Preparing HMSC model runs

## Species occurrence data processing

Starting from the raw data, I cleaned the data by selecting the native species that were present on both Crozet and Kerguelen islands.

As I explained in a previous document, the data are simple presence records that come from different sampling designs, such that the sites are not homogeneous in terms of surfaces. Since the target resolution of the JSDM predictions are to be of approx 20 x 20m, I discarded all the occurrences of species from sites with surfaces larger than 400m2. The median surface size is 100m2, since a lot of sites are standardized at 10 x 10m. I kept all the sites with inferior surfaces, because the risk of missing an observation decreases with plot size.

Each site is considered to have been visited only once, but in reality some sites overlap a lot (spatially, but then there are different sampling years).

These are all the observations since 2010.

library(tidyverse)  
library(Hmsc)  
library(V.PhyloMaker)  
library(brranching)  
library(raster)  
library(stars)  
  
theme\_set(theme\_bw())  
  
# this prepares the community matrices for Crozet and Kerguelen, but for now   
# let's look at Crozet  
crozet = TRUE  
source("process\_occurrences.R")

# look at matrix fill :   
  
# CROZET   
dim(cro\_com\_mat) # sites x species.

## [1] 1956 20

sum(cro\_com\_mat) # total number of observations

## [1] 11200

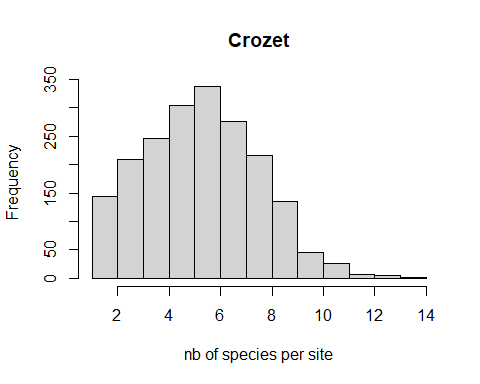
colSums(cro\_com\_mat) # nb of sites where each plant appears. Cf Limosella australis : remove it.

## Pringlea antiscorbutica Acaena magellanica   
## 488 1281   
## Agrostis magellanica Azorella selago   
## 1730 1277   
## Lycopodium magellanicum Uncinia compacta   
## 584 791   
## Galium antarcticum Phlegmariurus saururus   
## 545 243   
## Ranunculus biternatus Austroblechnum penna-marina   
## 1182 723   
## Juncus scheuchzerioides Crassula moschata   
## 527 84   
## Leptinella plumosa Poa cookii   
## 125 738   
## Montia fontana Callitriche antarctica   
## 354 107   
## Notogrammitis crassior Deschampsia antarctica   
## 87 286   
## Limosella australis Colobanthus kerguelensis   
## 2 46

colSums(cro\_com\_mat) / nrow(cro\_com\_mat) \*100 # proportion of sites where a plant is observed.

## Pringlea antiscorbutica Acaena magellanica   
## 24.9488753 65.4907975   
## Agrostis magellanica Azorella selago   
## 88.4458078 65.2862986   
## Lycopodium magellanicum Uncinia compacta   
## 29.8568507 40.4396728   
## Galium antarcticum Phlegmariurus saururus   
## 27.8629857 12.4233129   
## Ranunculus biternatus Austroblechnum penna-marina   
## 60.4294479 36.9631902   
## Juncus scheuchzerioides Crassula moschata   
## 26.9427403 4.2944785   
## Leptinella plumosa Poa cookii   
## 6.3905930 37.7300613   
## Montia fontana Callitriche antarctica   
## 18.0981595 5.4703476   
## Notogrammitis crassior Deschampsia antarctica   
## 4.4478528 14.6216769   
## Limosella australis Colobanthus kerguelensis   
## 0.1022495 2.3517382

rowSums(cro\_com\_mat) %>% hist(., main = "Crozet", xlab = "nb of species per site")



# KERGUELEN  
dim(ker\_com\_mat) # " 2450 sites, 20 species"

## [1] 2450 20

sum(ker\_com\_mat) # total number of observations

## [1] 9086

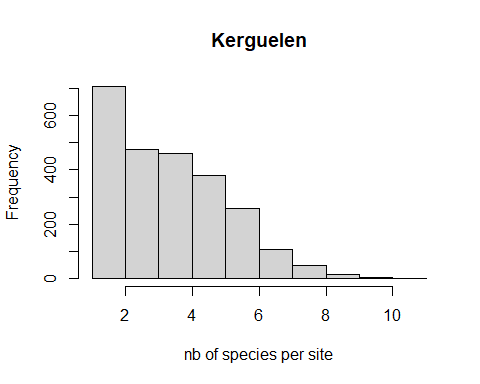
colSums(ker\_com\_mat) # nb of sites where each plant appears

## Crassula moschata Leptinella plumosa   
## 143 454   
## Ranunculus biternatus Montia fontana   
## 1147 572   
## Callitriche antarctica Acaena magellanica   
## 143 1790   
## Agrostis magellanica Juncus scheuchzerioides   
## 940 511   
## Azorella selago Galium antarcticum   
## 1161 447   
## Austroblechnum penna-marina Deschampsia antarctica   
## 312 365   
## Colobanthus kerguelensis Poa cookii   
## 156 126   
## Pringlea antiscorbutica Lycopodium magellanicum   
## 430 168   
## Limosella australis Uncinia compacta   
## 26 138   
## Phlegmariurus saururus Notogrammitis crassior   
## 44 13

colSums(ker\_com\_mat) / nrow(ker\_com\_mat) \*100 # proportion of sites where a plant is observed.

## Crassula moschata Leptinella plumosa   
## 5.8367347 18.5306122   
## Ranunculus biternatus Montia fontana   
## 46.8163265 23.3469388   
## Callitriche antarctica Acaena magellanica   
## 5.8367347 73.0612245   
## Agrostis magellanica Juncus scheuchzerioides   
## 38.3673469 20.8571429   
## Azorella selago Galium antarcticum   
## 47.3877551 18.2448980   
## Austroblechnum penna-marina Deschampsia antarctica   
## 12.7346939 14.8979592   
## Colobanthus kerguelensis Poa cookii   
## 6.3673469 5.1428571   
## Pringlea antiscorbutica Lycopodium magellanicum   
## 17.5510204 6.8571429   
## Limosella australis Uncinia compacta   
## 1.0612245 5.6326531   
## Phlegmariurus saururus Notogrammitis crassior   
## 1.7959184 0.5306122

rowSums(ker\_com\_mat) %>% hist(., main = "Kerguelen", xlab = "nb of species per site")



Questions :

* is there a threshold under which we consider a species to be too rare for the models to cope with ?

In HMSC school, they suggested 10-20 prevalence (=count occurrences in different sites), so I’ll go for that.

* should we decide to remove a species, we should remove it in both islands, even if it’s only rare in one of them ?

I haven’t so far.

## Traits

I tried to get some trait data from the TRY database, but I ran into a series of problems, and in the end I don’t think we’ll be able to incorporate traits in this way. The problem is the TRY database only allows to download 20 traits at a time for a given species, so you have to know which traits you want. I tried to download the traits that have been measured the most over the whole database and were meaningful for the list of native plants, but in our case, they don’t overlap very much. This means we would have to drop species out of the analysis. I’d rather try and compensate for this with the phylogeny data, rather than put too much weight on traits that aren’t even necessarily that relevant to the question we want to ask.

In any case, if we really need to incorporate traits, I’m going to need some help and talk with specialists from either TRY or ASICS plant traits, I can’t solve this alone…

## Phylogenies

The discussion I had with Claudine Montgelard a few moths ago was useful to help me understand some of the issues, but in the end the tools she suggested turned out to be unpractical : mostly online services, with options to tick based on her recommendations (and would have been tough for me to defend in a publication). Plus I did try and submit a request, but it never came back to me.

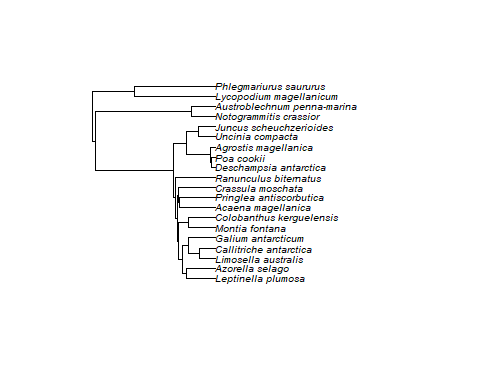
I have scripts to extract gene/protein sequences and build trees with those, but this requires to take many decisions I don’t really trust myself to do. This was what I went to see Claudine for in the 1st place, but she didn’t provide any straightforward answers to them.

In the end, I chose a method which seems to be the easiest way to do things, but it’s not ideal : I accessed a huge Tree of Life reference phylogeny for plants (Qian, H. & Jin, Y. (2016) Journal of Plant Ecology), it’s supposed to be the most up-to-date one there is. Then I simply pruned it to keep only the species I’m interested in. Claudine Montgelard said it should be OK…

Long story short : I have a tree, it’s probably somewhat rough but Hopefully it’ll do the job for exploratory model runs.

This is what it looks like :

load("../data/traits\_trees/phylomaker\_tree.RData")  
plot(result[[3]], cex = 0.6)



However, if you look at the “status” column in the following table, you’ll notice that 7/20 have a “prune” status, which means those plants found a match in the mega tree. The remaining 13 ones have “bind”. This means that a binding to the tree was possible, but it’s not a match. It’s probably not to be trusted.

result$species.list

## species genus family status  
## 1 acaena\_magellanica acaena rosaceae prune  
## 2 agrostis\_magellanica agrostis poaceae bind  
## 3 austroblechnum\_penna-marina austroblechnum blechnaceae bind  
## 4 azorella\_selago azorella apiaceae bind  
## 5 callitriche\_antarctica callitriche plantaginaceae bind  
## 6 colobanthus\_kerguelensis colobanthus caryophyllaceae bind  
## 7 crassula\_moschata crassula crassulaceae prune  
## 8 deschampsia\_antarctica deschampsia poaceae prune  
## 9 galium\_antarcticum galium rubiaceae bind  
## 10 juncus\_scheuchzerioides juncus juncaceae bind  
## 11 leptinella\_plumosa leptinella asteraceae bind  
## 12 limosella\_australis limosella scrophulariaceae prune  
## 13 lycopodium\_magellanicum lycopodium lycopodiaceae bind  
## 14 montia\_fontana montia montiaceae prune  
## 15 notogrammitis\_crassior notogrammitis polypodiaceae bind  
## 16 phlegmariurus\_saururus phlegmariurus lycopodiaceae bind  
## 17 poa\_cookii poa poaceae bind  
## 18 pringlea\_antiscorbutica pringlea brassicaceae prune  
## 19 ranunculus\_biternatus ranunculus ranunculaceae prune  
## 20 uncinia\_compacta uncinia cyperaceae bind

**My conclusion is that if we want to use a phylogeny, we’ll need to go back to the good old sequence comparison method.**

## Temperature layers and downscaling

We used Jonas’ downscaling method from the lapse rates. He sent me the lapse rate data used by CHELSA, ie monthly data spanning years 1979 to 2020. From this data, I calculated an annual mean.

This is one question I have: **wouldn’t it be better to average only over the months we’re interested in ?**

I’ll share what we did, because even though the output looks legit, I’d like your opinion on what we did.

For now let’s stick with Cro.

This is Jonas’ temperature downscaling formula : T = d-b\*(a-c), where:

a = the digital elevation model at the target resolution :

b = the lapserate (spline interpolated to the target resolution)

c = the 1km reference digital elevation model (spline interpolated to the target resolution). For CHELSA that GMTED2010

d = the 1km temperature layer (spline interpolated to the reference resolution).

The c layer is not clear to me…

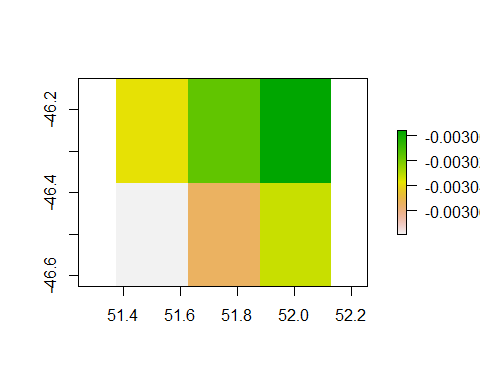
# 1. Load the mean annual laspe rates, crop over Crozet/Ker  
lapse <- raster("../data/Jonas/mean\_annual.tif") # 1 km mean annual lapse rates  
res(lapse) #

## [1] 0.25 0.25

# provide coordinates of crop window centered on Crozet:  
 box2 <- c(51.4, 52, -46.6, -46.2)   
   
 # crop   
 small\_lapse <- crop(lapse, box2)   
 plot(small\_lapse)

# this box provides a padding around the island : it's bigger than Crozet so   
# that I downscale from more than just 1 1km temperature pixel.   
box3 <- extent(small\_lapse) # the extent from box2 was extended to include all pixels

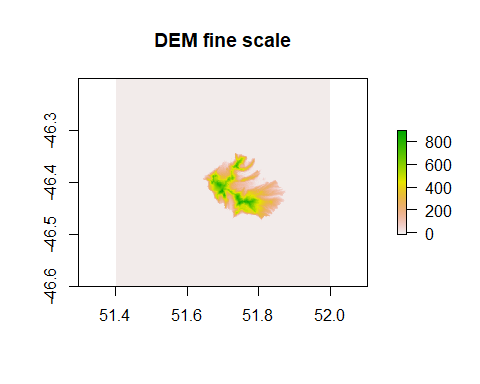
L’île de la Possession is somewhere on those middle giant pixels



# 2. Load high-resolution DEM, crop over Crozet/Ker  
 DEM <- raster("../data/DEM/cro/Cro\_DEM.tif")  
  
 # crop it to the same size as the lapse rates  
 small\_DEM <- crop(DEM, box2)  
 plot(small\_DEM, main="DEM fine scale")

res(small\_DEM)

## [1] 0.0002777778 0.0002777778



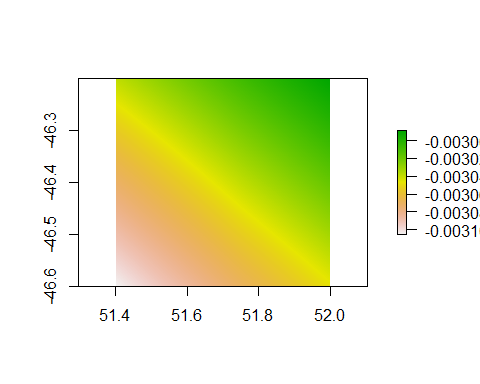
# 3. Spline-interpolate lapse rates by the high-res DEM  
s\_lapse\_dem <- resample(small\_lapse, small\_DEM, method = 'bilinear') # resample output  
res(s\_lapse\_dem)

## [1] 0.0002777778 0.0002777778

res(small\_DEM) # should be the same as s\_lapse\_rates

## [1] 0.0002777778 0.0002777778

plot(s\_lapse\_dem)



# now we have lapse rates at the super fine scale of the DEM...

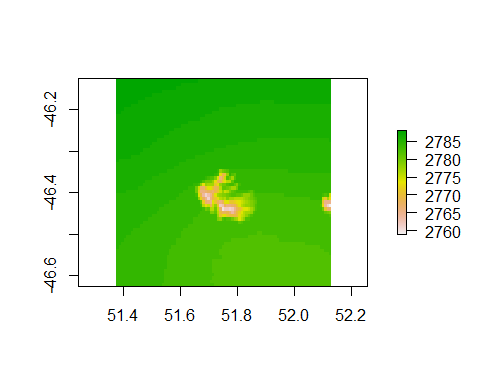
sunset paradise smoothie…

# 4. Load the CHELSA bioclim layer.   
# Here I used the monthly 1km temperature means, from 1981 to 2010.   
  
biolayer <- raster("../data/chelsa/CHELSA\_bio1\_1981-2010\_V.2.1.tif")  
res(biolayer)

## [1] 0.008333333 0.008333333

# crop it over Crozet :  
biolayer <- crop(biolayer, box3) # 1 km temp rates  
plot(biolayer)

I could have used the max or min temperatures too, I downloaded the rasters.



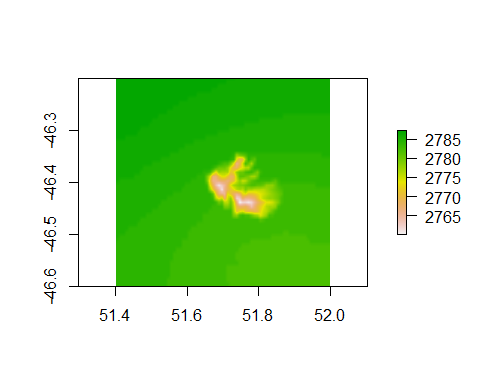
# 5. spline-interpolate lapse rates by the DEM  
s\_biolayer\_dem <- resample(biolayer, small\_DEM, method = 'bilinear') # resample output  
res(s\_biolayer\_dem)

## [1] 0.0002777778 0.0002777778

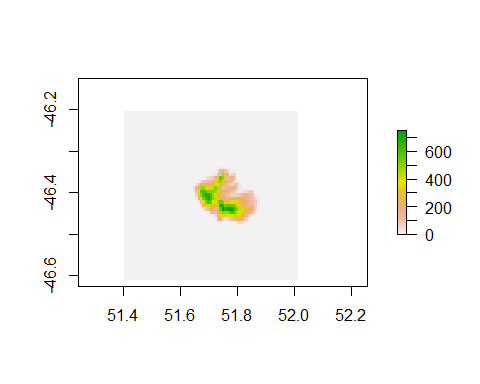
res(biolayer)

## [1] 0.008333333 0.008333333

# so now we also have the Chelsa temperatures downscaled at the very fine   
# resolution of the DEM  
  
plot(s\_biolayer\_dem)



# 6 : resample DEM at biolayer resolution.  
# this is because we don't have the c = the 1km reference digital elevation model (spline interpolated to the target resolution), so what Jonas did was to upscale the target resolution DEM up to 1km...  
s\_dem\_biolayer <- resample(small\_DEM, biolayer, method ='bilinear')  
plot(s\_dem\_biolayer)

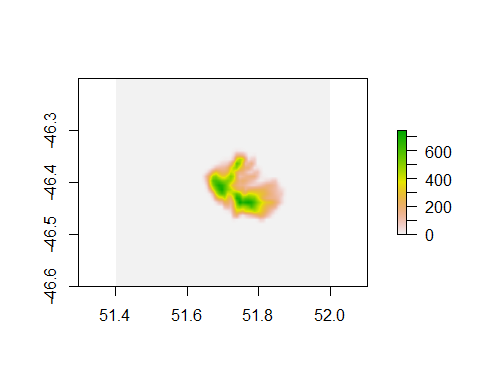


**Not sure this is kosher. I should perhaps use twhat CHELSA uses, GMTED2010 ? Not sure what this is but I’ll look.**

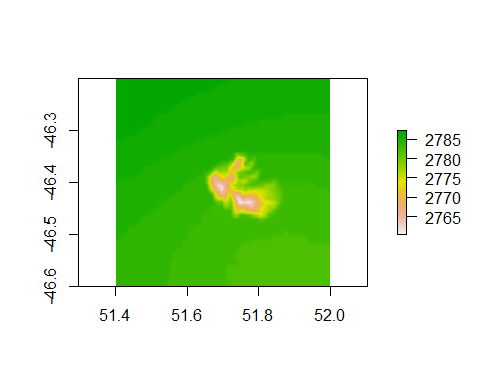
# 7. spline-interpolate the DEM at biolayer resolution at the DEM  
ss\_dem\_bio\_dem <- resample(s\_dem\_biolayer, small\_DEM, method='bilinear')  
res(ss\_dem\_bio\_dem)

## [1] 0.0002777778 0.0002777778

plot(ss\_dem\_bio\_dem)



# 8. Use the formula given by Jonas to calculate the downscaled temperature : T = d-b\*(a-c)  
# a = the digital elevation model at the target resolution  
# b = the lapserate (spline interpolated to the target resolution)  
# c = the 1km reference digital elevation model (spline interpolated to the target resolution). For CHELSA that GMTED2010  
# d = the 1km temperature layer (spline interpolated to the reference resolution).  
T\_downscaled <- s\_biolayer\_dem - s\_lapse\_dem \* (small\_DEM - ss\_dem\_bio\_dem)  
  
plot(T\_downscaled)  
  
  
small\_T <- crop(T\_downscaled, box2)  
plot(small\_T)



The much expected downscaled temperature raster! Looks alright, right?