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

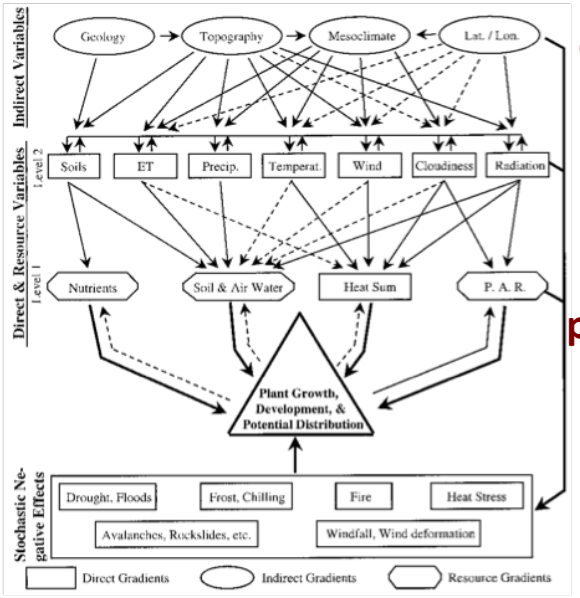


Fig. 3. Example of a conceptual model of relationships between resources, direct and indirect environmental gradients (see e.g. Austin and Smith, 1989), and their influence on growth, performance, and geographical distribution of vascular plants and vegetation.

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

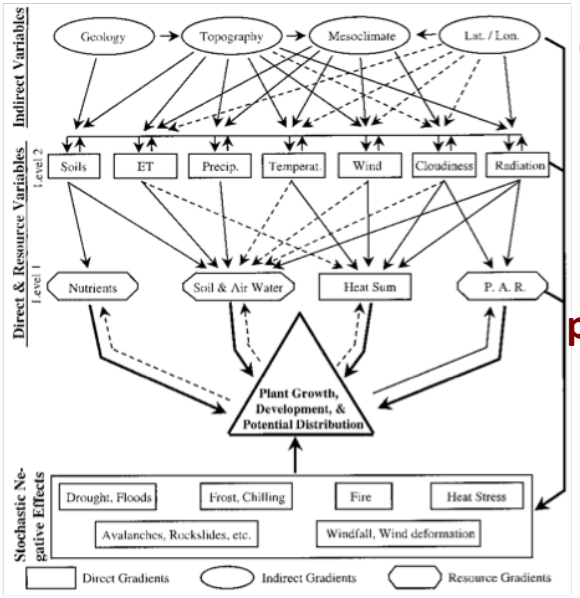


Fig. 3. Example of a conceptual model of relationships between resources, direct and indirect environmental gradients (see e.g. Austin and Smith, 1989), and their influence on growth, performance, and geographical distribution of vascular plants and vegetation.

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

3. Model considerations

- Categorical data?

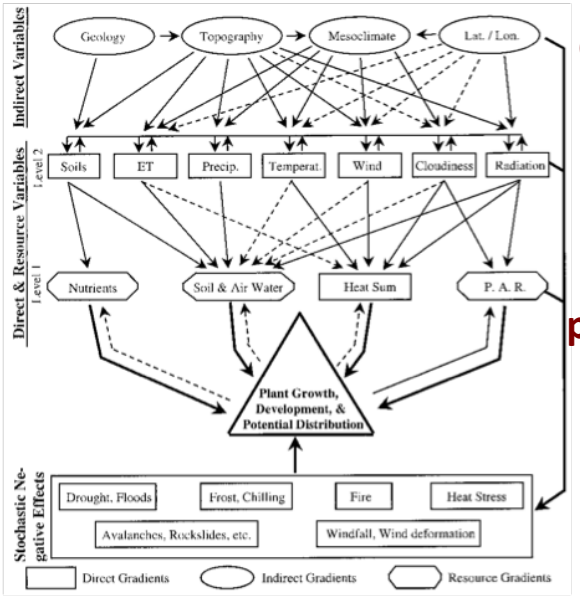
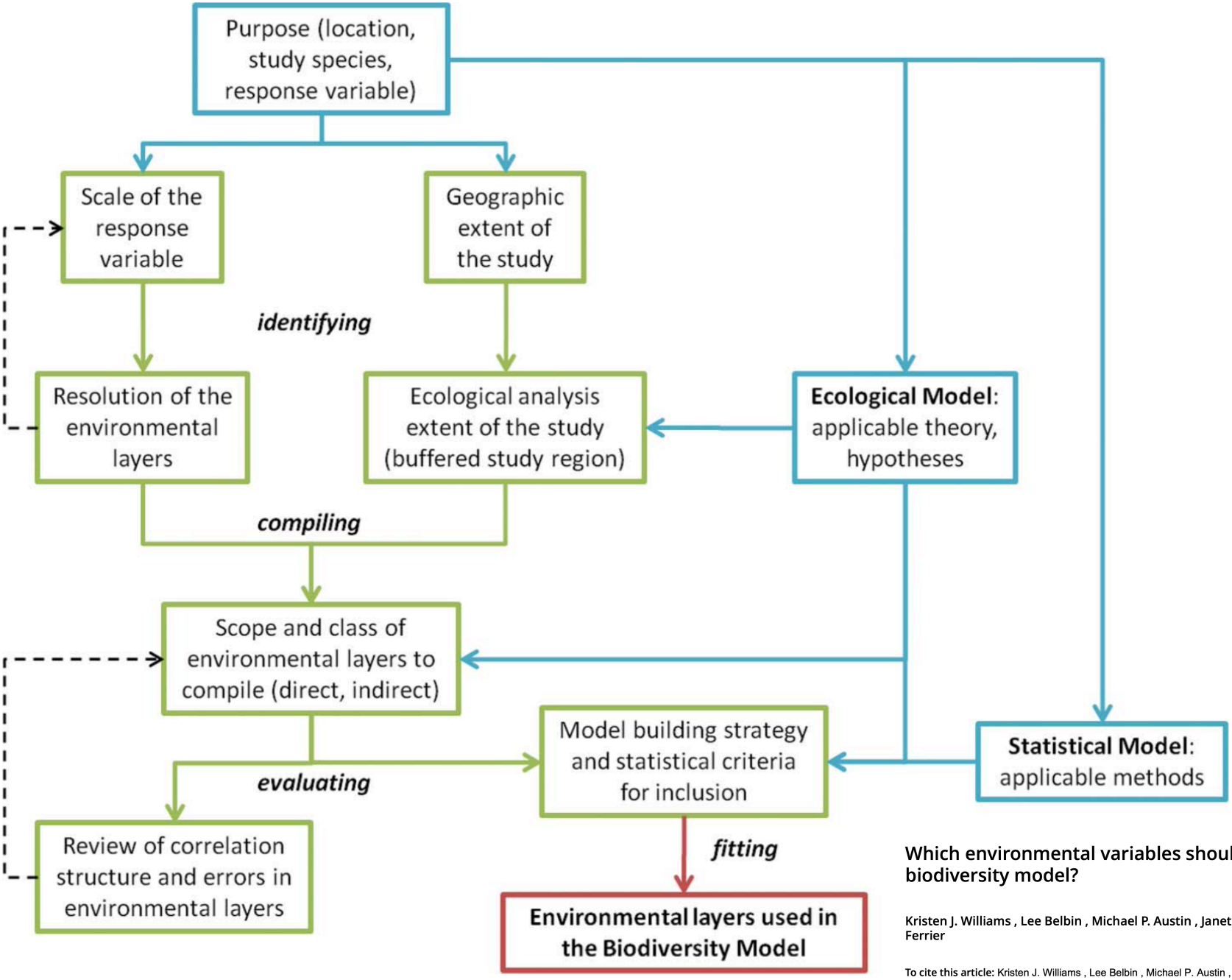


Fig. 3. Example of a conceptual model of relationships between resources, direct and indirect environmental gradients (see e.g. Austin and Smith, 1989), and their influence on growth, performance, and geographical distribution of vascular plants and vegetation.

Guisan & Zimmermann



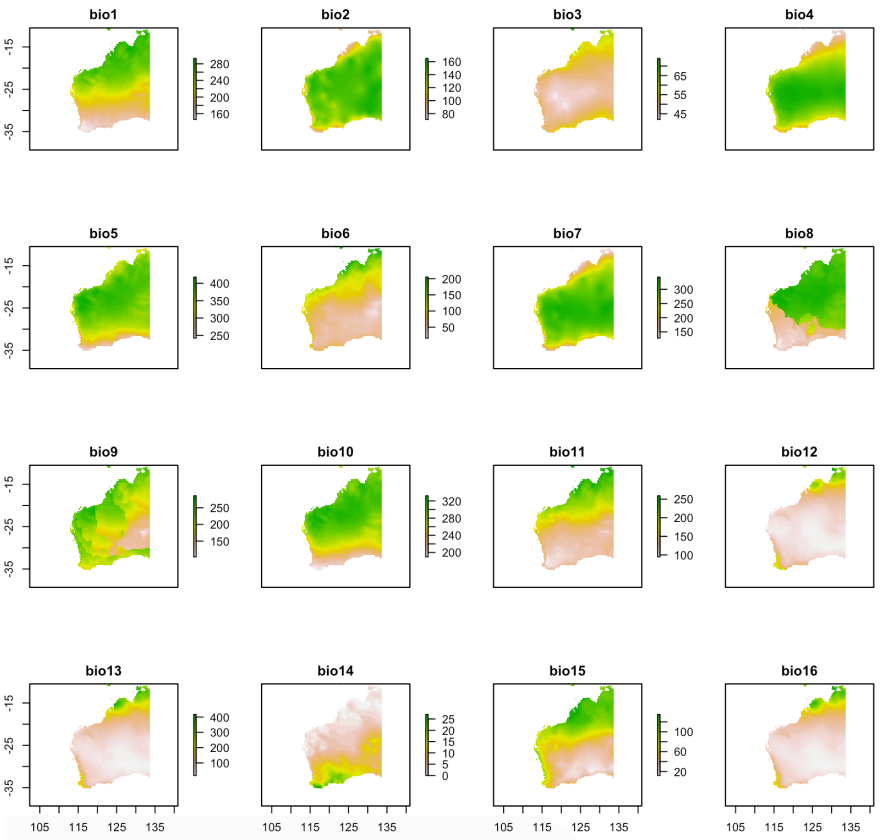
Which environmental variables should I use in my biodiversity model?

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

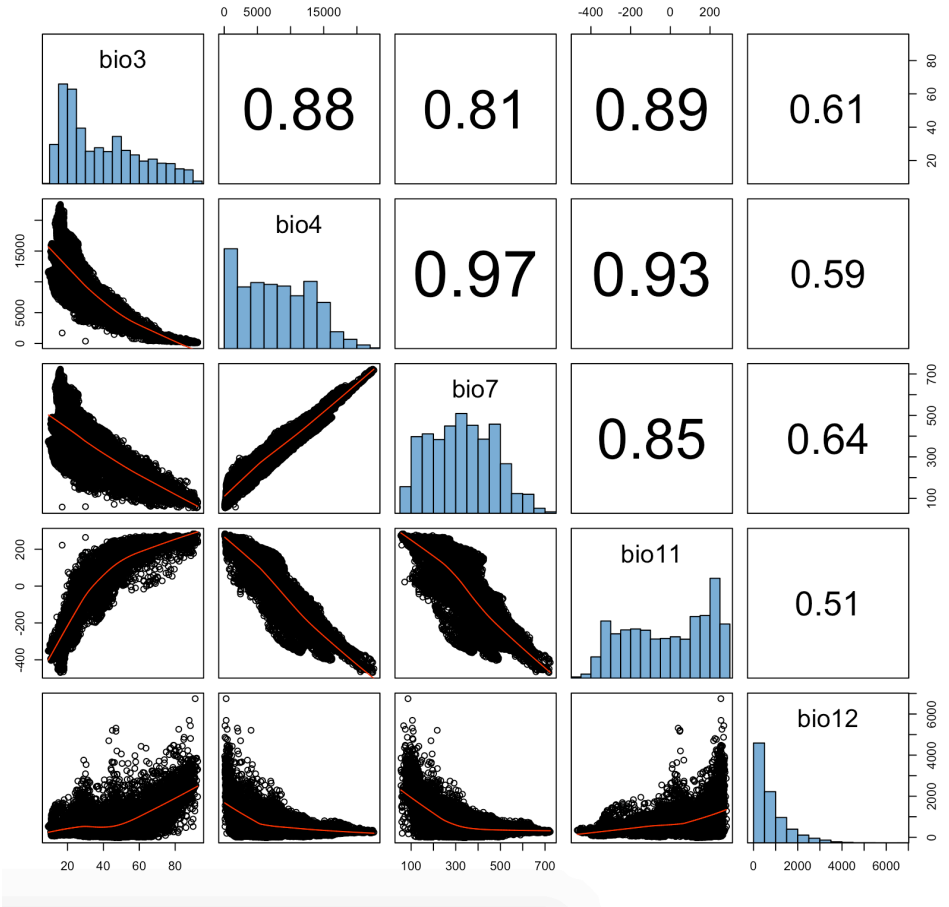
Correlation, collinearity & variance inflation

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



Correlation, collinearity & variance inflation

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

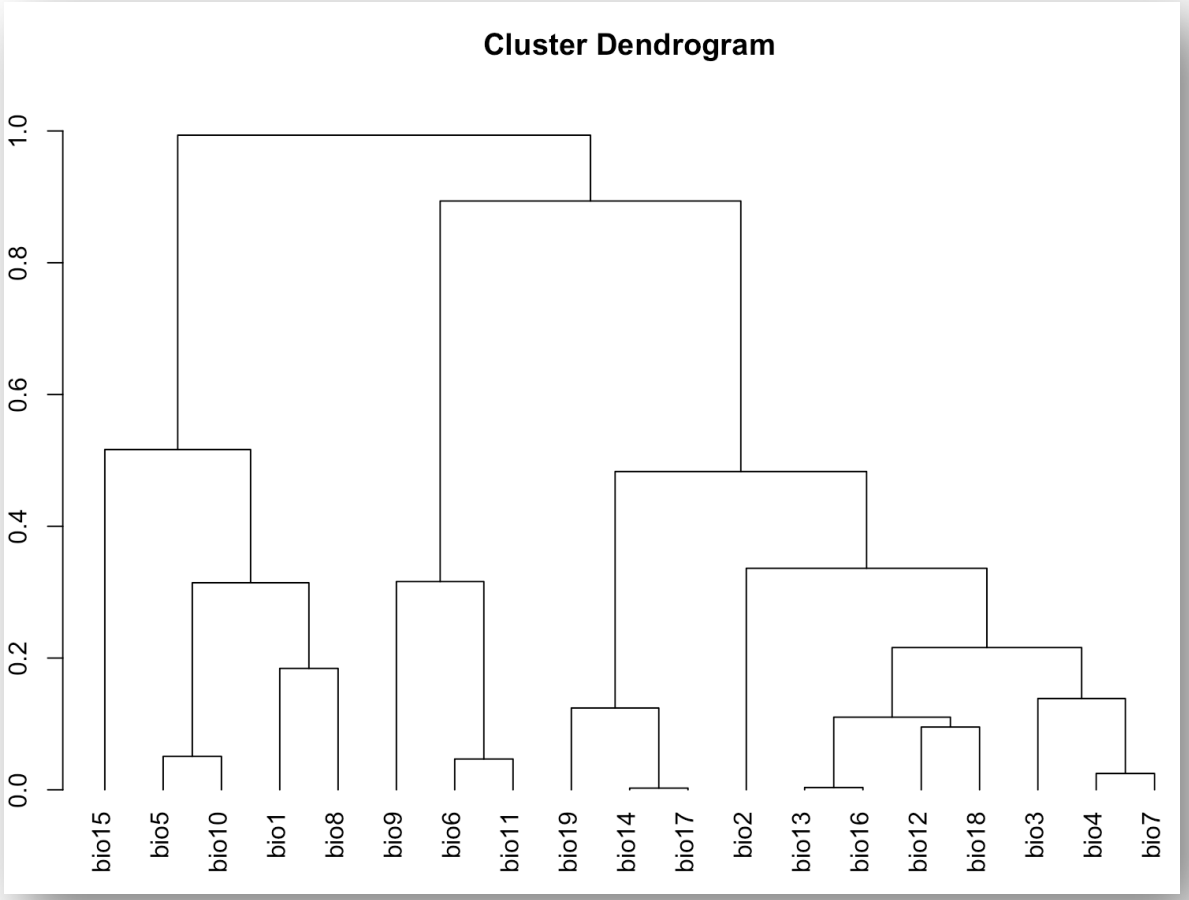


'ecospat.cor.plot'

in ecospat package

Correlation, collinearity & variance inflation

- Visualization
  - Determine correlations
  - Cluster
  - Plot dendrogram



Correlation, collinearity & variance inflation

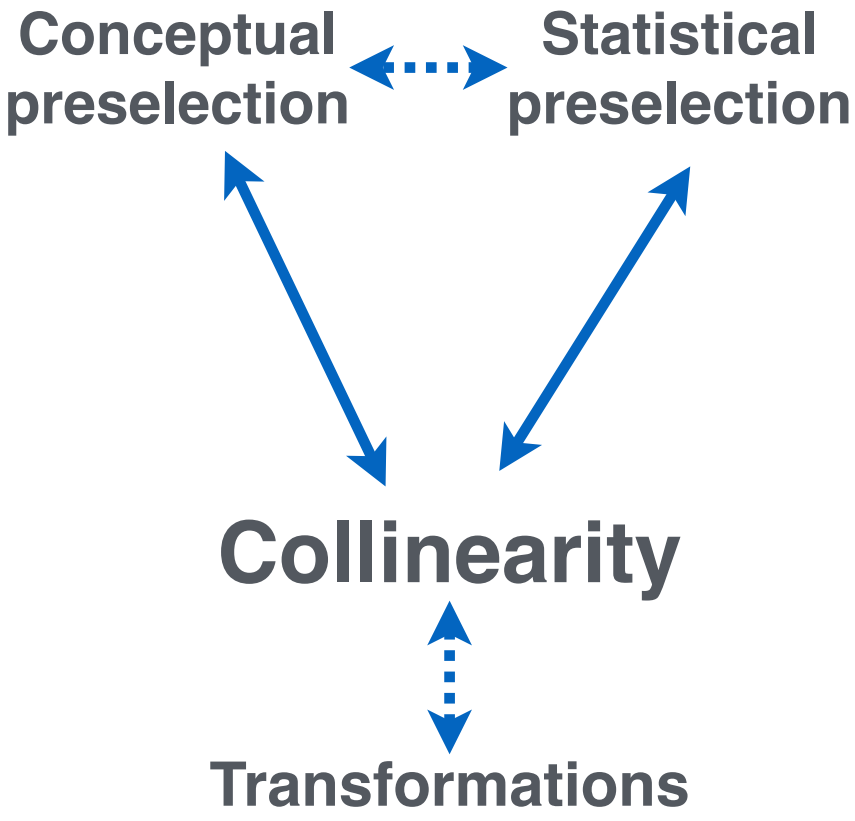
- Variance Inflation Factor (VIF)
  - Measures extent to which variance in a regression increases due to collinearity compared to when uncorrelated variables are used
- Values > 10 (~20 maybe) problematic
- ‘vif’ and related commands in ‘usdm’ package

```
> 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
```

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

bio4

After excluding the collinear variables, the linear correlation co
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
```

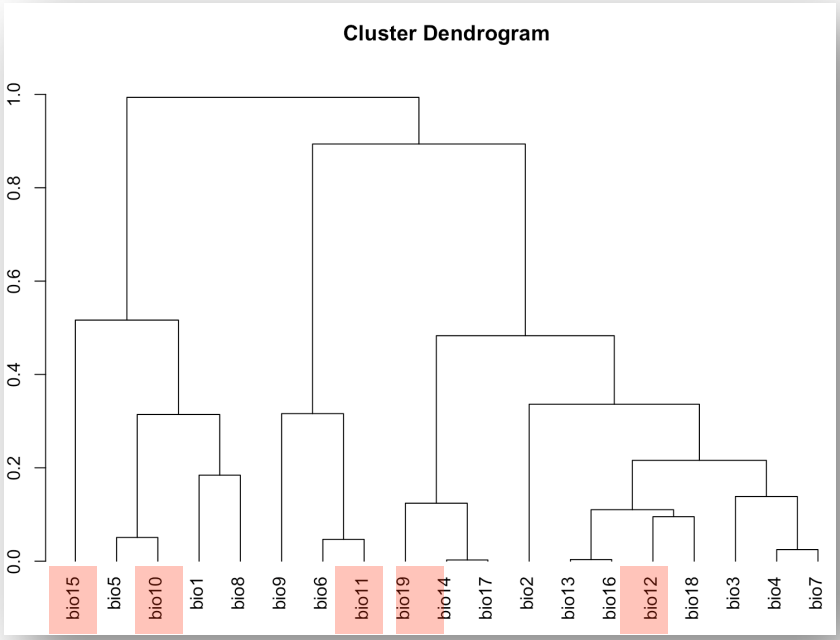
Variable selection - summary

- **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



Variable selection - summary

- **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?