Variable selection

Variable selection - general considerations

1. Ecological Framework / conceptual considerations

- What aspects of the environment should be important and why?
- Mechanistic explanation of predictors
- Direct vs. indirect
  - Avoid indirect unless:
    - study area is small
    - goal is a highly accurate model in that region only
    - No plans to project elsewhere

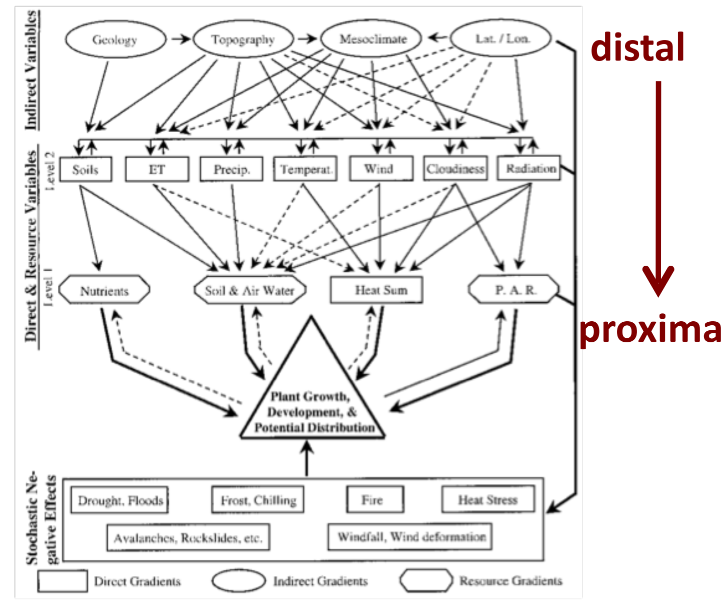


Fig. 3. Example of a conceptual model of relationships between resources, direct and indirect environmental gradients (see e.g. Austin and Smith, 1989), and their influence on growth, performance, and geographical distribution of vascular plants and vegetation.

Guisan & Zimmermann

Variable selection - general considerations

- Direct vs. indirect
  - Try to use direct, especially if:
    - Goal is to understand spatial patterns / drivers of distribution
    - Projecting to new places / times

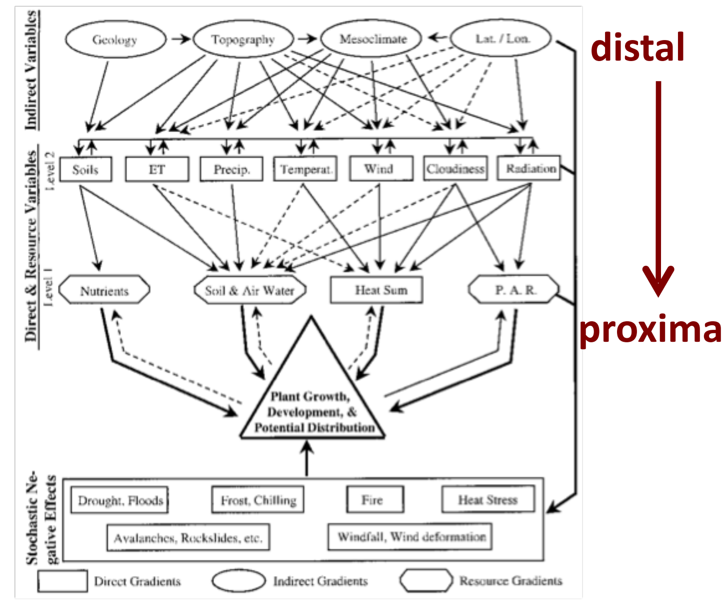


Fig. 3. Example of a conceptual model of relationships between resources, direct and indirect environmental gradients (see e.g. Austin and Smith, 1989), and their influence on growth, performance, and geographical distribution of vascular plants and vegetation.

Guisan & Zimmermann

Variable selection - general considerations

2. Data considerations

- Resolution and extent
  - What matches the occurrence data and the known distribution of the species?
  - Do not truncate using political boundaries
- Scope of available predictors

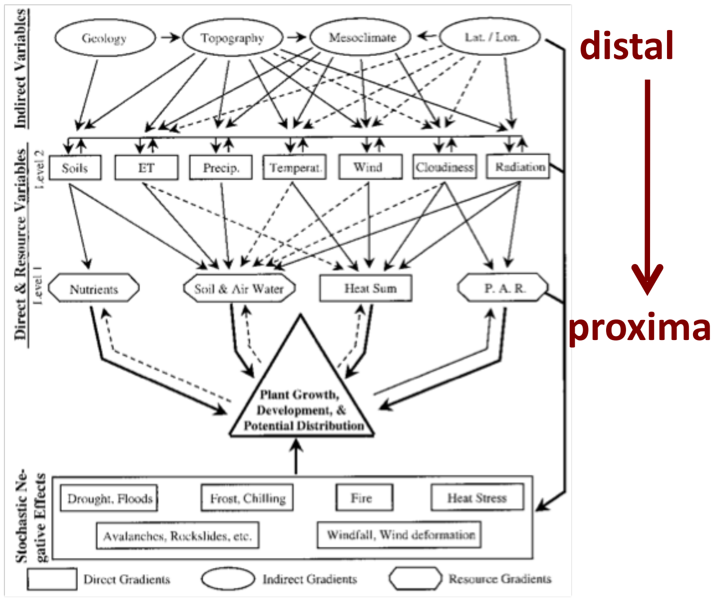


Fig. 3. Example of a conceptual model of relationships between resources, direct and indirect environmental gradients (see e.g. Austin and Smith, 1989), and their influence on growth, performance, and geographical distribution of vascular plants and vegetation.

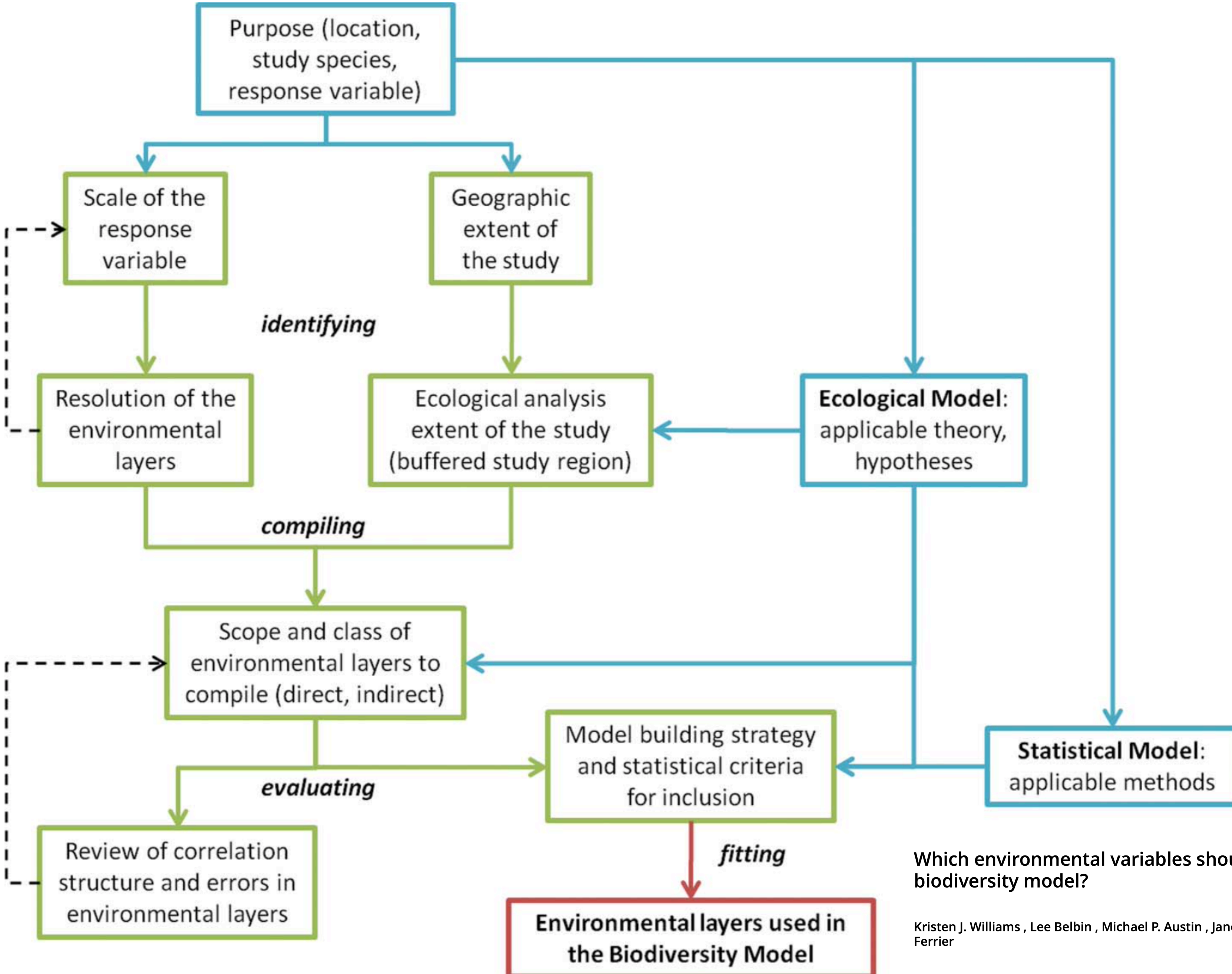
Guisan & Zimmermann

3. Model considerations

- Categorical data?

just because we can make a raster at a fine resolution, doesn't mean that it is accurate

categorical data will drive the algorithms available to you



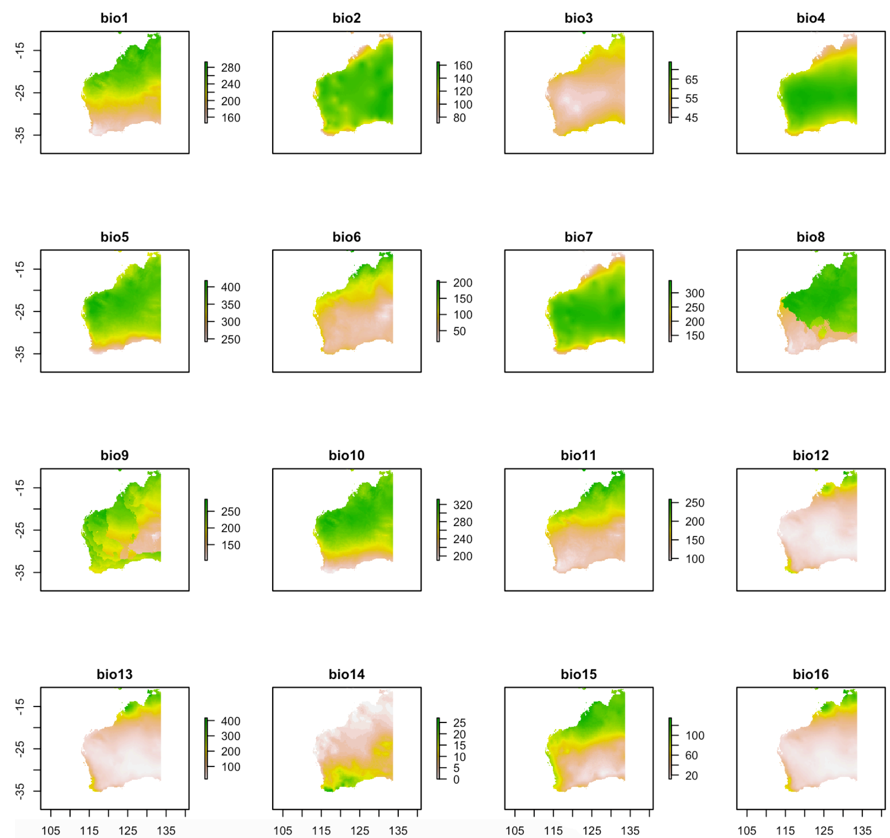
Which environmental variables should I use in my biodiversity model?

Kristen J. Williams , Lee Belbin , Michael P. Austin , Janet L. Stein & Simon Ferrier

To cite this article: Kristen J. Williams , Lee Belbin , Michael P. Austin , Janet L. Stein & Simon Ferrier (2012) Which environmental variables should I use in my biodiversity model?, International Journal of Geographical Information Science, 26:11, 2009-2047, DOI: 10.1080/13658816.2012.698015

Correlation, collinearity & variance inflation

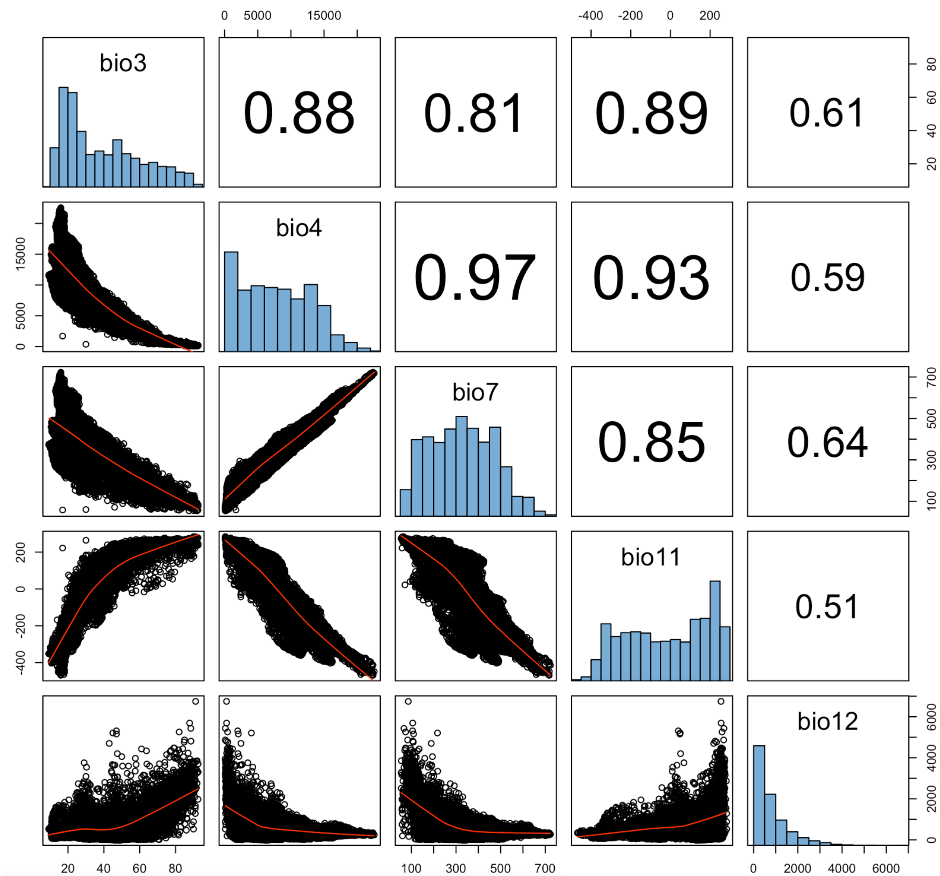
- Highly correlated variables will cause problems
  - Statistical inference
  - Interpretation
- Climate variables tend to be highly correlated
- Need to assess issues and remove problematic variables



variance inflation factor used in the homework

Correlation, collinearity & variance inflation

- Visualization
  - Pairwise correlations / plots
  - Remove if >0.7-0.8
  - May hide hidden structure



pick the variable to keep that is physiologically easier to understand

‘ecospat.cor.plot’

in ecospat package

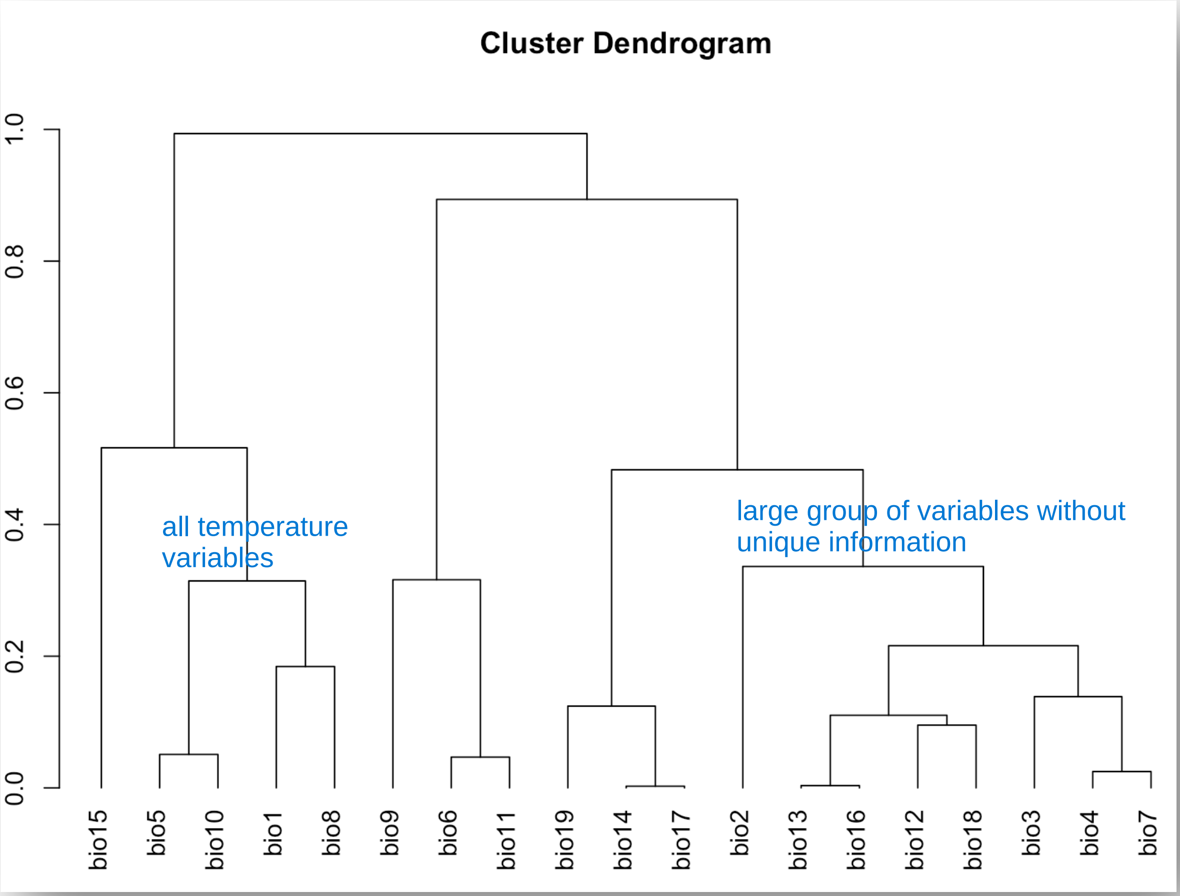
this is a univariate comparison that only compares pairs of variables at a time

univariate approaches can hide multicollinearity

Correlation, collinearity & variance inflation

- Visualization
  - Determine correlations
  - Cluster
  - Plot dendrogram

variables that tend to be correlated with one another tend to group



Correlation, collinearity & variance inflation

normally it's one variable that's multicollinear with all the other variables
VIFStep goes through this process sequentially removing the variables eith
highest collinearity

- Variance Inflation Factor (VIF) best way to do it
  - Measures extent to which variance in a regression increases due to collinearity compared to when uncorrelated variables are used
- Values > 10 (~20 maybe) problematic
- 'vif' and related commands in 'usdm' package

```
> library(usdm)
> vif(data[,4:8])
  Variables       VIF
1      bio3  6.873612
2      bio4 61.507510
3      bio7 31.857622
4     bio11 11.901976
5     bio12  2.151471
```

VIF score > 10 is problematic but can specify any variable

```
> vifstep(data[,4:8])
1 variables from the 5 input variables have collinearity problem:

bio4

After excluding the collinear variables, the linear correlation c
min correlation ( bio12 ~ bio11 ):  0.5183214
max correlation ( bio11 ~ bio3 ):  0.8917852

---------- VIFs of the remained variables --------
  Variables      VIF
1      bio3 5.848365
2      bio7 4.654822
3     bio11 6.975376
4     bio12 1.950816
```

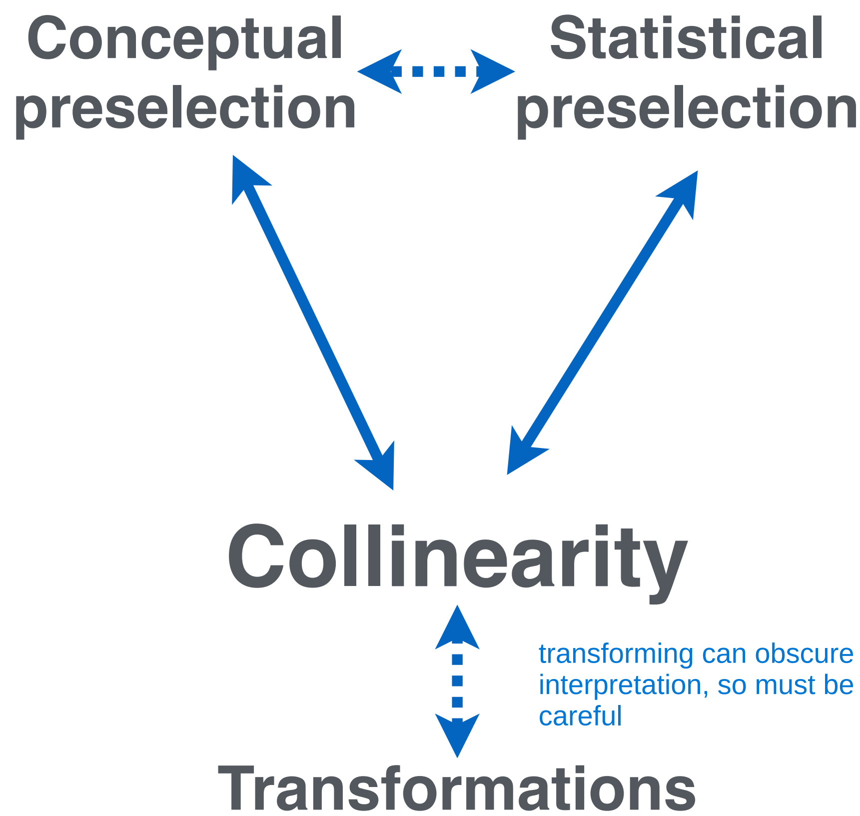
VIFStep continues until all variables are below 10

if using machine learning, these approaches are pretty insensitive to multicollinearity but the interpretation of importance is sensitive to multicollinearity

in machine learning, if you do a permutation test to see variable importance and which is driving the pattern, if there is collinearity the other variables will just stand in for the permuted variable, making it more difficult to see the impact of the permutation

Variable selection - summary

- **Use conceptual preselection to the extent possible**
    - What is important from:
        - Literature
        - Experiments
        - Expert knowledge
    - What can be removed from the outset as unimportant?
- **Use statistical preselection**
    - Methods like GBM
- **Check correlations using VIF**
- Blindly using all 19 bioclim variables not a good idea



sometimes people fit a PCA and then use the uncorrelated orthogonal to fit the model, but this can obscure interpretation

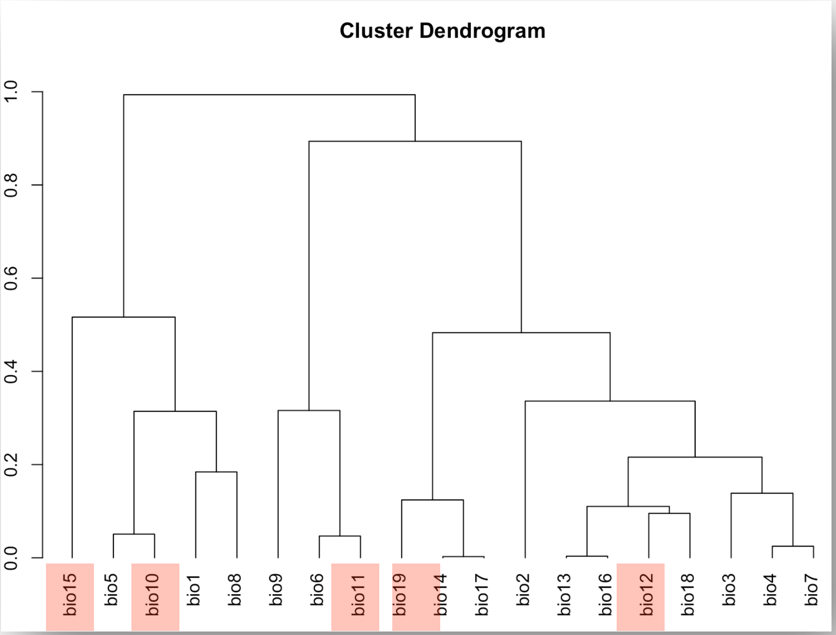
Variable selection - summary

how many variables should you use for parametric models (ie. GLM, GAM)?
You should have at least ten occurrence points per predictor variable
-- this rule is not for machine learning: can use as many as you want

ie if you have 20 points, only use 2 predictors

for rare species, people fit models with only 2-3 variables to avoid breaking the rule and then they combine the models into an ensemble model to have all the variables

- **Use conceptual preselection**
  - What is important from:
    - Literature
    - Experiments
    - Expert knowledge
  - What can be removed from the outset as unimportant?
- **Use statistical preselection**
  - Methods like GBM
- **Check correlations using VIF**
- Blindly using all 19 bioclim variables not a good idea



Questions?