10 Simple Rules for data analysis in the Biodiversity Exploratories

Synthesis Core Project 2021

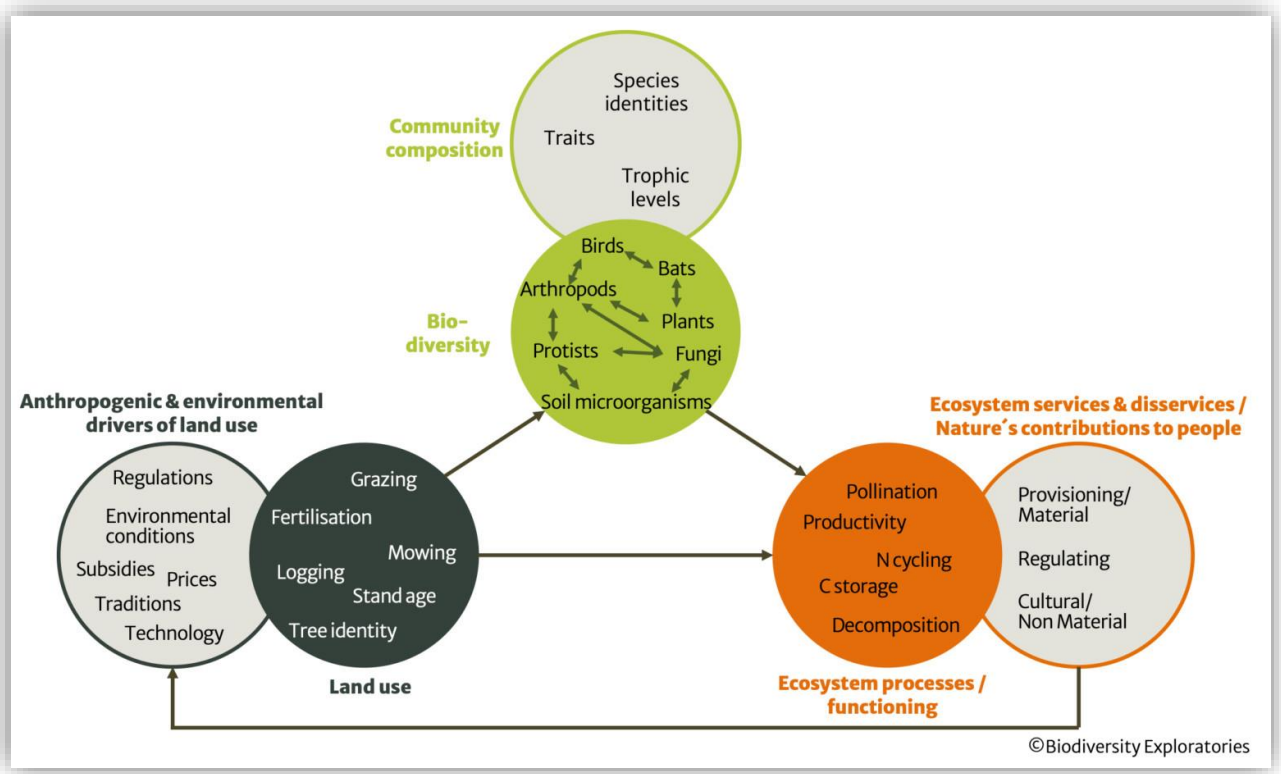
Why 10 simple «rules»?

- Share experience on common mistakes
- Common ground and starting point
- Make our studies more comparable
- Point out to resources
- Cover different steps of data analysis: from question to publication
- These are guidelines and food for thought, not hard rules

- Circulated to all explorers at next assembly: feedback welcome!

Note: the title and format is inspired by the nice PLOS Computational Biology collection (https://collections.plos.org/collection/ten-simple-rules/)

Rule 1: Think about how your question fits into the Biodiversity Exploratories framework and design



By design:

- **biodiversity**, **functions** and **services** (and their components) are response variables
- **LUI gradient** main explanatory variable
- (3 regions)
- (Multiple years)

- Where does your question fit in this framework?

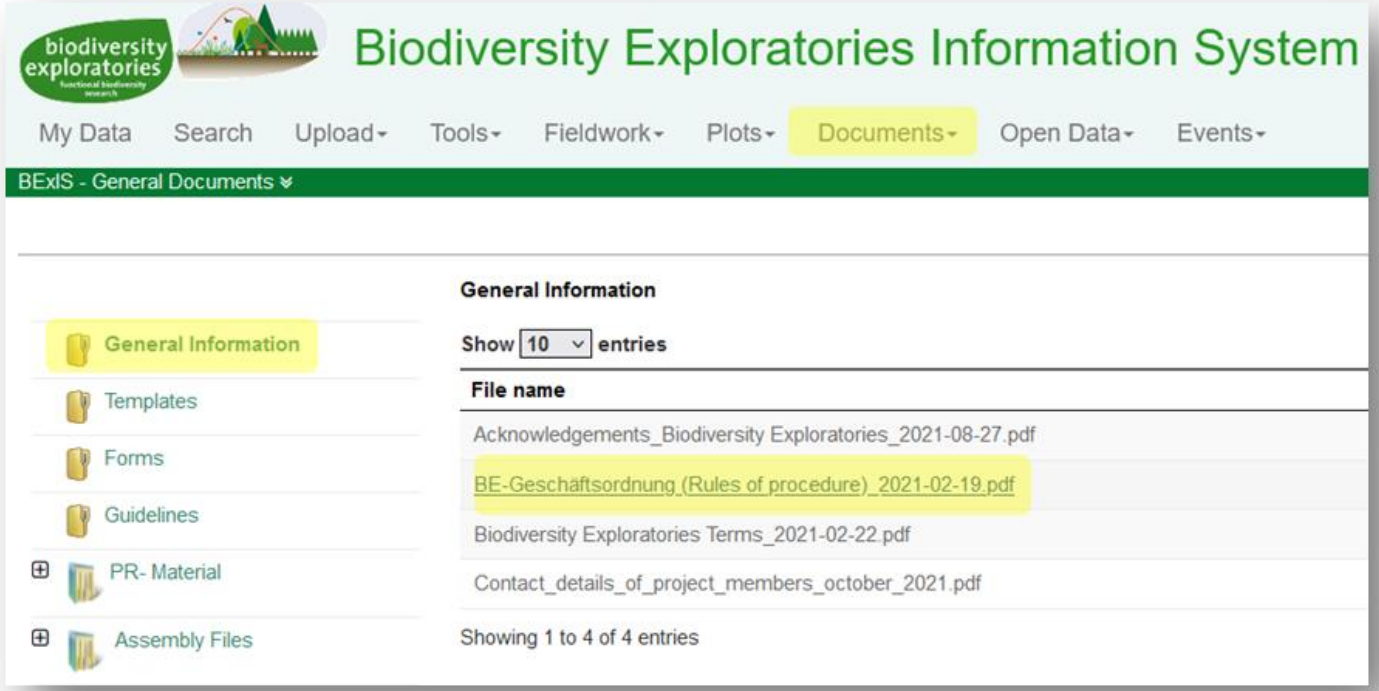
- What else is important in your study system?

- This will help identifying response and explanatory variables and types of analyses

A typical model will be: biodiversity or function or service ~ Region + LUI + covariates

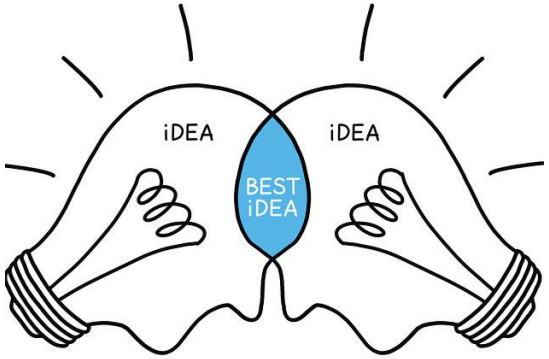
Rule 3: Read and follow the rules of procedure

- Data policy
  - Definitions
  - Data management
  - Creation and upload
  - Quality
  - Public / BE access
  - Synthesis datasets
  - Recommendations
  - Dataset release and DOI
- Publication
  - Acknowledgement of data suppliers
  - Co-authorship
  - Synthesis



https://www.bexis.uni-jena.de/FMT/GeneralFiles/Show?viewTitle=General%20Documents&viewName=GeneralFiles&rootMenu=BeoInformation

Rule 3: Involve the data owner(s)

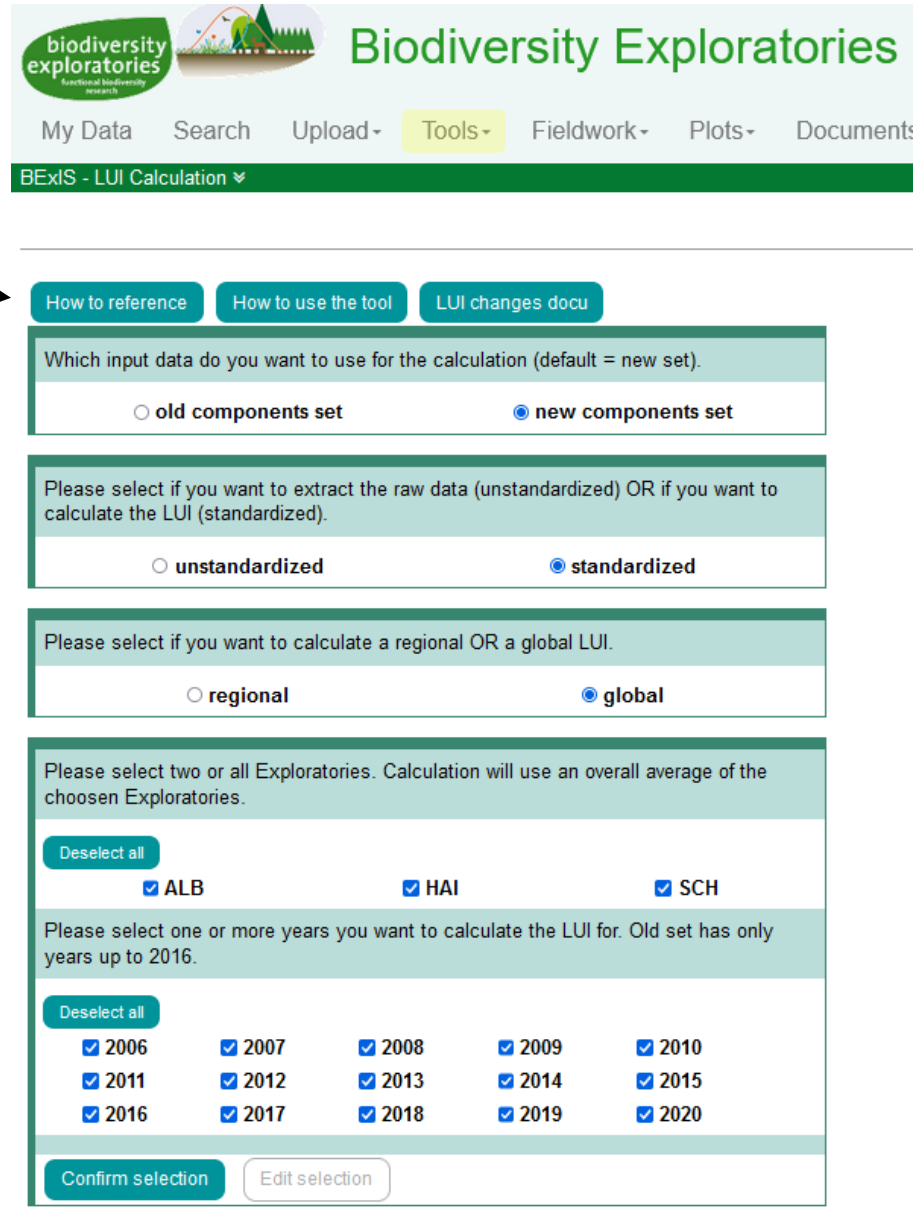


- Project designed to share data
- Almost always need other datasets
- Data owners/collectors = knowledge
  - Methods
  - Study system
  - Previous publications
  - Experience with the BE
- Involve them early to avoid surprises
- Co-authorship based on Rules of Procedure

Rule 4: Choose the right LUI calculation

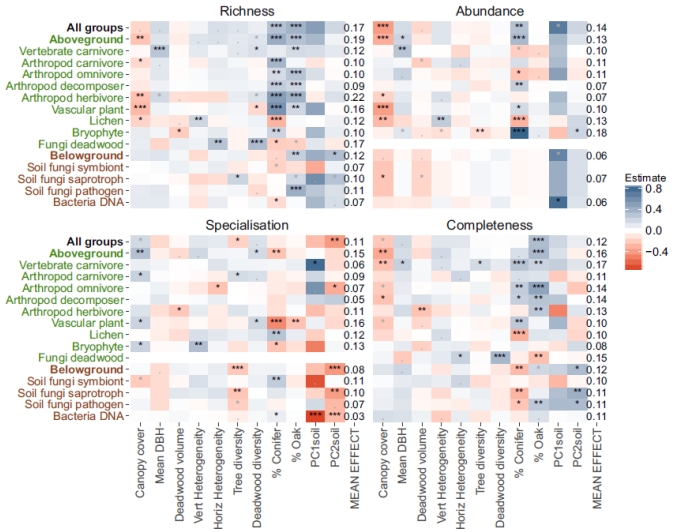
- Use the LUI tool in Bexis
- Read the documentation
- Use **new** components
- **Standardise**
- Global
- Regions: match your data
- Years: match your data (+ past?)
- Analyse LUI and components
- Mowing and fertilisation are correlated (by farmers practices)

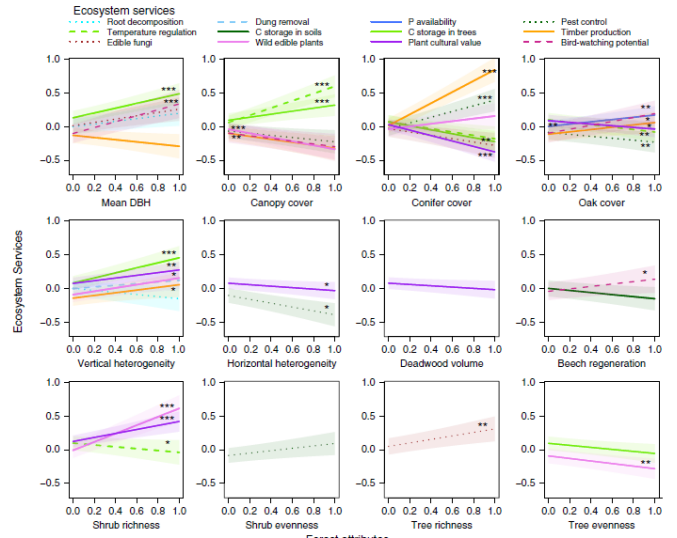
$$LUI(i) = \sqrt{\frac{G(i)}{G_{mean}} + \frac{M(i)}{M_{mean}} + \frac{F(i)}{F_{mean}}}$$



Rule 5: Understand well forest management and indexes

- Forest management is complex and multidimensional
- Forest Management Intensity ForMI: % of harvested tree volume (Iharv), % of non-native species (Inonat), % of dead wood with signs of saw cuts (Idwcut)
- Silvicultural management intensity indicator SMI: tree species, stand age and aboveground, living and dead wooden biomass
- ForMI and SMI results might be counterintuitive:
  - positive LUI effects because of conifers
- Most important variables for biodiversity and functions:
  - Composition (% beech, % oak, % conifers)
  - Canopy cover
  - Forest age
  - Vertical heterogeneity
- Depends on taxa, functions and considered dimensions





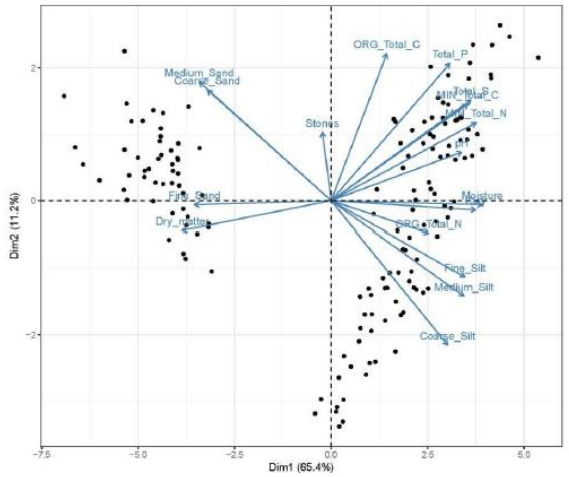
References: Penone et al 2019 Ecol. Lett. doi: 10.1111/ele.13182 and Felipe-Lucia et al 2018 Nat. Comm. https://doi.org/10.1038/s41467-018-07082-4.

Rule 6: Include the regions in your models but not as random factor

- Regions are important by design
- Several studies showed that results in Schorfeide-Chorin are different
- **Include region as a factor in your models**: it will be significant very often
- If you want to learn more about the study system or better understand or discuss some results: do separate models per region
- But be careful: 3 regions are not enough to study biogeographical patterns or make generalisations
- Do not include region as random effect in mixed models (unless you have a nested structure): random effects need many levels (>5) to be able to estimate a variance
- In SEMs: transform to two binary regions or use residuals (see slide at the end)

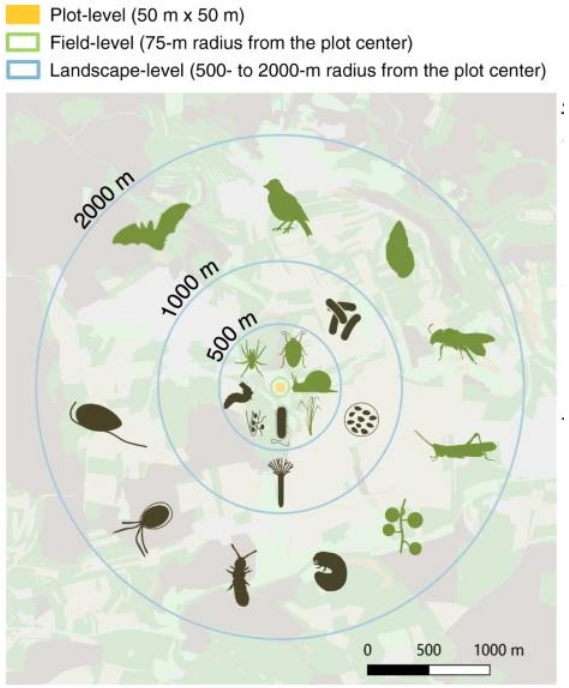
Rule 7: Consider including landscape and soil variables as covariates

- Soil is (obviously) key for soil organisms and functions but some aboveground groups also have a soil life stage
- If soil is not the focus it is possible to reduce its dimensionality by using PCA axes



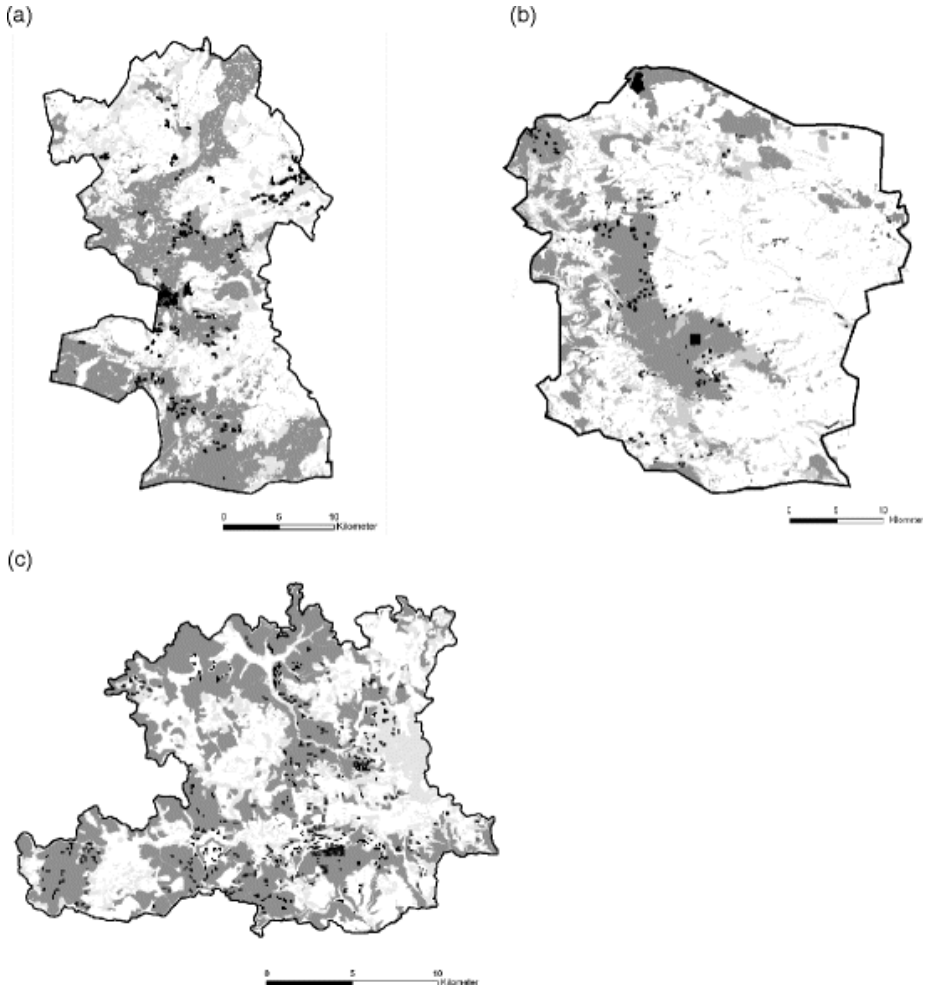
- Landscape composition and history are important for several trophic groups (and potentially functions)

See Le Provost et al 2021 Nat Comm
https://doi.org/10.1038/s41467-021-23931-1



Rule 8: Think about spatial autocorrelation

- Our plots are aggregated in three regions but also in smaller clusters
- Check for spatial autocorrelation in quantile **model residuals** (example here)
- Adding Region in models is also important to correct for spatial autocorrelation
- In general the *right* set of variables (e.g. soil, climate, altitude, landscape) will solve the problem
- If not: use gls or other models to address the issue, see Dormann et al 2007 https://doi.org/10.1111/j.2007.0906-7590.05171.x

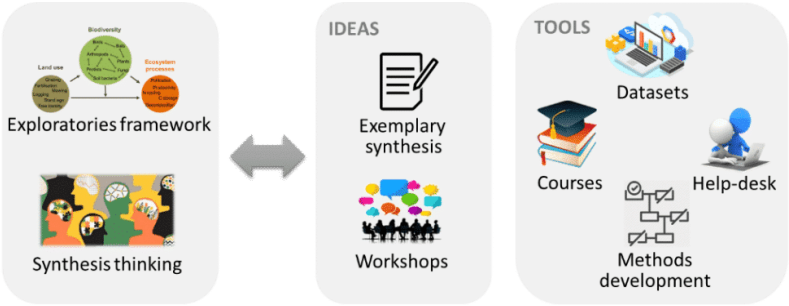


Rule 9: Use the different years wisely

- We have time series from 2006 to 2021 and time «points» (soil campaigns)
- Interest in change in time → analyse separately
  - Drop plots with incomplete data
- Interest in sample completeness → aggregate
  - Sum: sensitive to missing plots
  - Average: less sensitive to missing plots
- Include important temporal factors:
  - Changes in LUI
  - Climate!
- Think about temporal autocorrelation:
  - If < 5 years: include year as fixed effect
  - If > 5 years: include year as random effect

Rule 10: In doubt, ask the Synthesis helpdesk team!

- We have an overview on data / experts / stakeholders
- We do not know everything but we like new problems and are happy to help you figure things out!
- Service for the Exploratories: does not grant co-authorship



How:

- Approach us by email: noelle-schenk@ips.unibe.ch, caterina.penone@ips.unibe.ch, hugo.saizbustamante@ips.unibe.ch
- We can exchange code and data examples by email or Github
- We can do Zoom meetings
- If several people are interested: we can organise a course
- You can visit us

Further good general tips

- Don’t let the data decide: models should be hypothesis driven, not data driven
- Document your steps: from BExIS dataset to publication
- Use ordination analyses (PCA, NMDS, etc.) for data visualisation or reduction but robust analyses (e.g. linear models) for statistical tests
- Data transformation: scale if you want to compare effect sizes but keep units if you want interpretable results
- Continuous variables are preferred over factors/categorical ones
- If you use residuals (e.g. in SEM) correct both response and explanatory variables
- Use your network of colleagues, also outside of ecology

Examples of R code

- General types of models

```
mod <- lm(biomass ~ Region + LUI + plant_richness, data = BE dat)
mod <- lmer(biomass ~ Region + LUI + functional_group + (1|Plot) + (1|Species) + (1|Species:Plot), data = BEdat) #random group intercept
```

See here for model specification: http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#model-specification

- Soil PCA

```
PCAsoil <- ade4::dudi.pca(various_soil_variables, center = TRUE, scale = TRUE, scann=F)
PC1 <- PCAsoil$li[,1] #extract first Principal Component
```

- Residuals – **USE WITH CAUTION! (in most cases you won't need this and keep in mind that this decreases your statistical power)**

```
biomass_res <- residuals(lm(biomass ~ Region))
LUI_res <- residuals(lm(LUI ~ Region))
plant_richness _res <- residuals(lm(plant_richness ~ Region))
mod <- lm(biomass_res ~ LUI_res + plant_richness_res)
```

Examples of R code

- Test for spatial or temporal autocorrelation in model residuals

https://rdrr.io/cran/DHARMa/man/testSpatialAutocorrelation.html

https://rdrr.io/cran/DHARMa/man/testTemporalAutocorrelation.html

See general explanation here:

https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html

- Other resources:

https://www.highstat.com/index.php/beginner-s-guide-to-regression-models-with-spatial-and-temporal-correlation