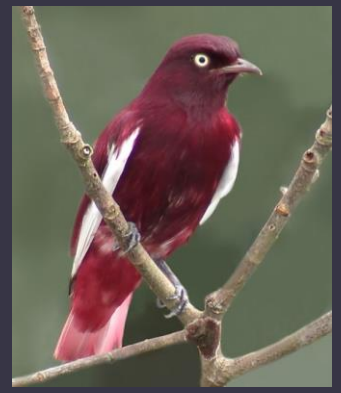
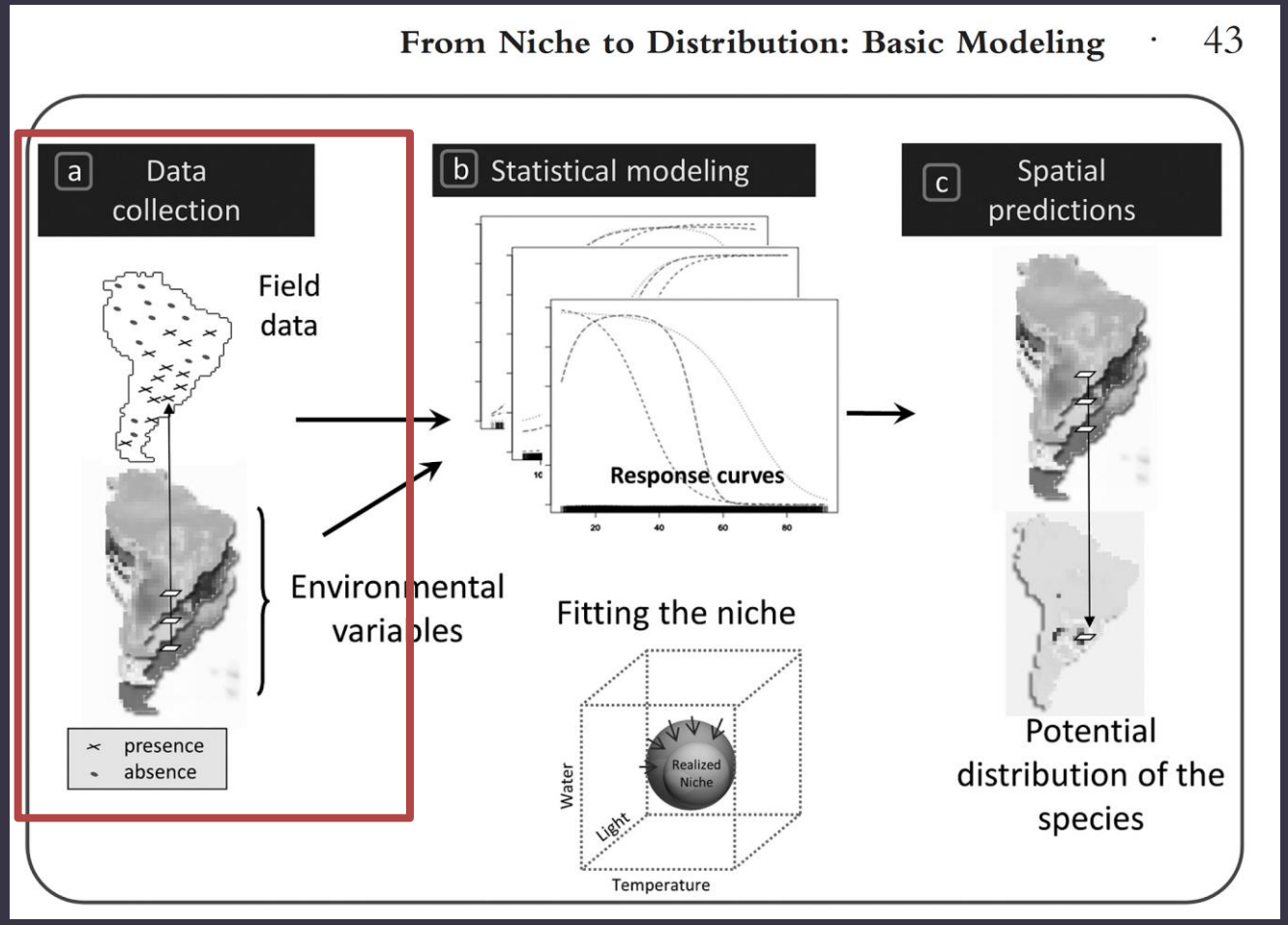
SDM Crash Course #2: Occurrence Cleaning & Background Points




What are we doing today?



SDM Data:

- Occurrences
- Background Points
- Predictor Data
  - Environmental
  - Physiological

Occurrence Data

Presence: Where a species is known to occur in space and time

- Needed to correlate presence data with predictor variables

Absence/Pseudo-Absence/Background: Where a species is *not* known to occur

- Unless used surveys include data about where species were searched for and not found, absence data is just a guess

  Serves two functions:

  1. Provide contrast to the presence/predictor correlation
  2. Allows for model evaluation: how well does the model predict presences compared to random data?

Occurrence Data

Presence: Where a species is known to occur in space and time

- Needed to correlate presence data with predictor variables

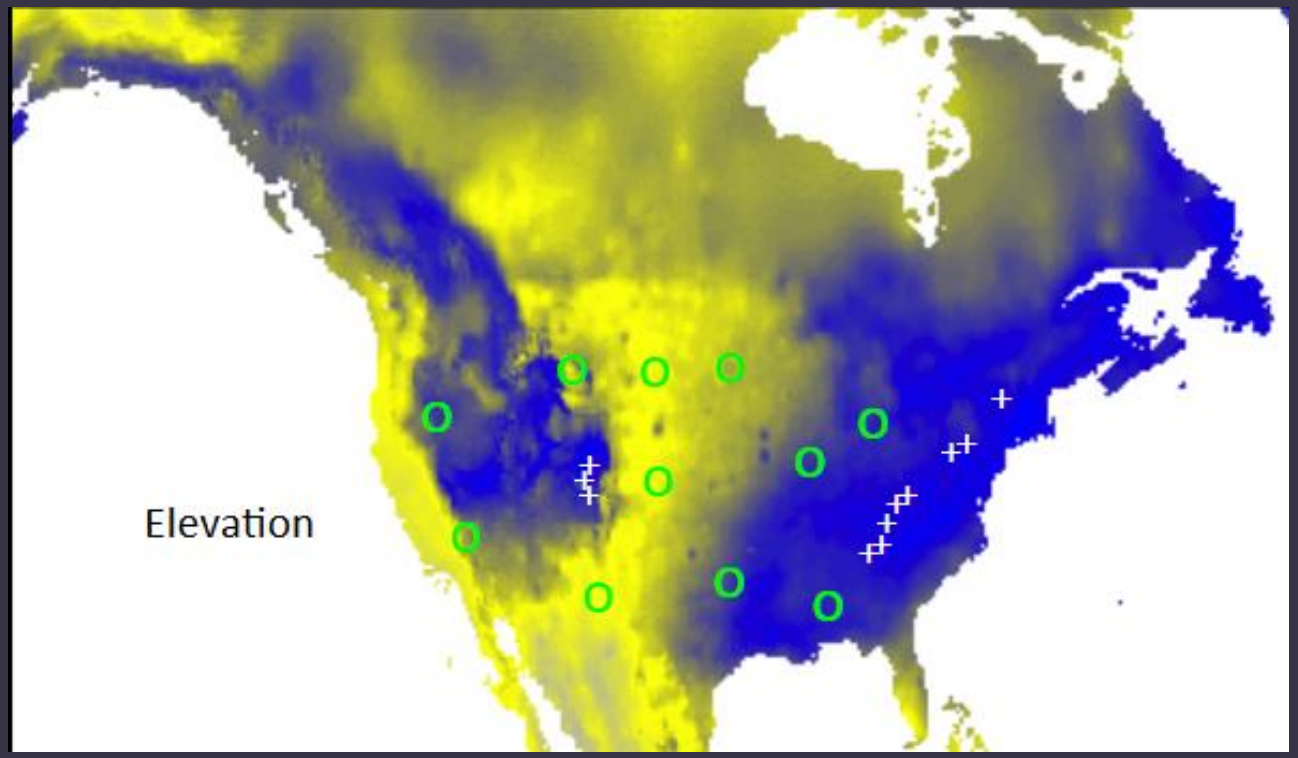
Absence/Pseudo-Absence/Background: Where a species is *not* known to occur

- Unless used surveys include data about where species were searched for and not found, absence data is just a guess

Serves two functions:

1. Provide contrast to the presence/predictor correlation
2. Allows for model evaluation: how well does the model predict presences compared to random data?

Absence Data to Gauge Correlation



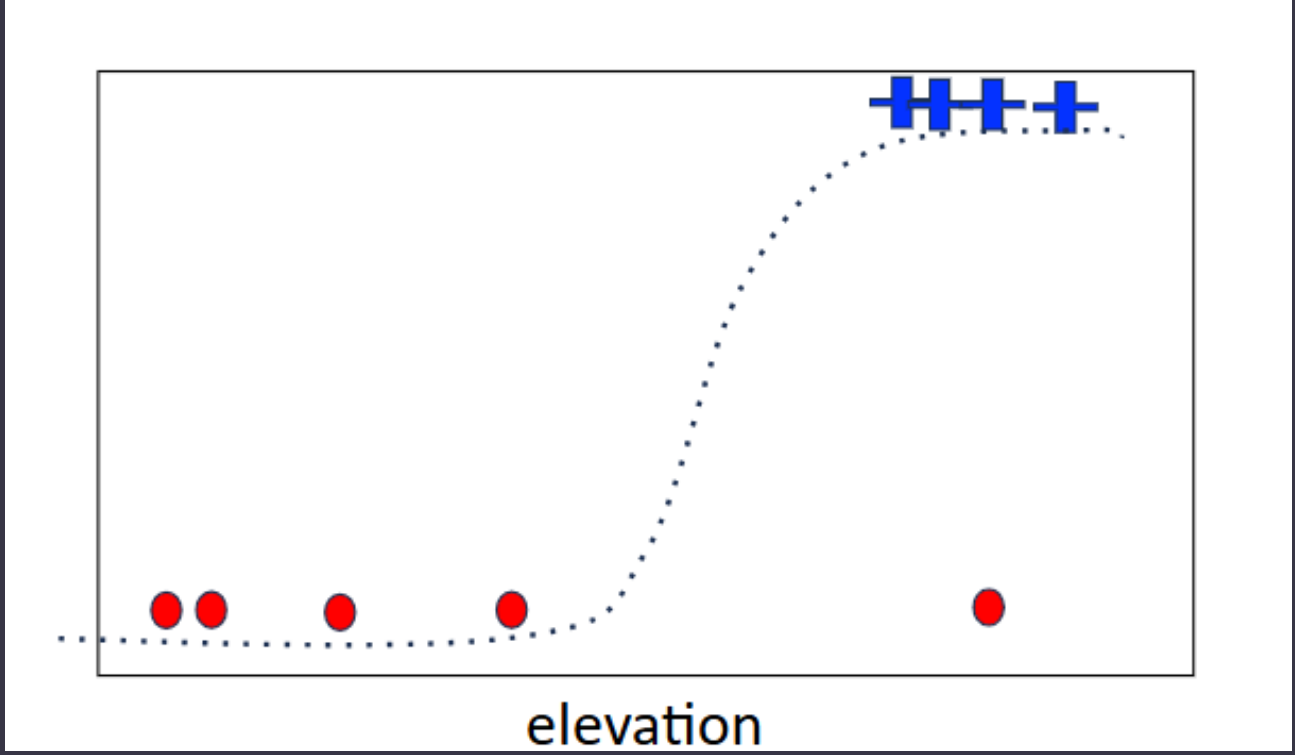
Crosses denote presence, circles are background points

How well does our presence data correlate with elevation?

It depends on the variation of elevation in the landscape we are looking at

Absence data samples that variation

Absence Data to Gauge Correlation



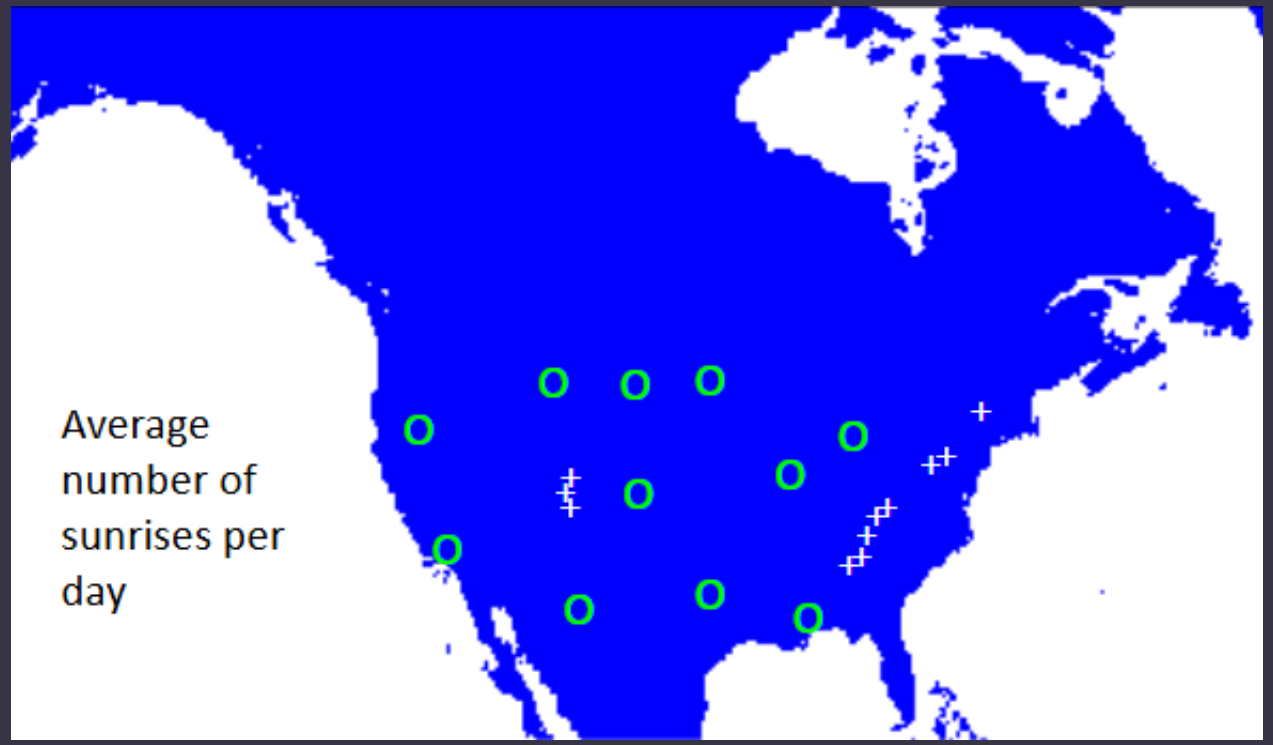
How well does our presence data correlate with elevation?

It depends on the variation of elevation in the landscape we are looking at

Absence data samples that variation

Presence data correlates with elevation

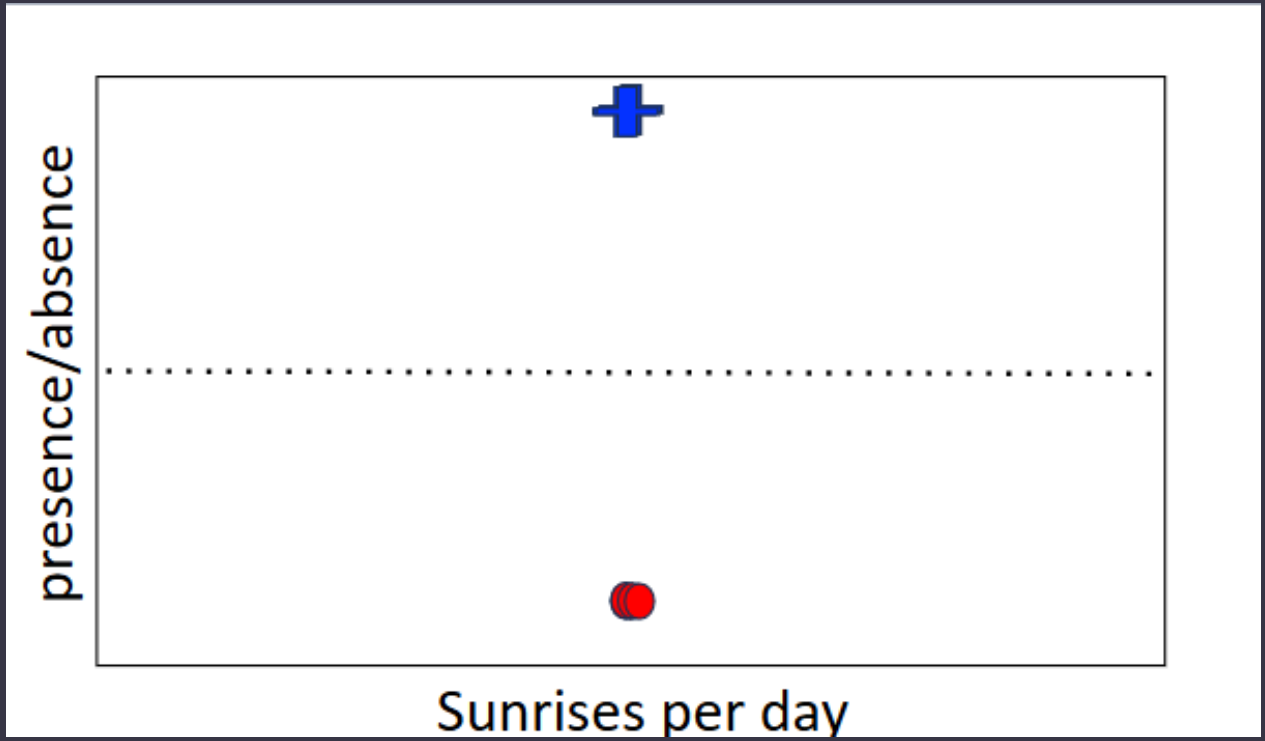
Absence Data to Gauge Correlation



Crosses denote presence, circles are background points

How about with average number of sunrises per day?

Absence Data to Gauge Correlation



Context is so common, with so little variation, as to be an essentially meaningless predictor

Presence data does not correlate with sunrise frequency

Properties of a Well-Behaved Sample

Random

- Probability of sampling an occurrence is proportional to the probability that the species occupies the space
- ***Not biased***

Independent

- Probability of sampling an occurrence is not influenced by other occurrences
- ***Not autocorrelated***

Sufficient

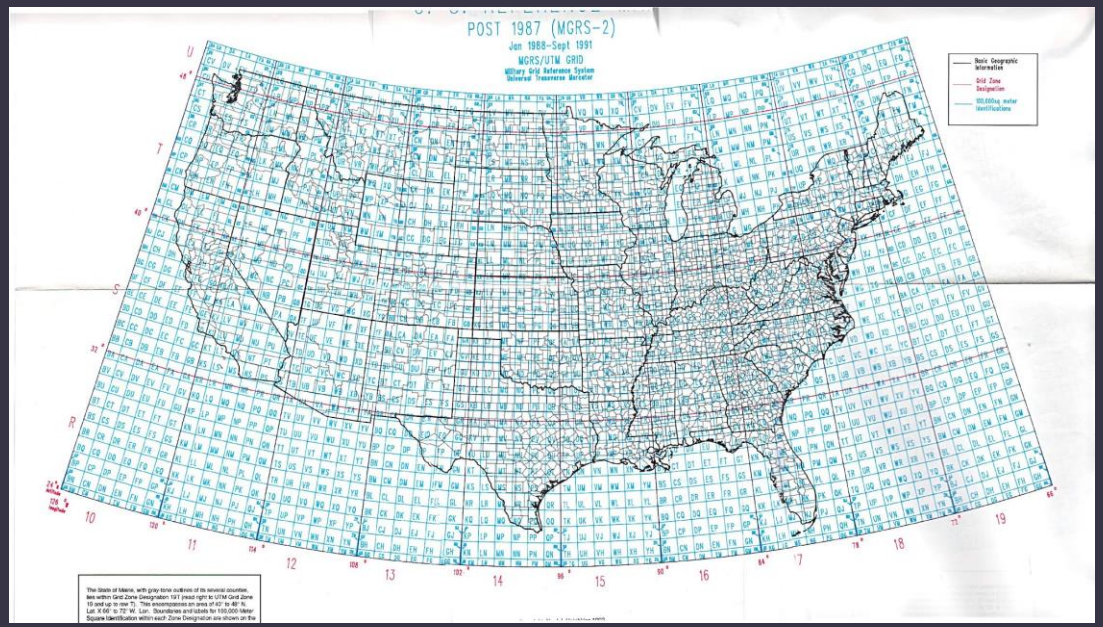
- Enough samples to approximate a population
- ***Adequate power***

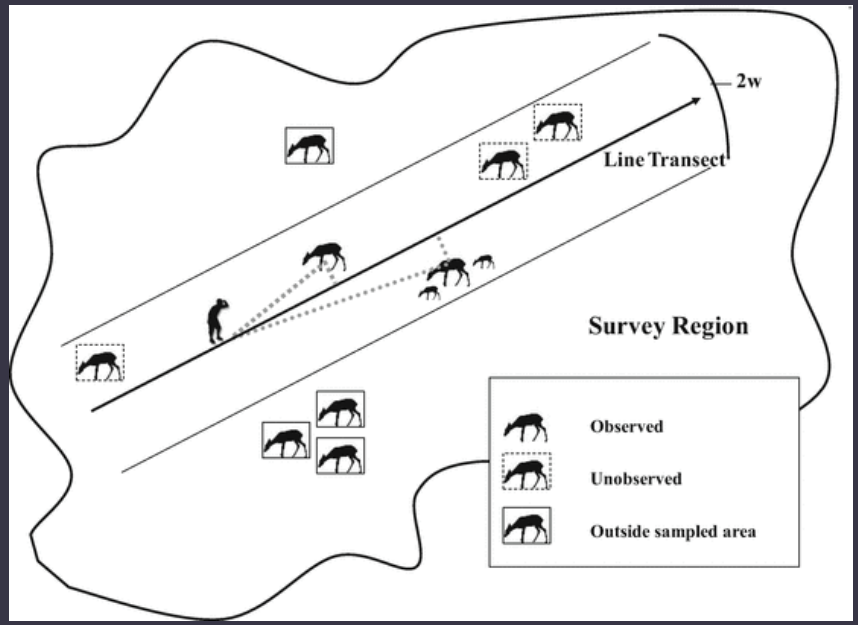
In a Perfect World: Random

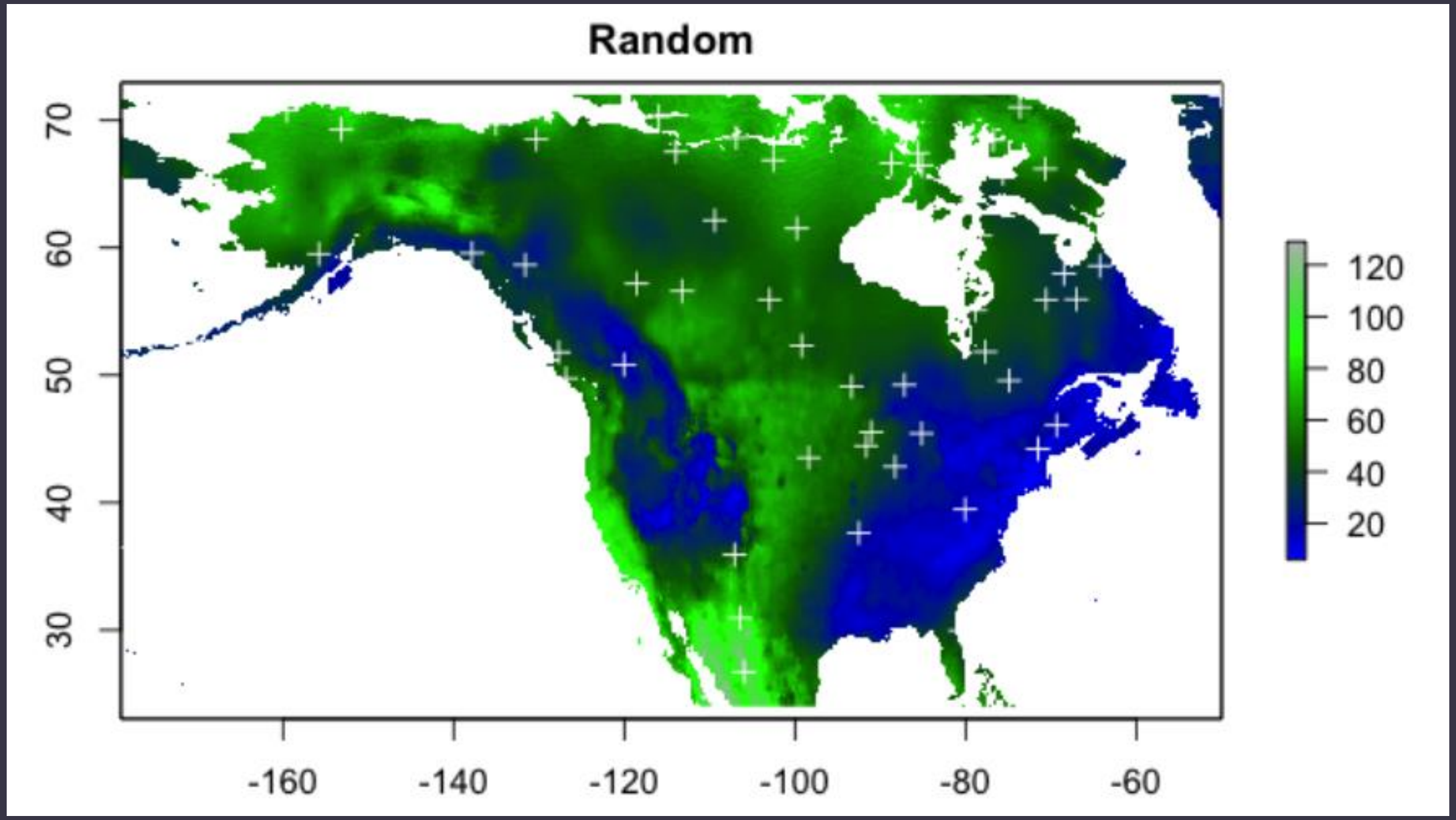
Randomly sampled:

-Grid (Cartesian)

-Transect (Vector)






Random
70
60
50
40
30
-160
-140
-120
-100
-80
-60
120
100
80
60
40
20

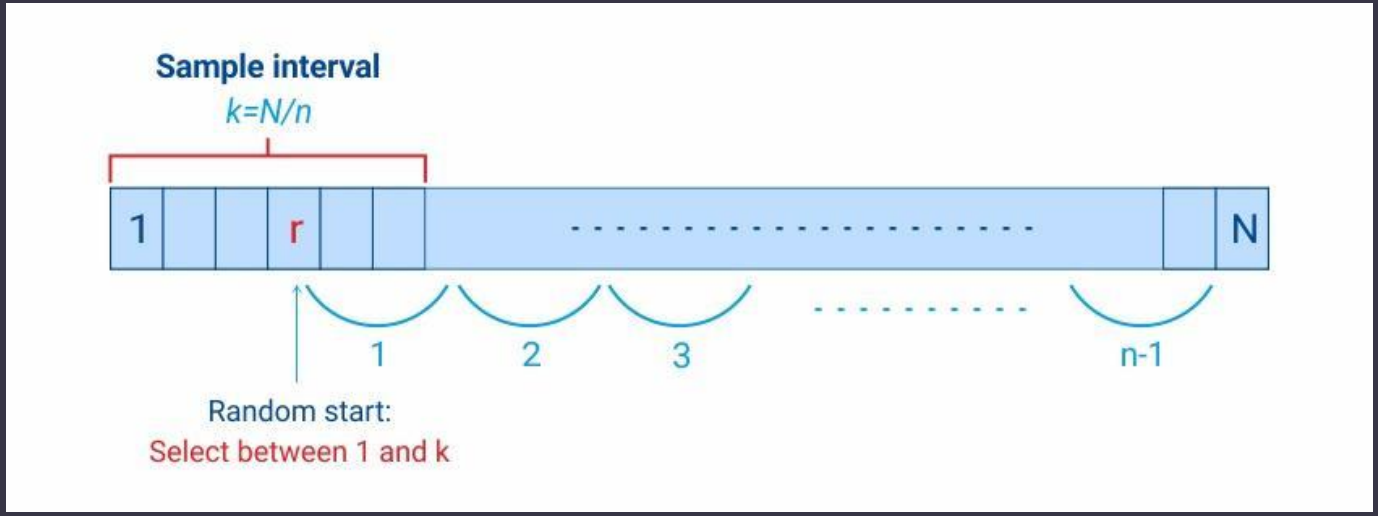
In a Perfect World: Independent

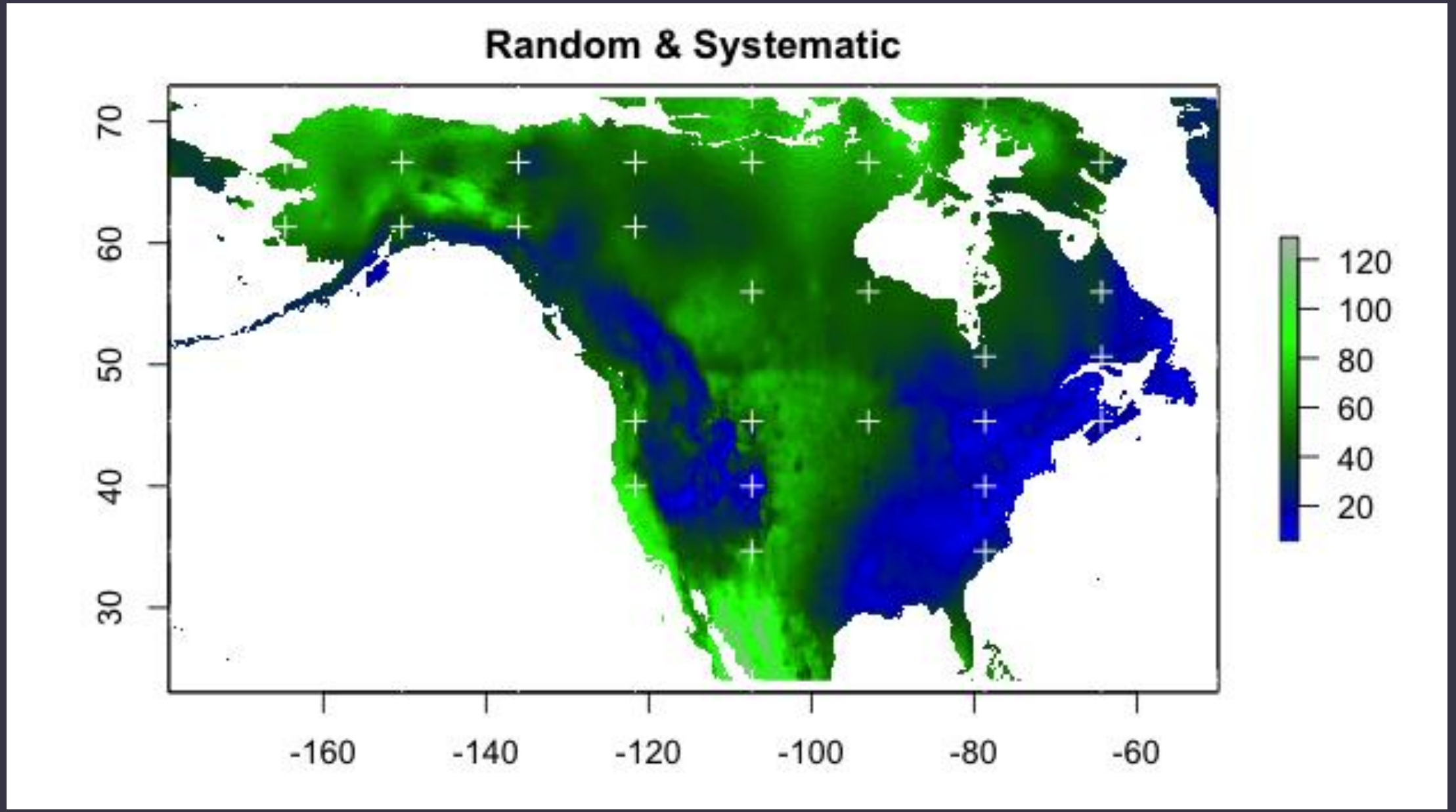
Independent Data:

-Maximize average nearest neighbor distance

-Sample at fixed intervals

-Utilize systematic or random uniform sampling




Random & Systematic
70
60
50
40
30
-160
-140
-120
-100
-80
-60
120
100
80
60
40
20

In a Perfect World: Sufficient

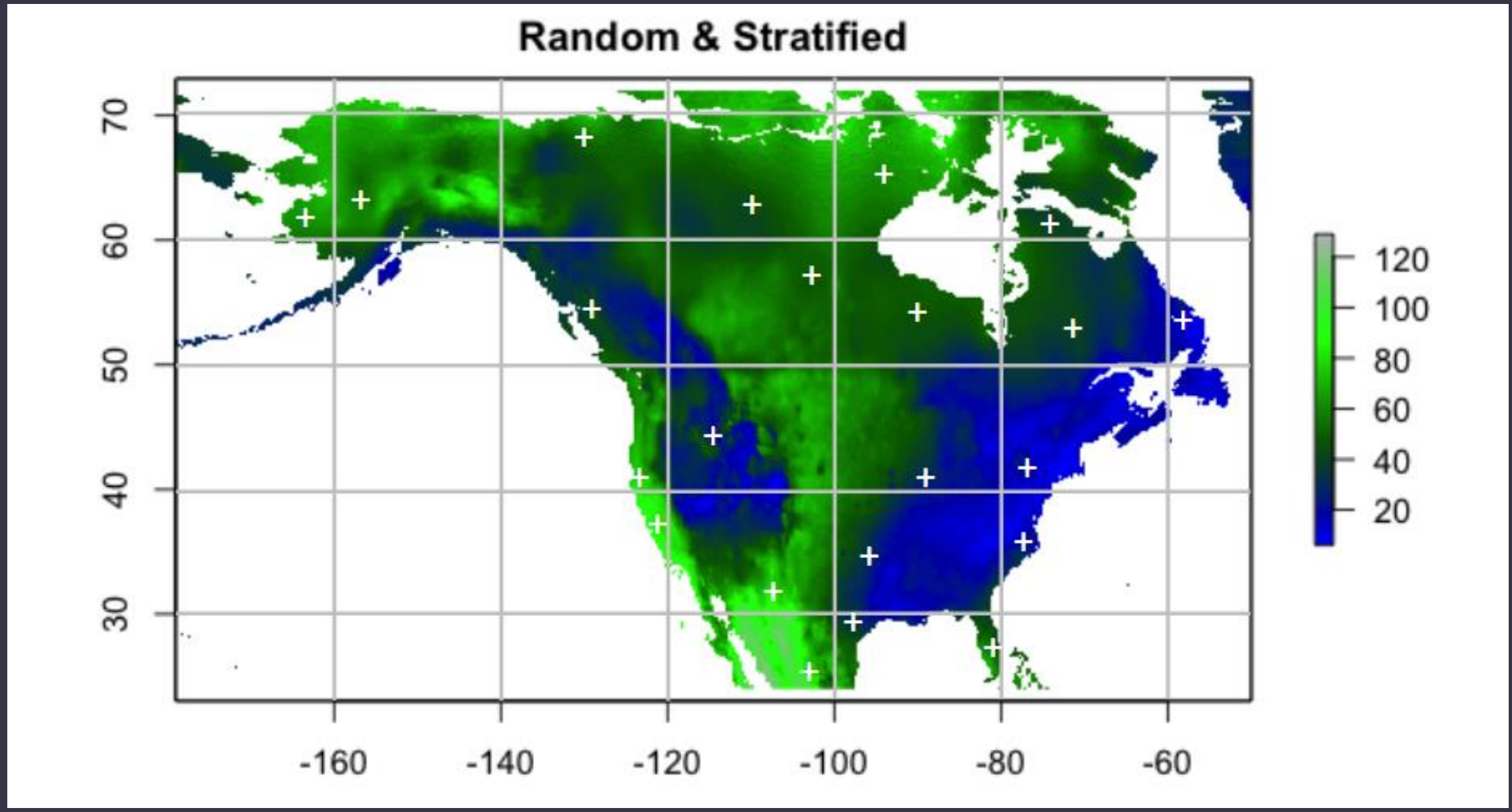
Sufficient Data:

-How many samples exactly will depend on the study species

-Coarse rule of thumb: 10-20x the number of predictors

-Coverage: samples should be taken from all over the map

Stratification: Divide samples into discrete areas, sample randomly from each to obtain coverage


Random & Stratified
70
60
50
40
30
-160
-140
-120
-100
-80
-60
120
100
80
60
40
20

Real Data is Imperfect

Table 7.1 *Examples of databases that store species distribution data.*

General	GBIF	www.gbif.org
General	Map of Life	https://mol.org/
General	LifeMapper	lifemapper.org/
General	IUCN Red List	www.iucnredlist.org/
Herps	HerpNET	herpnet.org/
Mammals	MaNIS	vertnet.org/
Marine species	OBIS	www.iobis.org/
Amphibians	AmphibiaWeb	http://amphibiaweb.org/
Birds	ORNIS	http://ornisnet.org
Birds	Bird Life	www.birdlife.org/
Plants	Atlas Flora Europaea	www.luomus.fi/en/database-atlas-florae-europaeae/
Plants	BIEN	http://bien.nceas.ucsb.edu/bien/
Central America	REMIB	www.conabio.gob.mx/remib_ingles/doctos/remibnodosdb.html?
Brazil	SpeciesLink	http://splink.cria.org.br/

Screenshot

- Collections & online repositories are used more than perfectly designed surveys
  - Often opportunistic

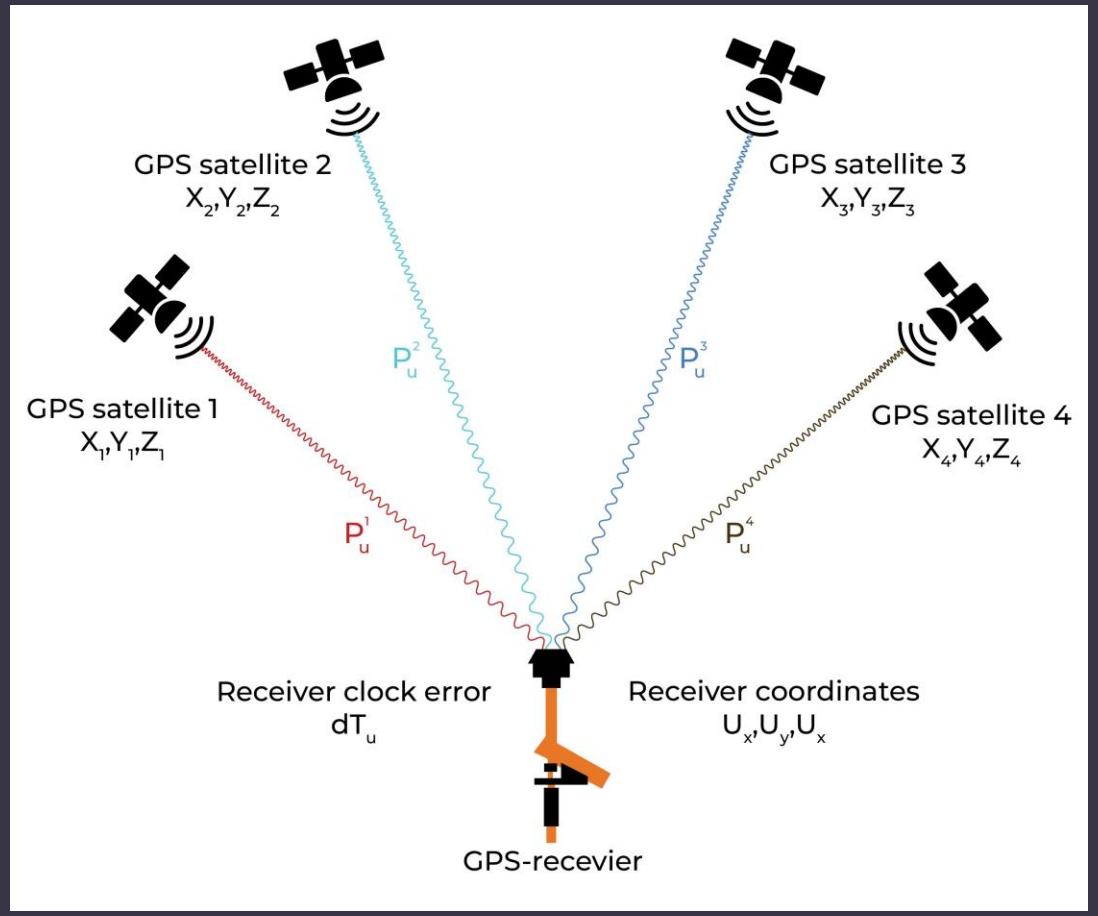
Sample Inaccuracy: Species ID

Lobster

Turkey

Frog (photo cred: Gracie)

Sample Inaccuracy: Spatial Inaccuracy



What can I do about record inaccuracy?

Visually inspect:

- Records that are far out of the range of other records
- Records that don't match administrative data
- Records associated with collections facilities
- Records in improbable locations
  - Physiologically, in middle of ocean, Antarctica

Solanum acaule
Wild potato found in the Andes

Violations

#1: Collector Bias

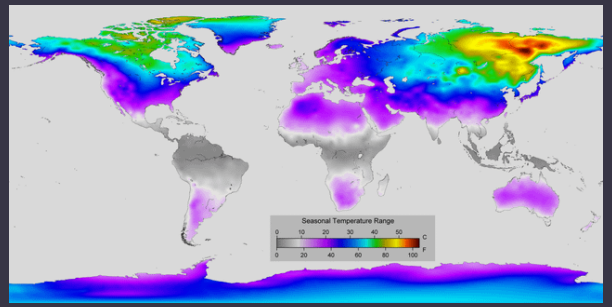
- People collect where it is easy to collect (roadside, near infrastructure, in safe regions, etc.)
  - ***Solution:*** Spatial thinning AND background point generation (more in a sec)

#2: Spatial Autocorrelation

- Species live in populations, members aren't randomly distributed. Env. Var. are also autocorrelated
  - ***Solution:*** Systematic or stratified random sampling to maximize nearest neighbor distance

#3: Temporal Variation

- Individuals sampled more than once, or one particular area has multiple collections
  - ***Solution:*** Remove spatial duplicates in your dataset (all but one)
- Some organisms vary in abiotic requirements by season
  - ***Solution:*** Filter by month/season/year



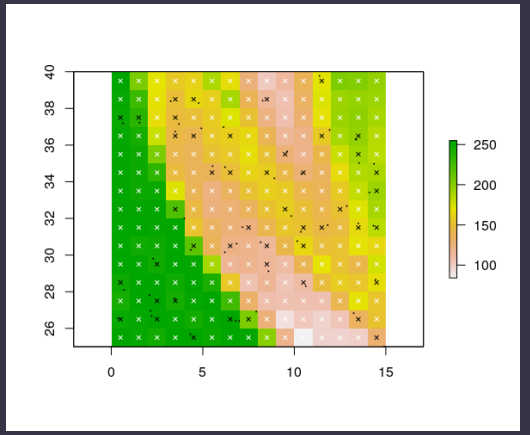
Absence Data

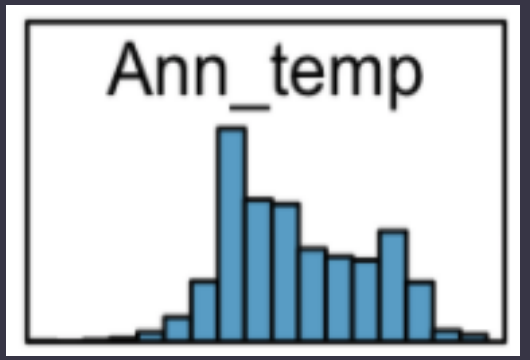
Absence/Pseudo-Absence/Background: Where a species is *not* known to occur

- Unless used surveys include data about where species were searched for and not found, absence data is just a guess

  Serves two functions:

  1. Provide contrast to the presence/predictor correlation
  2. Allows for model evaluation: how well does the model predict presences compared to random data?





History of Methods

Early Approach (1990-2006ish)

Can we distinguish habitats in which species occur from ones in which they don't?

-Identify where species aren't to some degree

Newer Approach (2006ish-Now)

Can we distinguish habitats in which species are known to occur from habitats in which they are not known to occur

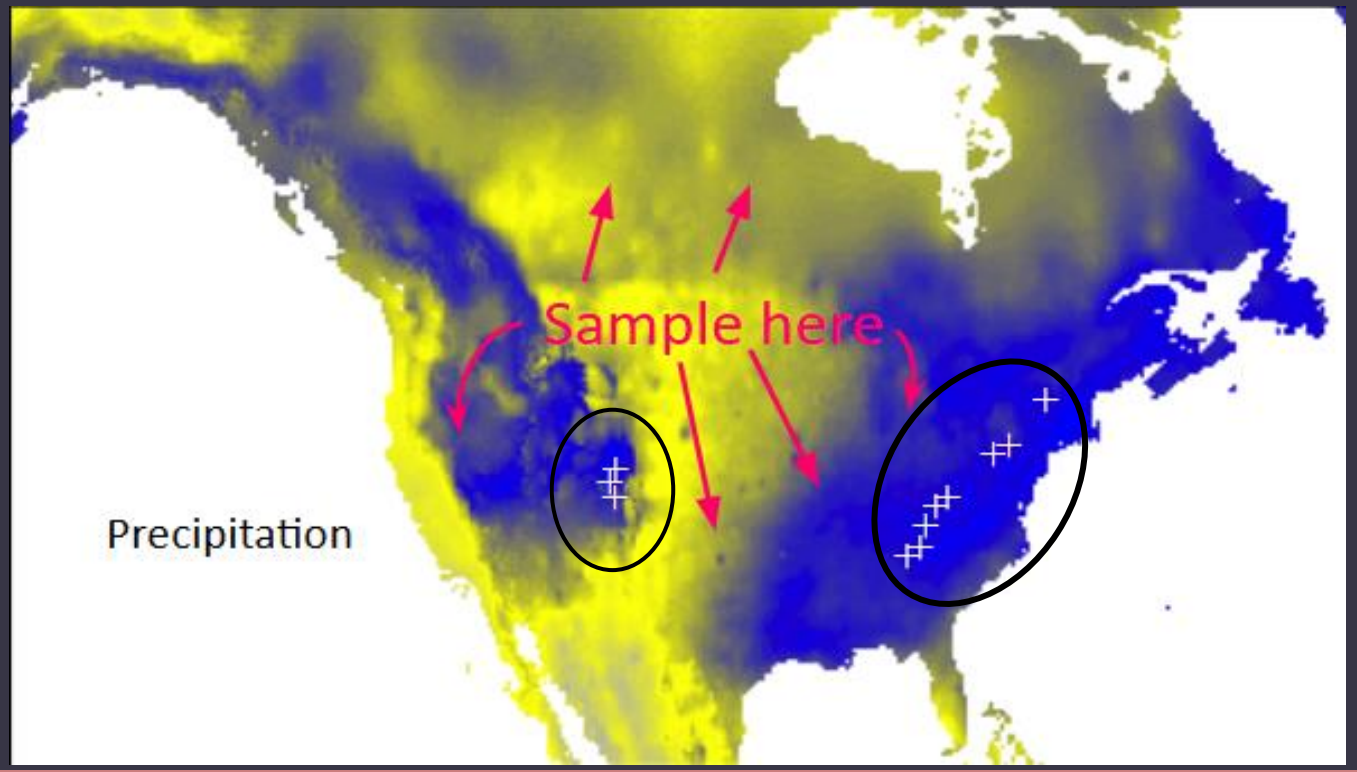
-Occurrence VS Random Sample

-More biologically realistic and conservative

Early Approach: Method #1

Without true absence data, how well can we really guess where a species *isn't*?

1. Define a species range (buffer around occurrence points) and randomly sample outside

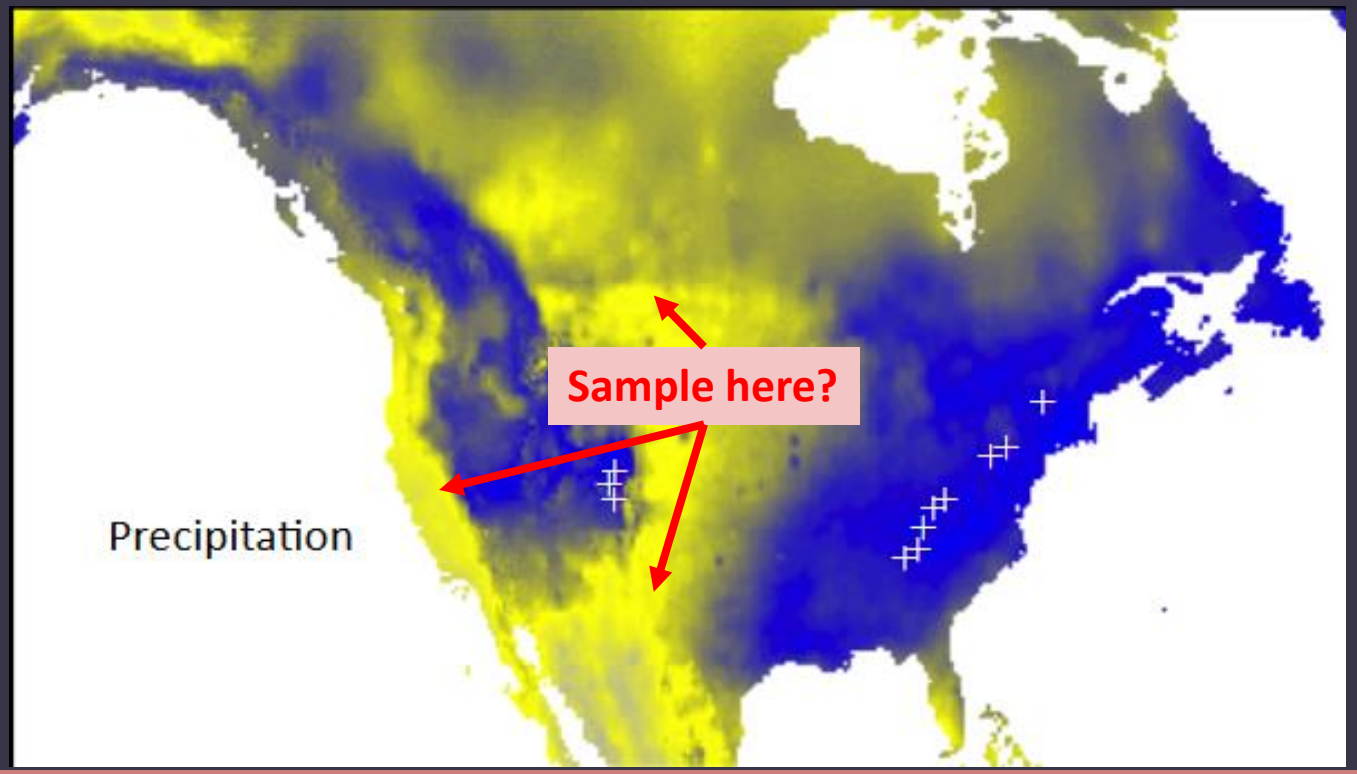


- Buffer is somewhat arbitrary
- How large of an area beyond the buffer do we sample?
  - A few degrees?
  - Within expected dispersal range
  - Continent?
  - Universe?

Early Approach: Method #2

Without true absence data, how well can we really guess where a species *isn't*?

2. Sample only in areas in which environmental conditions are different

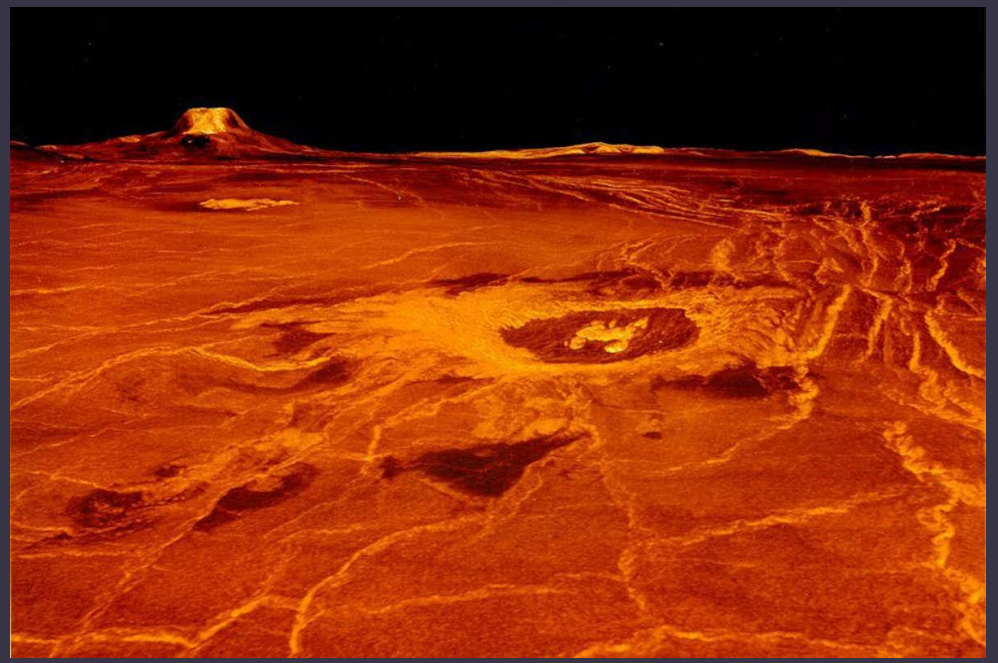


- Which conditions?
- What is considered different?
- How large of an area beyond the buffer do we sample?
  - A few degrees?
  - Within expected dispersal range
  - Continent?
  - Universe?

An Aside: Why the Scale Matters

We can 100% say that this species does not occur on any planet other in our solar system than Earth...

...But how useful is a model that can distinguish terrestrial habitats from those on Venus?



Refined Approach

Treat “absences” as background/pseudo-absence points

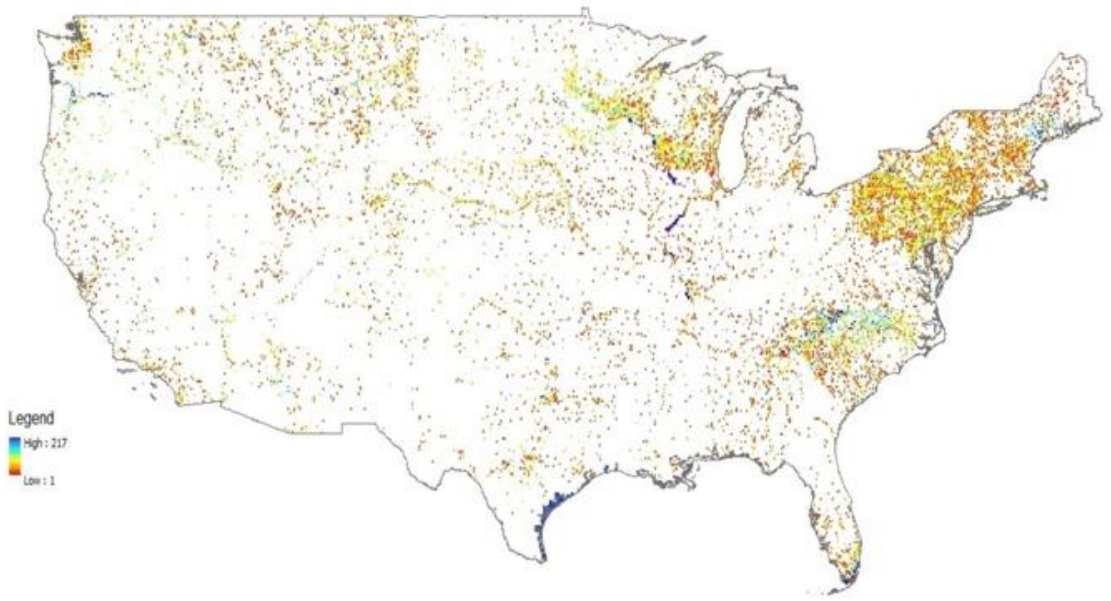
- Models are designed to distinguish presence points from random points

Background sample should have the same collector bias as occurrence sample

- Phillips et al. 2009, Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data

How do we implement this?

Refined Approach: Method #1



Sampling bias grid derived from GBIF data on freshwater fish. The grid represents sampling effort and is constructed according to the target-group approach of Phillips *et al.* (Ecological Applications, **19**, 181-197).

1. Replicate sample bias by using occurrence points from species with similar collection histories as your background/pseudo-absence data
   - However, issue of scale remains

Refined Approach: Method #2

2. Sample from within the range of occurrences
   - Can also do a “combined” manner of sampling

A bit more challenging for the model
   - Has to discriminate occurrence points from random points in an area with very similar environmental conditions

Scale issue remains
   - How big do you make the buffer?
   - If sampling in and out, what weighting scheme do you use?