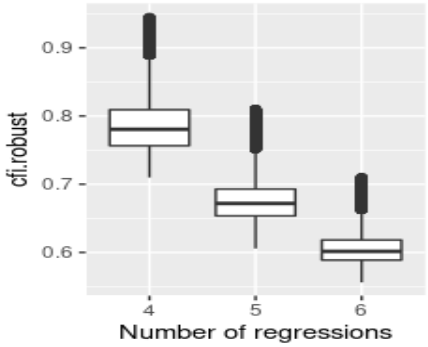
BRUTE FORCE SEM VARIABLE SELECTION

Important! Keep in mind that this checks initial model setup fit, not model fit that might happen when manually following the modifications function e.g. Therefore, this is a method to find a good starting point. Manual adjustments might still improve models, so consider also models with less good fit but more variables!

Also important: indirect paths make the model fit worse, filtering for goodness of fit is biased against including these paths/variables that cause inclusion(tra_sd+sea_sd). Check models with more indirect paths separately

Model fit is heavily affected by additional indirect paths

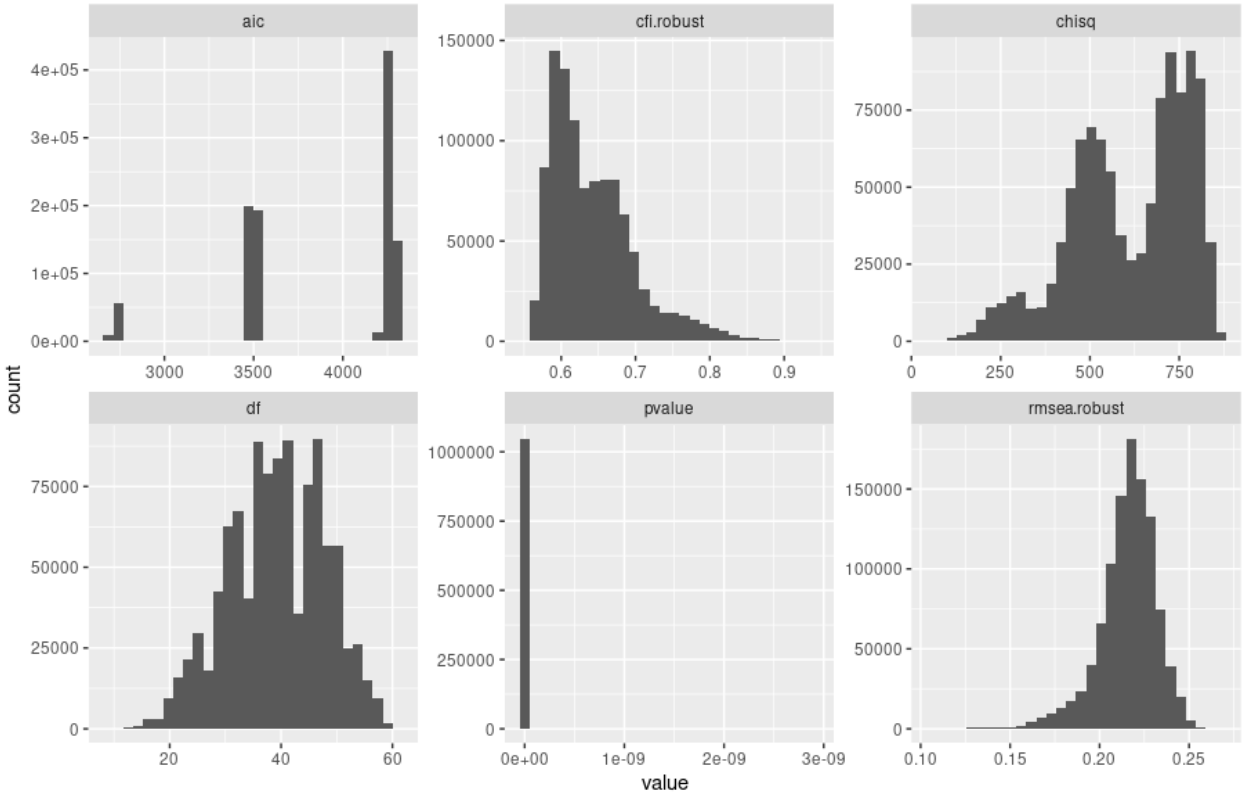


Possible TODO
- is there a way to incorporate modifications indices automatically that only uses modifications that make sense? Possible combinations would need to be defined beforehand, e.g. lists of parameters that can be affected by another list… maybe this goes too far.

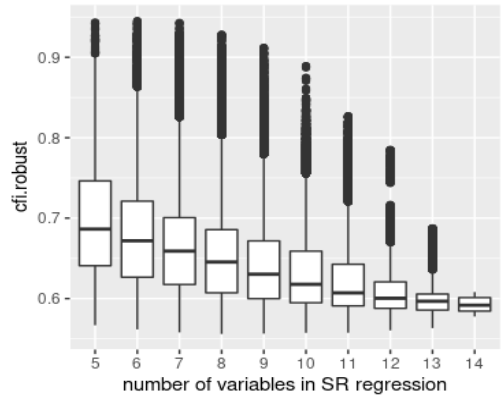
How it works

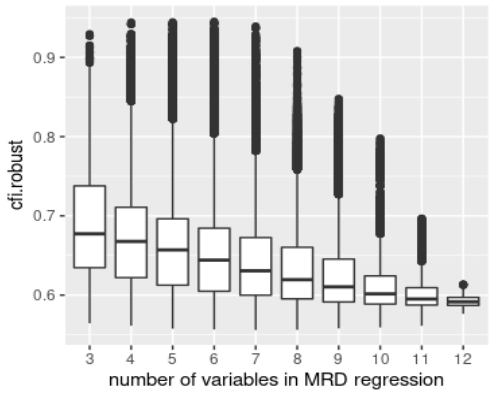
1. build fixed parts of regression formulas
   - sr_trans ~ soil + sub_trop_mbf + area + mrd
   - mrd ~ pre_m + sea_m
2. build vectors with potential variables to include in each regression separately
   - sr_variables <- c("sub_trop_dbf", "pre_sd", "mont_gs" , "pre_m", "sea_sd", "tra_m" , "mat_m" ,"tra_sd", "mat_sd" , "sea_m" , "medit_fws" , "temp_gss" , "tri" , "pet_sd" , "pet_m")
   - mrd_variables <- c("tra_m", "sea_sd", "soil" , "tri", "tra_sd" , "pre_sd" , "pet_m", "mat_m" , "area" , "mat_sd" , "temp_bmf" , "sub_trop_mbf" , "pet_sd" , "medit_fws" , "sub_trop_cf" , "deserts_x_shrub")
3. build all combinations for each regression formulas
4. build all combinations of all combinations each formula
   → you end up with 2^(sr variable number)*2^(mrd variable number)
5. define indirect path formulas based on correlations in correlation matrix + knowledge
   - dir_path_soil_area <- "soil ~ area" # always present
   - dir_path_sea_sd_area <- "sea_sd ~ area" # grepl("sea_sd", mod)
   - dir_path_tra_sd_area <- "tra_sd ~ area" # grepl("tra_sd", mod)
   - subtrop_path <- "sub_trop_mbf ~ pre_m + tra_m + pet_m" # always present
   - desert_path <- "deserts_x_shrub ~ pre_m + sea_m + pet_m" # if biome present
6. include secondary regression (indirect paths) according to if variables are present or not using grepl()
7. Collect goodness of fit parameters, Rquared for SR and MRD regressions, model formula
8. Run!

Some results from 2^10 * 2^10 combinations (first 10 most important variables)



- no p-value is non-significant
- 126 models (0.0019) have a robust CFI >0.9
- 0 models have an RMSEA <0.1 (22<0.11)
-





number variables in SR regression number variables in SR regression

4 best models (CFI>0.9 + maximal variable number)

> cat(fin[1])
sr_trans ~ soil + sub_trop_mbf + area + mrd + sub_trop_dbf+mont_gs+pre_m+tra_m+sea_m
mrd ~ pre_m + sea_m + tra_m+soil+tri+pet_m+area
soil ~ area
sub_trop_mbf ~ pre_m + tra_m + pet_m

> cat(fin[2])
sr_trans ~ soil + sub_trop_mbf + area + mrd + sub_trop_dbf+pre_m+tra_m+mat_m+sea_m
mrd ~ pre_m + sea_m + tra_m+soil+pet_m+mat_m+area
soil ~ area
sub_trop_mbf ~ pre_m + tra_m + pet_m
> cat(fin[3])
sr_trans ~ soil + sub_trop_mbf + area + mrd + mont_gs+pre_m+tra_m+mat_m+sea_m
mrd ~ pre_m + sea_m + tra_m+soil+pet_m+mat_m+area
soil ~ area
sub_trop_mbf ~ pre_m + tra_m + pet_m
> cat(fin[4])
sr_trans ~ soil + sub_trop_mbf + area + mrd + pre_m+tra_m+mat_m+sea_m
mrd ~ pre_m + sea_m + tra_m+soil+tri+pet_m+mat_m+area
soil ~ area
sub_trop_mbf ~ pre_m + tra_m + pet_m

More indirect paths
best model with max indirect paths (CFI=0.71)
sr_trans ~ soil + sub_trop_mbf + area + mrd + tra_m+tra_sd+sea_m
mrd ~ pre_m + sea_m + tra_m+sea_sd+soil+tra_sd+pet_m+area
soil ~ area
sub_trop_mbf ~ pre_m + tra_m + pet_m
sea_sd ~ area
tra_sd ~ area