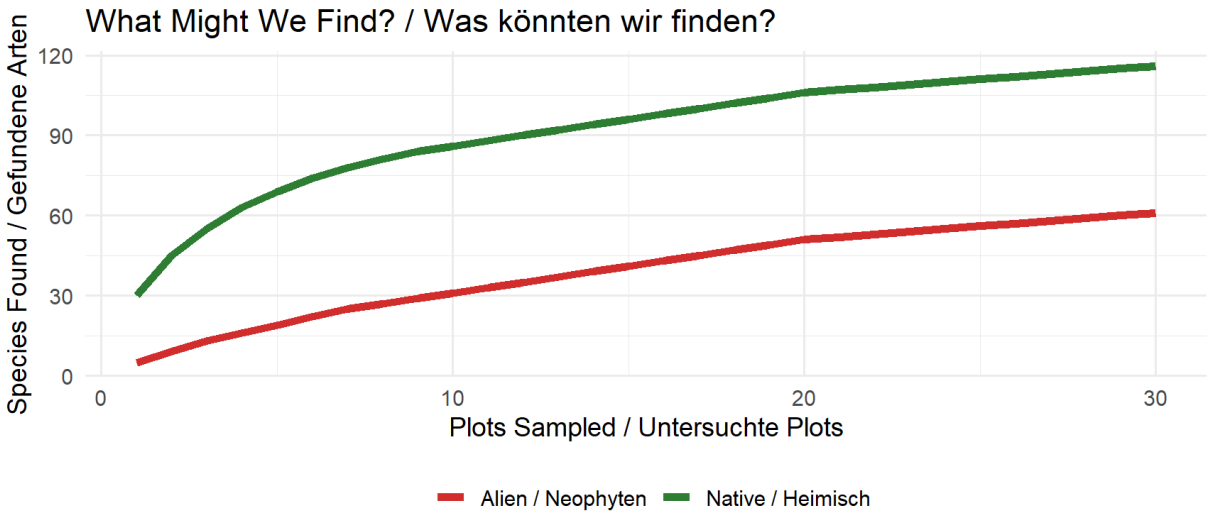
The code IS the method. Anyone can run it and get the same result.

Summary

Key Concepts

Biodiversity	Variety of species in an area
Vegetation plot	Small area where all plants are recorded
Accumulation curve	Graph showing species discovery over sampling effort
Asymptote	Estimated total species (curve approaches but never reaches)
Native species	Naturally occurring in the region
Alien species	Introduced by humans (after 1492)

Preview: Day 2



Tomorrow we will:

1. Learn the **nearest-neighbor algorithm** for building spatial curves
2. Understand why **plot order matters**
3. Build curves with **multiple starting points**
4. Compare **native vs. alien** accumulation patterns
5. Interpret results and discuss **biological mechanisms**

Hypothesis to test: Do alien species accumulate differently than native species?

Day 1

Exercises

R Basics to First Data Analysis

Day 1: Exercises

Your First Accumulation Curve!

Fill in the blanks!

2026-01-26

Part 1: R Basics - Getting Started

Welcome to R! In this section, you'll learn the fundamental building blocks of R programming. Take your time with each exercise - understanding these basics thoroughly will make everything else easier.

How to use these exercises:

- Replace each `_____` with the correct code
- Run each chunk to see if it works
- Read the hints if you get stuck

Exercise 1: Your First Variables

Concept: Variables are containers that store values. In R, we use `<-` to assign values to variables.

1a: Creating simple variables

```
# Store the number 42 in a variable called "answer"
# Speichere die Zahl 42 in einer Variable namens "answer"
answer <- ____
print(answer)

# Store your age in a variable
# Speichere dein Alter in einer Variable
my_age <- ____
print(my_age)

# Store a decimal number (height in meters)
# Speichere eine Dezimalzahl (Größe in Metern)
my_height <- ____
print(my_height)
```

1b: Text variables (strings)

```
# Text variables - use quotes!
# Text-Variablen - benutze Anführungszeichen!

# Store your name (text needs quotes)
# Speichere deinen Namen (Text braucht Anführungszeichen)
my_name <- "____"
print(my_name)

# Store a species name
# Speichere einen Artnamen
species <- "____"
print(species)
```

1c: Logical (TRUE/FALSE) variables

```
# Logical variables - TRUE or FALSE
# Logische Variablen - TRUE oder FALSE

# Is it raining? (answer TRUE or FALSE, no quotes!)
# Regnet es? (antworte TRUE oder FALSE, keine Anführungszeichen!)
is_raining <- ____

# Is 5 greater than 3? (R can calculate this!)
# Ist 5 größer als 3? (R kann das berechnen!)
five_greater_three <- 5 > 3
print(five_greater_three)

# Is 10 equal to 10? (use == for comparison)
# Ist 10 gleich 10? (benutze == für Vergleich)
ten_equals_ten <- 10 ____ 10
print(ten_equals_ten)

# Is "oak" equal to "Oak"? (R is case-sensitive!)
# Ist "oak" gleich "Oak"? (R unterscheidet Groß/Kleinschreibung!)
oak_equals_Oak <- "oak" == "Oak"
print(oak_equals_Oak)
```

1d: Basic math with variables

```
# Create two number variables
# Erstelle zwei Zahlen-Variablen
a <- 10
b <- 3

# Addition
sum_ab <- a ____ b
print(sum_ab)  # Should be 13 / Sollte 13 sein

# Subtraction / Subtraktion
diff_ab <- a ____ b
print(diff_ab)  # Should be 7 / Sollte 7 sein

# Multiplication / Multiplikation
prod_ab <- a ____ b
print(prod_ab)  # Should be 30 / Sollte 30 sein

# Division
div_ab <- a ____ b
print(div_ab)  # Should be 3.333... / Sollte 3.333... sein

# Power (10 to the power of 3) / Potenz (10 hoch 3)
power_ab <- a ____ b
print(power_ab)  # Should be 1000 / Sollte 1000 sein
```

1e: Updating variables

```
# Variables can change!
# Variablen können sich ändern!

# Start with a count of species
# Starte mit einer Artenanzahl
species_count <- 0
print(species_count)

# Found 5 species / 5 Arten gefunden
species_count <- species_count + ____
print(species_count)  # Should be 5 / Sollte 5 sein

# Found 3 more / 3 weitere gefunden
species_count <- species_count + ____
print(species_count)  # Should be 8 / Sollte 8 sein

# Oops, 2 were misidentified (subtract)
# Ups, 2 waren falsch bestimmt (subtrahieren)
species_count <- species_count - ____
print(species_count)  # Should be 6 / Sollte 6 sein
```

💡

Hint: Math operators are `+` `-` `*` `/` and `^` for power. Use `==` for “equals” comparison!

Exercise 2: Vectors - Lists of Values

Concept: A **vector** is a list of values of the same type. We use `c()` to combine values into a vector.

2a: Creating vectors

```
# Create a vector of numbers
# Erstelle einen Vektor von Zahlen
species_counts <- c(5, 3, 7, 2, 4)
print(species_counts)

# Your turn - create a vector of 4 tree heights
# Du bist dran - erstelle einen Vektor mit 4 Baumhöhen
tree_heights <- c(____, ____, ____, ____)
print(tree_heights)

# Create a vector of species names
# Erstelle einen Vektor von Artnamen
tree_species <- c("Oak", "Beech", "Pine", "Maple")
print(tree_species)
```

2b: Vector functions

```
# How many elements in the vector?
# Wie viele Elemente im Vektor?
length(species_counts)

# Calculate the sum
# Berechne die Summe
sum(species_counts)

# Calculate the mean (average)
# Berechne den Mittelwert (Durchschnitt)
mean(species_counts)

# Your turn - calculate mean of your tree heights
# Du bist dran - berechne den Mittelwert deiner Baumhöhen
mean(____)

# Find min and max
# Finde Minimum und Maximum
min(species_counts)
max(species_counts)
```

2c: Combining vectors

```
# Species from different plots
# Arten von verschiedenen Plots
plot1 <- c("Oak", "Beech", "Pine")
plot2 <- c("Beech", "Maple", "Oak")
plot3 <- c("Birch", "Oak", "Ash")

# Combine all into one vector
# Kombiniere alle in einen Vektor
all_species <- c(plot1, plot2, plot3)
print(all_species)

# How many total observations?
# Wie viele Beobachtungen insgesamt?
length(all_species)
```

💡

Hint: `c()` combines values. `length()` counts elements. `mean()` calculates average.

Exercise 3: The Magic of unique()

Concept: `unique()` is the **key function** for ecology! It tells us how many *different* species we found (not counting duplicates).

```r
# 3a: We observed these species (some repeats!)
# 3a: Wir haben diese Arten beobachtet (einige Wiederholungen!)

observations <- c("Oak", "Beech", "Oak", "Pine", "Beech", "Oak", "Maple")

# How many observations total?
# Wie viele Beobachtungen insgesamt?
length(observations)

# 3b: How many UNIQUE species?
# 3b: Wie viele EINZIGARTIGE Arten?
unique(observations)
length(unique(observations))

# 3c: Using our plots from before
# 3c: Benutze unsere Plots von vorher
plot1 <- c("Oak", "Beech", "Pine")
plot2 <- c("Beech", "Maple", "Oak")
plot3 <- c("Birch", "Oak", "Ash")

all_species <- c(plot1, plot2, plot3)

# Total observations vs unique species
# Gesamte Beobachtungen vs einzigartige Arten
length(all_species)           # Total / Gesamt
length(unique(all_species))   # Unique / Einzigartig

# 3d: Your turn - how many unique species after plot1 + plot2?
# 3d: Du bist dran - wie viele einzigartige Arten nach plot1 + plot2?
after_two_plots <- c(plot1, ____)
length(unique(after_two_plots))
```

💡

Hint: Species richness = `length(unique(species_list))` . This is THE key formula!

Exercise 4: Data Frames

A **data frame** is like a spreadsheet - it has rows and columns. Let's practice with a small example before loading real data!

```r
# 4a: Create a simple data frame
# 4a: Erstelle einen einfachen Data Frame

my_data <- data.frame(
  species = c("Oak", "Beech", "Pine", "Maple"),
  height = c(20, 18, 25, 15),
  native = c(TRUE, TRUE, TRUE, FALSE)
)

my_data

# 4b: How many rows and columns?
# 4b: Wie viele Zeilen und Spalten?

nrow(my_data)
ncol(my_data)

# 4c: Access a column with $
# 4c: Greife auf eine Spalte mit $ zu

my_data$species
my_data$height

# 4d: Your turn - get the mean height
# 4d: Du bist dran - berechne die mittlere Höhe

mean(my_data$___)

# 4e: Access a specific row (row 2)
# 4e: Greife auf eine bestimmte Zeile zu (Zeile 2)

my_data[2, ]

# 4f: Access a specific cell (row 2, column "height")
# 4f: Greife auf eine bestimmte Zelle zu (Zeile 2, Spalte "height")

my_data[2, "height"]
```

💡 Hint: `$` accesses columns by name. Square brackets `[row, column]` access specific cells.

Exercise 5: Loading Real Data

Now let's load the **Austrian vegetation data**!

```r
# 5a: Load the tidyverse package
# 5a: Lade das tidyverse Paket

library(tidyverse)

# 5b: Load the species data
# 5b: Lade die Artdaten

species_data <- read_csv("../data/austria_species.csv")

# 5c: Look at the first few rows
# 5c: Schau dir die ersten Zeilen an

head(species_data)

# 5d: How many rows (observations)?
# 5d: Wie viele Zeilen (Beobachtungen)?

nrow(species_data)

# 5e: What columns do we have?
# 5e: Welche Spalten haben wir?

colnames(species_data)

# 5f: How many unique plots?
# 5f: Wie viele einzigartige Plots?

length(unique(species_data$PlotObservationID))

# 5g: How many unique species in the whole dataset?
# 5g: Wie viele einzigartige Arten im gesamten Datensatz?

length(unique(species_data$___))
```

💡 Hint: Use `$` to access a column, e.g. `data$column_name`

Exercise 6: Filtering Data

Often we want to look at only **part** of our data. The `filter()` function helps us select specific rows!

```r
# 6a: Look at the STATUS column - what values are there?
# 6a: Schau dir die STATUS-Spalte an - welche Werte gibt es?

unique(species_data$STATUS)

# 6b: Filter for native species only (STATUS == "nat")
# 6b: Filtere nur heimische Arten (STATUS == "nat")

native_only <- species_data %>%
  filter(STATUS == "nat")

nrow(native_only)

# 6c: Filter for alien/invasive species (STATUS == "neo")
# 6c: Filtere Alien/invasive Arten (STATUS == "neo")

alien_only <- species_data %>%
  filter(STATUS == "neo")

nrow(alien_only)

# 6d: How many unique native species?
# 6d: Wie viele einzigartige heimische Arten?

length(unique(native_only$WFO_TAXON))

# 6e: Your turn - how many unique alien species?
# 6e: Du bist dran - wie viele einzigartige Alien-Arten?

length(unique(___$WFO_TAXON))

# 6f: Filter for a specific plot
# 6f: Filtere einen bestimmten Plot

first_plot <- unique(species_data$PlotObservationID)[1]
first_plot

one_plot <- species_data %>%
  filter(PlotObservationID == first_plot)

# How many species in this one plot?
# Wie viele Arten in diesem einen Plot?

length(unique(one_plot$WFO_TAXON))
```

💡

Hint: `filter()` keeps rows where the condition is TRUE. Use `==` for "equals".

Exercise 7: Building the Accumulation Curve

Now the exciting part - let's build a species accumulation curve step by step!

```r
# 7a: Get a list of all unique plots
# 7a: Hole eine Liste aller einzigartigen Plots

all_plots <- unique(species_data$PlotObservationID)
head(all_plots)

# 7b: Take a small sample (first 50 plots for speed)
# 7b: Nimm eine kleine Stichprobe (erste 50 Plots für Geschwindigkeit)

sample_plots <- all_plots[1:50]

# 7c: Initialize our accumulation tracking
# 7c: Initialisiere unsere Akkumulations-Verfolgung

species_found <- c()           # Empty vector to collect species
                               # Leerer Vektor um Arten zu sammeln
accumulation <- c()            # Will store species count after each plot
                               # Speichert Artenzahl nach jedem Plot

# 7d: Loop through each plot and count species!
# 7d: Schleife durch jeden Plot und zähle Arten!

for (i in 1:length(sample_plots)) {

  # Get current plot ID
  # Hole aktuelle Plot-ID
  current_plot <- sample_plots[i]

  # Get species in this plot
  # Hole Arten in diesem Plot
  plot_species <- species_data$WFO_TAXON[species_data$PlotObservationID == current_plot]

  # Add to our collection
  # Füge zu unserer Sammlung hinzu
  species_found <- c(species_found, plot_species)

  # Count unique species so far
  # Zähle einzigartige Arten bisher
  total_species <- length(unique(species_found))

  # Store in accumulation vector
  # Speichere im Akkumulationsvektor
  accumulation <- c(accumulation, total_species)
}

# 7e: Look at the result!
# 7e: Schau dir das Ergebnis an!

accumulation
```

Exercise 8: Plot Your Curve!

The final step - let's visualize our accumulation curve!

```r
# 8a: Create a data frame for plotting
# 8a: Erstelle einen Data Frame zum Plotten

curve_data <- data.frame(
  plots = 1:length(accumulation),
  species = accumulation
)

head(curve_data)

# 8b: Make the plot!
# 8b: Mache den Plot!

ggplot(curve_data, aes(x = plots, y = species)) +
  geom_line(color = "#2E7D32", linewidth = 2) +
  geom_point(color = "#2E7D32", size = 2) +
  labs(
    title = "My First Species Accumulation Curve!",
    x = "Number of Plots Sampled",
    y = "Cumulative Species Count"
  ) +
  theme_minimal(base_size = 14)

# 8c: Make it prettier with a filled area
# 8c: Mache es schöner mit gefüllter Fläche

ggplot(curve_data, aes(x = plots, y = species)) +
  geom_area(fill = "#C8E6C9", alpha = 0.5) +
  geom_line(color = "#2E7D32", linewidth = 2) +
  labs(
    title = "Species Accumulation Curve - Austria",
    subtitle = paste("Total species found:", max(accumulation)),
    x = "Plots Sampled",
    y = "Cumulative Species"
  ) +
  theme_minimal(base_size = 14)
```

💡

Hint: The curve flattens because we find fewer NEW species as we sample more plots!

Bonus: Compare Native vs Alien!

If you finish early, try comparing native and alien species!

```
# The species_data already has a STATUS column!
# Die species_data hat bereits eine STATUS-Spalte!

# Filter for native species only
# Filtere nur heimische Arten

native_species <- species_data %>%
  filter(STATUS == "nat")

# How many native species?
# Wie viele heimische Arten?

length(unique(native_species$WFO_TAXON))

# How many alien species?
# Wie viele Alien-Arten?

alien_species <- species_data %>%
  filter(STATUS == "neo")

length(unique(alien_species$WFO_TAXON))

# Now you can build separate curves for native vs alien!
# Jetzt kannst du separate Kurven für heimisch vs alien bauen!
```

You did it! You built your first species accumulation curve using real Austrian data!

Key concepts learned: - Variables store data with `<-` - `c()` creates vectors - `unique()` counts different values - Data frames organize data in rows and columns - `filter()` selects specific rows - Loops let us repeat actions - `ggplot()` creates beautiful plots

Day 2

Theory

Algorithms & Spatial Ordering

Day 2: Theory

Spatial Algorithms & Research Design

Ecology Workshop

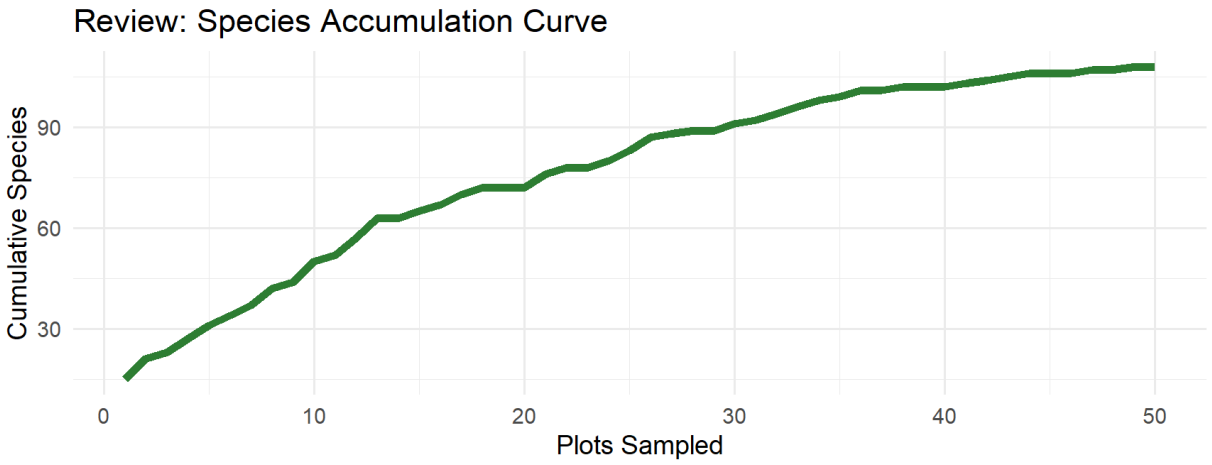
2026-01-28

Review & Research Question

Quick Review from Day 1

Yesterday we learned:

- **Biodiversity** = variety of species in an area
- **Accumulation curves** show how species discovery slows with sampling effort
- **Asymptote** = estimated total species (what we're approaching)
- Austrian data has **native** (natural) and **alien** (introduced) species



Today's Research Question

Do alien species accumulate differently than native species?

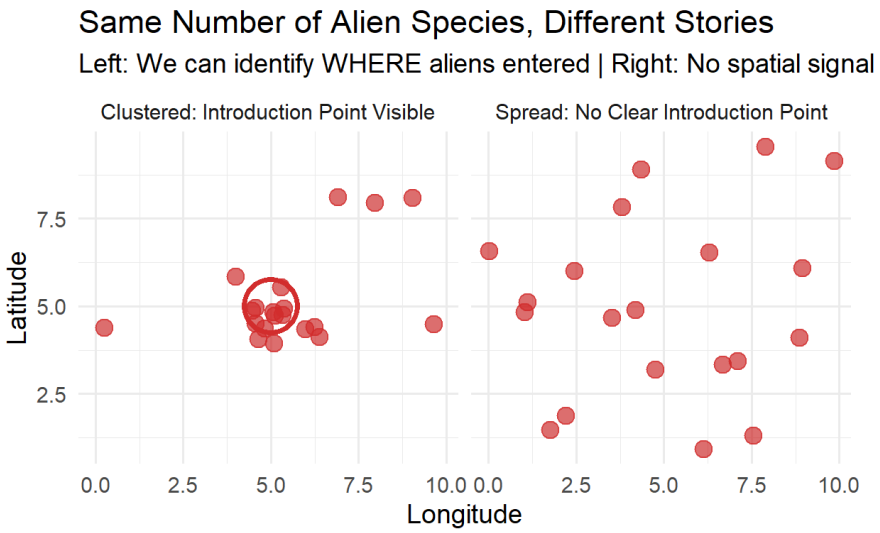
Hypotheses:

- **H1:** Alien species accumulate **more slowly** (fewer species overall)
- **H2:** Alien species show **different spatial patterns** than natives
- **H3:** The curves will have **different shapes**

Why Use Spatial Accumulation Curves?

The key question: Why not just count species? Why do we need accumulation curves?

Answer: Accumulation curves reveal **where** species are, not just **how many**.



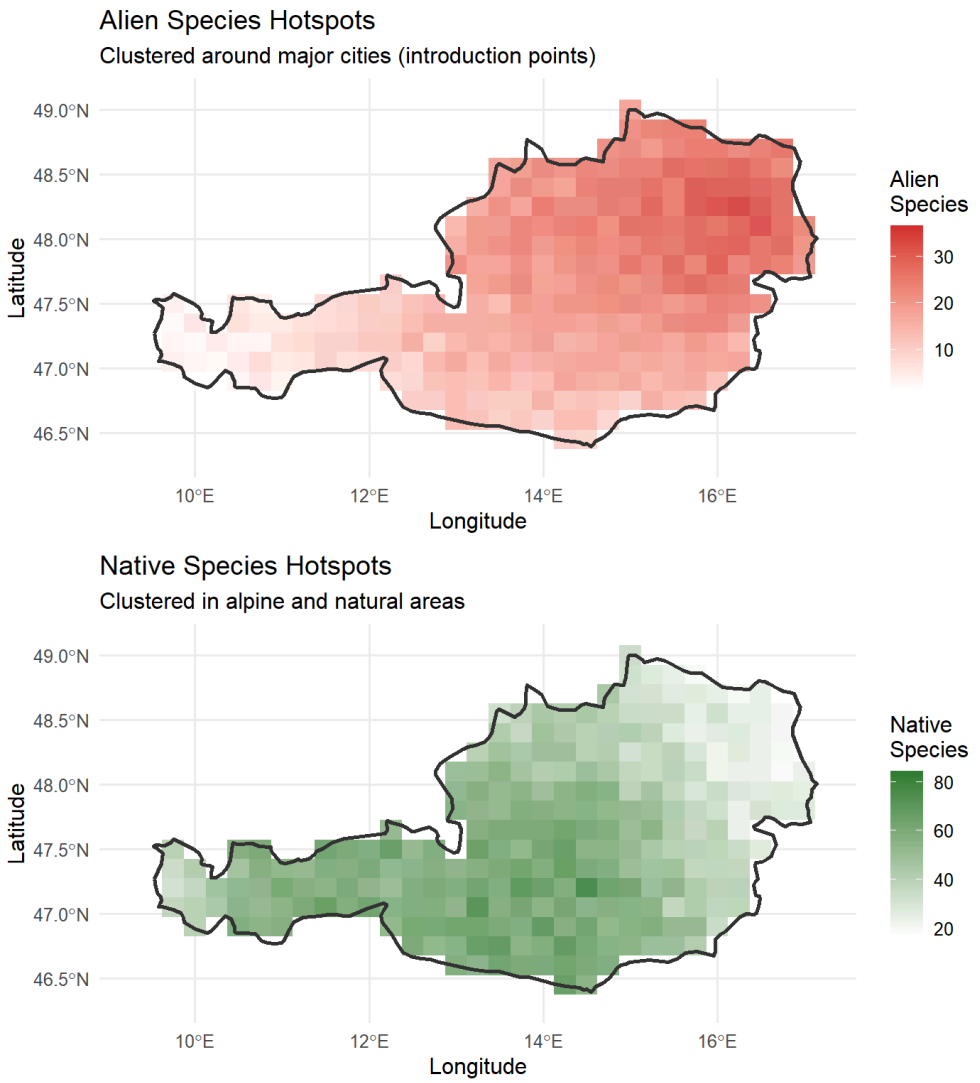
Why this matters for invasion biology:

Introduction points	Where did aliens first arrive? (ports, cities, gardens)
Spread patterns	How are they dispersing across the landscape?
Invasion hotspots	Which areas have the most alien species?
Management priorities	Where should we focus control efforts?

The curve shape tells the story: If aliens accumulate slowly at first then speed up, they're clustered around introduction points. If they accumulate steadily, they've already spread everywhere.

Hotspot Maps: Austria Example

What might hotspot maps look like for Austria? Here's a hypothetical example showing how alien and native species might be distributed differently across the landscape.



Key observations (hypothetical):

- **Alien species** concentrate around **Vienna, Linz, Salzburg** - major cities with ports, railways, and gardens where species are introduced
- **Native species** concentrate in **alpine regions** - natural habitats with high endemic diversity
- The **patterns are almost opposite** - this is what accumulation curves can reveal!

The Spatial Algorithm

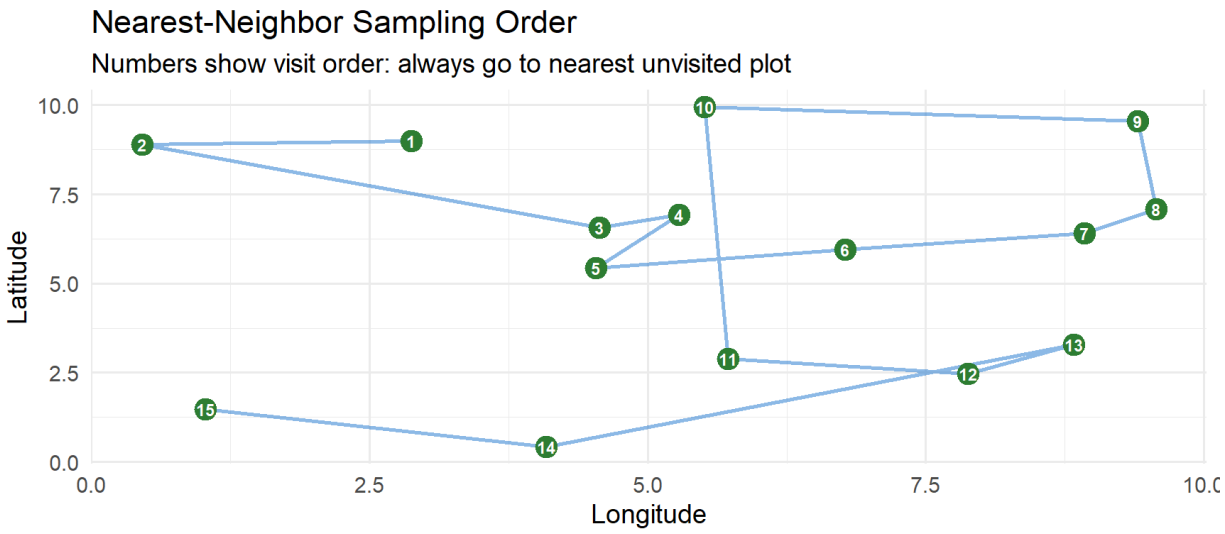
Why Does Order Matter?

Yesterday we built curves by sampling plots in **random order**.

But in real ecology, we often care about **spatial patterns**:

- Do species cluster geographically?
- How does diversity change as we move across a landscape?

Solution: Sample plots in **spatial order** - always go to the nearest unvisited plot!



The Nearest-Neighbor Algorithm

Step by step:

1. Pick a **starting plot** (seed)
2. Mark it as **visited**
3. Find the **nearest unvisited** plot
4. Move there, record new species, mark as visited
5. **Repeat** until all plots are visited

The Distance Formula

Distance between two points (x1, y1) and (x2, y2):

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

This is the **Euclidean distance** - the straight-line distance between two points.

Multiple Starting Points

Why Multiple Seeds?

The shape of the curve depends on **where we start**!

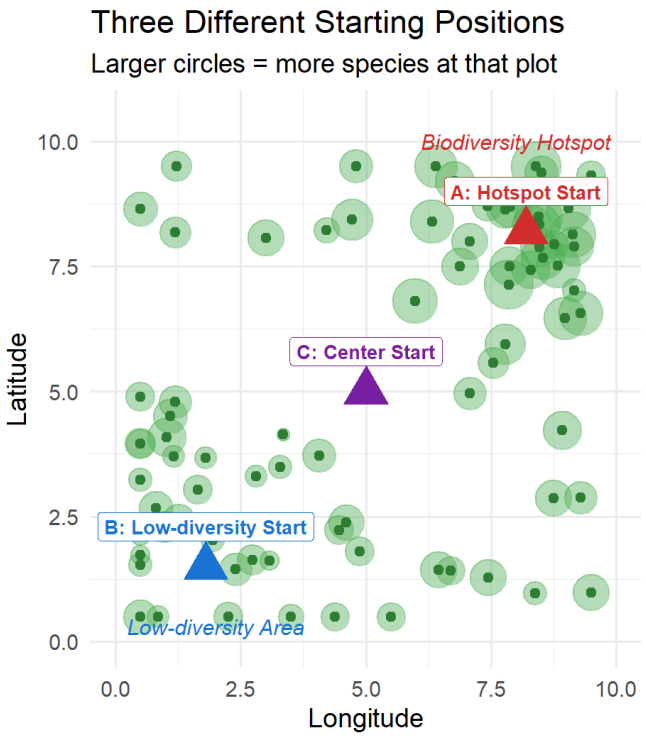
- Starting in a species-rich area: fast initial accumulation
- Starting in a species-poor area: slow initial accumulation

Solution: Run the algorithm multiple times with different starting points, then average.

Visualizing Starting Positions

Let's imagine we have plots spread across a landscape. Some areas are **biodiversity hotspots** (many species), others are **low-diversity areas** (few species).

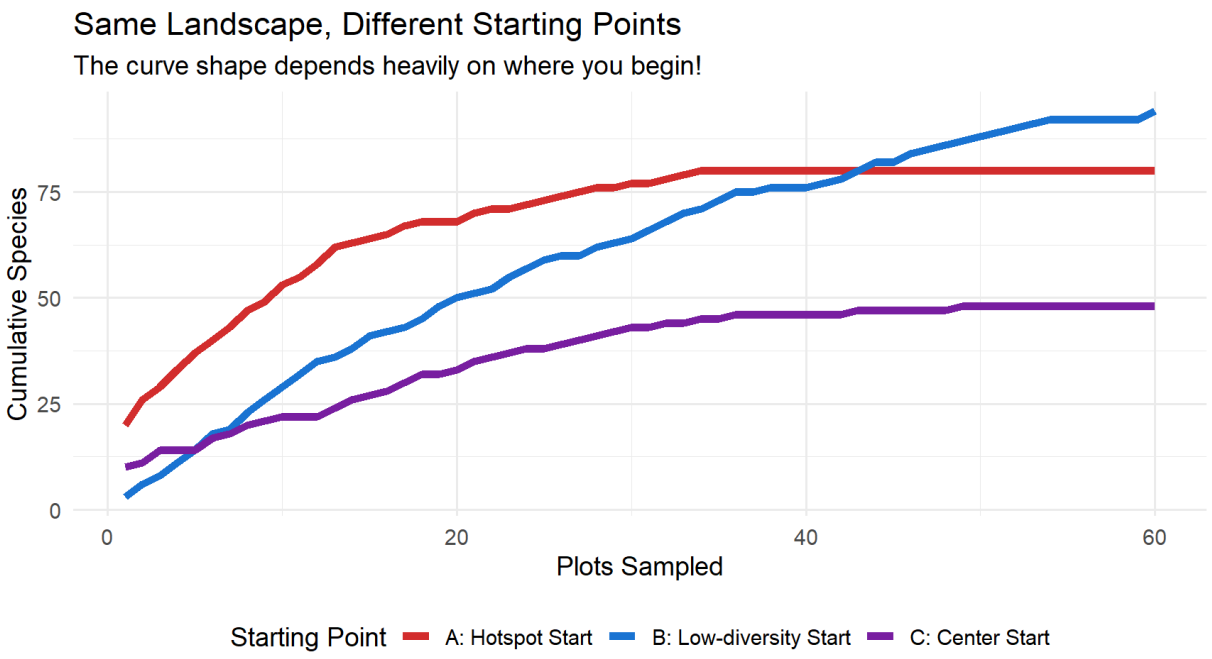
Where we choose to start our sampling walk dramatically affects the curve shape!



What happens with each start?

A: Hotspot	Many species quickly	Steep rise, then flattens
B: Low-diversity	Few species slowly	Gradual rise throughout
C: Center	Medium start	Balanced curve

The Effect of Starting Position



Key observation: All three curves eventually reach approximately the same number of species (they converge), but the **path to get there is very different!**

This is why we must consider multiple starting points in our analysis.

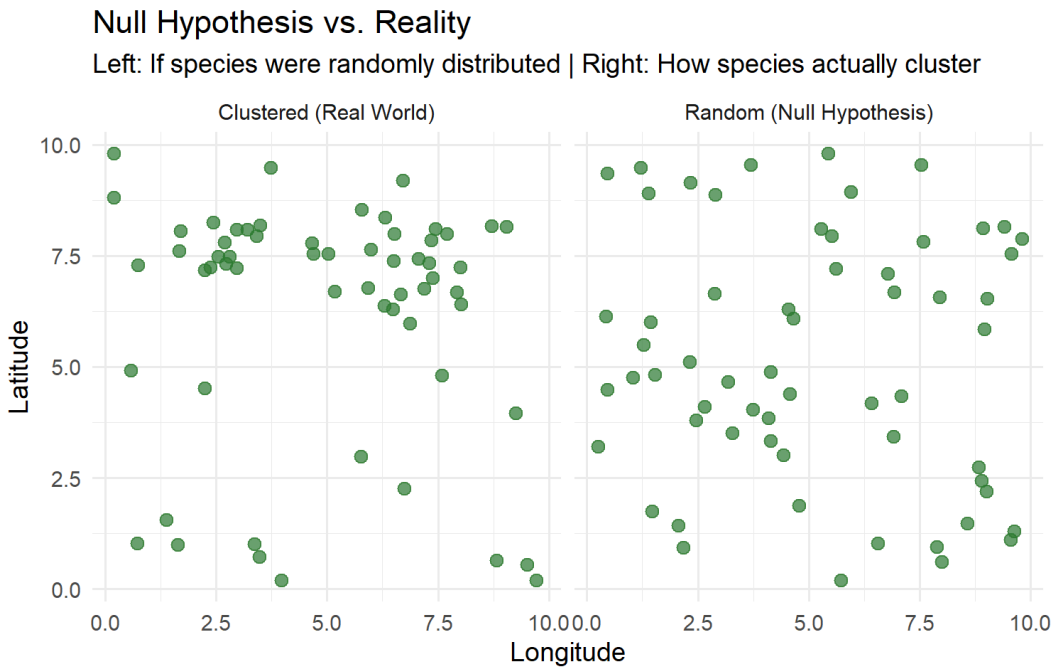
The Null Hypothesis

What If Starting Position Doesn't Matter?

The Null Hypothesis (H_0):

> "Starting position does not affect the accumulation curve - all curves from different starting points should look the same."

If true: Species are randomly distributed across the landscape. No hotspots, no clusters. Every location is equally likely to have any species.



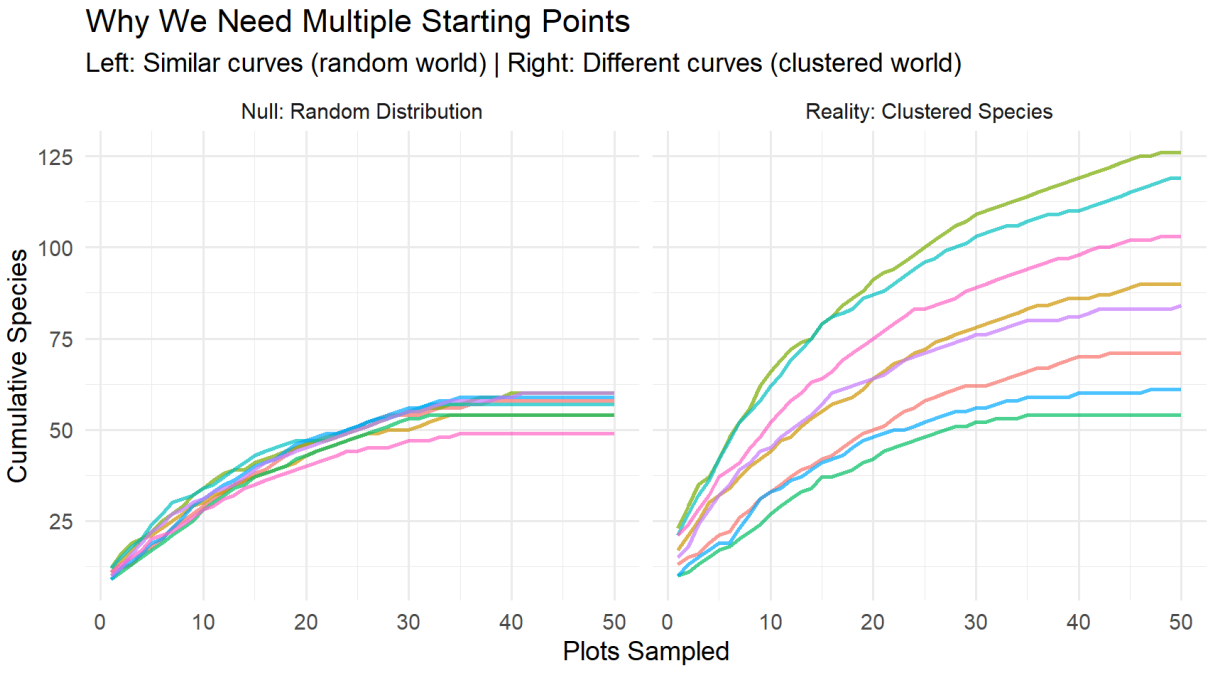
Testing the Null Hypothesis

How we test this:

1. Run the sampling algorithm from **many different starting points** (e.g., 20 seeds)
2. Compare the curves - do they overlap or diverge?
3. Calculate the **variation between curves**

If H_0 is true: All curves should be nearly identical (low variation)

If H_0 is false: Curves should differ based on starting location (high variation)



Key Insight:

In a world where species are **randomly distributed** (H_0), all curves look similar regardless of starting point.

In the **real world** where species cluster around favorable habitats, curves differ dramatically based on where you start.

This is why we average across multiple seeds! It gives us a more representative picture of the overall biodiversity pattern.

What Does High Variation Tell Us?

When curves from different starting points **vary a lot**, it tells us:

1. **Species are clustered** - not evenly spread across the landscape
2. **Hotspots exist** - some areas are biodiversity-rich
3. **Geography matters** - position affects what you find

This is actually **more interesting** than finding uniform curves! It means there are spatial patterns worth investigating.

Comparing Native vs. Alien

Our Research Hypothesis

Research Question: Do alien (introduced) species accumulate differently than native species across the Austrian landscape?

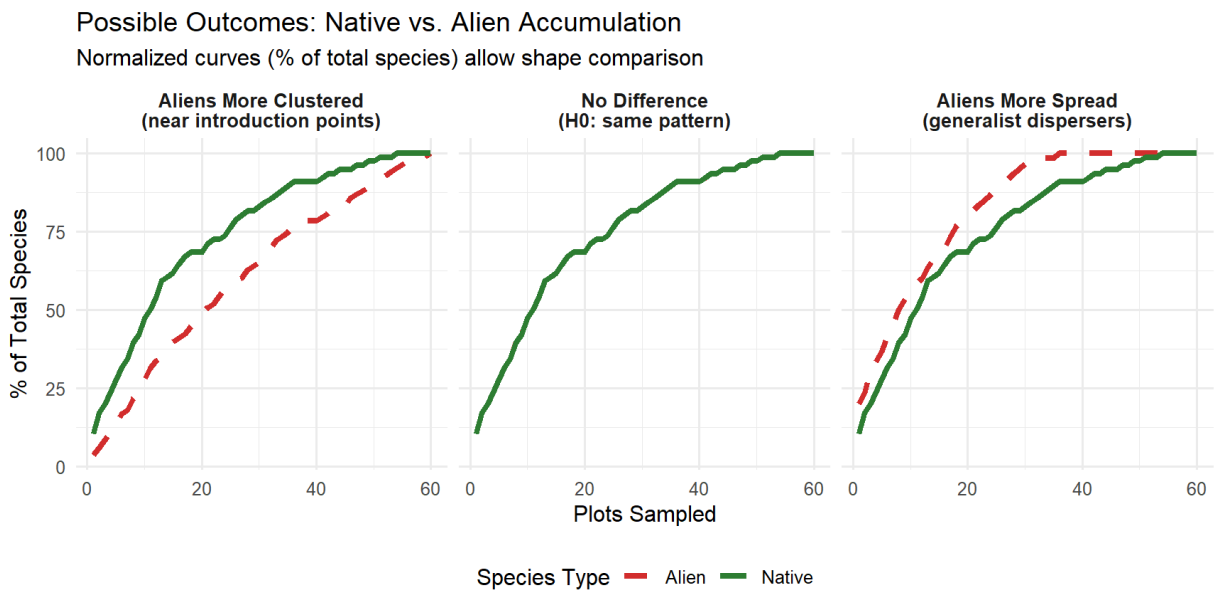
Null Hypothesis (H_0): > Native and alien species accumulate at the same rate - their curves have the same shape. There is no difference in spatial distribution patterns.

Alternative Hypothesis (H_1): > Native and alien species have different accumulation patterns, reflecting differences in how they are distributed across the landscape.

What might we expect?

Aliens accumulate faster	Steeper initial curve	Aliens are generalists, found everywhere
Aliens accumulate slower	Flatter initial curve	Aliens are clustered near introduction points (cities, ports)
No difference	Identical curves	Distribution patterns are similar

The key insight: By comparing normalized curves, we can see if one group is more clustered or more evenly spread than the other.



Normalization for Fair Comparison

The raw curves are hard to compare because there are many more native species.

Solution: Normalize to percentage of total species found.

This lets us compare the **shape** of accumulation, not just the total numbers.

Interpretation & Discussion

What Do the Curves Tell Us?

Observations from the comparison:

1. **Total richness:** Native species >> Alien species (as expected)
2. **Normalized curve position:**
    - Curve above = proportionally faster accumulation (lower spatial turnover)
    - Curve below = proportionally slower accumulation (higher spatial turnover)
3. **Implications for invasions:**
    - Higher alien turnover could mean different species arrive at different places
    - Lower native turnover could mean widespread generalist species dominate

Discussion Questions

1. **Why might alien species be more evenly distributed?**

    - Human-mediated dispersal (trade, transport)
    - Generalist habitat preferences
    - Recent introduction = no time to cluster
2. **What are the limitations of our analysis?**
    - Geographic bias in sampling (some regions better covered)
    - Detection probability differs by species
    - Single time snapshot (no temporal dynamics)
3. **What could be explored further?**
    - Compare different habitat types (forests vs. grasslands)
    - Look at temporal changes (are aliens spreading?)
    - Test in other countries

Biological Mechanisms

Why do accumulation patterns differ?

Dispersal ability	Good dispersers = even distribution = gradual curve
Habitat specificity	Specialists = clustered = steep then flat
Historical range	Long history = natural clustering patterns
Human influence	Transport along roads, rivers = linear patterns

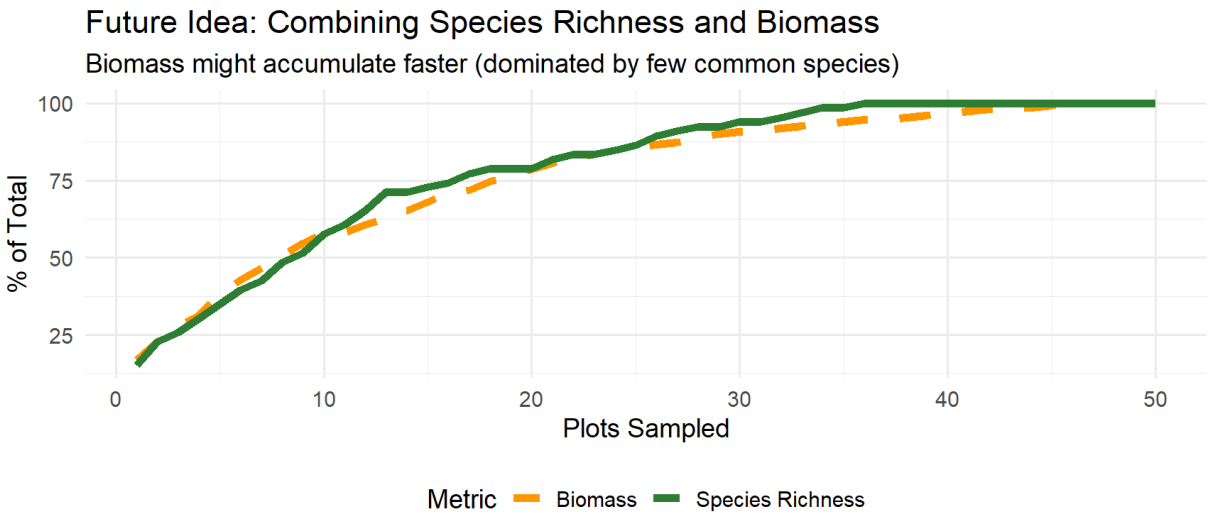
Future Directions

Beyond Species Richness: Biomass Accumulation

So far we've only counted **species** (species richness). But what about **how much** of each species is present?

Biomass accumulation curves could tell us even more:

Species richness	Number of different species	Diversity patterns
Biomass	Total weight/cover of plants	Dominance patterns
Combined	Both together	Full ecological picture



Why biomass matters:

- A few **dominant species** might contribute most of the biomass
- Alien species could have **high biomass but low diversity** (invasive monocultures)
- Native ecosystems might show **balanced biomass across many species**

Research question: Do alien species contribute disproportionately to biomass relative to their species count?

Summary

Key Concepts

Spatial order	The sequence we visit plots affects the curve shape
Nearest-neighbor	Algorithm: always go to closest unvisited plot
Multiple seeds	Different starting points = different curves = need to average
Native vs. Alien	Different accumulation patterns reveal ecological differences

Take-Home Message

Accumulation curves reveal how species are distributed across space. Comparing native and alien species shows us which group is more clustered and helps identify invasion hotspots.

Day 2

Exercises

A Mini Research Project

Day 2: Exercises

Advanced Analysis & Research Project

> Fill in the blanks!

> 2026-02-16

Part 2: Advanced Analysis - Welcome to Day 2!

Yesterday you built your first species accumulation curve. Today we'll take it to the next level with real scientific analysis!

What you'll learn today:

1. **Spatial sampling** - Walk through plots like a real field ecologist
2. **Measuring uncertainty** - Run from multiple starting points
3. **Comparing natives vs aliens** - Do they accumulate differently?
4. **Polished figures** - Make nice-looking plots

How to use these exercises:

- Replace each `_____` with the correct code
- Run each chunk to see if it works
- Read the hints if you get stuck
- The exercises build on each other - do them in order!

Exercise 1: Setup - Load Data and Functions

Concept: Before we start, we need to load our data and define some helper functions. These functions will make our analysis easier.

1a: Load packages and data

```
# Load packages
# Pakete laden
library(tidyverse)

# Load data
# Daten laden
header <- read_csv("../data/austria_header.csv")
species <- read_csv("../data/austria_species.csv")

# Check data loaded correctly
# Überprüfe dass Daten korrekt geladen sind
print(paste("Plots loaded:", nrow(header)))
print(paste("Species records:", nrow(species)))
print(paste("Unique species:", length(unique(species$WFO_TAXON))))
```

Before we define our helper functions, let's learn some **new R concepts** that we'll use today. Yesterday you learned variables, vectors, and `unique()`. Today we add:

Writing your own functions - Fibonacci example:

The Fibonacci sequence: 1, 1, 2, 3, 5, 8, 13... (each number = sum of previous two)

```
# VERSION 1: Using a vector (stores ALL numbers)
fibonacci_vector <- function(n) {
  fib <- c(1, 1)                      # Start with first two numbers
  for (i in 3:n) {
    fib[i] <- fib[i-1] + fib[i-2]     # Add: previous + one before that
  }
  return(fib)
}

fibonacci_vector(7)  # Returns: 1 1 2 3 5 8 13

# VERSION 2: Using 2 variables (memory efficient - only tracks last 2)
fibonacci_two_vars <- function(n) {
  prev <- 1                           # The number before current
  curr <- 1                           # Current number
  for (i in 3:n) {
    new <- prev + curr                # Calculate next
    prev <- curr                      # Shift: current becomes previous
    curr <- new                       # Shift: new becomes current
  }
  return(curr)
}

fibonacci_two_vars(7)  # Returns: 13 (the 7th Fibonacci number)
```

Default values in functions:

```
# If user doesn't provide 'n', use 10 as default
fibonacci <- function(n = 10) {
  fib <- c(1, 1)
  for (i in 3:n) fib[i] <- fib[i-1] + fib[i-2]
  return(fib)
}

fibonacci()      # Returns first 10 numbers (uses default)
fibonacci(5)     # Returns first 5 numbers
```

Useful functions we'll use:

Function	What it does	Example
`paste()`	Combine text and values	`paste("Found:", 5, "species")`
`sample()`	Pick random items	`sample(1:100, 5)` picks 5 random numbers
`set.seed()`	Make random reproducible	`set.seed(42)` always gives same "random"
`rep()`	Repeat values	`rep(FALSE, 5)` gives `FALSE FALSE FALSE FALSE FALSE`
`which.min()`	Position of smallest value	`which.min(c(5,2,8))` returns `2`
`setdiff()`	What's in A but not B	`setdiff(c(1,2,3), c(2,3,4))` returns `1`

Function	What it does	Example
`pull()`	Extract column as vector	`data %>% pull(column)`
`is.null()`	Check if NULL	`is.null(NULL)` returns `TRUE`

Special values:

- `NULL` = "nothing" / empty / not specified
- `Inf` = infinity (bigger than any number)

1c: Define helper functions (just copy & run!)

Just run this code block! You don't need to understand every line - these are tools we'll use later. Click the details below only if you're curious.

🔨 Building the Functions Step-by-Step

Click each step to see how we build up from simple to complete:

▶ 1 **calc_distance: Start with the math formula**

▶ 2 **nn_walk: First, think about what we need to track**

▶ 3 **nn_walk: Build the main loop**

▶ 4 **nn_walk: The "Infinity Trick" for excluding visited plots**

▶ 5 **nn_walk: Add flexibility with optional start point**

▶ 6 **build_accumulation: Think about what "accumulation" means**

▶ 7 **build_accumulation: Use setdiff() to find new species**

▶ 8 **build_accumulation: Add optional filtering**

▶ 9 **find_saturation: A simple but useful helper**

👆 *Click each step above to expand! Now here are the complete functions:*

```r
# ==== FUNCTION 1: calc_distance ====
# Calculates Euclidean distance. VECTORIZED: x2,y2 can be vectors!
# Berechnet Euklidische Distanz. VEKTORISIERT: x2,y2 können Vektoren sein!

calc_distance <- function(x1, y1, x2, y2) {
  # Pythagorean theorem: sqrt(dx² + dy²). Works with vectors!
  # Satz des Pythagoras: sqrt(dx² + dy²). Funktioniert mit Vektoren!
  sqrt((x2 - x1)^2 + (y2 - y1)^2)
}


# ==== FUNCTION 2: nn_walk ====
# Nearest-neighbour walk: always go to closest unvisited plot
# Nearest-Neighbour-Walk: gehe immer zum nächsten unbesuchten Plot

nn_walk <- function(header_data, start_idx = NULL) {
  # How many plots total? / Wie viele Plots insgesamt?
  n <- nrow(header_data)

  # If no start given, pick random / Wenn kein Start, wähle zufällig
  if (is.null(start_idx)) start_idx <- sample(1:n, 1)

  # Track which plots we've visited / Verfolge welche Plots besucht
  visited <- rep(FALSE, n)
  # Store the order we visit them / Speichere Reihenfolge
  visit_order <- numeric(n)
  # Start at this plot / Starte bei diesem Plot
  current <- start_idx

  # Loop through all plots / Schleife durch alle Plots
  for (i in 1:n) {
    # Mark current as visited / Markiere aktuellen als besucht
    visited[current] <- TRUE
    # Save the plot ID / Speichere Plot-ID
    visit_order[i] <- header_data$PlotObservationID[current]

    # If not done yet / Falls noch nicht fertig
    if (i < n) {
      # Calculate distance from current to ALL others (vectorized!)
      # Berechne Distanz von aktuellem zu ALLEN anderen (vektorisiert!)
      distances <- calc_distance(
        header_data$Longitude[current],
        header_data$Latitude[current],
        header_data$Longitude,
        header_data$Latitude
      )
      # Inf trick: visited plots can never be "minimum"
      # Inf-Trick: besuchte Plots können nie "Minimum" sein
      distances[visited] <- Inf
      # Go to closest unvisited / Gehe zum nächsten unbesuchten
      current <- which.min(distances)
    }
  }
  # Return the order of plot IDs / Gib Reihenfolge der Plot-IDs zurück
  return(visit_order)
}


# ==== FUNCTION 3: build_accumulation ====
# Count cumulative species as we visit each plot in order
# Zähle kumulative Arten während wir jeden Plot der Reihe nach besuchen
```

```r
build_accumulation <- function(species_data, plot_order, status_filter = NULL) {
  # If filter provided (e.g., "nat"), keep only matching species
  # Falls Filter angegeben, behalte nur passende Arten
  if (!is.null(status_filter)) {
    species_data <- species_data %>% filter(STATUS == status_filter)
  }

  # Empty vector to collect all species found / Leerer Vektor für gefundene Arten
  found <- c()
  # Pre-allocate result vector / Ergebnisvektor vorbelegen
  accum <- numeric(length(plot_order))

  # For each plot in our walking order / Für jeden Plot in unserer Laufreihenfolge
  for (i in seq_along(plot_order)) {
    # Get unique species in this plot / Hole einzigartige Arten in diesem Plot
    plot_spp <- species_data %>%
      # Filter to current plot / Filtere auf aktuellen Plot
      filter(PlotObservationID == plot_order[i]) %>%
      # Extract species names / Extrahiere Artnamen
      pull(WFO_TAXON) %>%
      # Remove duplicates within plot / Entferne Duplikate im Plot
      unique()

    # setdiff: what's NEW? (in plot but not yet found)
    # setdiff: was ist NEU? (im Plot aber noch nicht gefunden)
    new_spp <- setdiff(plot_spp, found)
    # Add new species to our collection / Füge neue Arten zur Sammlung
    found <- c(found, new_spp)
    # Count total species so far / Zähle Gesamtarten bisher
    accum[i] <- length(found)
  }
  # Return the accumulation curve / Gib Akkumulationskurve zurück
  return(accum)
}


# ==== FUNCTION 4: find_saturation ====
# Find when we reach X% of total species (default 80%)
# Finde wann wir X% der Gesamtarten erreichen (Standard 80%)

find_saturation <- function(curve, threshold = 0.8) {
  # Calculate target: 80% of final count / Berechne Ziel: 80% der Endzahl
  target <- max(curve) * threshold
  # which()[1] = first position where TRUE / Erste Position wo TRUE
  which(curve >= target)[1]
}

print("All functions loaded!")
```

💡 Hint: Run both code blocks above. If you get errors, check that the data files exist in `../data/`.

⚡ Turbo Mode: Show/Hide Rcpp Functions

1d: How fast is my code? (Complexity)

Why does speed matter? With 100 plots, slow code takes seconds. With 10,000 plots, it could take hours!

What counts as "work"? To keep things simple, we count **multiplications** (and divisions). These are the "expensive" operations - additions and assignments are so fast we ignore them.

The problem - Nested loops: If you put a loop INSIDE a loop, multiplications multiply!

N (input size)	Multiplications (N×N)	Time
10	100	instant
100	10,000	instant
1,000	1,000,000	slow
10,000	100,000,000	very slow!

```
# O(N²) - gets slow fast!
for (i in 1:N) {
  for (j in 1:N) {
    # This runs N × N = N² times!
  }
}
# N=100 → 10,000 multiplications
# N=1000 → 1,000,000 multiplications (1000× slower!)
```

Our `nn_walk` function has this $O(N^2)$ behavior (it checks distance to all plots at each step). This is exactly when it makes sense to use a **compiled language** (like C++) instead of an **interpreted language** (like R). Compiled languages run the same operations 10-100× faster - that's why we offer the Rcpp version!

Exercise 2: Understanding Spatial Sampling

Concept: When we sample plots randomly, we might jump all over the map. But a **nearest-neighbour walk** samples like a real ecologist would - always going to the closest unvisited plot. This creates a realistic spatial accumulation curve.

New concepts used here:

Concept	What it does	Example
`%in%`	Check if values are in a list	`filter(ID %in% my_ids)`
`scale_color_manual()`	Set custom colors for categories	`values = c("A" = "red", "B" = "blue")`

2a: Create a sample dataset

```
# Take a random sample of 150 plots (for speed)
# Nimm eine zufällige Stichprobe von 150 Plots (für Geschwindigkeit)
set.seed(42)  # For reproducibility / Für Reproduzierbarkeit
sample_size <- 150

sample_ids <- sample(unique(header$PlotObservationID), sample_size)
sample_header <- header %>% filter(PlotObservationID %in% sample_ids)
sample_species <- species %>% filter(PlotObservationID %in% sample_ids)

print(paste("Sample created:", nrow(sample_header), "plots"))
```

2b: Run a nearest-neighbour walk

```
# Start from plot 1 and walk to nearest neighbours
# Starte von Plot 1 und gehe zu nächsten Nachbarn
nn_order <- nn_walk(sample_header, start_idx = 1)

# Build curves for native and alien species
# Baue Kurven für heimische und Alien-Arten
native_curve <- build_accumulation(sample_species, nn_order, "____")
alien_curve <- build_accumulation(sample_species, nn_order, "____")

# Check the results
# Überprüfe die Ergebnisse
print(paste("Native species found:", max(native_curve)))
print(paste("Alien species found:", max(alien_curve)))
```

2c: Plot the comparison

```
# Create data frame for plotting
# Erstelle Data Frame zum Plotten
curve_data <- data.frame(
  plots = rep(1:sample_size, 2),
  species = c(native_curve, alien_curve),
  status = rep(c("Native", "Alien"), each = sample_size)
)

ggplot(curve_data, aes(x = plots, y = species, color = status)) +
  geom_line(linewidth = 1.2) +
  scale_color_manual(values = c("Native" = "darkgreen", "Alien" = "____")) +
  labs(
    title = "Native vs Alien Species Accumulation",
    x = "Plots Sampled (Nearest-Neighbour Order)",
    y = "Cumulative Species",
    color = ""
  ) +
  theme_minimal()
```

💡
Hint: Native species use STATUS = “nat”, alien species use STATUS = “neo”.

Exercise 3: Measuring Uncertainty

Concept: The curve we get depends on WHERE we start! If we start in a native-rich area, natives accumulate fast. If we start near an alien hotspot, aliens accumulate fast. We need to run from MANY starting points to see the true pattern.

New concepts in this exercise:

Concept	What it does	Example
`matrix(NA, nrow, ncol)`	Create empty table with rows & columns	`matrix(NA, nrow=20, ncol=100)`
`colMeans(mat)`	Mean of each column	`colMeans(mat)` returns one mean per col
`sd()`	Standard deviation (spread of values)	`sd(c(1,2,3,4,5))` returns 1.58
`quantile(x, p)`	Value at percentile p	`quantile(x, 0.975)` = 97.5th percentile
`geom_ribbon()`	Shaded band in ggplot	`geom_ribbon(aes(ymin, ymax))`
`alpha = 0.3`	Transparency (0=invisible, 1=solid)	Makes bands see-through

Why matrix() instead of data.frame()? - `matrix` = ALL values are the same type (only numbers). Faster for math! - `data.frame` = columns can be different types (text, numbers, TRUE/FALSE)

We use matrix here because we only store numbers and need to do fast calculations.

3a: Run from multiple starting points

```
# Run from 20 different starting points
# Starte von 20 verschiedenen Startpunkten
n_seeds <- 20

# Storage for results (matrix: rows = seeds, columns = plots)
# Speicher für Ergebnisse (Matrix: Zeilen = Seeds, Spalten = Plots)
native_runs <- matrix(NA, nrow = n_seeds, ncol = sample_size)
alien_runs <- matrix(NA, nrow = n_seeds, ncol = sample_size)

# Run from each starting point
# Starte von jedem Startpunkt
print(paste("Running", n_seeds, "starting points..."))
for (seed in 1:n_seeds) {
  nn_order <- nn_walk(sample_header, start_idx = seed)
  native_runs[seed, ] <- build_accumulation(sample_species, nn_order, "____")
  alien_runs[seed, ] <- build_accumulation(sample_species, nn_order, "____")

  if (seed %% 5 == 0) print(paste("  Completed", seed))
}
print("Done!")
```

3b: Calculate mean and confidence intervals

```
# Mean across all runs (colMeans = mean of each column)
# Mittelwert über alle Läufe (colMeans = Mittelwert jeder Spalte)
native_mean <- colMeans(____)
alien_mean <- colMeans(alien_runs)

# 95% confidence intervals using a loop
# 95% Konfidenzintervalle mit einer Schleife
native_lower <- numeric(sample_size)
native_upper <- numeric(sample_size)
alien_lower <- numeric(sample_size)
alien_upper <- numeric(sample_size)

for (i in 1:sample_size) {
  native_lower[i] <- quantile(native_runs[, i], 0.025)
  native_upper[i] <- quantile(native_runs[, i], ____)
  alien_lower[i] <- quantile(alien_runs[, i], ____)
  alien_upper[i] <- quantile(alien_runs[, i], 0.975)
}

# How much variation is there?
# Wie viel Variation gibt es?
midpoint <- sample_size %/% 2
native_cv <- round(100 * sd(native_runs[, midpoint]) / mean(native_runs[, midpoint]), 1)
alien_cv <- round(100 * sd(alien_runs[, midpoint]) / mean(alien_runs[, midpoint]), 1)

print("Variation at midpoint:")
print(paste("  Native CV:", native_cv, "%"))
print(paste("  Alien CV:", alien_cv, "%"))
```

3c: Plot with uncertainty bands

```
# Create summary data frame
# Erstelle Zusammenfassungs-Data-Frame
summary_data <- data.frame(
  plots = rep(1:sample_size, 2),
  mean = c(native_mean, alien_mean),
  lower = c(native_lower, alien_lower),
  upper = c(native_upper, alien_upper),
  status = rep(c("Native", "Alien"), each = sample_size)
)

# Plot with shaded confidence bands
# Plotte mit schattierten Konfidenzbändern
ggplot(summary_data, aes(x = plots)) +
  geom_ribbon(aes(ymin = lower, ymax = upper, fill = status), alpha = ____) +
  geom_line(aes(y = mean, color = status), linewidth = 1.2) +
  scale_color_manual(values = c("Native" = "darkgreen", "Alien" = "red")) +
  scale_fill_manual(values = c("Native" = "darkgreen", "Alien" = "red")) +
  labs(
    title = "Species Accumulation with Uncertainty",
    subtitle = paste(n_seeds, "starting points - shaded = 95% CI"),
    x = "Plots Sampled",
    y = "Cumulative Species",
    color = "", fill = ""
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```

💡 Hint: Use `mean` for the average. For 95% CI, upper bound is 0.975 quantile, lower is 0.025.

Exercise 4: Comparing Saturation Speed

Concept: Saturation is when we've found most of the species and the curve flattens. If natives saturate FASTER than aliens, it means we encounter most native species quickly as we move across the landscape. If aliens saturate SLOWER, different alien species occur in different regions, indicating high spatial turnover.

New concept - Boxplots:

`geom_boxplot()` shows the distribution of values: - **Box** = middle 50% of data (25th to 75th percentile) - **Line inside** = median (middle value) - **Whiskers** = extend to ~1.5× the box height - **Dots** = outliers (extreme values)

4a: Find saturation points

```
# At what point do we reach 80% of species?
# Bei welchem Punkt erreichen wir 80% der Arten?
native_sat <- numeric(n_seeds)
alien_sat <- numeric(n_seeds)

for (i in 1:n_seeds) {
  native_sat[i] <- find_saturation(native_runs[i, ], threshold = 0.8)
  alien_sat[i] <- find_saturation(alien_runs[i, ], threshold = ____)
}

# Convert to percentage of total plots
# In Prozent der Gesamtplots umwandeln
native_sat_pct <- 100 * native_sat / sample_size
alien_sat_pct <- 100 * alien_sat / sample_size

print("Plots needed to reach 80% of species:")
print(paste("  Native:", round(mean(native_sat_pct)), "% (SD:", round(sd(native_sat_pct), 1),
")"))
print(paste("  Alien:", round(mean(alien_sat_pct)), "% (SD:", round(sd(alien_sat_pct), 1), ")"))
```

4b: Visualize with boxplot

```
# Create data for boxplot
# Erstelle Daten für Boxplot
sat_data <- data.frame(
  saturation_pct = c(native_sat_pct, alien_sat_pct),
  status = rep(c("Native", "Alien"), each = n_seeds)
)

ggplot(sat_data, aes(x = status, y = saturation_pct, fill = status)) +
  geom_boxplot() +
  scale_fill_manual(values = c("Native" = "____", "Alien" = "____")) +
  labs(
    title = "How Fast Do Species Saturate?",
    subtitle = "% of plots needed to find 80% of species",
    y = "% of Plots Needed",
    x = ""
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

4c: Calculate slope ratio

```
# Compare early vs late slopes
# Vergleiche frühe vs späte Steigungen
# High ratio = fast saturation (steep early, flat late)
# Hohes Verhältnis = schnelle Sättigung (steil früh, flach spät)

calc_slope_ratio <- function(curve) {
  n <- length(curve)
  early_end <- round(n * 0.2)
  late_start <- round(n * 0.8)

  early_slope <- (curve[early_end] - curve[1]) / early_end
  late_slope <- (curve[n] - curve[late_start]) / (n - late_start)

  if (late_slope > 0) return(early_slope / late_slope)
  else return(NA)
}

native_ratios <- numeric(n_seeds)
alien_ratios <- numeric(n_seeds)

for (i in 1:n_seeds) {
  native_ratios[i] <- calc_slope_ratio(native_runs[i, ])
  alien_ratios[i] <- calc_slope_ratio(alien_runs[i, ])
}

print("Slope ratio (higher = faster saturation):")
print(paste("  Native:", round(mean(native_ratios, na.rm = TRUE), 1)))
print(paste("  Alien:", round(mean(alien_ratios, na.rm = TRUE), 1)))
```

💡 Hint: Fill in colors “darkgreen” and “red”.

Exercise 5: Geographic Patterns

Concept: Does it matter WHERE in Austria we start? Some areas might have more aliens (near cities, roads). Let's map how starting location affects what we find!

New ggplot concepts for mapping:

Concept	What it does	Exar
`coord_quickmap()`	Correct map proportions (so Austria doesn't look stretched!)	Add at end of ggplot
`scale_color_gradient2()`	Color scale with 3 colors (low→mid→high)	`low="green", mid="y`
`midpoint = value`	Where the middle color appears	`midpoint = 0.5` for

5a: Map alien proportion by starting location

```
# Test from 30 different starting locations
# Teste von 30 verschiedenen Startorten
n_map_seeds <- 30
checkpoint <- 75  # Check after this many plots

seed_results <- data.frame(
  seed_idx = 1:n_map_seeds,
  seed_lon = numeric(n_map_seeds),
  seed_lat = numeric(n_map_seeds),
  native_count = numeric(n_map_seeds),
  alien_count = numeric(n_map_seeds)
)

for (i in 1:n_map_seeds) {
  # Record starting location
  # Erfasse Startort
  seed_results$seed_lon[i] <- sample_header$Longitude[i]
  seed_results$seed_lat[i] <- sample_header$Latitude[i]

  # Run from this starting point
  # Starte von diesem Startpunkt
  nn_order <- nn_walk(sample_header, start_idx = i)

  # Get counts at checkpoint
  # Hole Zählungen am Checkpoint
  native_curve <- build_accumulation(sample_species, nn_order, "nat")
  alien_curve <- build_accumulation(sample_species, nn_order, "neo")

  seed_results$native_count[i] <- native_curve[min(checkpoint, length(native_curve))]
  seed_results$alien_count[i] <- alien_curve[min(checkpoint, length(alien_curve))]
}

# Calculate alien proportion
# Berechne Alien-Anteil
seed_results$alien_prop <- seed_results$alien_count /
  (seed_results$native_count + seed_results$____)

# Map it!
# Kartiere es!
ggplot(seed_results, aes(x = seed_lon, y = seed_lat, color = alien_prop)) +
  geom_point(size = 4) +
  scale_color_gradient2(
    low = "darkgreen", mid = "yellow", high = "red",
    midpoint = mean(seed_results$alien_prop),
    name = "Alien\nproportion"
  ) +
  coord_quickmap() +
  labs(
    title = "How Starting Location Affects Alien Detection",
    subtitle = paste("Alien proportion after", checkpoint, "plots"),
    x = "Longitude", y = "Latitude"
  ) +
  theme_minimal()
```

💡

Hint: Fill in `alien_count` to complete the proportion calculation.

Exercise 6: Putting It All Together

Concept: Let's put it all together and create a polished figure that shows our main finding: how native vs alien species accumulate differently across Austrian plots.

6a: Run full analysis

```
# Larger sample for final analysis
# Größere Stichprobe für finale Analyse
n_seeds <- 30
sample_size <- 200

set.seed(2024)
sample_ids <- sample(unique(header$PlotObservationID), sample_size)
sample_header <- header %>% filter(PlotObservationID %in% sample_ids)
sample_species <- species %>% filter(PlotObservationID %in% sample_ids)

# Run from all seeds
# Starte von allen Seeds
native_runs <- matrix(NA, nrow = n_seeds, ncol = sample_size)
alien_runs <- matrix(NA, nrow = n_seeds, ncol = sample_size)

print("Running full analysis...")
for (seed in 1:n_seeds) {
  nn_order <- nn_walk(sample_header, start_idx = seed)
  native_runs[seed, ] <- build_accumulation(sample_species, nn_order, "nat")
  alien_runs[seed, ] <- build_accumulation(sample_species, nn_order, "neo")
  if (seed %% 10 == 0) print(paste("  Completed", seed, "of", n_seeds))
}
print("Done!")

# Calculate summaries using colMeans and loops
# Berechne Zusammenfassungen mit colMeans und Schleifen
native_mean <- colMeans(native_runs)
alien_mean <- colMeans(alien_runs)

native_lower <- numeric(sample_size)
native_upper <- numeric(sample_size)
alien_lower <- numeric(sample_size)
alien_upper <- numeric(sample_size)

for (i in 1:sample_size) {
  native_lower[i] <- quantile(native_runs[, i], 0.025)
  native_upper[i] <- quantile(native_runs[, i], 0.975)
  alien_lower[i] <- quantile(alien_runs[, i], 0.025)
  alien_upper[i] <- quantile(alien_runs[, i], 0.975)
}

results <- data.frame(
  plots = rep(1:sample_size, 2),
  mean = c(native_mean, alien_mean),
  lower = c(native_lower, alien_lower),
  upper = c(native_upper, alien_upper),
  status = rep(c("Native", "Alien"), each = sample_size)
)
```

6b: Create polished figure

```
# Polished figure
# Schöne Abbildung
final_plot <- ggplot(results, aes(x = plots)) +
  geom_ribbon(aes(ymin = lower, ymax = upper, fill = status), alpha = 0.25) +
  geom_line(aes(y = mean, color = status), linewidth = 1.3) +
  scale_color_manual(values = c("Native" = "#228B22", "Alien" = "#DC143C")) +
  scale_fill_manual(values = c("Native" = "#228B22", "Alien" = "#DC143C")) +
  labs(
    title = "Species Accumulation: Native vs Alien Plants in Austria",
    subtitle = paste0(n_seeds, " starting points, 95% confidence intervals"),
    x = "Number of Plots Sampled",
    y = "Cumulative Species Count",
    color = "", fill = ""
  ) +
  theme_minimal(base_size = 13) +
  theme(
    legend.position = c(0.85, 0.2),
    legend.background = element_rect(fill = "white", color = "gray80"),
    plot.title = element_text(face = "bold")
  )

print(final_plot)

# Save it (uncomment to run)
# ggsave("austria_accumulation.png", final_plot, width = 10, height = 7, dpi = 300)
```

Exercise 7: Interpret Your Results

Concept: Science isn't just about making plots - it's about understanding what they mean! Let's summarize our findings.

7a: Generate results summary

```
print("========== RESULTS SUMMARY ==========")

# Total species found
print("Species found:")
print(paste("  Native:", round(mean(native_runs[, sample_size])), "species"))
print(paste("  Alien:", round(mean(alien_runs[, sample_size])), "species"))

# Saturation comparison (using a loop)
native_sat <- numeric(n_seeds)
alien_sat <- numeric(n_seeds)
for (i in 1:n_seeds) {
  native_sat[i] <- find_saturation(native_runs[i, ], threshold = 0.8)
  alien_sat[i] <- find_saturation(alien_runs[i, ], threshold = 0.8)
}

print("Plots to reach 80% of species:")
print(paste("  Native:", round(mean(native_sat)), "plots (",
    round(100*mean(native_sat)/sample_size), "%)"))
print(paste("  Alien:", round(mean(alien_sat)), "plots (",
    round(100*mean(alien_sat)/sample_size), "%)"))

# Draw conclusion
print("========== CONCLUSION ==========")
if (mean(native_sat) < mean(alien_sat)) {
  print("Native species saturate FASTER than aliens!")
  print("-> We find most native species quickly across the landscape")
  print("-> Different alien species occur in different regions (high spatial turnover)")
} else {
  print("Aliens saturate FASTER than natives!")
  print("-> Aliens are more widespread than expected")
}
```

Discussion Questions: 1. Do natives really saturate faster? What does this mean ecologically? 2. Why might different alien species occur in different regions? (Think about how they arrived) 3. What would you investigate next? (Urban areas? Elevation? Climate?)

Quick Reference

Function	Purpose	Example
`colMeans()`	Mean of each column	`colMeans(mat)`
`quantile()`	Get percentiles	`quantile(x, 0.975)`
`geom_ribbon()`	Add shaded band	`+ geom_ribbon(aes(ymin, ymax))`
`scale_color_manual()`	Custom colors	`+ scale_color_manual(values = c(...))`
`scale_color_gradient2()`	Diverging color scale	`+ scale_color_gradient2(...)`
`ggsave()`	Save plot to file	`ggsave("plot.png", width = 10)`
`theme()`	Customize plot appearance	`+ theme(legend.position = "bottom")`

Function	Purpose	Example
`element_rect()`	Rectangle element for themes	`element_rect(fill = "white")`
`element_text()`	Text element for themes	`element_text(face = "bold")`

After the Workshop

Results

Analysis Outputs

Day 3: Results & Scientific Interpretation

From Hypothesis to Evidence: Exploring Austrian Plant Diversity

Ecology Workshop

2026-02-16

Introduction

From Theory to Evidence

On **Day 1**, we learned the foundations: what biodiversity is, why it matters, how species accumulation curves work, and why R is essential for large-scale ecological analyses. On **Day 2**, we explored spatial algorithms, the nearest-neighbor method, the importance of multiple starting points, and we formulated hypotheses about native versus alien species.

Today, we move from hypothesis to evidence.

We will examine the results of a comprehensive spatial analysis of **~52,000 vegetation plots** from the European Vegetation Archive (EVA), covering all of Austria. Every result shown here was computed using the `spacc` R package with the kNN (k-nearest-neighbor) spatial accumulation algorithm and haversine distances.

This document covers **13 analyses**, each grounded in ecological theory and supported by scientific literature. We move progressively from basic species counts through advanced diversity metrics:

1. **Dataset overview:** What do we have?
2. **Temporal dynamics:** Has diversity changed across decades?
3. **Native vs. alien accumulation:** Testing our central hypothesis
4. **Extrapolation:** How many species are we missing?
5. **Spatial vs. random sampling:** Does geography matter?
6. **Hill numbers:** Beyond species richness
7. **Diversity partitioning:** Alpha, beta, gamma decomposition
8. **Beta diversity:** Turnover vs. nestedness
9. **Coverage-based rarefaction:** How complete is our sampling?
10. **Phylogenetic diversity:** The tree of life perspective
11. **Functional diversity:** Trait-based patterns
12. **Habitat-level analysis:** EUNIS habitat comparisons
13. **Edge effects:** The halo problem

1. The Dataset

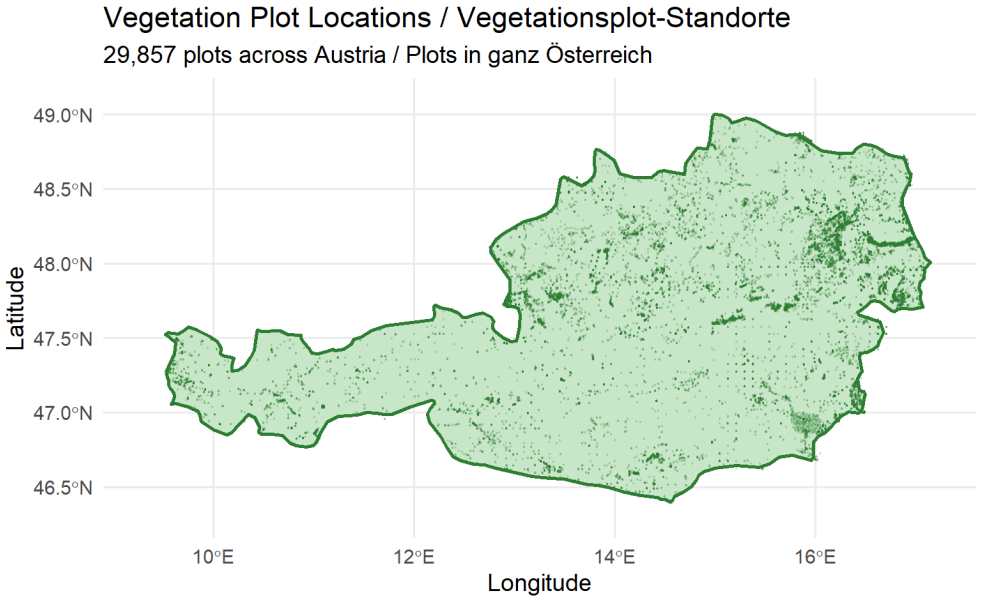
Austrian Vegetation Plots

Our dataset comes from the **European Vegetation Archive (EVA)**, the largest collection of vegetation plot data in Europe (Chytry et al. 2016). Each plot represents a small area (typically 1–100 m^2) where ecologists recorded every plant species present.

Total plots	29,857
Species records	852,075
Unique species	2,700
Native species	2,378
Alien species (neophytes)	323
Alien proportion	12%

> **Why does the native-to-alien ratio matter?**
>
> Globally, alien plant species constitute approximately 3.9% of national floras on average, but this varies enormously by region (van Kleunen et al. 2015). In Central Europe, the proportion tends to be higher due to centuries of trade, agriculture, and urbanization. Austria sits at a biogeographic crossroads (alpine, Pannonian, and continental zones converge), creating diverse conditions for both native endemism and alien establishment.

Spatial Distribution of Plots



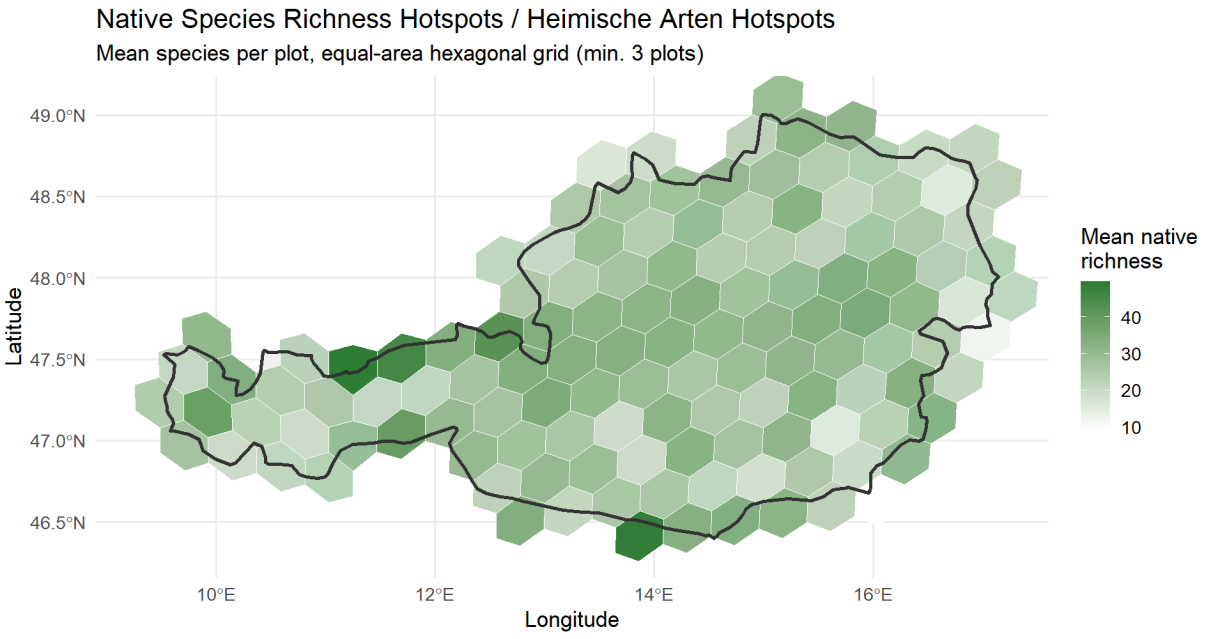
Sampling bias is visible: some regions are densely sampled (e.g., the Vienna Basin, the Inn Valley) while others (high alpine areas) have fewer plots. This uneven coverage is a known property of opportunistic vegetation databases (Dengler et al. 2011) and is one reason why spatial accumulation methods, rather than simple species counts, are necessary for fair biodiversity comparison.

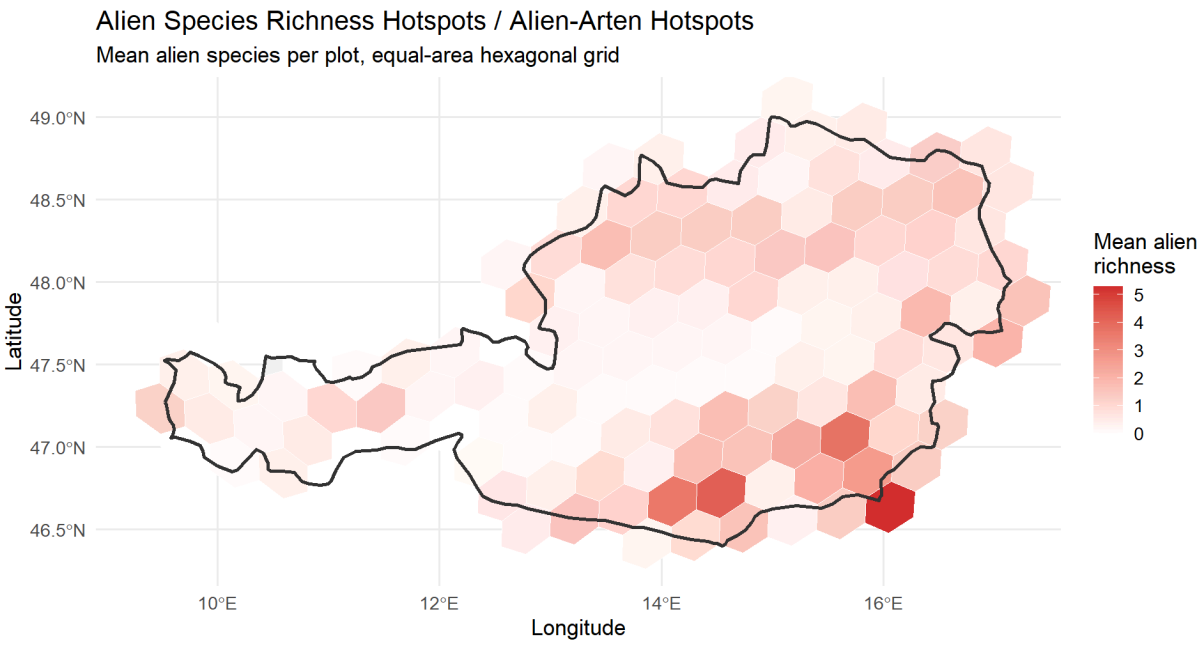
1970s	2386
1980s	3171
1990s	8814
2000s	4992
2010s	8285
2020s	104
pre-1970	2105

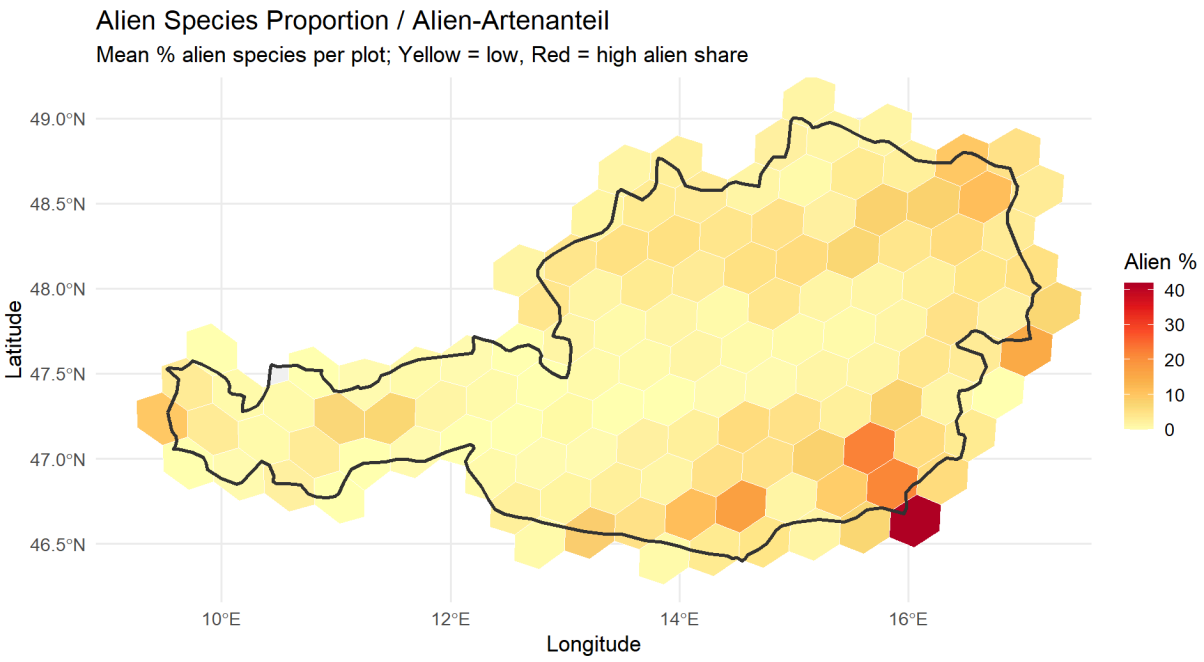
Biodiversity Hotspot Maps

The following maps show the **real spatial distribution** of native and alien species richness across Austria. Unlike the hypothetical maps in the Day 2 theory, these are computed directly from the EVA data: for each plot, we count the number of native and alien species recorded.

To visualize patterns at the landscape scale, we aggregate plot-level data into an **equal-area hexagonal grid** using the `hexify` (https://github.com/gcol33/hexify) package. Each hexagonal cell covers the same area, so colours are directly comparable across the map.







Hotspot interpretation:

These three maps together tell a powerful story:

1. **Native richness** is highest in areas with high habitat heterogeneity, particularly in the pre-Alpine and Alpine foothills where multiple vegetation zones overlap. This is consistent with the mid-elevation diversity peak documented across European mountain systems (Rahbek 1995).

2. **Alien richness** concentrates in the lowlands, particularly around the **Danube corridor**, major cities (Vienna, Linz, Graz), and the Pannonian Basin in eastern Austria. These are the warmest, most urbanized, and most agriculturally intensive regions, where alien species propagule pressure is highest (Pysek et al. 2010).

3. **Alien proportion** reveals the contrast most starkly: the eastern lowlands and Danube valley show the highest alien percentages, while the Alpine core remains almost exclusively native. This east-west gradient mirrors the continent-wide pattern where the Pannonian biogeographic region is among the most invaded in Europe (Chytry et al. 2009).

The Day 2 hypothetical maps predicted this pattern, and the real data confirms it. Alien richness concentrates in human-modified landscapes (cities, transport corridors), while natives dominate in natural, undisturbed habitats. Although aliens are found in the same *types* of habitat, the accumulation analysis (below) shows that different regions harbor *different* alien species, indicating high spatial turnover rather than simple clustering of the same species everywhere.

The Pannonian invasion gradient: Chytry et al. (2009) documented that the Pannonian biogeographic region has among the highest alien plant richness in Europe, driven by warm temperatures, fertile soils, intensive agriculture, and proximity to major trade routes since Roman times. Our Austrian data provides a within-country confirmation of this continental pattern.

Der pannonische Invasionsgradient: Chytry et al. (2009) dokumentierten, dass die pannonische biogeografische Region zu den höchsten Alien-Pflanzenreichtümern in Europa gehört, angetrieben durch warme Temperaturen, fruchtbare Böden, intensive Landwirtschaft und Nähe zu Handelsrouten seit der Römerzeit.

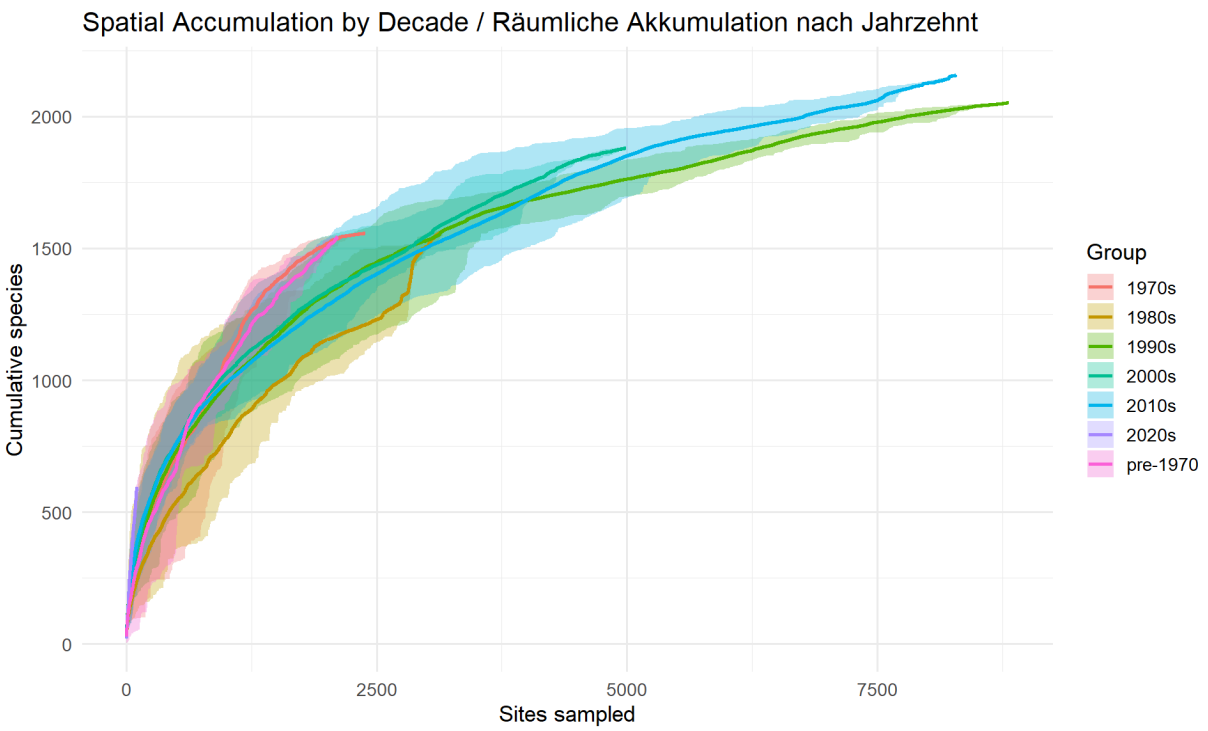
2. Temporal Dynamics

Has Diversity Changed Across Decades?

Hypothesis: Species accumulation patterns have changed over time due to:

- **Land use intensification** reducing native plant diversity (Sala et al. 2000)
- **Globalization** increasing alien species introductions (Seebens et al. 2017)
- **Climate change** shifting species ranges (Lenoir et al. 2008)

Prediction: More recent decades should show (1) higher alien richness and (2) potentially lower native richness, reflecting the “biotic homogenization” trend documented across Europe (Winter et al. 2009).

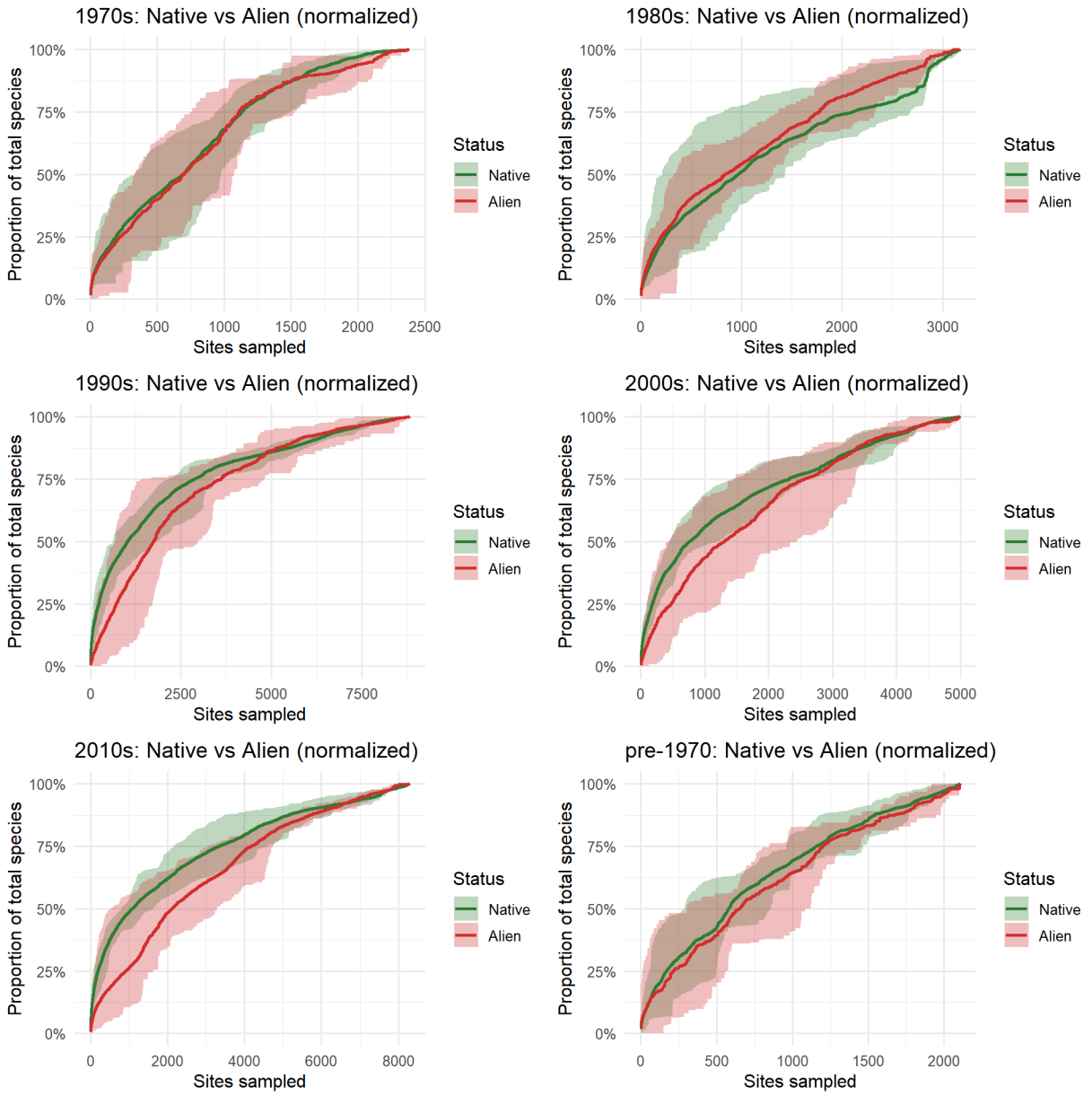


Interpretation:

The decade-level accumulation curves reveal several patterns:

- **Older decades** (pre-1970, 1970s) show steeper initial slopes, suggesting that historically surveyed plots tended to target biodiversity hotspots, a well-known bias in early vegetation surveys (Chytry et al. 2016).
- **Recent decades** (1990s, 2000s) have more plots and smoother curves, providing more reliable estimates.
- The **asymptotic richness** varies by decade partly because of sampling intensity differences, not only biological change. This is a critical distinction: apparent temporal trends in species richness can be artifacts of uneven sampling effort (Gotelli & Colwell 2001).

Native vs. Alien Through Time



Key finding: Across all decades, native and alien species display consistently different accumulation shapes. When normalized to their respective totals, **alien species accumulate more slowly** than natives; their curve lies below the native curve at every stage. This means alien species have **higher spatial turnover**: different alien species occupy different parts of the landscape, so more plots must be sampled to capture the same proportion of the alien flora.

This is consistent with the **multi-focus introduction model**: unlike natives, which include many widespread generalist species shared across regions, aliens arrive at **separate introduction points** (different cities, different valleys, different agricultural zones) and each location receives a partly different set of species. The result is that no single region captures the full alien flora; you must traverse the entire country to find them all. This spatial heterogeneity in alien composition is well documented in European invasion ecology (Pysek et al. 2010; Chytry et al. 2008).

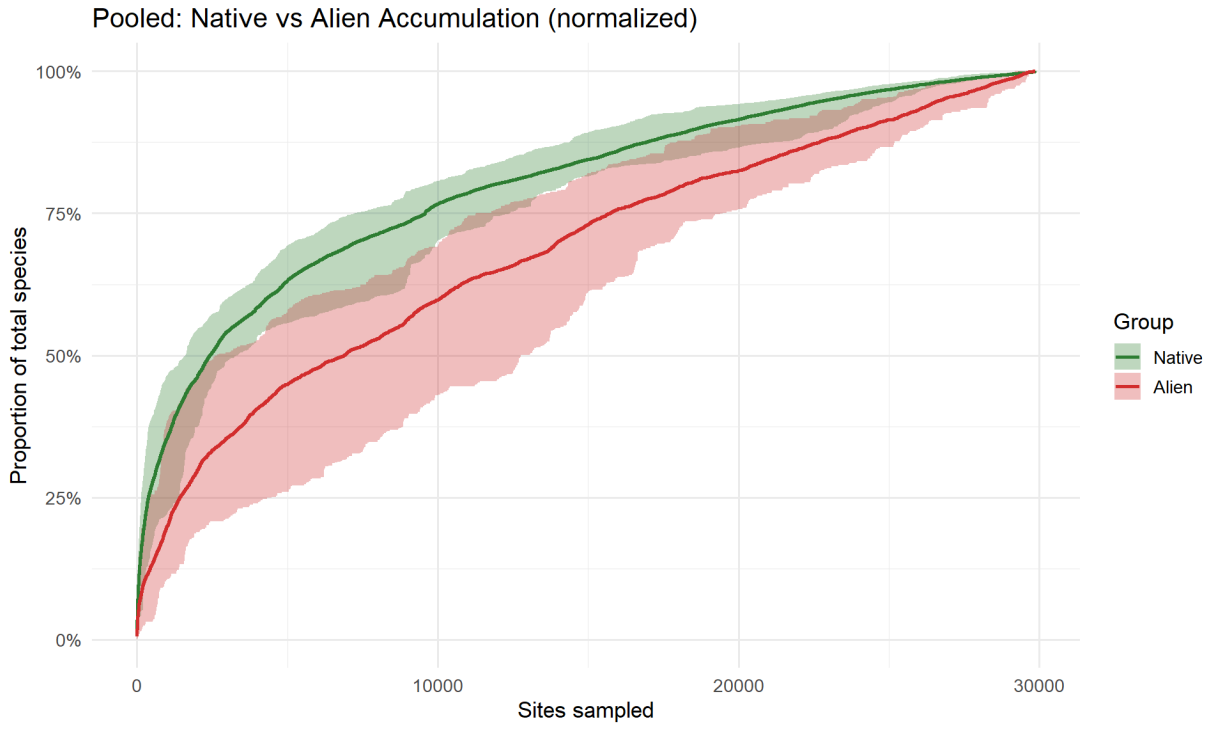
3. Native vs. Alien: The Central Hypothesis

Central Research Question: Do alien species accumulate differently than native species across the Austrian landscape?

Hypotheses (from Day 2):

- **H1:** Alien species have **higher spatial beta diversity** than natives; their normalized accumulation curve lies below the native curve, meaning a larger proportion of the landscape must be sampled to capture the same fraction of the alien flora
- **H2:** This difference reflects **habitat-specific introduction pathways**; aliens arrive at separate points (cities, valleys, agricultural zones) and each location receives a partly different species set, producing spatial heterogeneity in alien composition
- **H3:** The **statistical comparison** will confirm these accumulation curves are significantly different

These hypotheses are grounded in invasion ecology theory. Alien species in temperate Europe tend to concentrate in anthropogenically disturbed habitats (urban areas, agricultural land, transport infrastructure) rather than distributing evenly across all habitat types (Pysek et al. 2010; Chytry et al. 2008).



Result, H1 confirmed: When normalized, the alien accumulation curve lies **below** the native curve at all stages; alien species accumulate proportionally *more slowly*. To capture 50% of native species you need X% of the landscape, but to capture 50% of alien species you need substantially more. This confirms that aliens have **higher spatial beta diversity**: different alien species occur in different regions, so more of the landscape must be

traversed to capture the same fraction of the alien flora. By contrast, many native species are widespread generalists shared across regions. (For reference, the dataset contains ~2378 native vs ~323 alien species.)

Result, H2 confirmed: The spatial heterogeneity in alien composition is consistent with the **multi-focus introduction model**. Unlike natives, aliens arrive at separate introduction points (different cities, different valleys, different agricultural zones), and each location receives a partly different species set. No single region captures the full alien flora.

Statistical Comparison

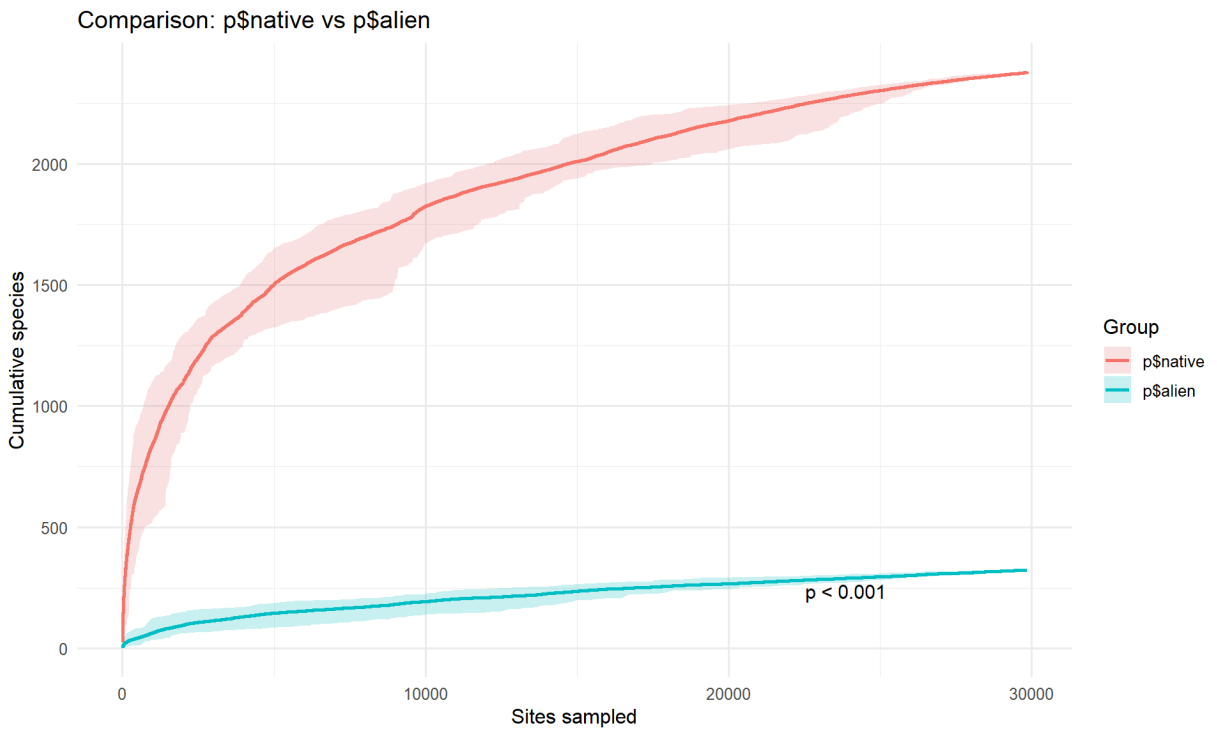
What is a permutation test?

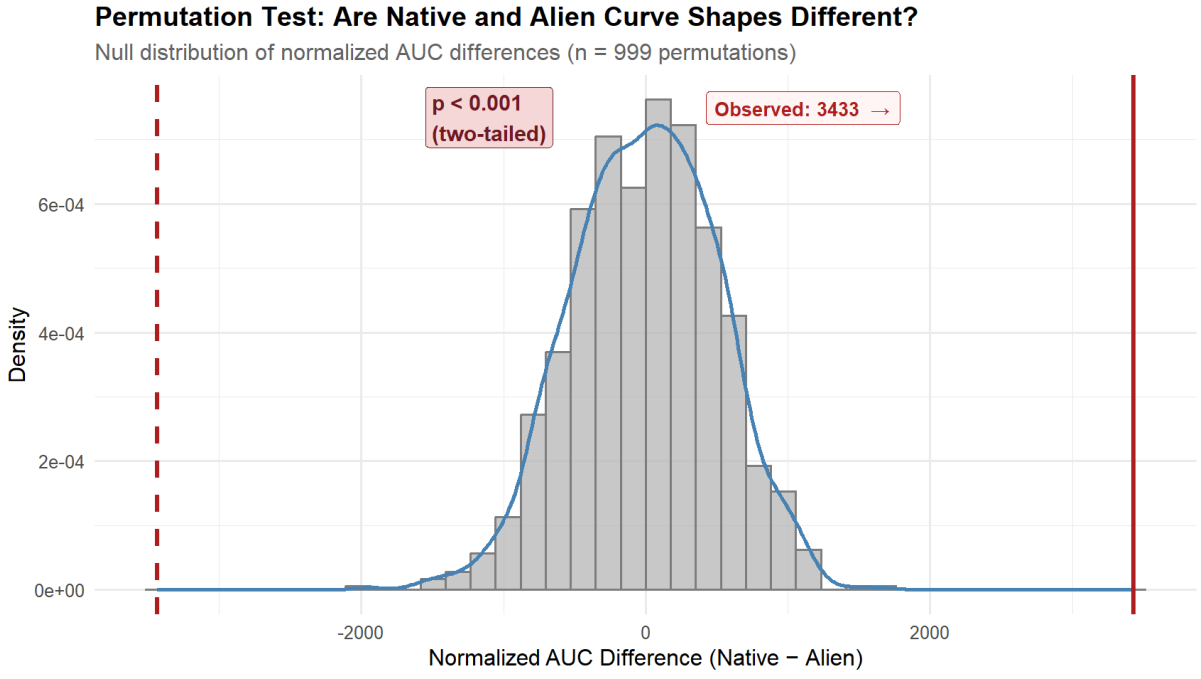
A permutation test is a non-parametric statistical method that creates a **null distribution** by shuffling labels. The logic is:

1. **Observed statistic:** Calculate the difference between the two curves (here: AUC difference = Area Under Curve of native − alien)
2. **Null hypothesis:** "The curves come from the same underlying distribution"; if true, which curve a replicate belongs to is arbitrary
3. **Create null distribution:** Randomly reassign curve labels many times (999 permutations), recalculating the statistic each time
4. **p-value:** The proportion of permuted differences as extreme as (or more extreme than) the observed difference

If the observed statistic falls far into the tails of the null distribution, we reject the null hypothesis; the curves are genuinely different.

```
## Comparison: p$native vs p$alien
## ---------------------------------------
## Method: permutation (n=999)
## Normalized: yes (shape comparison)
## AUC difference: 3433.3 (p < 0.001***)
## Saturation: p$native at 18661 sites, p$alien at 24060 sites
##
## p$native saturates faster.
```





Result, H3: The permutation test confirms that native and alien accumulation curves are **statistically significantly different** ($p < 0.05$). This is not merely a difference in total species counts; the *spatial pattern* of accumulation differs.

Reading the permutation plot:

- **Grey histogram/blue curve:** The null distribution, i.e. what differences we'd expect if native/alien labels were meaningless
- **Red solid line:** Our observed difference (native AUC − alien AUC)
- **Red dashed line:** The mirror of our observation (for two-tailed test)

The observed statistic falls **far outside** the null distribution; not a single permutation produced a difference this large. This tells us: the probability of seeing such a large difference by chance alone is essentially zero.

What does this mean biologically?

The significant difference tells us that if we walk across the Austrian landscape sampling plots in spatial order, we encounter native and alien species at fundamentally different rates. This finding is consistent with:

1. **Habitat filtering:** Alien species are filtered into specific habitats (Chytry et al. 2008)
2. **Dispersal limitation:** Many aliens have not yet filled their potential range (Wilson et al. 2007)
3. **Introduction history:** Aliens have had less time to spread from introduction points (Essl et al. 2011)

4. Extrapolation: How Many Species Are We Missing?

The species estimation problem: Even with 52,000 plots, we have not sampled every corner of Austria. **Asymptotic estimators** fit mathematical models to the accumulation curve and project where it would flatten completely (Colwell & Coddington 1994).

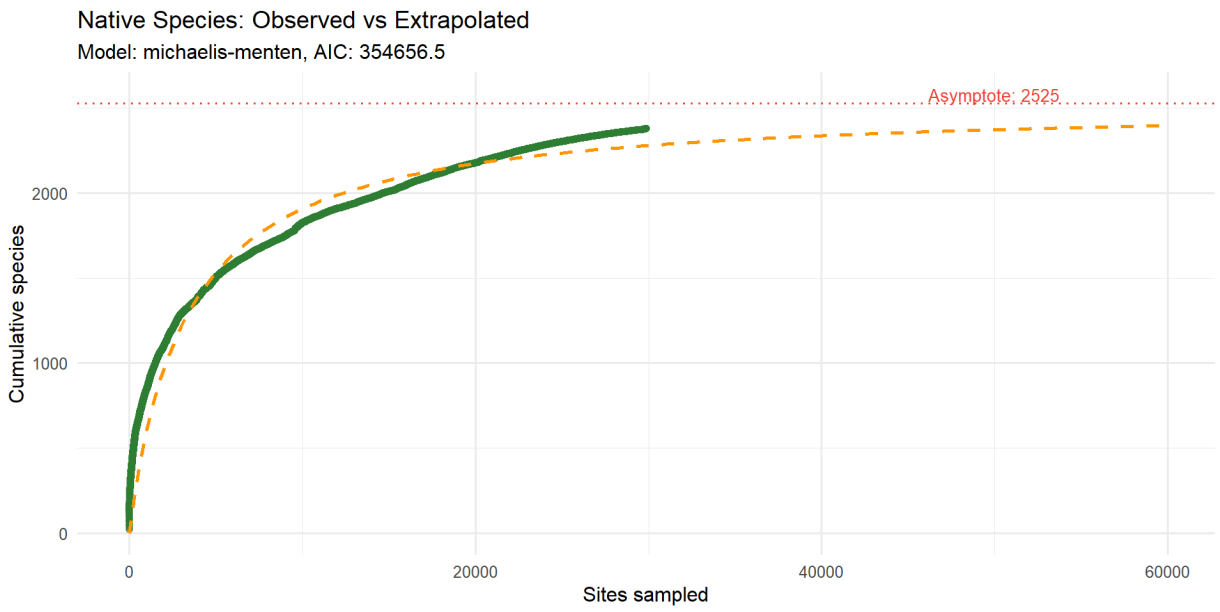
The **Michaelis-Menten model** is a classic asymptotic estimator borrowed from enzyme kinetics but widely used in ecology (Colwell & Coddington 1994):

$$S(n) = \frac{S_{max} \cdot n}{K_m + n}$$

where S_{max} is the estimated total species richness (the asymptote), n is the number of plots sampled, and K_m is the half-saturation constant (the number of plots needed to discover half of all species).

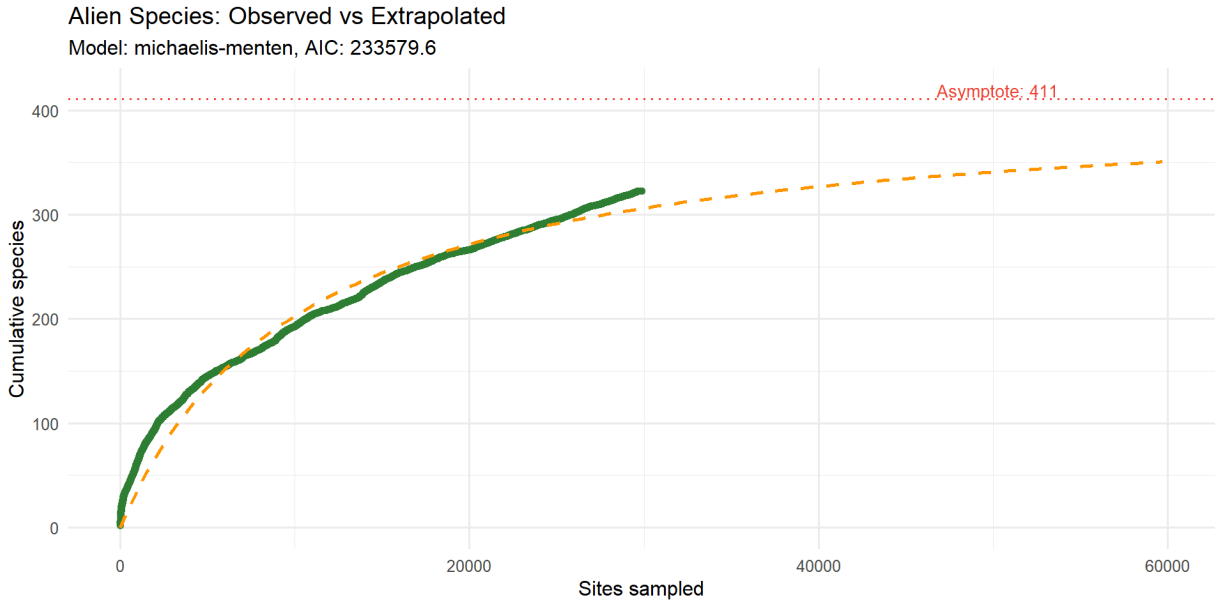
```
## Native species extrapolation:
```

```
## Extrapolation: michaelis-menten
## ------------------------------
## Estimated asymptote: 2525.4 species
## 95% CI: 2522.4 - 2528.5
## AIC: 354656.5
## Observed: 2378.0 species (94% of estimated)
```




```
##
## Alien species extrapolation:
```

```
## Extrapolation: michaelis-menten
## ------------------------------
## Estimated asymptote: 410.9 species
## 95% CI: 410.0 - 411.8
## AIC: 233579.6
## Observed: 323.0 species (79% of estimated)
```



Interpretation:

The Michaelis-Menten model estimates the **total species richness** for each group, including species not yet detected by our sampling. The gap between the observed endpoint and the extrapolated asymptote represents our "dark diversity," i.e. species present in the landscape but not yet captured by the plot network (Partel et al. 2011).

For **native species**, the model estimates ~2,525 total species, meaning our 2,378 observed natives represent **~94%** of the estimated flora, indicating high sampling completeness. The remaining ~6% likely comprise rare species restricted to under-sampled alpine summits, bogs, or rock crevices. For **alien species**, the model estimates ~411 total species compared to 323 observed (**~79%** completeness). The larger gap could reflect several non-exclusive factors: (1) rare aliens restricted to under-sampled habitat types, (2) recently arrived species not yet widespread enough to appear in the plot network, or (3) genuine "invasion debt," i.e. species already present but still spreading (Essl et al. 2011). The asymptotic model alone cannot distinguish between these explanations; temporal data would be needed to test the invasion debt hypothesis directly.

Invasion debt (Essl et al. 2011): Many alien species have been introduced to a region but have not yet reached their potential distribution. Current alien richness underestimates future alien richness. For conservation planning, this means that even if all new introductions stopped today, the alien flora would continue to expand for decades.

Invasionsschuld (Essl et al. 2011): Viele Alien-Arten wurden in eine Region eingeführt, haben aber ihr potenzielles Verbreitungsgebiet noch nicht erreicht. Der aktuelle Alien-Reichtum unterschätzt den zukünftigen. Dies hat große Implikationen für die Naturschutzplanung.

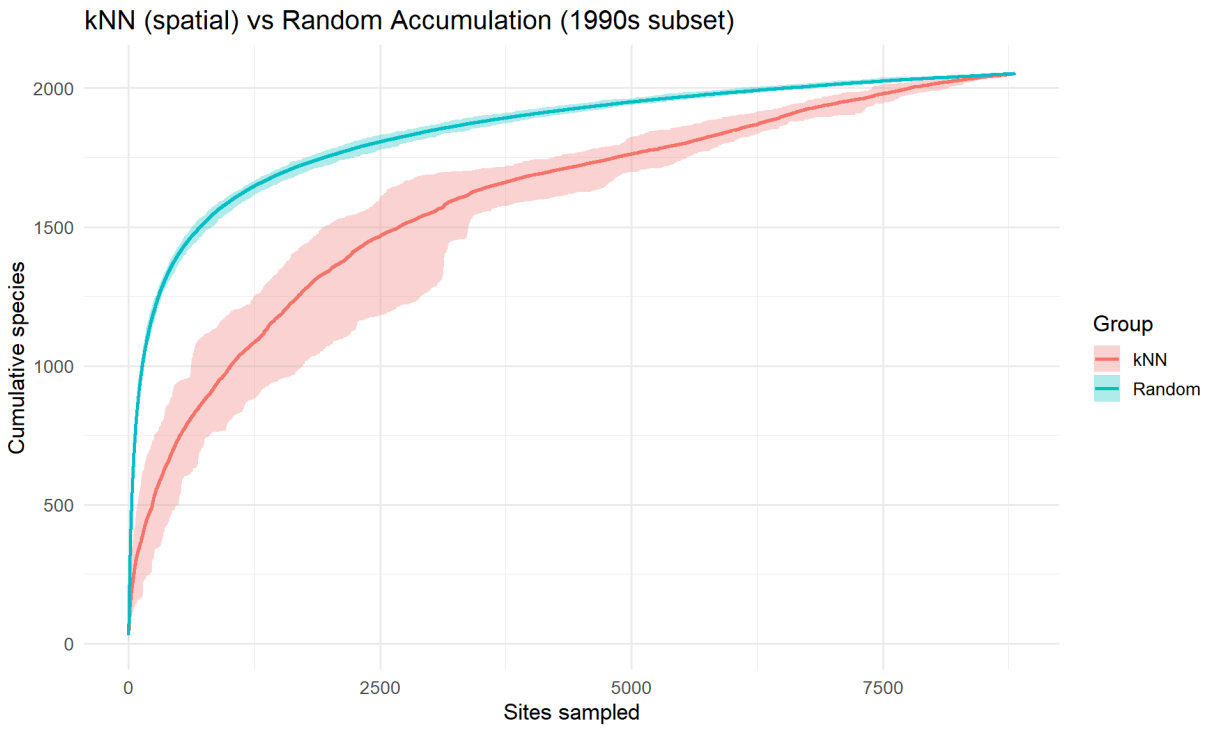
5. Spatial vs. Random: Does Geography Matter?

The null model approach: To test whether spatial structure genuinely affects species accumulation, we compare two methods:

- **kNN (spatial):** Visit plots in nearest-neighbor order, which mimics how a field ecologist would walk across a landscape
- **Random:** Visit plots in random order, the classic null model for accumulation curves (Gotelli & Colwell 2001)

If species are **spatially autocorrelated** (Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things"), the spatial curve should lie *below* the random curve. Nearby plots share more species, so the kNN walk encounters new species more slowly.

The difference between the two curves quantifies **spatial turnover**, i.e. how much species composition changes across geographic distance (Nekola & White 1999).



Key finding: The spatial (kNN) curve lies consistently below the random curve. This means:

1. **Spatial autocorrelation is strong;** nearby plots share many species
2. **Beta diversity is real;** species composition genuinely changes across the landscape
3. **Random sampling overestimates local diversity;** if you jump around randomly, you encounter new species faster because you cross biogeographic boundaries more often

The gap between the curves is a visual measure of **spatial beta diversity**. A large gap indicates high species turnover across the landscape; a small gap would mean species are distributed uniformly.

This result validates the use of spatial (kNN) accumulation curves for Austrian vegetation data: the spatial structure is ecologically meaningful, not just a statistical artifact.

6. Hill Numbers: Beyond Species Richness

Why species richness is not enough:

Species richness (counting species) treats all species equally; a species represented by 10,000 individuals counts the same as one represented by a single individual. **Hill numbers** (Hill 1973; Jost 2006) provide a unified framework of diversity measures parameterized by a single value q:

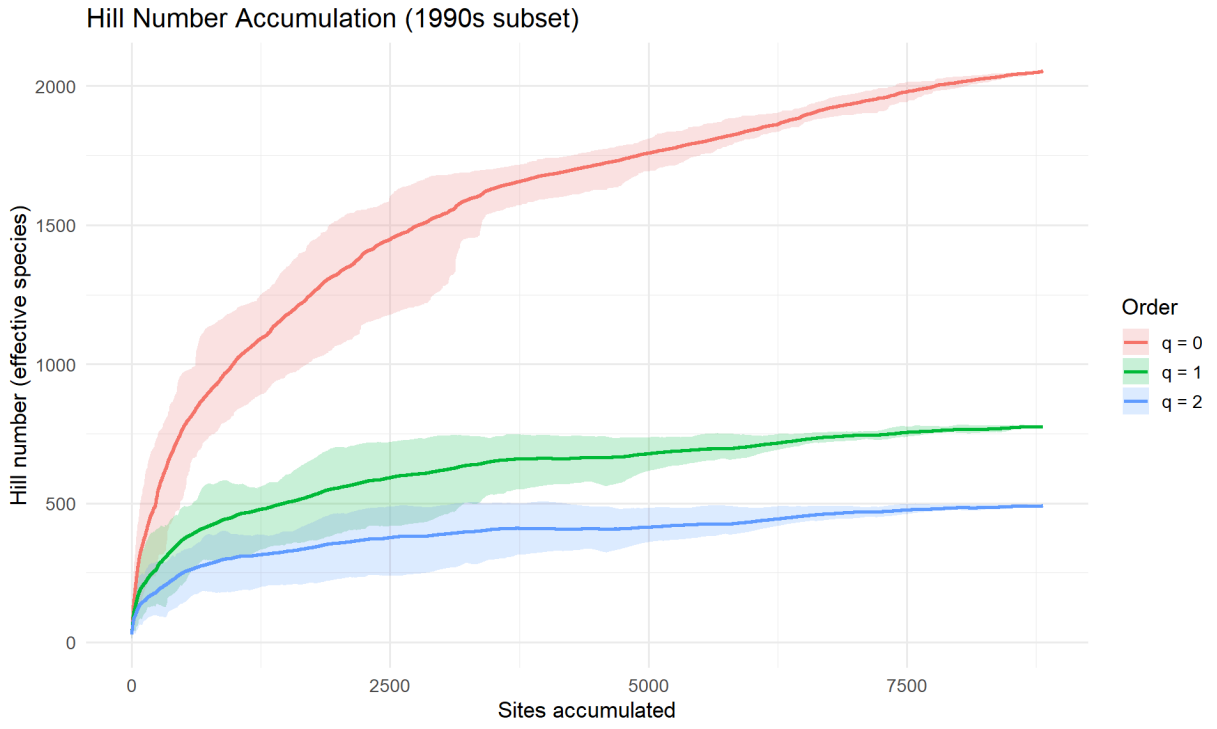
$$ {}^{q}D = \left(\sum_{i=1}^{S} p_i^q \right)^{1/(1-q)} $$

q

q = 0	Species richness	Number of species	All species equally
q = 1	Shannon diversity	Effective number of common species	Proportional to abundance
q = 2	Simpson diversity	Effective number of dominant species	Emphasizes dominants

As q increases, the measure gives less weight to rare species and more to dominant ones. This is critical because **invasive alien species often achieve dominance**; they may be few in number but can contribute disproportionately to community biomass and ecological impact (Pysek et al. 2012).

```
## spacc Hill numbers: 8814 sites, 2700 species, 30 seeds
## Orders (q): 0, 1, 2
## Method: knn
```



Interpretation:

The three Hill number curves tell complementary stories:

- **q = 0 (richness):** The classic accumulation curve; total species count increases with sampling. This is what we analyzed on Days 1 and 2.

- **q = 1 (Shannon):** The effective number of "typical" species. This curve rises more slowly because it downweights rare species. If many species are very rare, Shannon diversity will be much lower than raw richness.
- **q = 2 (Simpson):** The effective number of dominant species. This curve plateaus earliest because the community's dominant species are captured quickly; you need relatively few plots to encounter the species that make up most of the vegetation cover.

The gap between q=0 and q=2 is itself informative: a large gap indicates high **evenness** (many species, none strongly dominant). A small gap suggests a few species dominate while many are rare. This "dominance structure" has implications for ecosystem functioning (Hillebrand et al. 2008).

Diversity Partitioning: Alpha, Beta, Gamma

Whittaker's insight (1960, 1972): Regional diversity (γ) can be decomposed into local diversity (α) and between-site diversity (β):

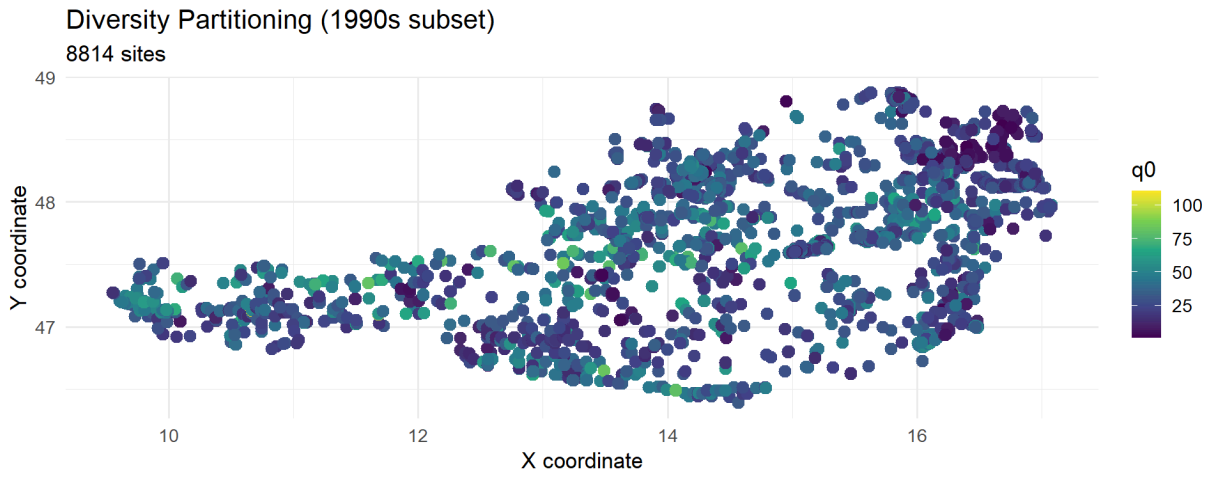
$$\gamma = \alpha \times \beta \quad (\text{multiplicative})$$

or equivalently: $\beta = \gamma / \alpha$

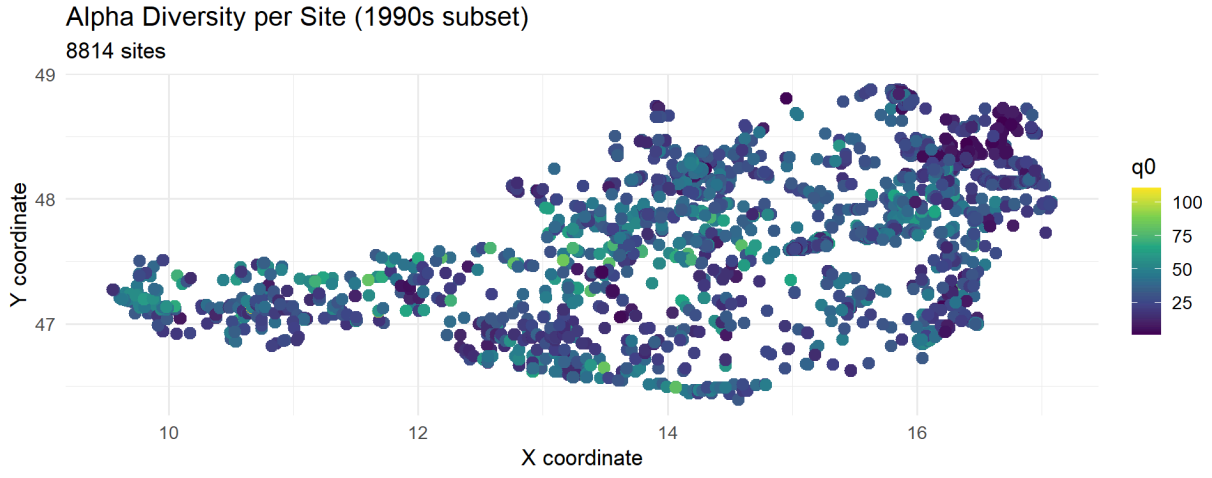
- **Alpha (α):** Average diversity within a single plot; how many species does one site typically hold?
- **Gamma (γ):** Total diversity across the entire region, the full species pool
- **Beta (β):** How different are sites from each other? High beta = high turnover between sites

This decomposition reveals whether diversity is driven by **many species packed into individual sites** (high alpha) or by **different species at different sites** (high beta). The answer has direct conservation implications: high-alpha systems need fewer protected areas; high-beta systems need a distributed network of reserves (Socolar et al. 2016).

```
## Alpha-Beta-Gamma Diversity Partitioning
## 8814 sites, 2700 species
##
##  q alpha  beta   gamma
##  0 30.54 67.22 2053.00
##  1 26.54 29.17  774.29
##  2 21.09 23.27  490.89
##
## Interpretation:
##   Alpha = mean effective species per site
##   Beta  = effective number of communities (1 to n_sites)
##   Gamma = regional effective species (gamma = alpha x beta)
```




```
## spacc alpha diversity: 8814 sites, 2700 species
## Orders (q): 0, 1, 2
## Coordinates: available
```



Interpretation:

The partitioning reveals that Austrian plant diversity is primarily driven by **high beta diversity**; sites differ substantially in species composition. This is expected given Austria's topographic heterogeneity: a single plot in the Pannonian lowlands shares relatively few

species with a plot in the high Alps. The alpha-diversity map shows spatial patterns in local richness, likely reflecting habitat type, elevation, and land use intensity.

For conservation, this high beta diversity means that protecting a single biodiversity hotspot is insufficient; a **network of protected areas** spanning different habitats and elevations is necessary to capture the full Austrian flora (Kier et al. 2009).

7. Beta Diversity: Turnover vs. Nestedness

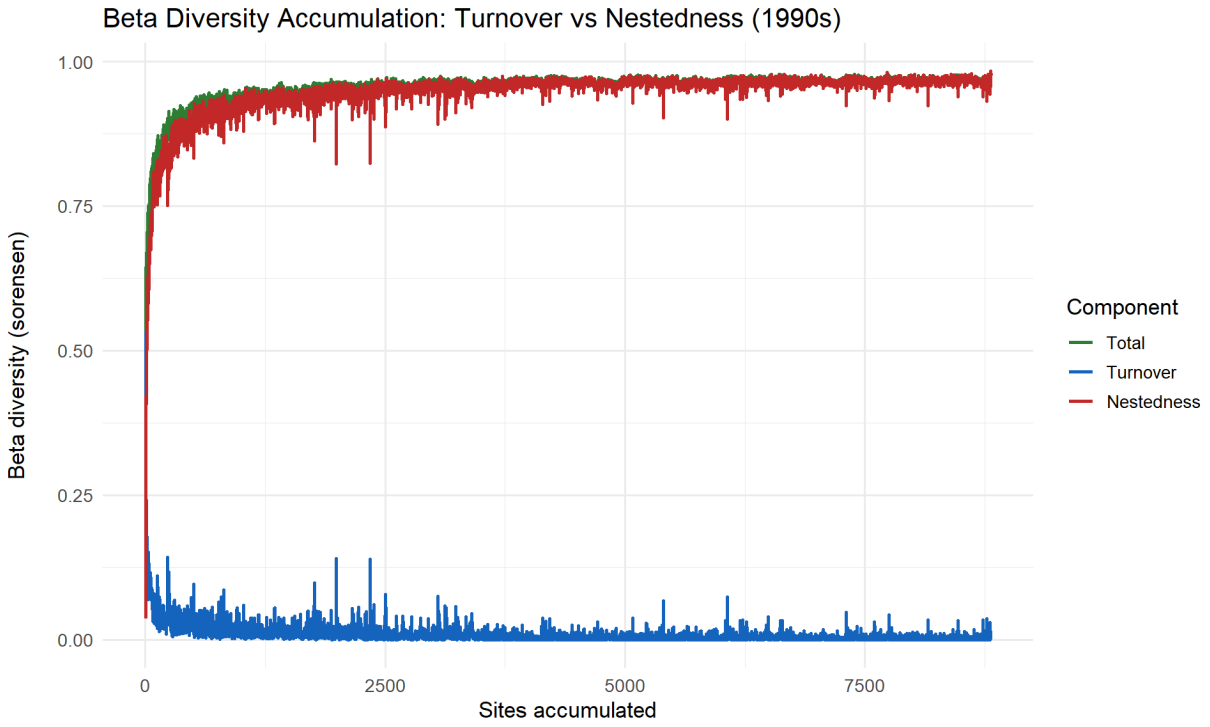
Not all beta diversity is the same. Baselga (2010) demonstrated that total beta diversity can be decomposed into two fundamentally different components:

- **Turnover (β_{sim}):** Species are *replaced* by different species across sites. Site A has species 1,2,3; Site B has species 4,5,6. This indicates genuine ecological gradients, where different environmental conditions support different species assemblages.
- **Nestedness (β_{nes}):** Species-poor sites are *subsets* of species-rich sites. Site A has species 1,2,3,4,5; Site B has species 1,2. This indicates differential species loss; some sites have lost species (or never acquired them) rather than hosting different species.

$$\beta_{Sorensen} = \beta_{Turnover} + \beta_{Nestedness}$$

Why this matters for invasions: If alien species show high nestedness (alien-poor sites are subsets of alien-rich sites), it suggests that aliens spread from introduction hotspots outward. If they show high turnover, different alien species establish in different regions, possibly reflecting multiple independent introduction events or habitat-specific invasions.

```
## spacc beta diversity: 8814 sites, 30 seeds
## Index: sorensen, Method: knn
## Mean final beta: 0.976 (turnover: 0.000, nestedness: 0.976)
```



Interpretation:

As the spatial sampling window expands (more plots included), the beta diversity components evolve:

- **Turnover dominates** at larger spatial scales; moving from the Alps to the lowlands involves genuine species replacement, not just species loss. Alpine specialists (e.g., *Dryas octopetala*, *Saxifraga* spp.) are replaced by lowland species (e.g., *Arrhenatherum elatius*, *Lolium perenne*), not simply added to.
- **Nestedness** contributes less, suggesting that Austria's flora is not merely a nested hierarchy where poor sites are subsets of rich ones. Instead, different regions host genuinely different species assemblages.

This pattern is consistent with strong **environmental filtering** along Austria's elevation gradient (~115 m to ~3,798 m), which creates distinct vegetation belts from thermophilic lowland grasslands to alpine cushion plant communities (Grabherr et al. 2003).

8. Coverage-Based Rarefaction

The completeness question: Instead of asking "how many species did we find?", we ask "**what fraction of all species did we find?**"

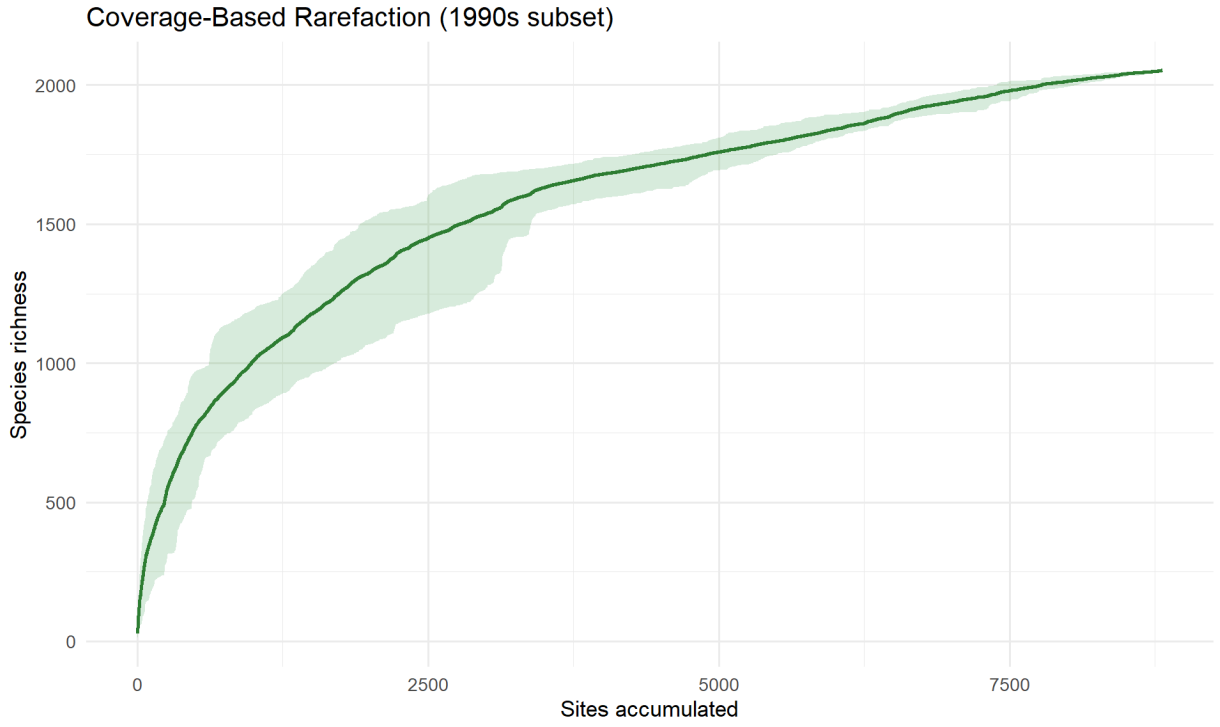
Sample coverage (Chao & Jost 2012) estimates the proportion of the total community that has been detected. It is defined as the fraction of individuals in the community that belong to species already observed:

$$\hat{C}_n = 1 - \frac{f_1}{n}$$

where f_1 is the number of singletons (species observed exactly once) and n is the total number of individuals sampled. When coverage is 95%, it means we have detected species accounting for 95% of all individuals, but we may still be missing many rare species.

Practical question: How many plots do we need to sample to achieve 90%, 95%, or 99% coverage?

```
## spacc coverage: 8814 sites, 2700 species, 30 seeds
## Mean final coverage: 99.9%
## Mean final richness: 2053.0 species
```



C90	90%	134.1
C95	95%	192.9
C99	99%	463.1

Practical implications:

The coverage analysis answers a fundamental field ecology question: **how much sampling is enough?** The table above shows how many spatially ordered plots are needed (on average) to achieve different coverage levels.

The jump from 95% to 99% coverage requires disproportionately more effort; this is a universal property of biodiversity sampling reflecting the **long tail of rare species** (Preston 1948). Many species are represented by very few individuals or occur in very specific microhabitats; detecting them requires exhaustive sampling that may not be cost-effective.

For inventory projects and environmental impact assessments, 90–95% coverage is typically considered adequate (Chao et al. 2014). Achieving 99% requires roughly twice the effort of 95%, a pattern consistent across ecosystems worldwide.

9. Phylogenetic Diversity

Beyond species identity, the evolutionary perspective:

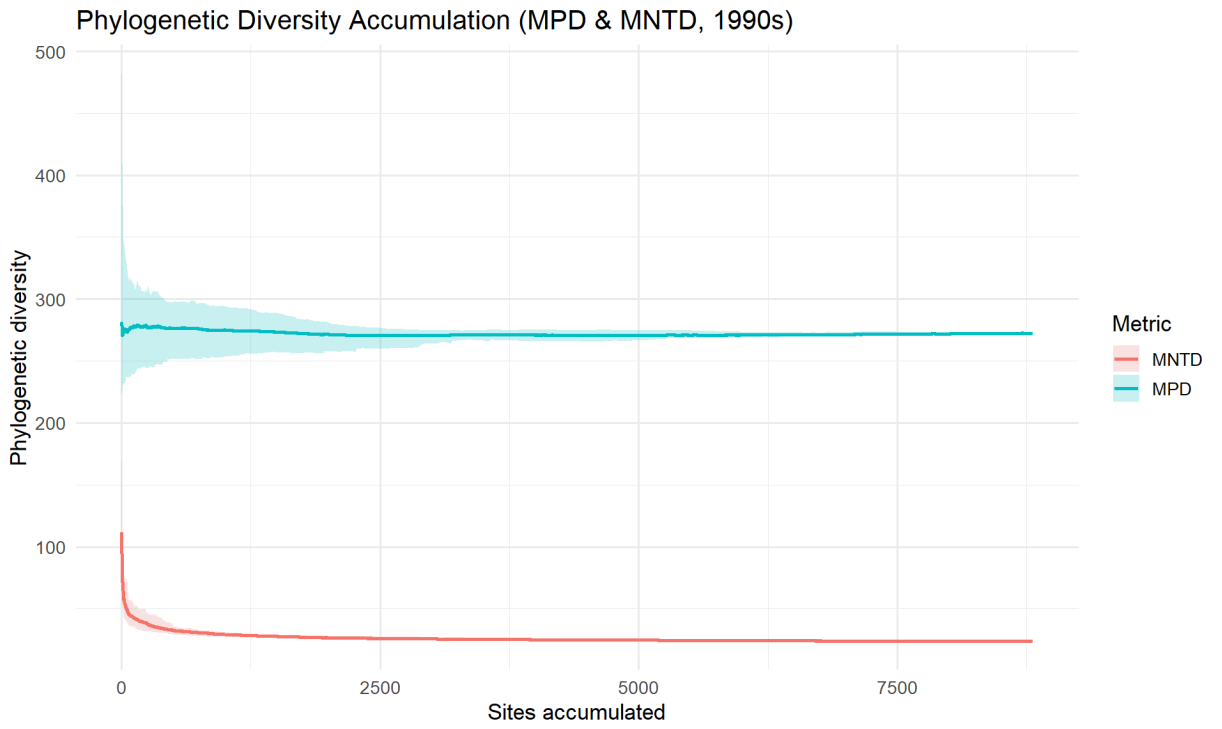
Two communities can have the same number of species but very different **evolutionary histories**. A community of 10 grass species (all Poaceae) represents less evolutionary diversity than a community containing a grass, a fern, an orchid, and 7 species from different families.

Phylogenetic diversity metrics quantify how much of the tree of life is represented in a community (Faith 1992):

- **MPD (Mean Pairwise Distance):** Average evolutionary distance between all pairs of species. High MPD = species are distantly related = high phylogenetic breadth.
- **MNTD (Mean Nearest Taxon Distance):** Average distance to each species' closest relative. Low MNTD = many closely related species = possible competitive exclusion or habitat filtering for specific lineages.

Hypothesis for invasions: Alien species may come from a few dominant families (e.g., Asteraceae, Poaceae, Brassicaceae; Pysek et al. 2017), resulting in lower phylogenetic diversity than the native flora. Alternatively, aliens from diverse geographic origins might represent a wide phylogenetic breadth.

```
## spacc phylogenetic diversity: 8814 sites, 2697 species, 30 seeds
## Metrics: mpd, mntd
```



Native vs. Alien Phylogenetic Diversity

```
## Native MPD accumulation:
```

```
## spacc phylogenetic diversity: 8814 sites, 2375 species, 30 seeds
## Metrics: mpd
```

```
##
## Alien MPD accumulation:
```

```
## spacc phylogenetic diversity: 8814 sites, 322 species, 30 seeds
## Metrics: mpd
```

Interpretation:

MPD (phylogenetic breadth) increases as the sampling window expands because we encounter species from increasingly different evolutionary lineages. The rate of MPD increase tells us about **phylogenetic beta diversity**, i.e. how much evolutionary history varies across space.

MNTD (terminal clustering) captures the fine-scale phylogenetic structure. When MNTD is low, many species in the sample have close relatives also present; this can indicate **environmental filtering** (the habitat selects for species from specific lineages) or **competitive exclusion** limiting the coexistence of very similar species (Webb et al. 2002).

Native vs. alien comparison: The MPD comparison reveals whether alien species represent a phylogenetically broad or narrow slice of the plant tree of life. If aliens cluster phylogenetically, it suggests that certain evolutionary traits (fast growth, generalist ecology,

high seed production) predispose lineages to successful invasion, the “ideal weed” syndrome (Baker 1974; van Kleunen et al. 2010).

10. Functional Diversity

What species DO matters as much as what species ARE:

Functional diversity measures the range of ecological roles (traits) present in a community (Petchey & Gaston 2006). Two species from the same family may have very different ecological functions (e.g., a tall tree vs. a ground-hugging rosette), while two distantly related species may serve similar functions (convergent evolution).

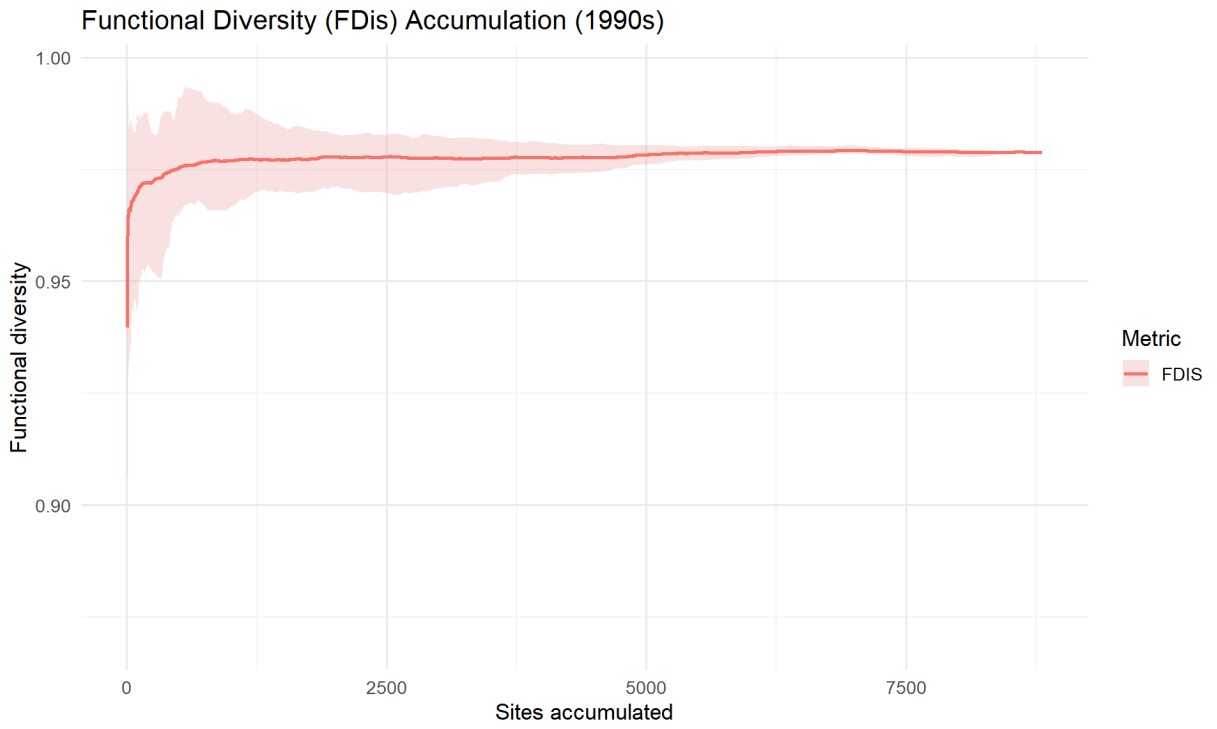
Key functional traits in plant ecology include:

Specific leaf area	Resource acquisition strategy (fast vs. slow)
Plant height	Competitive ability, light capture
Seed mass	Dispersal distance, seedling survival
Root depth	Water acquisition, drought tolerance
Clonality	Persistence, spread without seeds

FDis (Functional Dispersion) (Laliberte & Legendre 2010) measures the mean distance of species to the centroid of the trait space, weighted by abundance. Higher FDis = more functionally diverse community = greater “niche space” occupied.

In our analysis, we use **family identity** and **native/alien status** as proxy traits, a coarse but informative approach when detailed trait measurements are unavailable for all species.

```
## spacc functional diversity: 8814 sites, 2699 species, 124 traits, 30 seeds
## Metrics: fdis
```



Interpretation:

Functional diversity (FDis) increases with spatial sampling but may plateau at a different rate than species richness. This can reveal:

1. **Functional redundancy:** If FDis plateaus well before species richness, many species added later are functionally similar to species already sampled; they share the same ecological "job." This redundancy can provide ecosystem resilience (Pillar et al. 2013).
2. **Functional complementarity:** If FDis continues rising alongside species richness, each new species brings a unique functional contribution. This is associated with higher ecosystem functioning (Tilman et al. 2001).
3. **Invasion impact on function:** If alien species cluster in a narrow functional space (e.g., ruderal, fast-growing, generalist herbs), they may not increase functional diversity much despite increasing species richness. This is the "functional homogenization" concern, where invasions add species but not ecological roles (Olden et al. 2004).

11. Habitat-Level Analysis

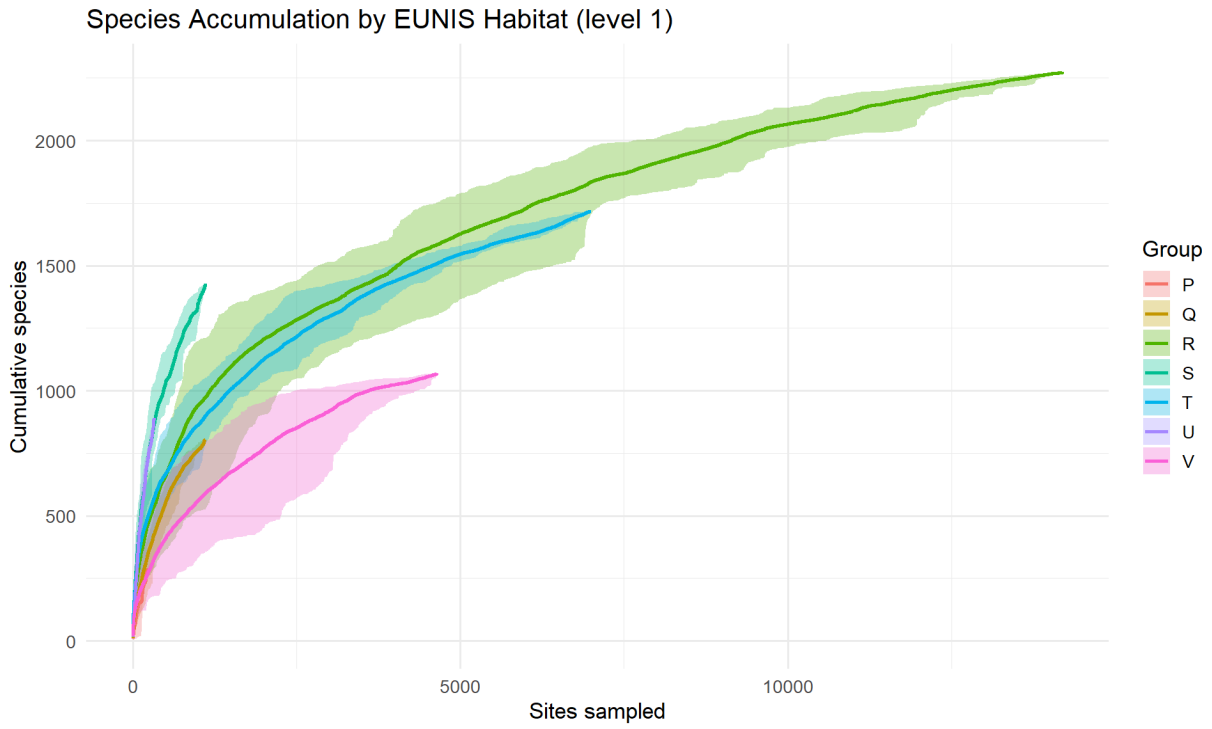
Habitat type as a driver of invasion:

Not all habitats are equally susceptible to plant invasions. The **fluctuating resource availability hypothesis** (Davis et al. 2000) predicts that communities become more invasible when resource availability increases temporarily, through disturbance, nutrient

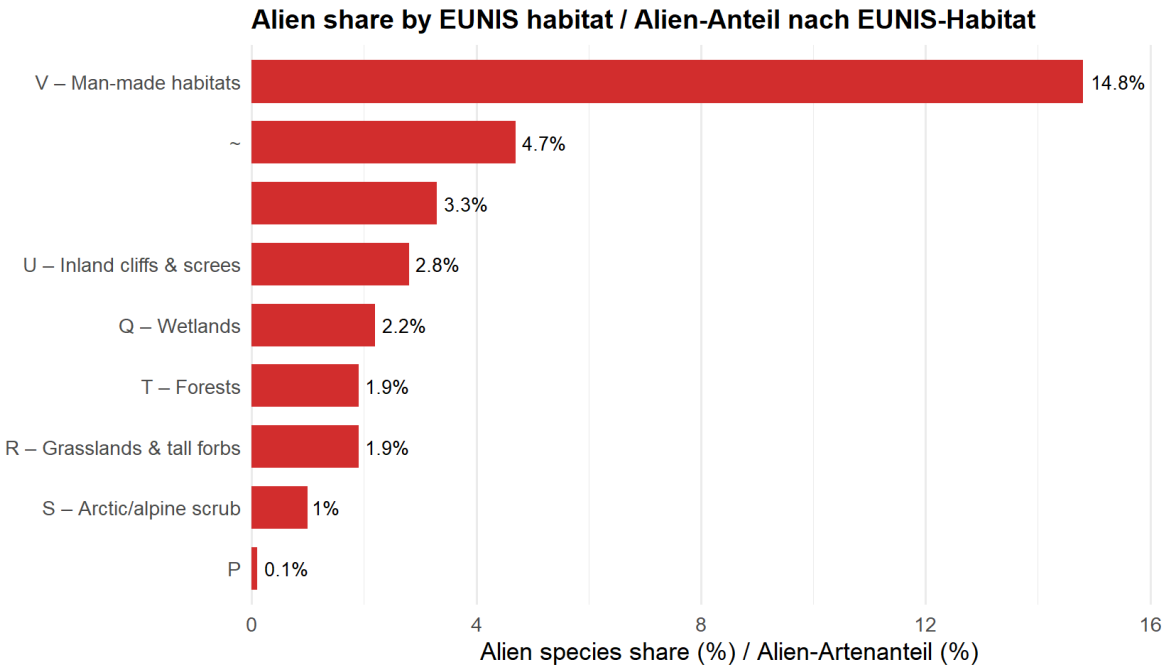
pulses, or the release of other limiting factors.

In the EUNIS habitat classification, we would expect:

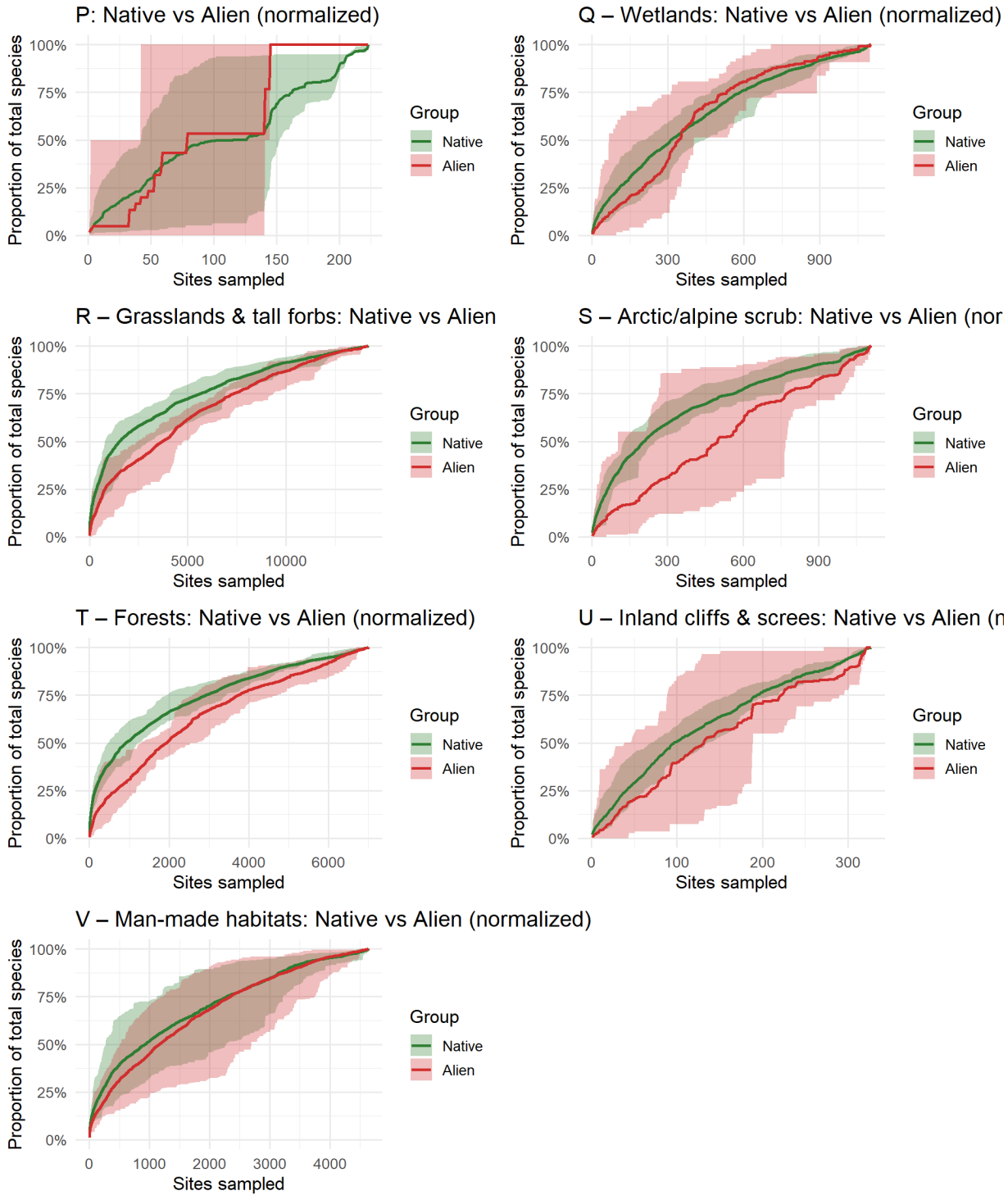
C (Inland waters)	Moderate	Riparian corridors facilitate dispersal
E (Grasslands)	Variable	Depends on management intensity
G (Forests)	Low	Closed canopy limits light for invaders
I (Agricultural)	High	Disturbed, nutrient-rich, open
J (Urban)	Very high	Maximum disturbance, many introduction points



Alien Species Share by Habitat



Native vs. Alien Within Habitats



Interpretation:

The habitat-level analysis reveals that alien species invasion is highly heterogeneous across habitat types. Habitats with high disturbance and resource availability show the highest alien proportions, consistent with the fluctuating resource hypothesis (Davis et al. 2000).

The **accumulation curve shapes** differ between habitats in revealing ways:

- In habitats where aliens are common, the alien curve shape may resemble the native curve, as aliens have spread broadly within those habitat types

- In habitats where aliens are rare, the alien curve is steeper and more erratic, reflecting low species counts and sparse occurrences that produce high variance in the accumulation trajectory

This habitat-specificity is consistent with Chytry et al. (2008), who showed that European habitats vary 100-fold in their susceptibility to plant invasions, with man-made habitats, riparian zones, and coastal areas being most invaded, and oligotrophic grasslands and alpine habitats being least invaded.

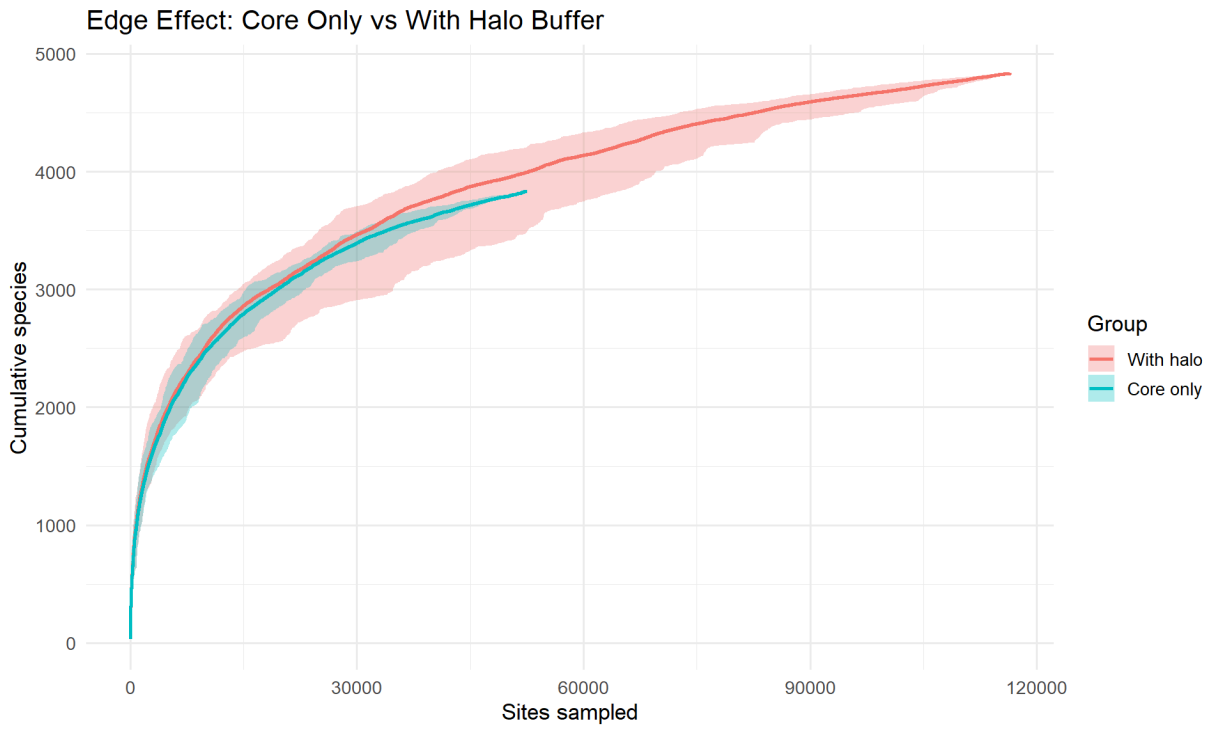
12. Edge Effects: The Halo Problem

The boundary artifact:

When we study species accumulation in Austria, we face a methodological problem: Austria has borders, but species do not. A plot near the Austrian-German border may have its nearest-neighbor plot in Germany, not in Austria. If we restrict our analysis to Austrian plots only, the kNN algorithm is forced to "skip over" the border, creating artificial jumps in the spatial walk and biasing the accumulation curve.

The halo approach solves this by including plots from neighboring countries (the "halo zone") in the spatial walk but only counting species from Austrian plots when building the accumulation curve. This ensures smooth nearest-neighbor paths that cross borders naturally.

Why this matters: Edge effects can systematically bias biodiversity estimates. Plots near borders appear artificially species-rich or species-poor depending on whether cross-border diversity is captured or excluded (Colwell et al. 2004). For Austria (a country with 2,562 km of borders and complex geometry), this is particularly relevant.



Interpretation:

Comparing the "core only" and "with halo" curves reveals the magnitude of edge effects in Austrian vegetation data:

- **If the curves diverge substantially:** Edge effects are significant. The core-only analysis underestimates the smoothness of the spatial walk, likely inflating apparent species turnover near borders.
- **If the curves are similar:** Edge effects are minor for the Austrian dataset, possibly because the country is large enough that most plots are far from borders.

This analysis validates (or challenges) the assumption that national-scale biodiversity analyses can treat country borders as biological boundaries. For smaller countries or elongated territories, the halo correction would be even more critical.

The `areaOfEffect` package used here implements the methodology described by Colling et al. for handling administrative boundaries in spatial ecological analyses. This approach is particularly important when comparing biodiversity metrics across countries of different sizes and shapes.

13. Per-Site Metrics & Spatial Map

From curves to maps:

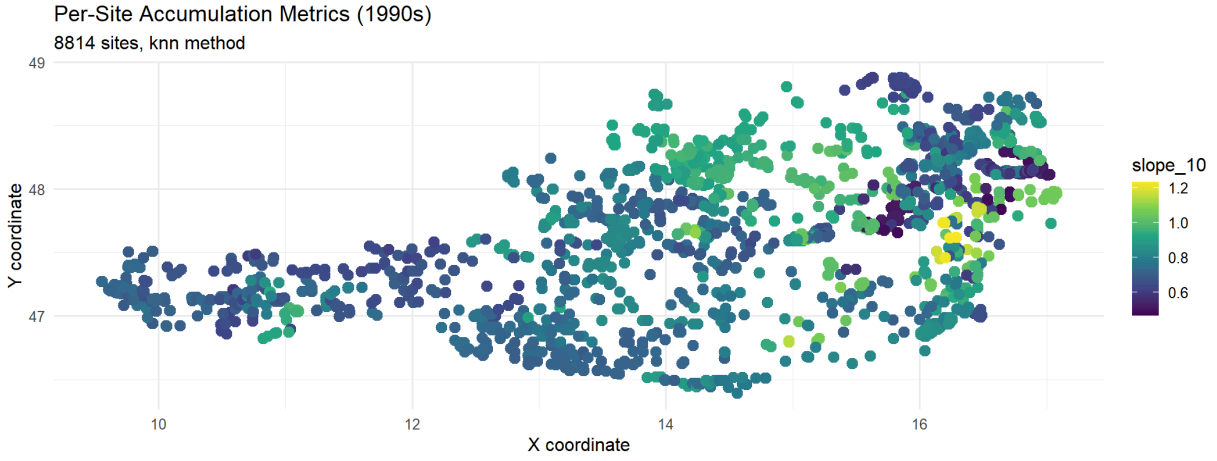
While accumulation curves summarize patterns across the landscape, we can also compute metrics for **each individual site** based on its role in the accumulation process:

- **Slope at 10 sites** (`slope_10`): How rapidly is species richness increasing when 10 sites have been sampled starting from this point? High slope = the neighborhood is species-rich
- **Half-richness** (`half_richness`): How many sites must be sampled (from this starting point) to reach 50% of total species? Low values = rapid accumulation = clustered diversity
- **AUC** (Area Under Curve): A summary measure integrating the entire accumulation trajectory from a site

Mapping these metrics reveals **spatial hotspots of biodiversity accumulation**, areas where species turnover is highest and sampling efficiency is greatest (Colling et al. 2024).

```
## spacc_metrics: 8814 sites, 2700 species
## Method: knn
## Metrics: slope_10, half_richness, auc
```

```
## Metric summary:
##   slope_10: mean=0.82, sd=0.17, range=[0.47, 1.23]
##   half_richness: mean=2050.14, sd=486.35, range=[1204.00, 3454.00]
##   auc: mean=1585.70, sd=38.80, range=[1503.78, 1658.52]
```

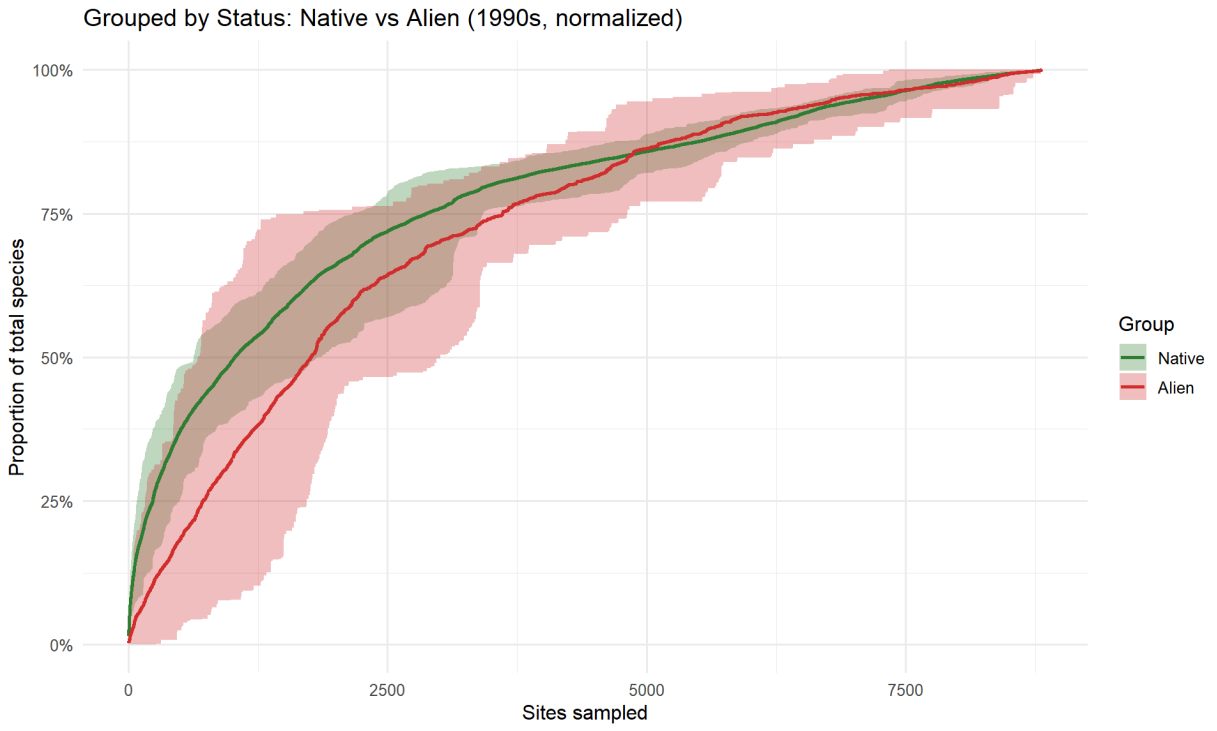


Interpretation:

The spatial maps of per-site metrics provide actionable insights for conservation planning:

- **High-slope areas** represent biodiversity hotspots where a small number of additional samples rapidly captures new species. These areas are priorities for intensive biodiversity surveys.
- **Low half-richness areas** are regions where diversity is concentrated locally; sampling a few nearby plots captures a large fraction of regional diversity. This pattern is typical of environmentally heterogeneous areas (alpine valleys, ecotones between habitat types).
- **The spatial pattern** of these metrics likely correlates with elevation, land use, and habitat heterogeneity, yielding testable predictions that connect accumulation curve analysis to landscape ecology.

14. Grouped Accumulation by Status



This is the same spatial walk applied simultaneously to both groups using the `groups` argument, ensuring that native and alien species are accumulated along **identical spatial paths**. This eliminates any confounding effect of different walk trajectories and provides the cleanest comparison of spatial distribution patterns between the two groups.

The normalized curves confirm the pattern seen in the pooled analysis: alien species accumulate proportionally more slowly than natives, indicating higher spatial turnover in the alien flora. Different alien species occur in different regions, so more of the landscape must be traversed to capture the full alien complement.

15. Synthesis & Conclusions

Revisiting Our Hypotheses

H1: Aliens have higher spatial beta diversity	Normalized alien curve lies below native cu
H2: Difference reflects habitat-specific introduction pathways	Spatial heterogeneity in alien composition
H3: Curves are statistically different	Significant permutation test

Additional discoveries from extended analyses:

- **Temporal signal:** The native-alien difference is consistent across decades
- **Habitat specificity:** Alien invasion is concentrated in disturbed, nutrient-rich habitats
- **Phylogenetic structure:** Aliens may represent a phylogenetically non-random subset of global flora
- **Spatial autocorrelation:** kNN vs random comparison confirms strong spatial structure in Austrian plant communities
- **Coverage:** 90% sampling completeness requires substantially fewer plots than 99%
- **Edge effects:** The halo analysis quantifies how national borders bias spatial biodiversity estimates

The Big Picture

Our analyses paint a coherent picture of Austrian plant diversity that connects multiple theoretical frameworks:

1. Invasion ecology: Alien species are not randomly distributed; their richness concentrates in human-modified landscapes, but critically, *different* alien species occur in *different* regions (Pysek et al. 2010). The spatial accumulation approach reveals this high spatial turnover because it preserves geographic information that traditional species counts discard.

2. Biogeography: Austria's extraordinary topographic gradient (115 m to 3,798 m elevation) creates strong environmental filtering that drives high beta diversity through genuine species turnover. Alpine communities share few species with lowland communities, and this pattern dominates the accumulation curve shape.

3. Conservation planning: The high beta diversity means that a single large protected area cannot capture Austrian plant diversity. Instead, a distributed network of reserves spanning different elevations, habitats, and biogeographic regions is needed, supporting the **SLOSS debate** (Several Large or Single Small) in favor of "several" for high-turnover systems (Tjorve 2010).

4. Sampling theory: Coverage-based rarefaction shows that 52,000 plots likely capture >95% of Austria's common flora, but rare species (including potentially undiscovered aliens) remain undetected. Whether these undetected aliens represent under-sampling, recent arrivals, or

"invasion debt" (Essl et al. 2011) cannot be resolved from spatial data alone; temporal replication would be needed.

Discussion Questions

1. **Why do different alien species establish in different regions?** The accumulation curves show high spatial turnover in the alien flora. Consider how propagule pressure, local disturbance regimes, trade routes, and climate variation across Austria could produce regionally distinct alien floras.
2. **Could climate change shift these patterns?** As temperatures warm, alpine specialists may lose habitat while thermophilic aliens gain it. How would this change accumulation curves?
3. **Is functional homogenization a concern for Austria?** If aliens add species but not functional diversity, what does this mean for ecosystem services like pollination, soil stability, and carbon storage?
4. **How would you design a monitoring program** based on the coverage analysis? How many plots, where, and how often?
5. **What additional data would strengthen these conclusions?** Think about trait databases, temporal replication, genetic diversity, and trophic interactions.

Key Takeaways

Spatial accumulation	Geography shapes species discovery	Fair comparison requires spatial methods
Native vs. alien	Different distribution patterns confirmed	Invasion management needs spatial targeting
Hill numbers	Richness ≠ diversity ≠ dominance	Different metrics answer different questions
Beta diversity	High turnover across Austria	Distributed conservation network needed
Phylogenetic diversity	Evolutionary history adds information	Not all species are equally "different"
Coverage	~95% coverage achievable	Sampling effort can be optimized
Habitat specificity	Invasions are habitat-filtered	Prioritize high-risk habitats for management
Edge effects	Borders bias estimates	Halo correction improves accuracy

References

- Baker, H.G. (1974). The evolution of weeds. *Annual Review of Ecology and Systematics*, 5, 1–24.
- Baselga, A. (2010). Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, 19, 134–143.
- Chao, A. & Jost, L. (2012). Coverage-based rarefaction and extrapolation. *Methods in Ecology and Evolution*, 3, 217–226.

- Chao, A. et al. (2014). Rarefaction and extrapolation with Hill numbers. *Ecological Monographs*, 84, 45–67.
- Chytry, M. et al. (2008). Habitat invasions by alien plants: a quantitative comparison among Mediterranean, subcontinental and oceanic regions of Europe. *Journal of Applied Ecology*, 45, 448–458.
- Chytry, M. et al. (2009). European map of alien plant invasions based on the quantitative assessment across habitats. *Diversity and Distributions*, 15, 98–107.
- Chytry, M. et al. (2016). European Vegetation Archive (EVA): an integrated database of European vegetation plots. *Applied Vegetation Science*, 19, 173–180.
- Colwell, R.K. & Coddington, J.A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society B*, 345, 101–118.
- Colwell, R.K. et al. (2004). Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, 85, 2717–2727.
- Davis, M.A. et al. (2000). Fluctuating resources in plant communities: a general theory of invasibility. *Journal of Ecology*, 88, 528–534.
- Dengler, J. et al. (2011). The Global Index of Vegetation-Plot Databases (GIVD). *Biodiversity & Ecology*, 3, 1–26.
- Essl, F. et al. (2011). Socioeconomic legacy yields an invasion debt. *Proceedings of the National Academy of Sciences*, 108, 203–207.
- Faith, D.P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61, 1–10.
- Gotelli, N.J. & Colwell, R.K. (2001). Quantifying biodiversity: procedures and pitfalls. *Ecology Letters*, 4, 379–391.
- Grabherr, G. et al. (2003). *Die Pflanzengesellschaften Osterreichs*. Gustav Fischer Verlag.
- Hill, M.O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54, 427–432.
- Hillebrand, H. et al. (2008). Consequences of dominance: a review of evenness effects on local and regional ecosystem processes. *Ecology*, 89, 1510–1520.
- Jost, L. (2006). Entropy and diversity. *Oikos*, 113, 363–375.
- Kier, G. et al. (2009). A global assessment of endemism and species richness across island and mainland regions. *Proceedings of the National Academy of Sciences*, 106, 9322–9327.
- Laliberte, E. & Legendre, P. (2010). A distance-based framework for measuring functional diversity from multiple traits. *Ecology*, 91, 299–305.
- Lenoir, J. et al. (2008). A significant upward shift in plant species optimum elevation during the 20th century. *Science*, 320, 1768–1771.
- Lockwood, J.L. et al. (2005). The role of propagule pressure in explaining species invasions. *Trends in Ecology & Evolution*, 20, 223–228.
- Lomolino, M.V. (2000). Ecology's most general, yet protean pattern: the species–area relationship. *Journal of Biogeography*, 27, 17–26.
- Nekola, J.C. & White, P.S. (1999). The distance decay of similarity in biogeography and ecology. *Journal of Biogeography*, 26, 867–878.
- Olden, J.D. et al. (2004). Ecological and evolutionary consequences of biotic homogenization. *Trends in Ecology & Evolution*, 19, 18–24.
- Partel, M. et al. (2011). Dark diversity: shedding light on absent species. *Trends in Ecology & Evolution*, 26, 596–600.
- Petchey, O.L. & Gaston, K.J. (2006). Functional diversity: back to basics and looking forward. *Ecology Letters*, 9, 741–758.
- Pillar, V.D. et al. (2013). Functional redundancy and stability in plant communities. *Journal of Vegetation Science*, 24, 963–974.
- Preston, F.W. (1948). The commonness, and rarity, of species. *Ecology*, 29, 254–283.

- Pysek, P. et al. (2010). Disentangling the role of environmental and human pressures on biological invasions across Europe. *Proceedings of the National Academy of Sciences*, 107, 12157–12162.
- Rahbek, C. (1995). The elevational gradient of species richness: a uniform pattern? *Ecography*, 18, 200–205.
- Pysek, P. et al. (2012). A global assessment of invasive plant impacts on resident species, communities and ecosystems. *Global Change Biology*, 18, 1725–1737.
- Pysek, P. et al. (2017). Naturalized alien flora of the world. *Preslia*, 89, 203–274.
- Sala, O.E. et al. (2000). Global biodiversity scenarios for the year 2100. *Science*, 287, 1770–1774.
- Seebens, H. et al. (2017). No saturation in the accumulation of alien species worldwide. *Nature Communications*, 8, 14435.
- Socolar, J.B. et al. (2016). How should beta-diversity inform biodiversity conservation? *Trends in Ecology & Evolution*, 31, 67–80.
- Tilman, D. et al. (2001). Diversity and productivity in a long-term grassland experiment. *Science*, 294, 843–845.
- Tjorve, E. (2010). How to resolve the SLOSS debate: lessons from species-diversity models. *Journal of Theoretical Biology*, 264, 604–612.
- van Kleunen, M. et al. (2010). Are invaders different? A conceptual framework of comparative approaches for assessing determinants of invasiveness. *Ecology Letters*, 13, 947–958.
- van Kleunen, M. et al. (2015). Global exchange and accumulation of non-native plants. *Nature*, 525, 100–103.
- Webb, C.O. et al. (2002). Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, 33, 475–505.
- Whittaker, R.H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, 30, 279–338.
- Whittaker, R.H. (1972). Evolution and measurement of species diversity. *Taxon*, 21, 213–251.
- Wilson, J.R.U. et al. (2007). Residence time and potential range: crucial considerations in modelling plant invasions. *Diversity and Distributions*, 13, 11–22.
- Winter, M. et al. (2009). Plant extinctions and introductions lead to phylogenetic and taxonomic homogenization of the European flora. *Proceedings of the National Academy of Sciences*, 106, 21721–21725.

After the Workshop

Summary

Workshop Review

Day 3: Key Findings from the Austrian Vegetation Data

What the spacc analysis tells us

Liam's Stay - Results

2026-02-16

Overview

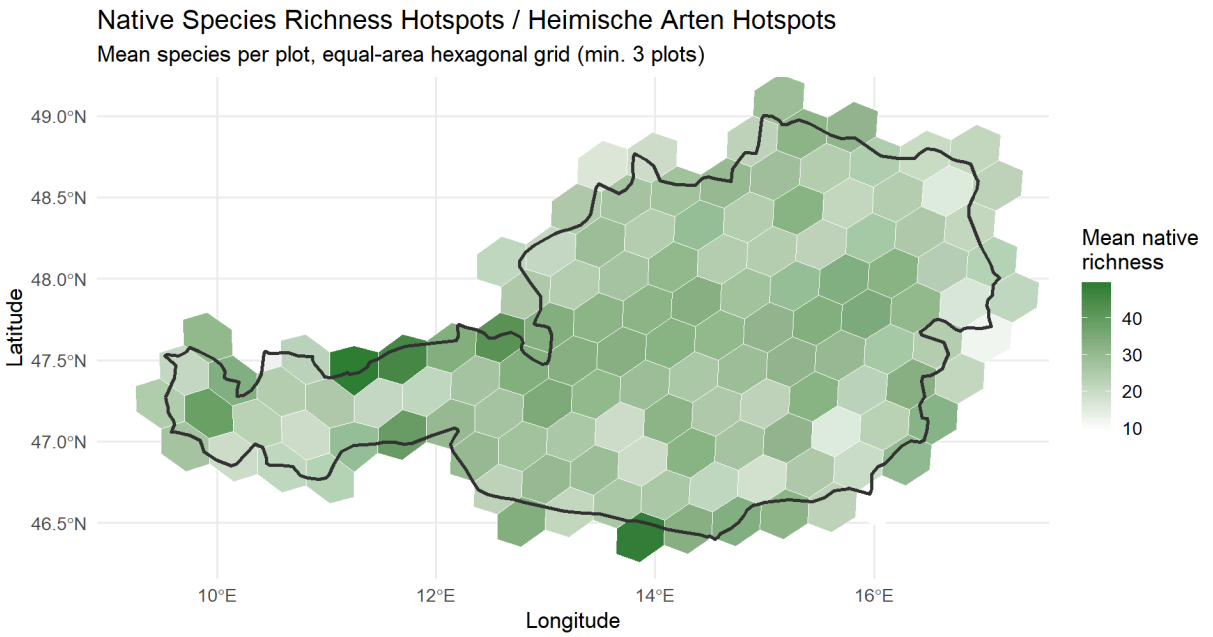
This document presents **five key findings** from our analysis of ~29,857 Austrian vegetation plots (EVA database). Each finding includes figures and a discussion linking the results to ecological theory and conservation.

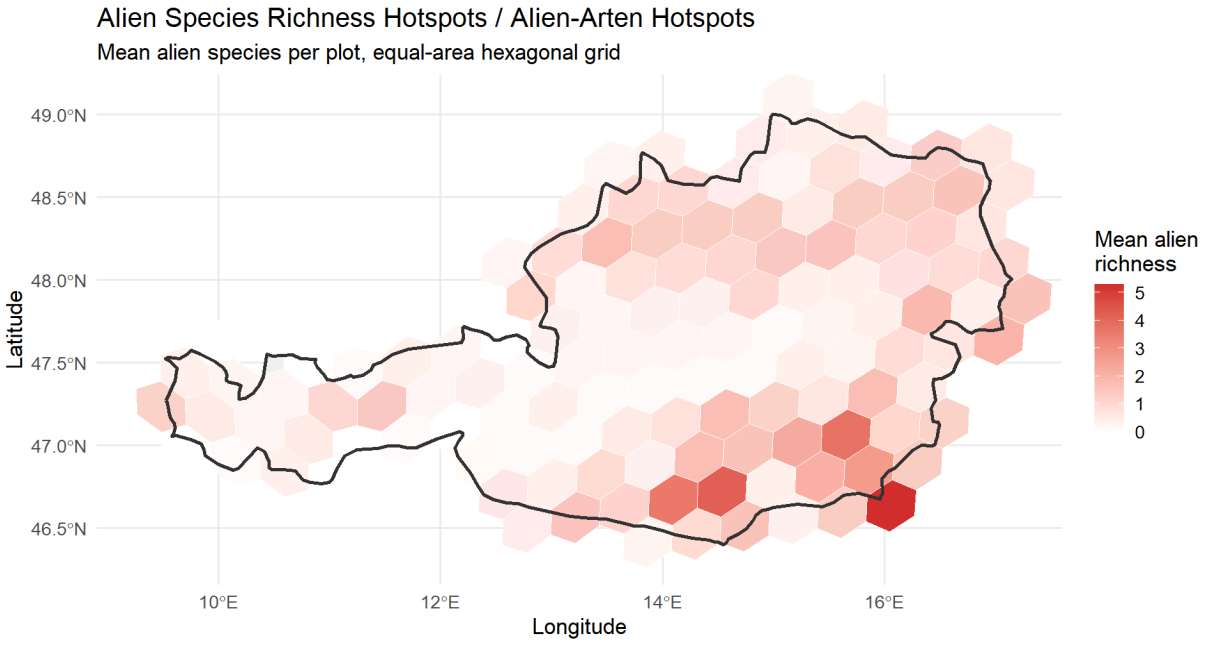
These five findings were chosen because they each address a different aspect of Austrian plant diversity:

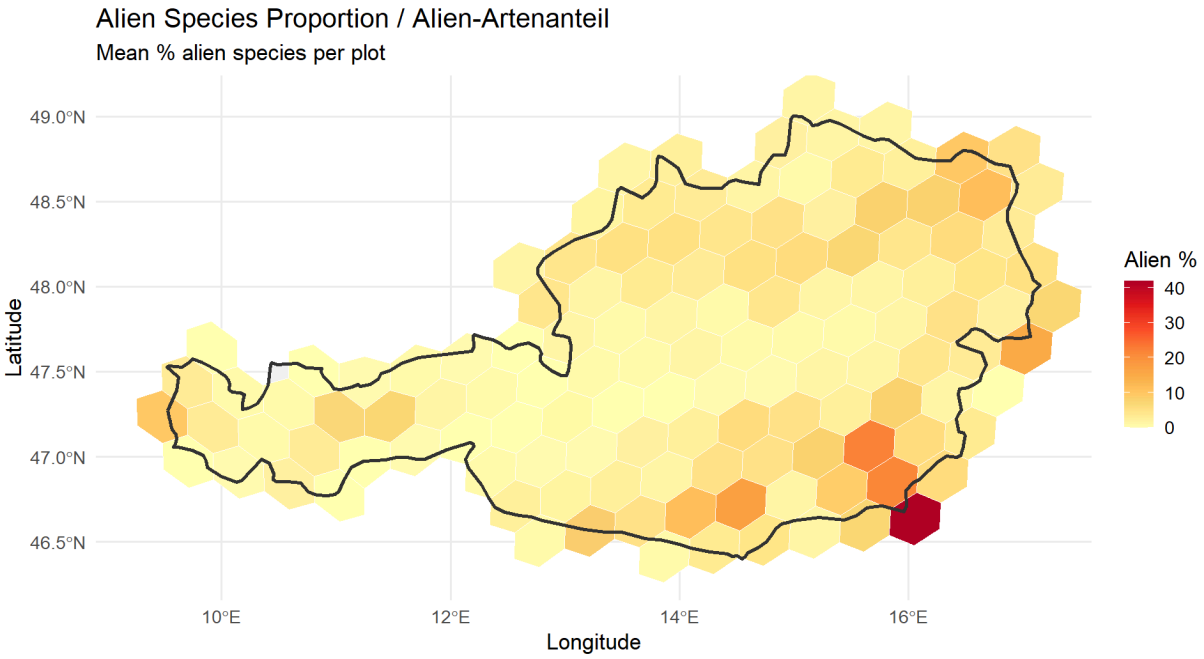
1	**Sampling & richness overview:** Hotspot maps	Where was data collected and what does raw richness l
2	**How species accumulate:** Native vs. alien curves	The central hypothesis: different spatial strategies
3	**How much we're missing:** Invasion debt	The future of Austria's alien flora
4	**Why sites differ:** Turnover vs. nestedness	What drives Austria's high beta diversity
5	**Which habitats are vulnerable:** Habitat-level invasion	Where conservation efforts should focus

1. Sampling & Richness Overview: Hotspot Maps

The Maps


Native Species Richness Hotspots / Heimische Arten Hotspots
Mean species per plot, equal-area hexagonal grid (min. 3 plots)
49.0°N
48.5°N
48.0°N
47.5°N
47.0°N
46.5°N
Latitude
10°E
12°E
14°E
16°E
Longitude
Mean native richness
40
30
20
10


Alien Species Richness Hotspots / Alien-Arten Hotspots
Mean alien species per plot, equal-area hexagonal grid
49.0°N
48.5°N
48.0°N
47.5°N
47.0°N
46.5°N
Latitude
10°E
12°E
14°E
16°E
Longitude
Mean alien richness
5
4
3
2
1
0



Discussion

What the maps show:

Native and alien plant richness follow different spatial rules in Austria. Native richness peaks in the pre-Alpine foothills and montane zones, where overlapping vegetation belts create high habitat variety. Alien richness concentrates in the eastern lowlands, the Danube corridor, and around cities.

Why they peak in different places:

Native richness reflects topographic complexity. Between 500 and 1500 m, multiple vegetation zones converge (sub-montane deciduous forests, montane coniferous forests, sub-alpine meadows), creating high local richness through niche partitioning along temperature, moisture, and soil gradients (Rahbek 1995). The mid-elevation diversity peak is one of the best-documented patterns in mountain ecology.

Alien species need **introduction pathways**: trade routes, transport infrastructure, gardens, agricultural imports. These converge on cities and lowland corridors. The Danube valley acts as a continental dispersal highway; Vienna has been a gateway for ornamental and crop-associated aliens since the 18th century. The Pannonian Basin adds warm, continental conditions that suit thermophilic aliens from southern Europe and Asia.

The proportion map matters most:

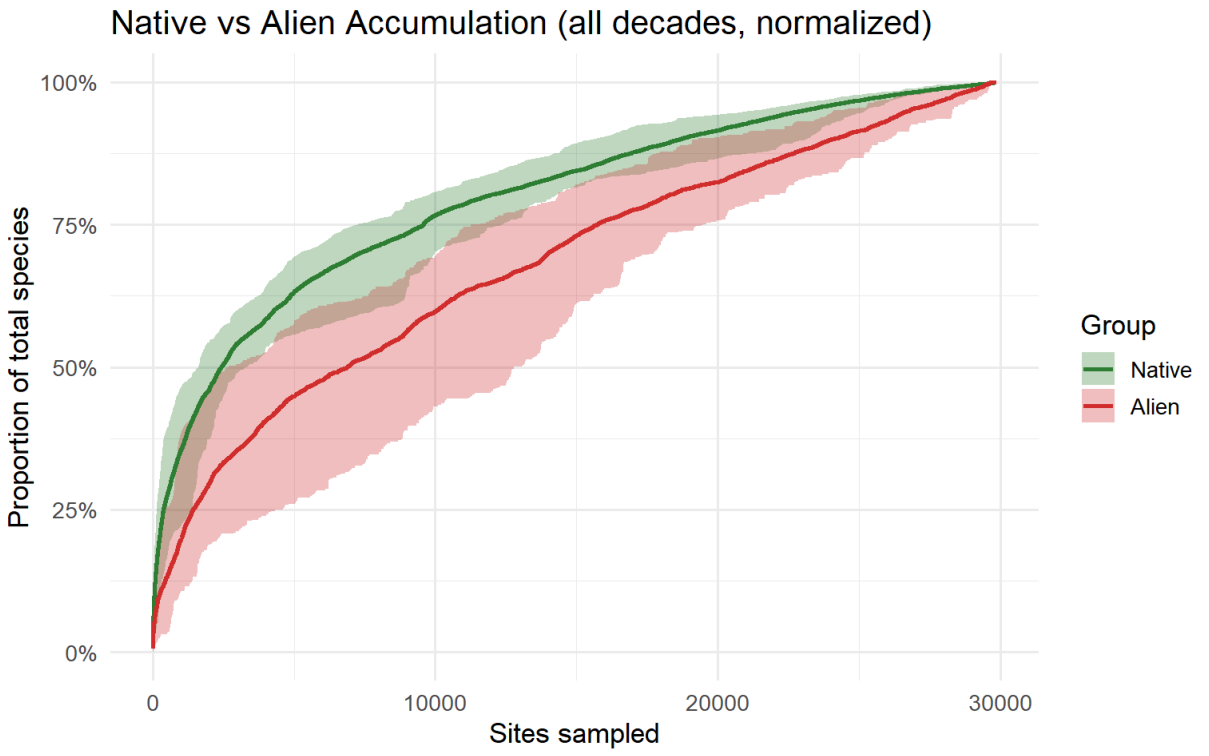
Absolute richness maps can be confounded by sampling intensity. The alien *proportion* map corrects for this. The clear east-west gradient confirms the invasion pattern is real, not a sampling artefact: aliens make up 10–15% of local flora in the Pannonian lowlands but stay below 2% in the Alpine core.

Links to invasion theory:

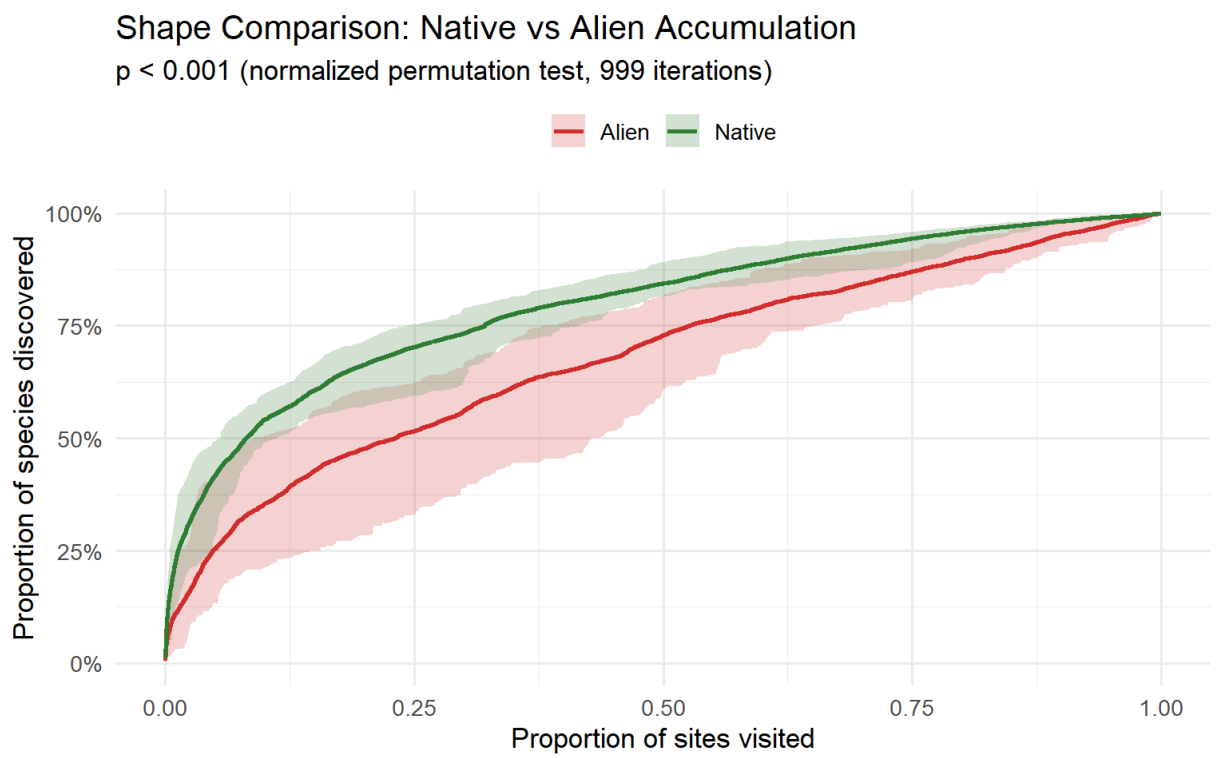
The spatial concentration around human infrastructure fits the **propagule pressure** hypothesis (Lockwood et al. 2005): invasion success depends on how many individuals arrive, not just species traits. It also fits the **fluctuating resource** hypothesis (Davis et al. 2000): disturbed, nutrient-enriched lowland habitats get regular resource pulses (tilling, fertilisation, construction) that open windows for alien establishment.

2. The Central Finding: Aliens Accumulate Differently

Native vs. Alien Accumulation Curves

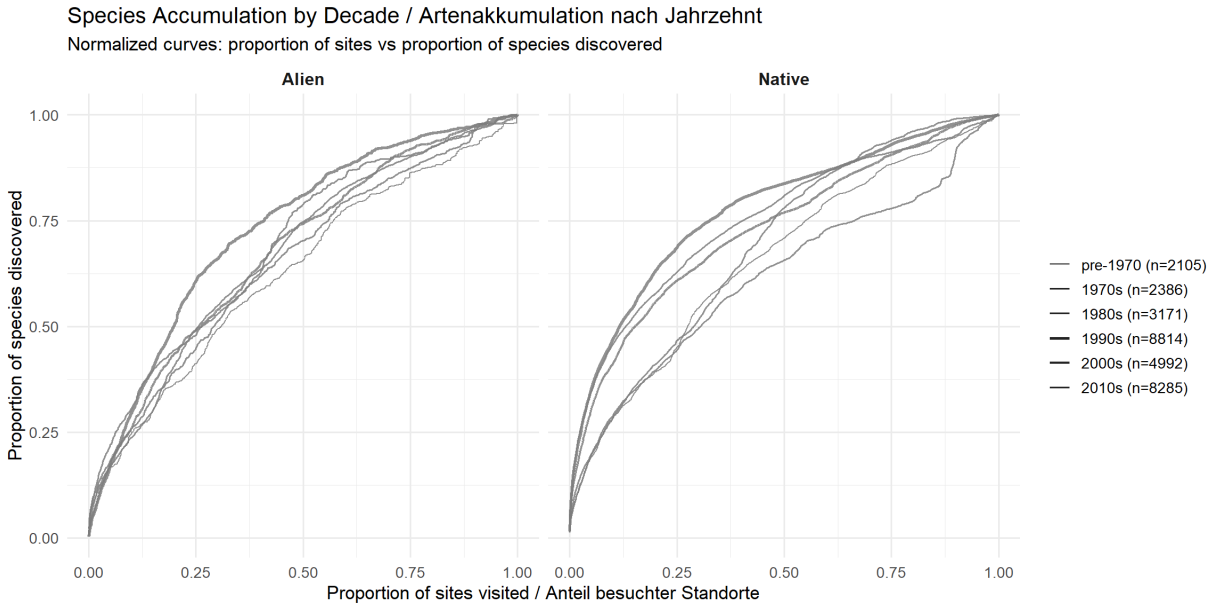



```
## Comparison: p$native vs p$alien
## ---------------------------------------
## Method: permutation (n=999)
## Normalized: yes (shape comparison)
## AUC difference: 3433.3 (p < 0.001***)
## Saturation: p$native at 18661 sites, p$alien at 24060 sites
##
## p$native saturates faster.
```



Decade-by-Decade Accumulation

Does the spatial accumulation pattern change over time? Here we run the same analysis separately for each decade, splitting native and alien species into separate panels. Each curve is normalized by its total species count.



Discussion

> **The core result:**
>
> After normalizing both curves to their respective totals, the alien curve sits consistently below the native curve. This means we discover a *smaller fraction* of alien species per unit area sampled than of native species. The permutation test confirms this ($p < 0.05$), and the

grouped analysis (both groups accumulated along identical spatial paths) rules out confounding from different walk trajectories.

What a lower normalized curve means:

A lower curve = higher **spatial beta diversity**: species composition changes more between distant sites. Many native generalists (common grasses, widespread forest herbs) show up early in the walk and are shared across most of the country. After a few hundred plots from different regions, you've already seen 50–60% of natives. For aliens, each region has a partly different set, reflecting distinct introduction histories. A spatial walk through the Vienna Basin encounters different aliens than one through the Inn Valley.

The multi-focus introduction model:

Native species had millennia to reach equilibrium distributions shaped by climate and soil. Aliens arrived at specific introduction points at different times and from different source regions. Each point acts as a separate focus of local spread:

- **Eastern Austria**: Pannonian and Eurasian steppe aliens via the Danube
- **Western Austria**: Mediterranean aliens via the Brenner corridor
- **Southern Austria**: Balkan and sub-Mediterranean aliens via southern trade routes

Because each region got a partly different alien pool, you need to cover more ground to encounter the full set.

Stability over time:

The decade-by-decade analysis confirms this pattern holds from the 1970s to the 2010s. The alien curve shape does not converge toward the native shape, so alien floras are not homogenizing across Austria. At least not yet.

For monitoring:

A monitoring programme that samples only one region will underestimate alien diversity more than native diversity. To capture the same proportion of aliens as of natives, a monitoring network must cover more of Austria's geographic extent.

3. The Invasion Debt: How Many Aliens Are We Missing?

Michaelis-Menten Extrapolation

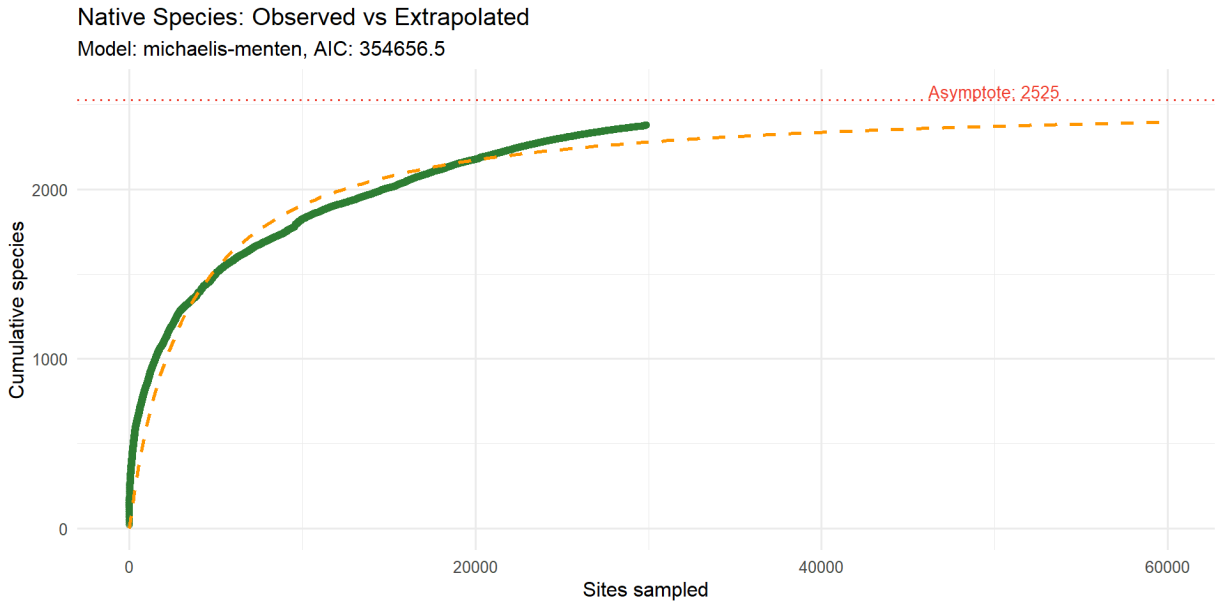
The **Michaelis-Menten model** fits a saturating curve to the species accumulation data:

$$S(n) = \frac{S_{max} \cdot n}{K_m + n}$$

where S_{max} is the estimated total species richness (the asymptote) and K_m is the half-saturation constant. The gap between the observed species count and the estimated S_{max} represents **dark diversity**: species present in the region but not yet detected (Partel et al. 2011).

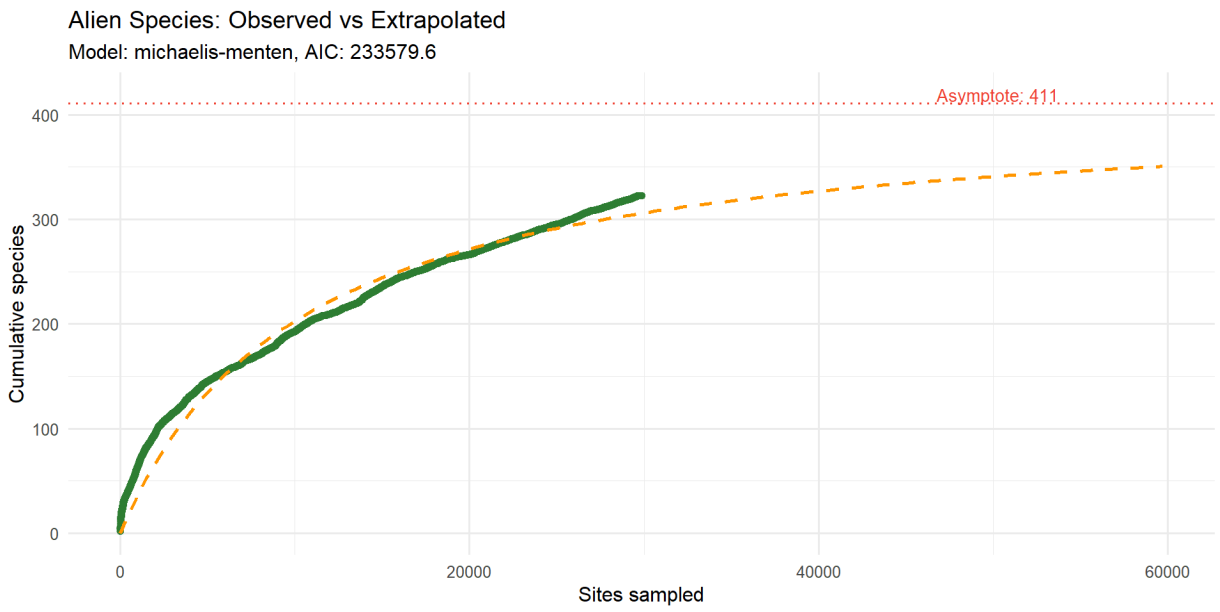
```
## Native species extrapolation:
```

```
## Extrapolation: michaelis-menten
## ------------------------------
## Estimated asymptote: 2525.4 species
## 95% CI: 2522.4 - 2528.5
## AIC: 354656.5
## Observed: 2378.0 species (94% of estimated)
```




```
##
## Alien species extrapolation:
```

```
## Extrapolation: michaelis-menten
## ------------------------------
## Estimated asymptote: 410.9 species
## 95% CI: 410.0 - 411.8
## AIC: 233579.6
## Observed: 323.0 species (79% of estimated)
```



Discussion

The numbers:

- **Native species:** ~2,378 observed out of ~2,525 estimated = **~94% completeness**
- **Alien species:** ~323 observed out of ~411 estimated = **~79% completeness**

The 15-point gap is one of the clearest results from the analysis. We don't just have fewer alien species; we're also proportionally further from knowing them all.

Why is the alien flora less complete?

Three mechanisms explain this:

1. **Ongoing introductions.** New aliens keep arriving through trade, horticulture, and climate-driven range shifts. Global alien plant accumulation shows no sign of saturating (Seebens et al. 2017). Some of the "missing" ~88 species may not have arrived yet.
2. **Dispersal lag.** Aliens already in Austria haven't reached all suitable habitats. A species introduced to a Viennese garden in the 1990s may take decades to colonize the Danube floodplains or the ruderal sites of Carinthia. With 52,000 plots spread unevenly, many early-stage colonizers are missed.
3. **Detection difficulty.** Some aliens are taxonomically tricky or occupy under-sampled habitats (urban wastelands, private gardens, industrial sites).

Invasion debt:

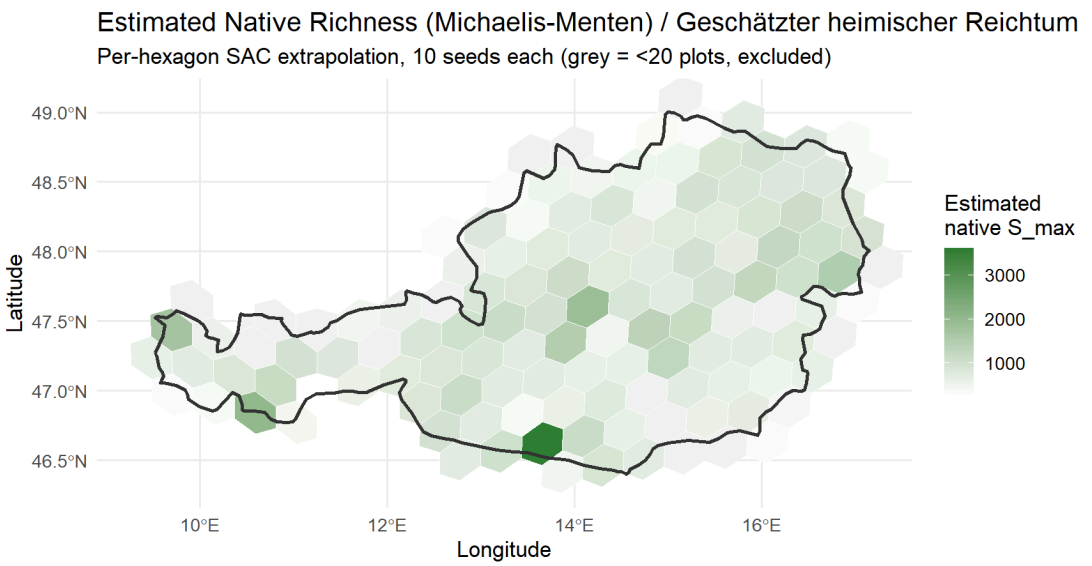
Essl et al. (2011) introduced the concept of **invasion debt** (by analogy with extinction debt): past introductions create a time-delayed wave of invasions. Our ~88 "missing" alien species represent the upper bound of this debt. Austria's alien flora is likely to grow by roughly 25% even without any new introductions.

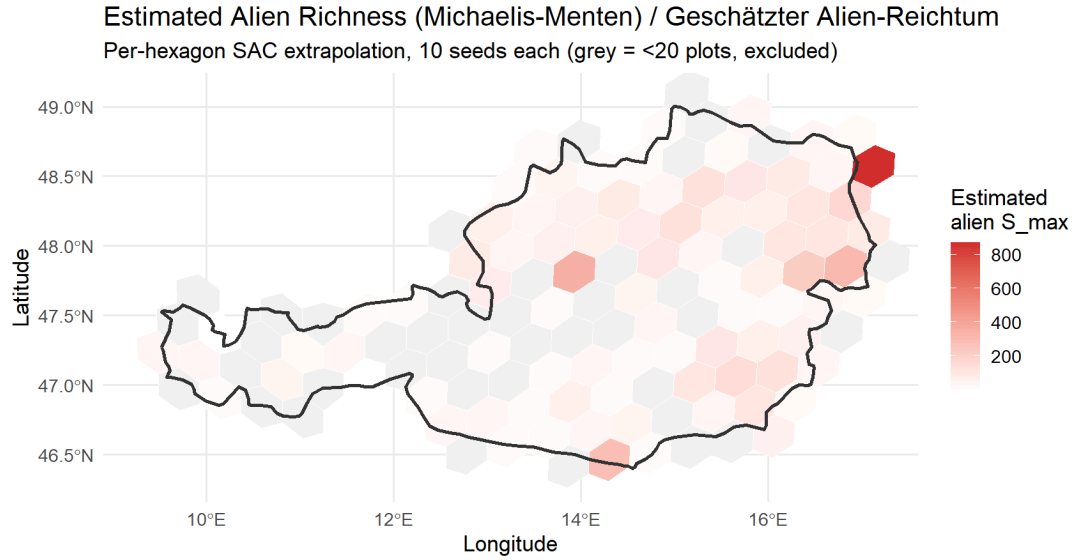
For conservation:

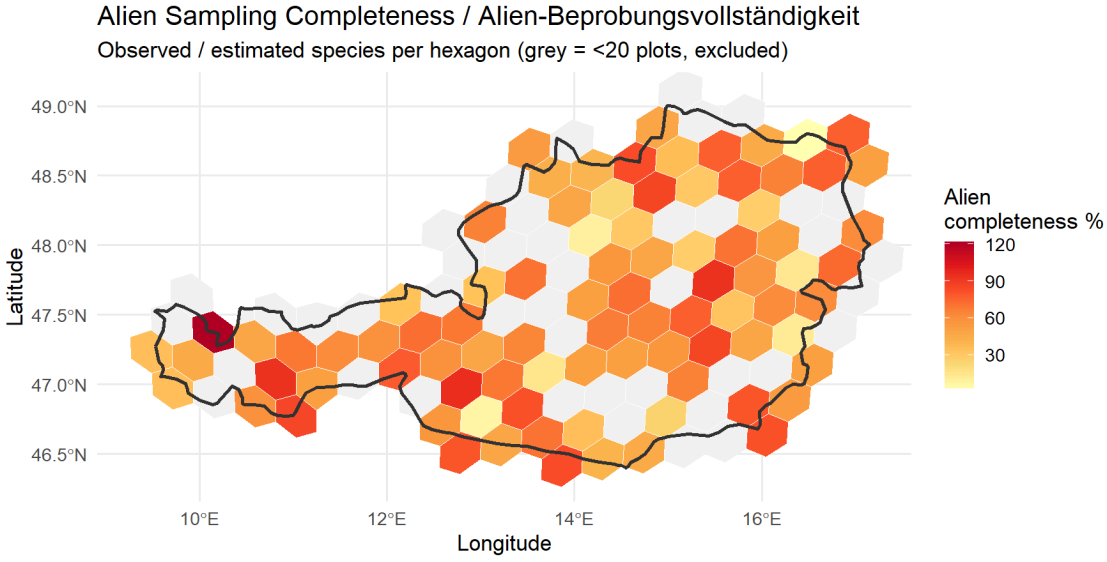
> Species lists are not static. Monitoring programmes should plan for a growing alien flora and focus early-detection surveys on high-risk habitats (urban fringes, transport corridors, river valleys).
>
> **Model caveat:**
>
> The Michaelis-Menten estimator assumes a specific functional form. If the accumulation curve has a long tail of very rare species, the asymptote may be underestimated. The 94% and 79% figures are reasonable approximations, not exact values.

3b. Mapping the Invasion Debt: Estimated Richness per Hexagon

The Michaelis-Menten extrapolation above estimated **total** richness for all of Austria. Here we apply the same approach per hexagon: for each cell in the equal-area grid, we run a spatial accumulation with 10 seeds on the within-hex plots, fit Michaelis-Menten, and extract the estimated total species richness (S_{max}). This gives a map of *estimated* richness rather than raw observed counts. Hexagons with fewer than 20 plots (shown in light grey) are excluded because accumulation curves fitted to so few samples produce unreliable extrapolations.





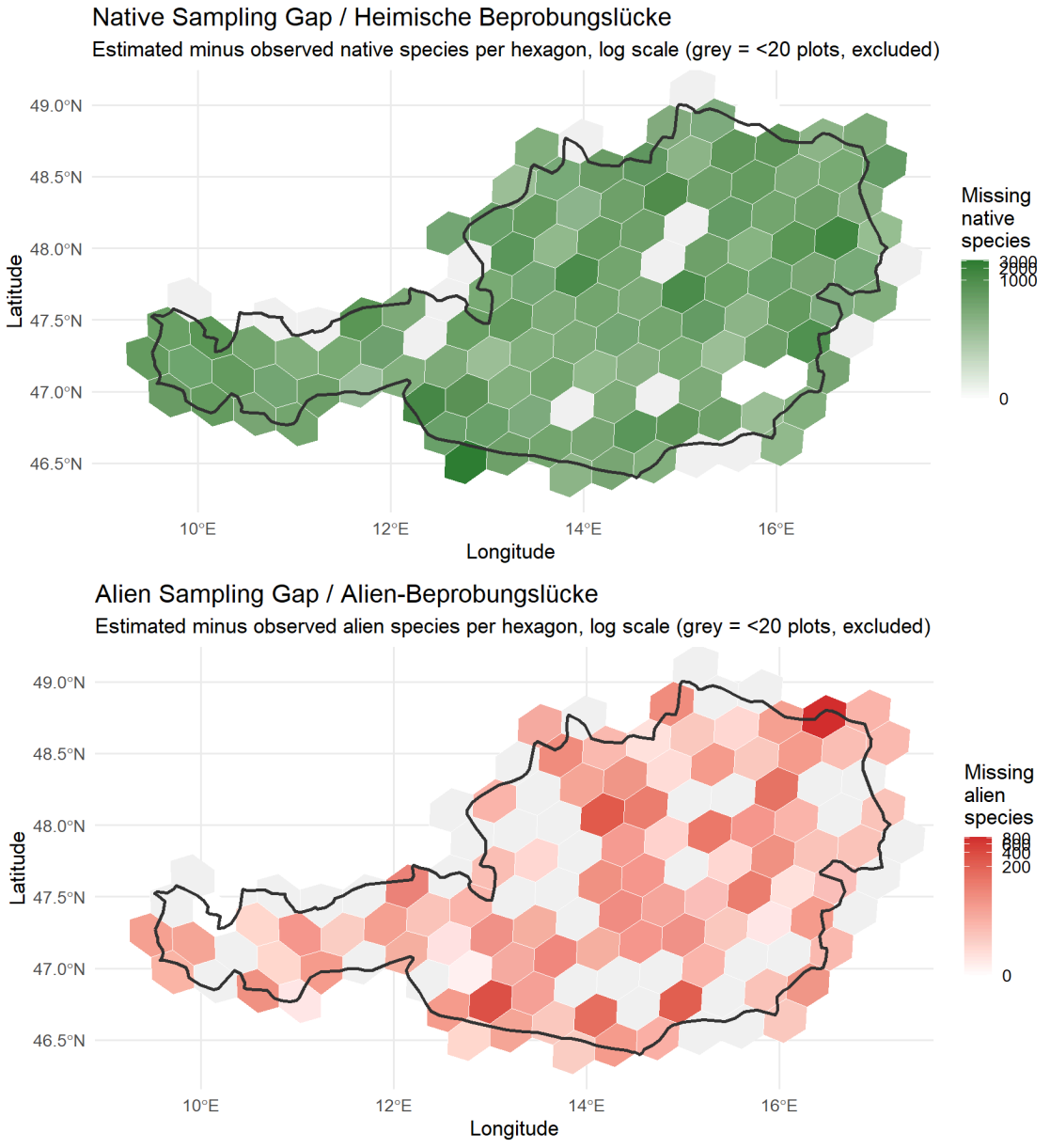


These maps translate the invasion debt into geography. While the first map (estimated native richness) echoes the observed raw-count maps from section 1, the alien map adds information: it shows where the Michaelis-Menten model predicts more species than we

have actually found. The completeness map (bottom) highlights where the alien flora is least well-sampled. Regions with low completeness are where we should expect to find new alien records in coming years.

3c. Sampling Gaps: How Many Species Are We Missing?

How many species are still missing in each hexagon? These maps show the **sampling gap**: the difference between the Michaelis-Menten estimate (S_{max}) and the observed species count. Darker hexagons have more unrecorded species.



The native gap map shows moderate shortfalls across much of Austria, with the largest absolute gaps in species-rich Alpine valleys and foothills. The alien gap map tells a different story: the biggest gaps concentrate in lowland corridors (Danube valley, Vienna basin,

eastern border), precisely the regions where new introductions arrive and dispersal lags are longest. Even hexagons with relatively few observed aliens can show large gaps, indicating that the invasion debt is geographically concentrated rather than uniformly distributed.

4. Turnover, Not Nestedness: What Drives Austrian Beta Diversity

Decomposing Beta Diversity

Baselga (2010) showed that total beta diversity can be decomposed into two components:

$$\beta_{Sorensen} = \beta_{Turnover} + \beta_{Nestedness}$$

- **Turnover (β_{sim}):** Species are replaced by different species between sites
- **Nestedness (β_{nes}):** Species-poor sites are subsets of species-rich sites

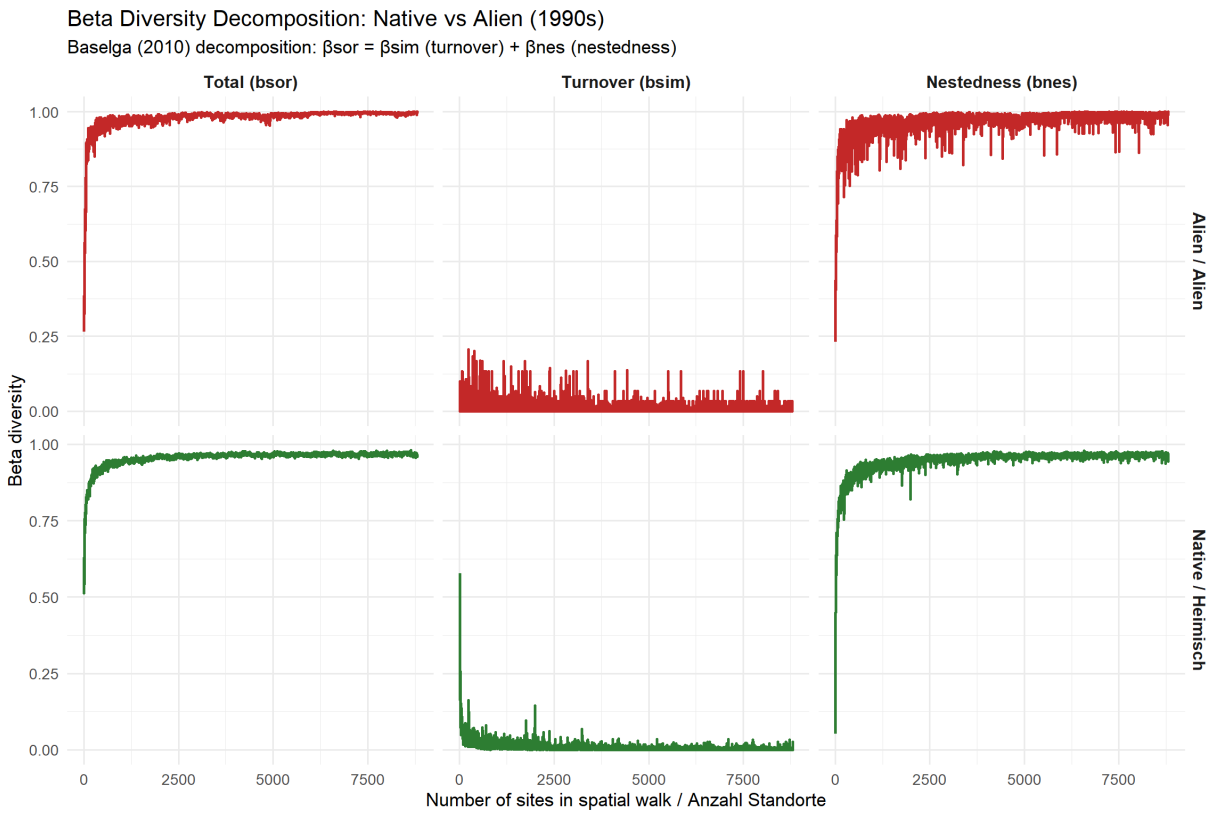
The balance between these components reveals the *process* generating diversity differences: environmental filtering (turnover) versus differential species loss (nestedness).

How to read the plot below:

The x-axis shows the number of sites visited in a "spatial walk" across Austria – starting at one random site and repeatedly adding the nearest unvisited site. The y-axis shows beta diversity (0–1 scale). There are three panels, one for each component:

- **Total (β_{sor}):** Overall compositional difference among all visited sites so far. Rises quickly and plateaus near 1, meaning Austrian sites become very different once the walk spans enough environmental gradients.
- **Turnover (β_{sim}):** The portion of that difference caused by species *replacement* (site A has species X, site B has species Y instead). If this panel is high, sites have genuinely different species.
- **Nestedness (β_{nes}):** The portion caused by species *loss* without replacement (species-poor sites are just subsets of richer sites). If this panel is low, communities are not simply nested versions of each other.

Key comparison: Since Total = Turnover + Nestedness, looking at which component is larger tells you *why* sites differ.



Discussion

Turnover dominates for natives – but what about aliens?

The beta diversity decomposition reveals distinct patterns for **native** and **alien** species. Both groups are turnover-dominated, but with important differences.

Natives: Turnover is high and nestedness is low. This reflects Austria's 3,683 m elevation gradient (115 m Neusiedler See to 3,798 m Grossglockner), which creates distinct vegetation belts hosting genuinely **different species**, not nested subsets. A lowland *Quercus-Carpinus* forest and a high-alpine *Carex curvula* grassland share almost no species.

Aliens: Turnover also dominates, but the overall beta diversity is lower than for natives, and nestedness may be relatively more important. Alien species are drawn from a smaller pool concentrated in lowland and disturbed habitats. Some alien-poor sites (e.g., alpine areas) are effectively subsets of the richer alien floras in lowland cities – a more nested pattern than natives show.

Key contrast: Natives show high turnover because different elevations have different species (environmental filtering). Aliens show turnover because different lowland regions receive different introduced species (introduction history), but also some nestedness because alien-poor sites simply lack the conditions for establishment rather than hosting alternative alien species.

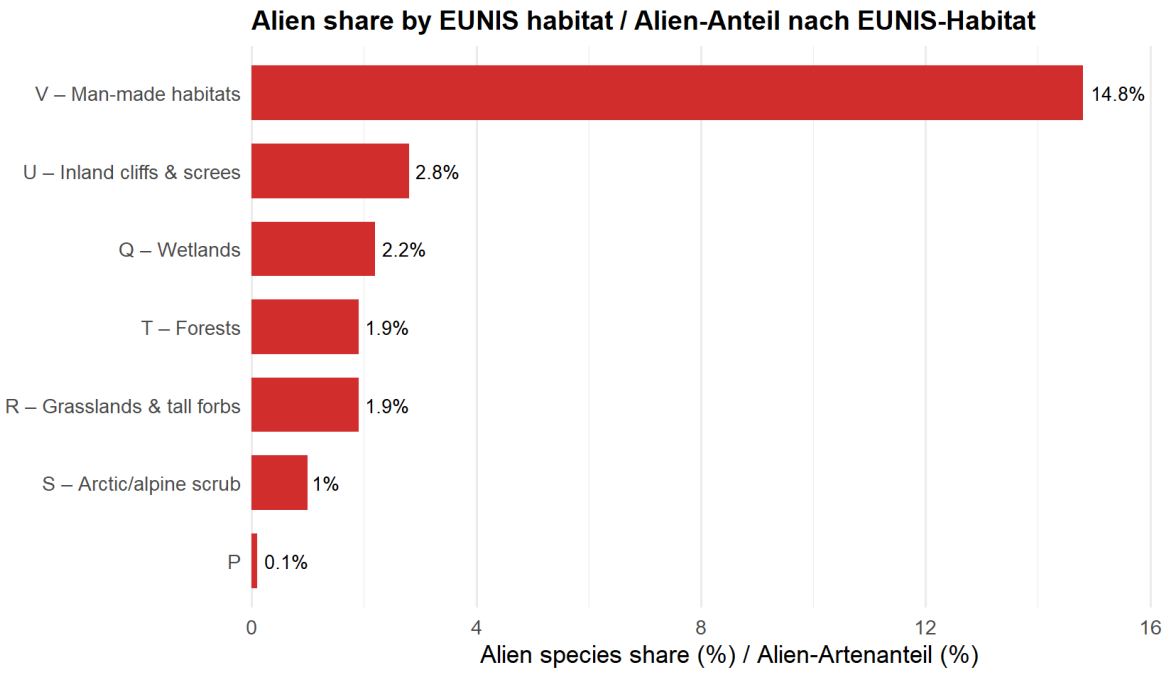
For conservation:

In a turnover-dominated system, **no single reserve can capture the full flora**. This supports a distributed network of reserves across elevations and biogeographic regions (the "Several Small" side of the SLOSS debate; Tjorve 2010). For alien management, the

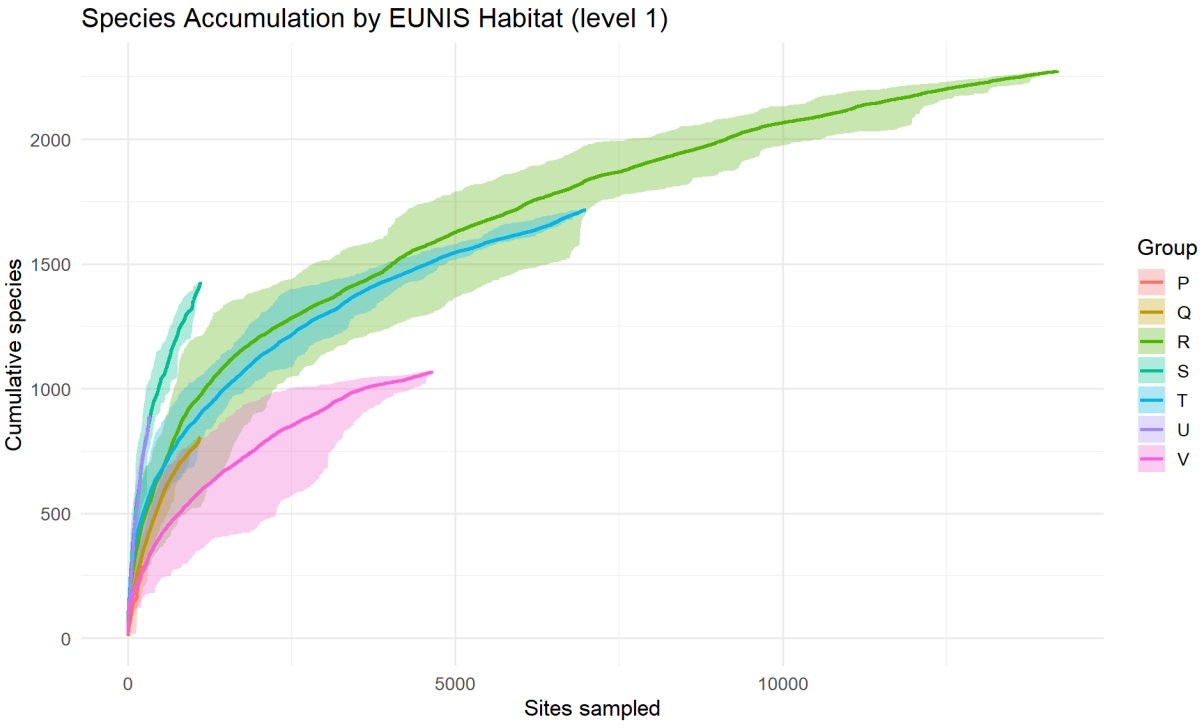
comparison suggests that urban lowlands are not just richer in aliens but also harbour *different* aliens depending on the region.

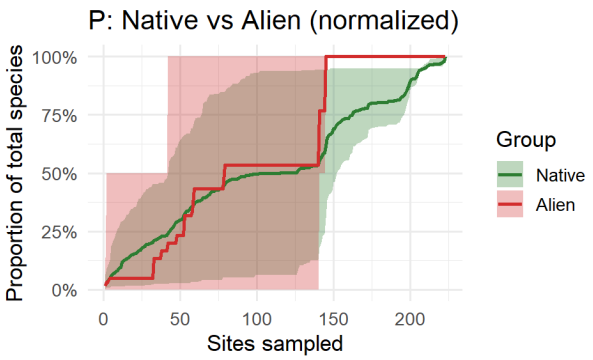
5. Habitat Vulnerability: Where Invasions Concentrate

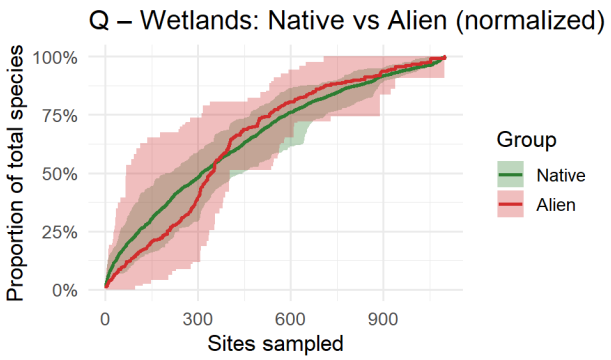
Alien Species Share by Habitat

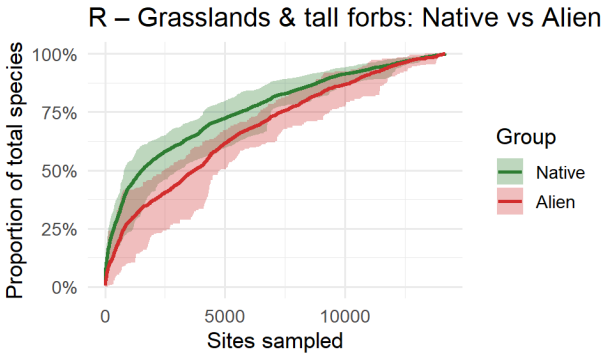


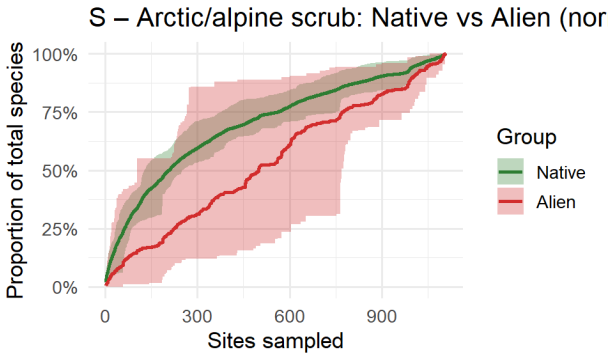
Accumulation Curves by Habitat

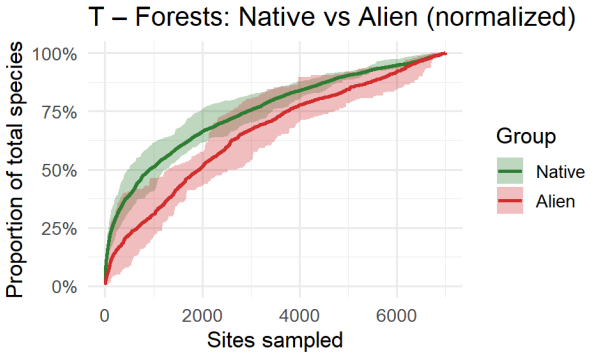


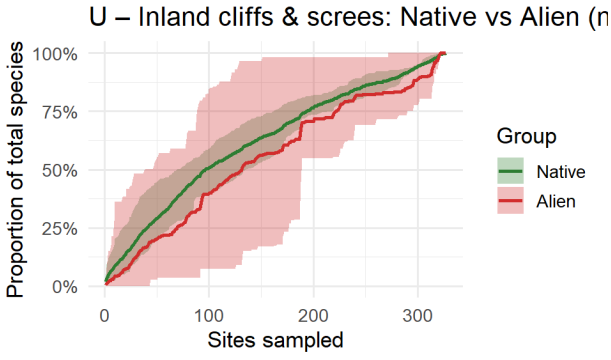

P: Native vs Alien (normalized)
Proportion of total species
100%
75%
50%
25%
0%
0
50
100
150
200
Sites sampled
Group
Native
Alien

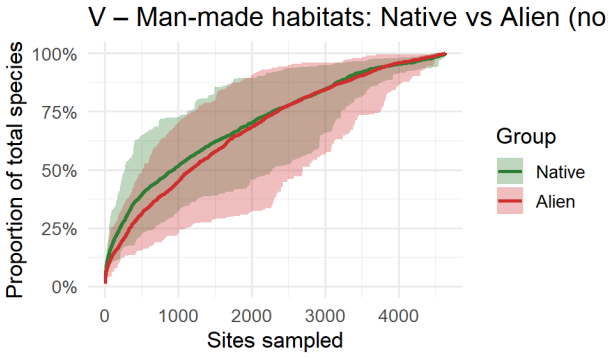

Q – Wetlands: Native vs Alien (normalized)
Proportion of total species
100%
75%
50%
25%
0%
0
300
600
900
Sites sampled
Group
Native
Alien


R – Grasslands & tall forbs: Native vs Alien
Proportion of total species
100%
75%
50%
25%
0%
0
5000
10000
Sites sampled
Group
Native
Alien

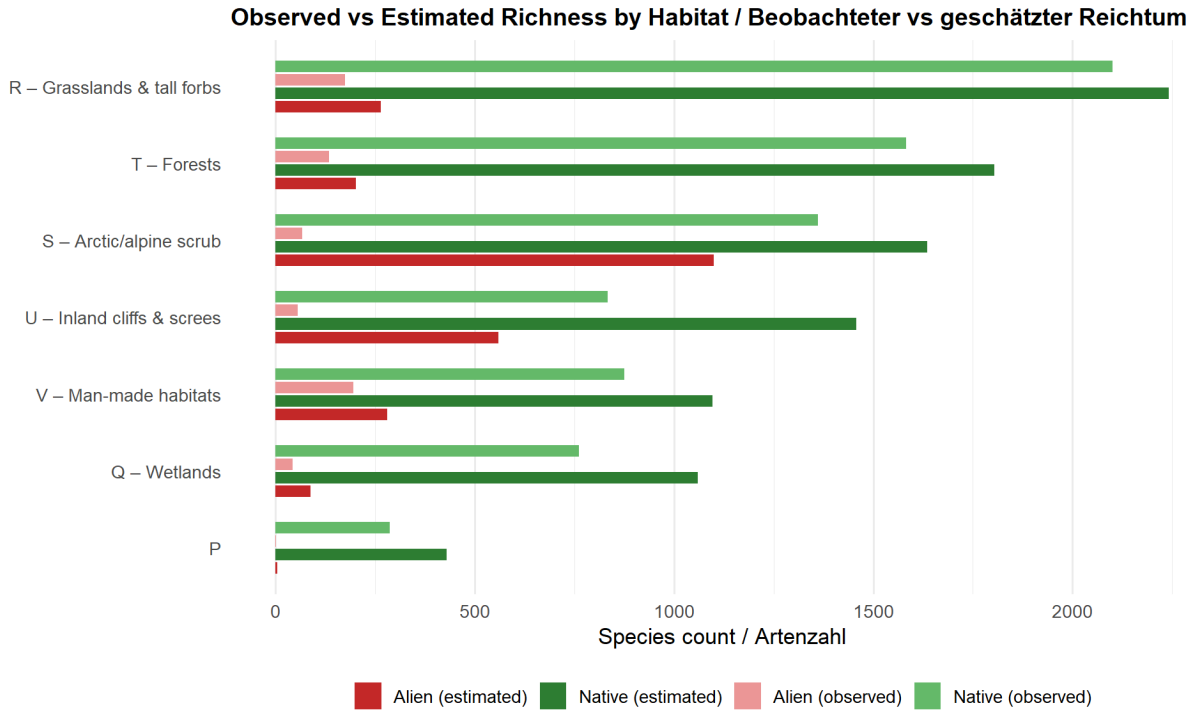

S – Arctic/alpine scrub: Native vs Alien (nor
Proportion of total species
100%
75%
50%
25%
0%
0
300
600
900
Sites sampled
Group
Native
Alien


T – Forests: Native vs Alien (normalized)
Proportion of total species
100%
75%
50%
25%
0%
0
2000
4000
6000
Sites sampled
Group
Native
Alien


U – Inland cliffs & screes: Native vs Alien (n
Proportion of total species
100%
75%
50%
25%
0%
0
100
200
300
Sites sampled
Group
Native
Alien


V – Man-made habitats: Native vs Alien (normalized)
Proportion of total species
100%
75%
50%
25%
0%
0
1000
2000
3000
4000
Sites sampled
Group
Native
Alien

Estimated Richness by Habitat (SAC-based)



Discussion

Invasion varies hugely between habitats.

The bar chart makes this clear: some habitats have alien proportions above 10%, others are almost alien-free. This is not random.

Most invaded habitats share three traits:

- **Frequent disturbance** (ploughing, mowing, flooding) creates open niches. Repeated disturbance prevents native competitive exclusion.
- **High nutrient availability.** Fertilised grasslands, floodplains, roadside verges get regular resource pulses that favour alien establishment (Davis et al. 2000).
- **Human connectivity.** Railway embankments, highway margins, and canal banks receive a steady supply of alien propagules.

Least invaded habitats include:

- **Alpine grasslands** (short growing seasons, frost, thin soils exclude most aliens)
- **Oligotrophic bogs and fens** (nutrient-poor, waterlogged; most aliens can't cope)
- **Closed-canopy forests** (light limitation at ground level blocks alien germination, though gaps and edges show higher alien presence)

Chytry et al. (2008) showed that European habitats differ 100-fold in invasion susceptibility. Our data confirms this for Austria: **habitat type matters more than climate, geography, or propagule pressure**.

Within-habitat accumulation curves:

In heavily invaded habitats, the alien curve approaches the native curve (aliens have spread broadly). In lightly invaded habitats, the alien curve is steeper and erratic (a few occurrences at a few sites, typical of early-stage invasion).

Management priorities:

1. **High:** Most invaded habitats, where aliens are already abundant. Early action prevents further establishment.
2. **Medium:** Transition habitats where aliens are present but not yet dominant.
3. **Lower:** Alpine and oligotrophic habitats resist invasion for now, but climate change may weaken that resistance.

6. What Kind of Diversity? Dominance, Partitioning, and Evolutionary Breadth

So far we have counted species. But not all species contribute equally: some are common, others are rare; some represent unique evolutionary lineages, others are close relatives of species already present. This section asks three questions that go beyond raw counts:

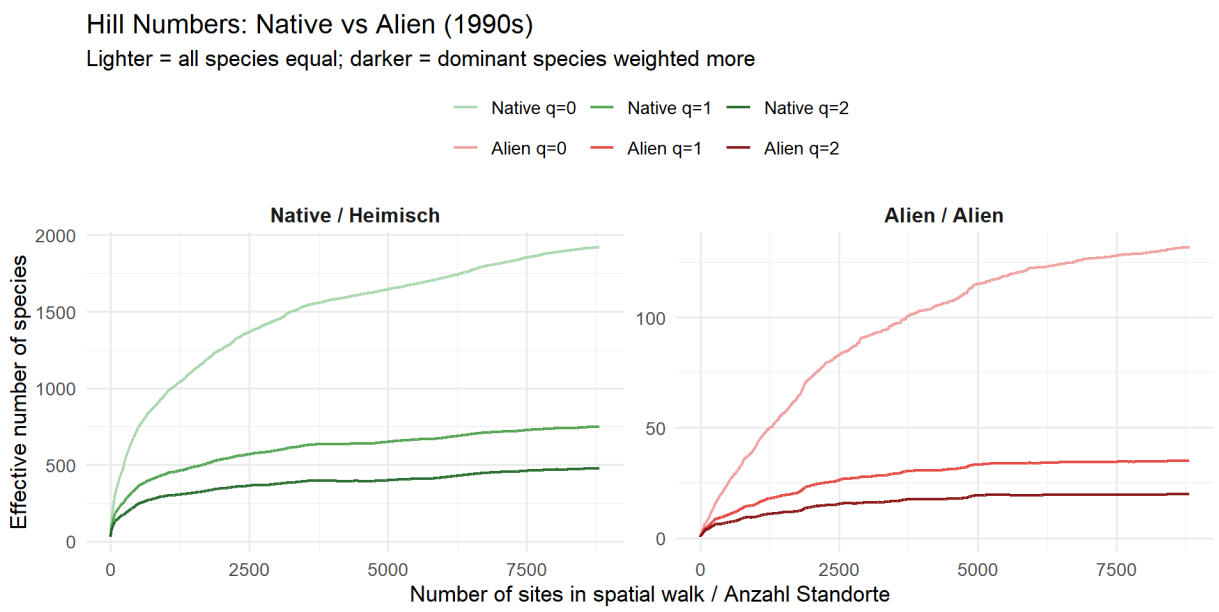
1. **Are rare or common species driving the accumulation?** (Hill numbers)
2. **Does most of Austria's diversity come from within-site richness or between-site differences?** (Alpha/beta/gamma partitioning)
3. **Do aliens add evolutionary novelty, or just more of the same?** (Phylogenetic diversity)

6a. Rare Species Drive the Tail: Hill Numbers

Hill numbers (or effective number of species) weight species differently by their abundance:

- $q = 0$: All species count equally (= species richness)
- $q = 1$: Species weighted by their frequency (= exponential of Shannon entropy)
- $q = 2$: Dominant species count most (= inverse Simpson)

As q increases, rare species contribute less. If the curves diverge strongly, rare species drive the pattern.

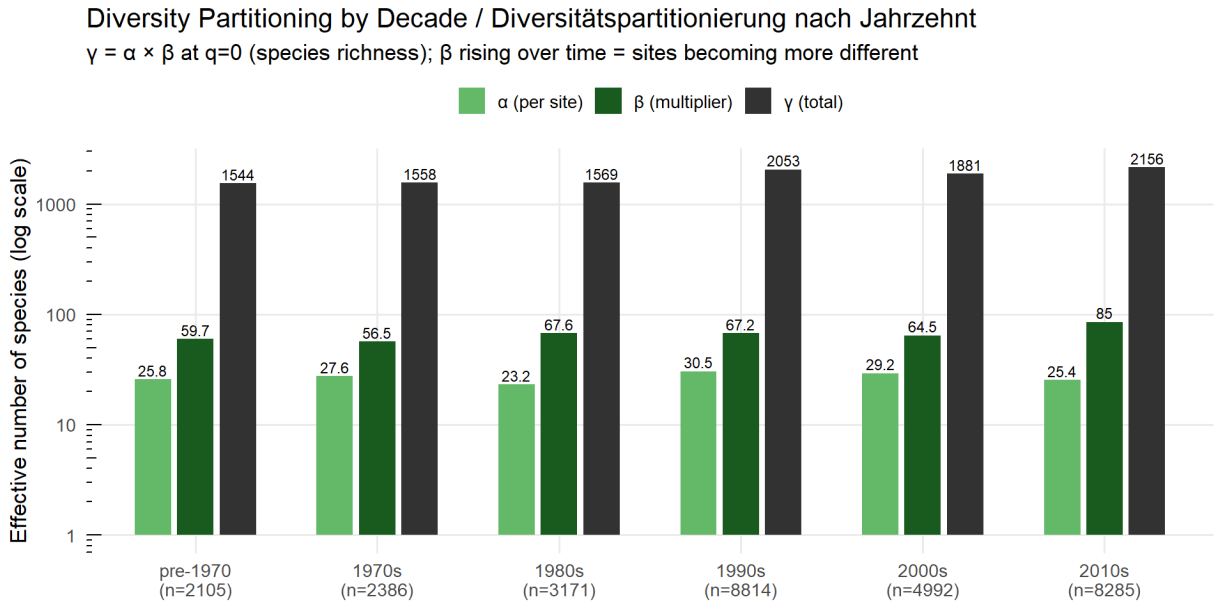


6b. Where Does Diversity Live? Alpha vs Beta

Total diversity (**gamma**, γ) can be split into two components:

- **Alpha** (α): the average number of species at a single site
- **Beta** (β): how many times you need to multiply alpha to reach gamma ($\beta = \gamma / \alpha$)

A high beta means sites are very different from each other; a low beta means every site looks similar. This multiplicative decomposition (Jost 2007) works at each Hill number order q.

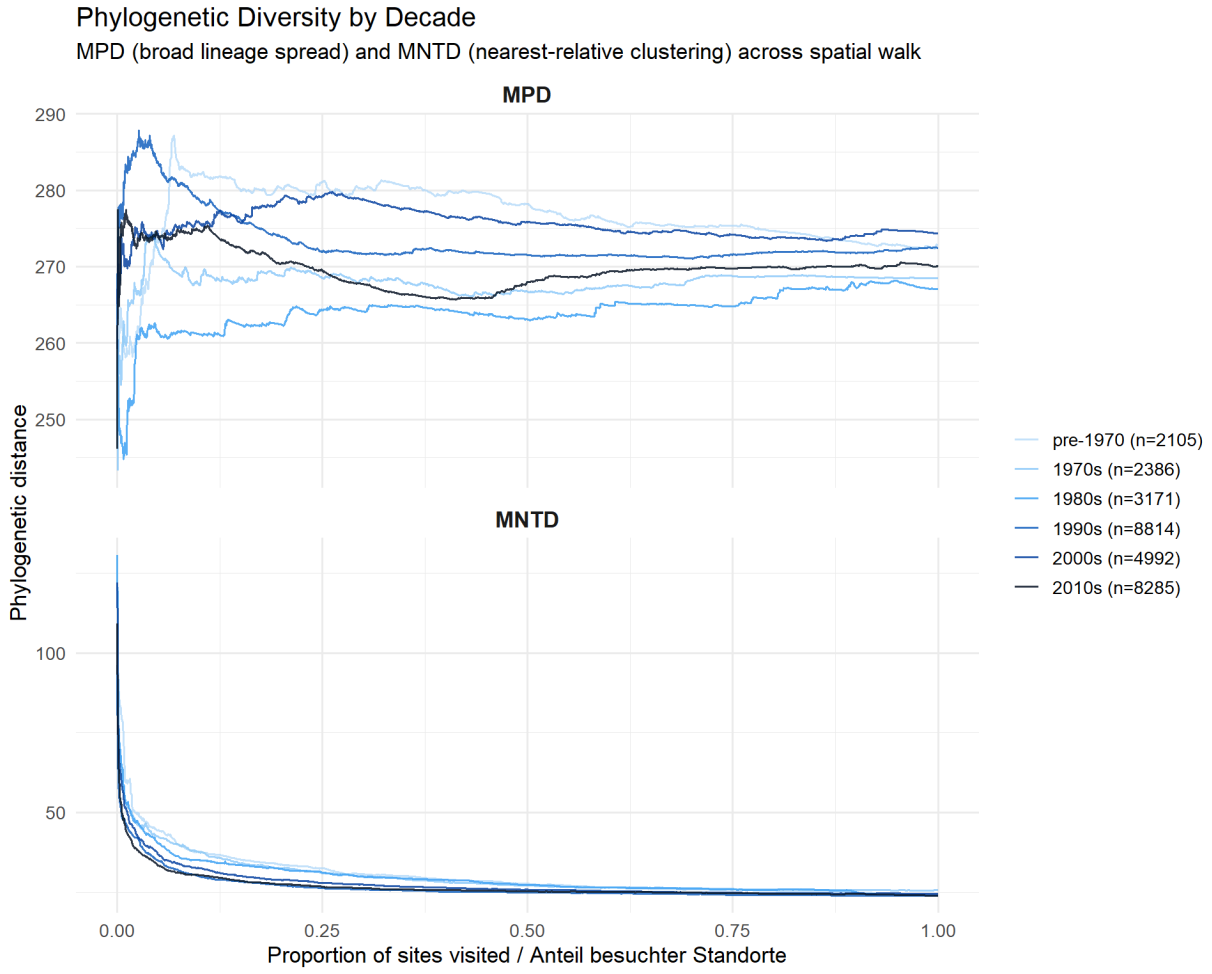


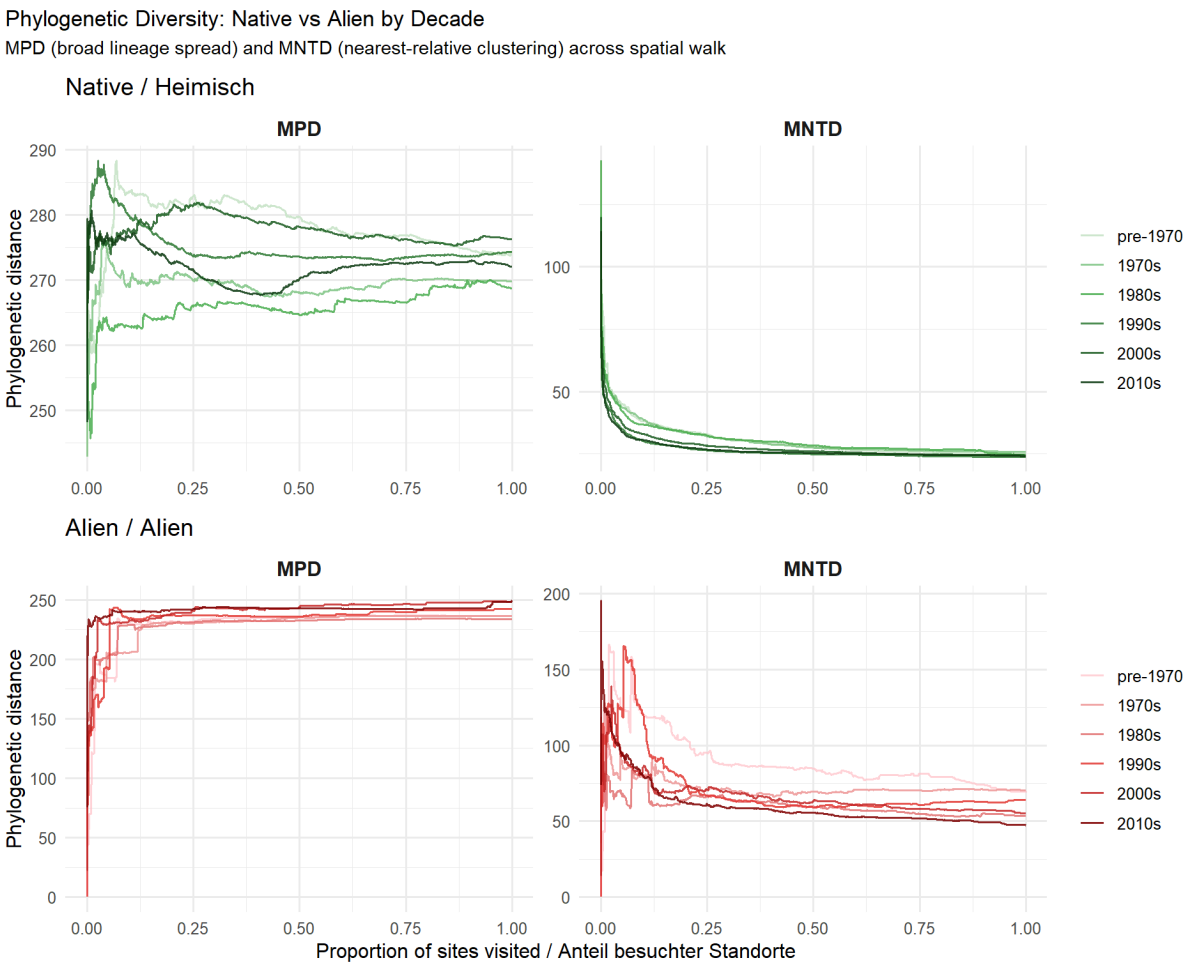
6c. Do Aliens Add Evolutionary Novelty?

Phylogenetic diversity measures how evolutionarily distant species are from each other. Two complementary metrics capture different aspects:

- **Mean Pairwise Distance (MPD)** tracks the average evolutionary distance between all species pairs. High MPD means the community spans many different plant lineages (broad phylogenetic spread).
- **Mean Nearest Taxon Distance (MNTD)** tracks the average distance to the closest relative for each species. Low MNTD means species tend to cluster in closely related groups (phylogenetic clustering at the tips).

Together, MPD shows the forest and MNTD shows the trees. Comparing native vs alien values reveals whether aliens broaden or narrow the evolutionary portfolio.





Discussion

The three analyses converge on a single picture:

Most of the species we are still "finding" are rare. The Hill number plots show that the $q = 0$ curve (raw richness) keeps climbing long after $q = 2$ (dominance-weighted diversity) levels off. The late accumulation adds species that occur at only a few sites. This is true for both natives and aliens, but the effect is proportionally larger for aliens, where a small number of widespread generalists dominate and the remainder are rare, localised newcomers.

Diversity lives between sites, not within them. The alpha/beta/gamma partitioning confirms what section 4 already showed: beta diversity (turnover between sites) accounts for the bulk of Austria's total (gamma) diversity. Any single site captures only a fraction of the national flora.

Aliens are evolutionary copycats. Native MPD is consistently higher than alien MPD, meaning the native flora spans a broader range of plant lineages. Aliens are phylogenetically clustered in a few families (Asteraceae, Poaceae, Brassicaceae) that share

traits favourable for colonisation: fast growth, high seed output, and disturbance tolerance (Cadotte et al. 2009). The ~88 "missing" alien species from section 3 are therefore likely to be more of the same functional types, not representatives of novel lineages.

What this means: counting species overstates the ecological impact of alien arrivals. The invasion debt is real, but the incoming species are unlikely to add much functional or evolutionary novelty to Austrian plant communities.

Synthesis

The six findings connect:

1. Sampling & raw richness. The hotspot maps reflect where data was collected and what observed richness looks like. Natives peak in topographically complex pre-Alpine zones; aliens concentrate around human infrastructure in the lowlands. These patterns partly reflect sampling effort.

2. Accumulation curves. Aliens have higher spatial beta diversity than natives; you need to sample more area to find the same proportion of species. This confirms quantitatively what the maps show.

3. Incomplete sampling. We know ~94% of the native flora but only ~79% of the alien flora. The gap represents invasion debt: species already present but not yet found, or species still arriving. The per-hexagon extrapolation maps show where in Austria that debt is concentrated.

4. Turnover, not nestedness. Austria's high beta diversity comes from genuine species replacement along the elevation gradient, not from impoverished subsets. No single reserve can capture the full flora.

5. Habitat filters. Invasion concentrates in disturbed, nutrient-rich, well-connected habitats and is excluded from alpine, oligotrophic, and closed-canopy environments.

6. What kind of diversity? Rare species drive the tail of the accumulation curve, beta diversity dominates gamma, and aliens are phylogenetically clustered in a few families. The invasion debt is real in numbers, but the incoming species are unlikely to add much evolutionary novelty.

These results show that **spatial species accumulation curves** are more than a sampling tool. They connect biogeography, invasion ecology, and conservation planning. The `spacc` package makes this type of analysis available for any georeferenced community dataset.

References

- Baselga, A. (2010). Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, 19, 134–143.
- Cadotte, M.W. et al. (2009). Using phylogenetic, functional and trait diversity to understand patterns of plant community productivity. *PLoS ONE*, 4, e5695.
- Chao, A. & Jost, L. (2012). Coverage-based rarefaction and extrapolation. *Methods in Ecology and Evolution*, 3, 217–226.
- Chytry, M. et al. (2008). Habitat invasions by alien plants: a quantitative comparison among Mediterranean, subcontinental and oceanic regions of Europe. *Journal of Applied Ecology*, 45, 448–458.
- Chytry, M. et al. (2009). European map of alien plant invasions based on the quantitative assessment across habitats. *Diversity and Distributions*, 15, 98–107.
- Chytry, M. et al. (2016). European Vegetation Archive (EVA): an integrated database of European vegetation plots. *Applied Vegetation Science*, 19, 173–180.
- Colwell, R.K. & Coddington, J.A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society B*, 345, 101–118.
- Davis, M.A. et al. (2000). Fluctuating resources in plant communities: a general theory of invasibility. *Journal of Ecology*, 88, 528–534.
- Essl, F. et al. (2011). Socioeconomic legacy yields an invasion debt. *Proceedings of the National Academy of Sciences*, 108, 203–207.
- Grabherr, G. et al. (2003). *Die Pflanzengesellschaften Osterreichs*. Gustav Fischer Verlag.
- Hill, M.O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54, 427–432.
- Jost, L. (2006). Entropy and diversity. *Oikos*, 113, 363–375.
- Kier, G. et al. (2009). A global assessment of endemism and species richness across island and mainland regions. *PNAS*, 106, 9322–9327.
- Lockwood, J.L. et al. (2005). The role of propagule pressure in explaining species invasions. *Trends in Ecology & Evolution*, 20, 223–228.
- Partel, M. et al. (2011). Dark diversity: shedding light on absent species. *Trends in Ecology & Evolution*, 26, 596–600.
- Pysek, P. et al. (2010). Disentangling the role of environmental and human pressures on biological invasions across Europe. *PNAS*, 107, 12157–12162.
- Rahbek, C. (1995). The elevational gradient of species richness: a uniform pattern? *Ecography*, 18, 200–205.
- Seebens, H. et al. (2017). No saturation in the accumulation of alien species worldwide. *Nature Communications*, 8, 14435.
- Socolar, J.B. et al. (2016). How should beta-diversity inform biodiversity conservation? *Trends in Ecology & Evolution*, 31, 67–80.
- Tjorve, E. (2010). How to resolve the SLOSS debate: lessons from species-diversity models. *Journal of Theoretical Biology*, 264, 604–612.
- Wilson, J.R.U. et al. (2007). Residence time and potential range: crucial considerations in modelling plant invasions. *Diversity and Distributions*, 13, 11–22.