Analyses for 2nd draft

Cape vs SWA

Ruan van Mazijk

2019-07-26

Preamble/outline

Here I layout the "new", second incarnation of the analyses as discussed over the course of May/June 2019, following the first draft of the manuscript.

To reiterate that manuscript, we hypothesise that the greater vascular plant species richness of the GCFR compared to that of the SWAFR is explained by the regions' difference in environmental heterogeneity.

The proposed "story" of questions for the analyses is as follows:

1. Is the GCFR more heterogeneous environmentally than the SWAFR, and does the scale of that heterogeneity differ to that of the SWAFR?
2. Do the regions differ w.r.t. the species richness of both HDS and QDS cells, and, for HDS cells' richness (S_{HDS}), does the explanatory power of mean QDS richness (S_{QDS}) and turnover (T_{QDS}) differ between the regions?
3. Does heterogeneity explain differences in richness and turnover between the regions?

1. Environmental heterogeneity & scale

Is the GCFR more heterogeneous environmentally than the SWAFR, and does the scale of that heterogeneity differ to that of the SWAFR?

In order to determine which region is more environmentally heterogeneous, and what scales heterogeneity is most pronounced, we calculated a measure of environmental heterogeneity at various spatial scales (namely: the base data resolution (0.05º x 0.05º), eighth- (EDS), quarter- (QDS), half- (HDS) and three-quarter-degree-squares (3QDS)).

Environmental "roughness" in both regions was calculated, in moving 3 x 3 cell windows, as the average absolute difference between cells and their (usually) 8 neighbours. Alternatively, for a focal cell x^*, the roughness is based on $x_1, x_2, \ldots, x_i, \ldots, x_8$ neighbour cells as:

$$Roughness(x^*) = f\begin{pmatrix} x_1 & x_2 & x_3 \\ x_4 & x^* & x_5 \\ x_6 & x_7 & x_8 \end{pmatrix} = \frac{1}{8}\sum_i |x^* - x_i|$$

In R, this is implemented this as follows:

```
roughness <- function(x) {
  raster::focal(x, matrix(1, nrow = 3, ncol = 3), function(x) {
    focal_cell <- x[5]
    focal_exists <- (!is.na(focal_cell)) & (!is.nan(focal_cell))
    if (focal_exists) {
      neighbour_exists <- (!is.na(x)) & (!is.nan(x)) & (x != focal_cell)
      neighbour_cells <- x[neighbour_exists]
      return(mean(abs(focal_cell - neighbour_cells)))
    } else {
      return(NA)
    }
  })
}
```

Following this, the various forms environmental heterogeneity were ordinated using principal component analysis (PCA), following $log(x+1)$-transformations to ensure normality, to summarise a major axis of heterogeneity in each region (Figure 1). Portions of the data matrices for each scale for these PCAs are shown in Table 1.

Biplots like those shown in Figure 1 are often used visualise a summary of the multidimensional environmental space encompassed by a geographic area. However, as our data are "roughness" values, it would be more accurate to refer to the biplots in Figure 1 as the heterogeneity spaces or heterogeneity envelopes of each geographic region, at five spatial scales. We can see clear separation between the GCFR and SWAFR in their heterogeneity envelopes across all five spatial scales. Of most importance here is the first principle component (PC1) of these spaces, where the GCFR is almost always as rough or more roughn than the SWAFR (Figure 1). W

Both the actual environmental heterogeneity values and the principal component of heterogeneity were then compared between the GCFR and SWAFR using common language effect sizes ($CLES$). The $CLES$ of GCFR vs SWAFR heterogeneity values was regressed against the spatial scale at which it was calculated using simple linear regression (Figure 2, Table 2).

We can see that PDQ, NDVI, pH and, arguably, elevation (Figure 2a,c,e,i) are all consistently more heterogeneous in the GCFR than in the SWAFR, regardless of spatial scale (Figure 2). The GCFR is more heterogeneous at finer scales in terms of MAP, surface temperature, CEC and soil carbon (Figure 2b,d,f,h). Notably bucking the trend, the GCFR is more pronouncedly heterogeneous at broad scales in terms of clay (Figure 2)—perhaps something to do with the Succulent Karoo vs CFR?. In general (i.e. regarding PC1; Figure 2j), the GCFR is more environmentally heterogeneous than the SWAFR, and particularly so at fine spatial scales.

We can conclude, then, that the GCFR is more environmentally heterogeneous than the SWAFR, across multiple environmental axes. Generally, the GCFR is more finely scaled in its heterogeneity, though some variables show no scale-dependence, and heterogeneity in clay is greatest in the GCFR at broad scales.

Table 1: Portions of the data matrices used in the PCA for this section of the analysis, where roughness values were $log(x+1)$-transformed to ensure normality.

region	Elevation	MAP	PDQ	Surface.T	NDVI	CEC	Clay	Soil.C	pH
GCFR	5.19	2.52	0.72	1.32	15.13	1.14	1.2	2.46	1.36
GCFR	5	2.7	0.61	1.16	15.01	1.11	1.11	1.74	1.83
GCFR	4.86	2.55	0.72	1.17	15.08	1.18	1.4	1.79	1.65
...	...	...	...	...	...	...	...	...	...
SWAFR	3.27	2.77	1.1	0.71	14.91	0.31	1.19	1.59	0.48
SWAFR	2.36	2.41	1.15	0.7	14.28	0.67	1.29	2.03	1.3
SWAFR	2.86	1.98	1.17	1.09	13.58	0.73	2.27	2.4	2.58

Table 2: Slopes and associated P-values from simple linear regressions of $CLES$ against scale for each form of environmental roughness (Figure 2).

Variable	Slope	P	
Elevation	0.044	0.016	*
MAP	-0.313	0.020	*
PDQ	0.010	0.387	
Surface.T	-0.330	0.026	*
NDVI	0.032	0.459	
CEC	-0.126	0.063	.
Clay	0.243	0.013	*
Soil.C	-0.298	0.003	*
pH	-0.010	0.756	
PC1	-0.172	0.010	*

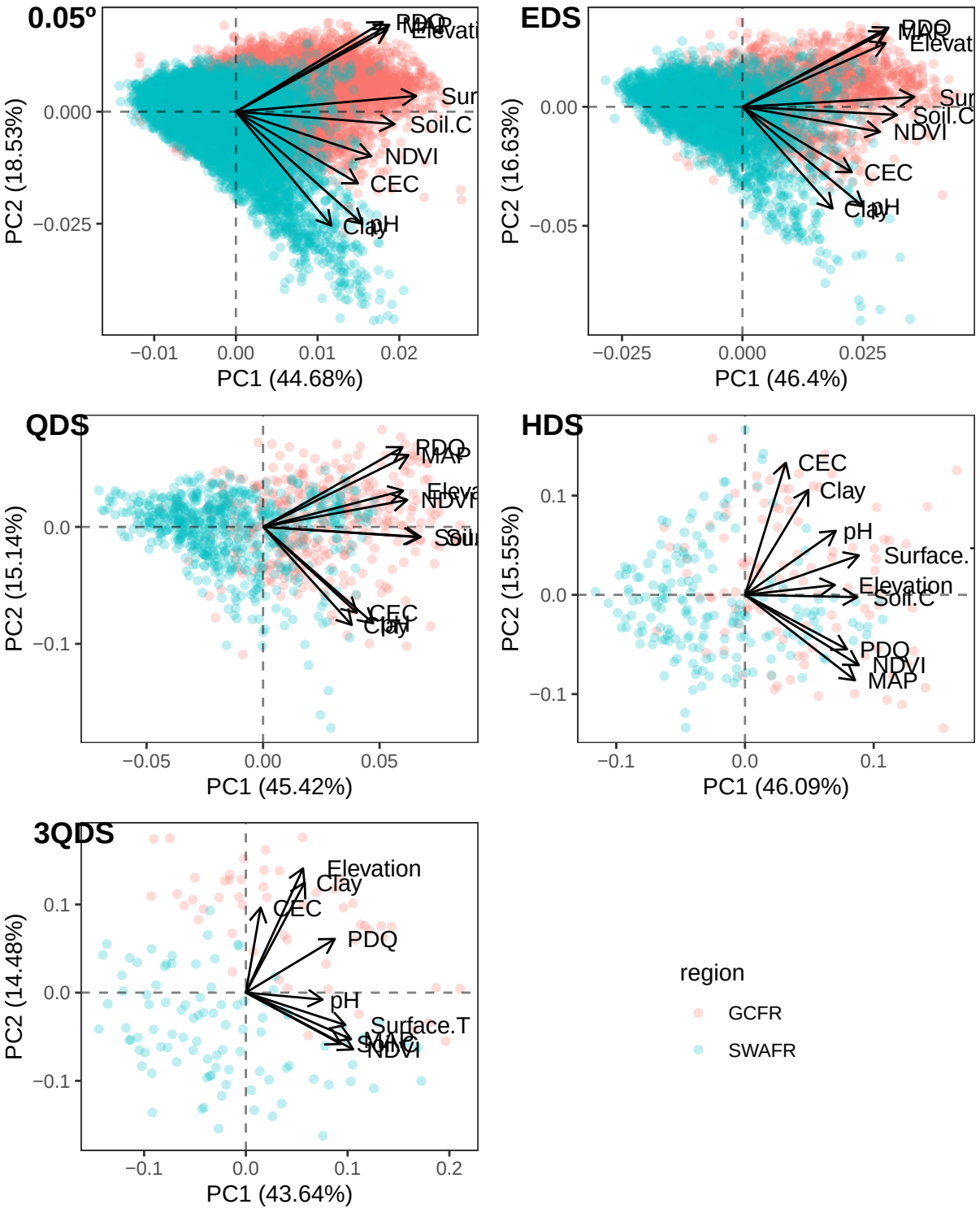


Figure 1: Scatter plots of the first and second principal components (PC1, PC2) of environmental heterogeneity following principal components analyses (PCA) of the various forms of environmental heterogeneity, $log(x+1)$-transformed (Table 1), repeated at the five spatial scales. The proportion of variation accounted for by each axis is denoted in parentheses. Arrows (labelled) denote the rotational loading of a given form of environmental heterogeneity. Note, the signs of loadings on PC1 have been forced to be positive, while the signs of loadings on PC2 are arbitrary. Here, greater values along PC1 represent greater overall environmental heterogeneity in an area. An area's PC2 value distinguishes the environmental axes most responsible for its overall observed heterogeneity.

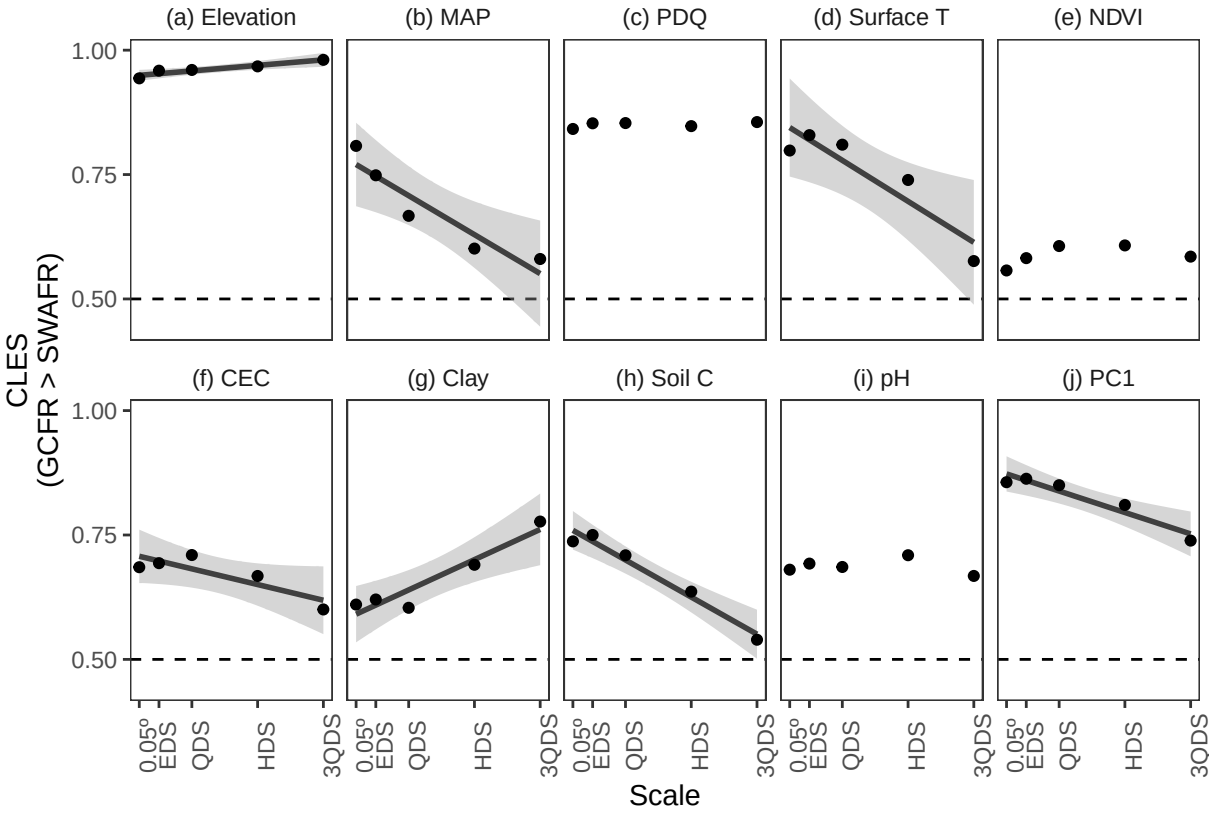


Figure 2: Simple linear regressions of the common language effect size ($CLES$) of various forms of environmental heterogeneity (a–i), and the first principal component of heterogeneity (j; see Figure 1), where the $CLES$ is treated as the effect of GCFR relative to SWAFR values. Only significant or marginally significant fits are plotted (Table 2). Grey bands denote 95% confidence intervals about the fitted lines. Across spatial scales, all $CLES$ values differed significantly from zero following two-sided t-tests ($P < 0.001$).

2. Species richness & turnover

Do the regions differ w.r.t. the species richness of both HDS and QDS cells, and, for HDS cells' richness (S_{HDS}), does the explanatory power of mean QDS richness (S_{QDS}) and turnover (T_{QDS}) differ between the regions?

To tackle this question, I compare measures of species richness and turnover between the regions. Species richness at the HDS-scale (S_{HDS}) can be partitioned into the average richness of the constituent QDS in HDS ($\overline{S}_{\text{QDS}}$) and species turnover ($T_{\text{QDS}}$) defined[1] as:

$$T_{\text{QDS}} = S_{\text{HDS}} - \overline{S}_{\text{QDS}}$$

The distributions of these data are presented in Figure 3b. To test for significant differences between GCFR and SWAFR values, I use Mann-Whitney U-tests and $CLES$ (Table 3), as most of the variables deviate significantly from normality (Shapiro-Wilk normality test; $P < 0.05$).

Additionally, a visualisation of how S_{HDS} is partitioned into $\overline{S}_{\text{QDS}}$ and T_{QDS} is presented in Figure 3a.

We can conclude that broad scale species richness (i.e. that at the HDS scale) is more strongly driven by turnover between areas (i.e. QDS) than so in the SWAFR.

[1] following Whittaker's original additive definition: $\gamma = \alpha + \beta$

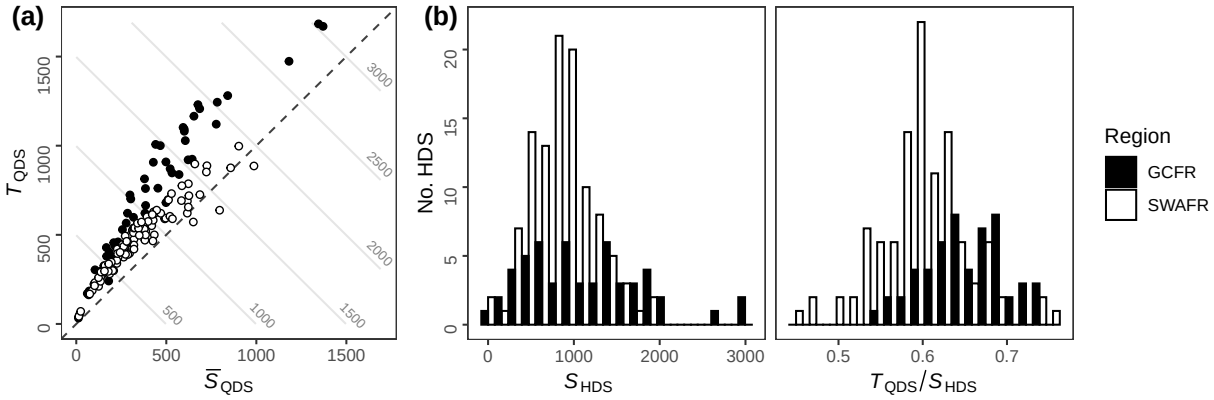


Figure 3: (a) Scatter plot of mean QDS-scale richness ($\overline{S}_{QDS}$) and turnover (T_{QDS}) with contour lines denoting the S_{HDS} that would arise as their sum (i.e. increasing from lower-left to upper-right). Distributions of (a) HDS-scale species richness (S_{HDS}) and (b) the turnover partition of that richness (T_{QDS}/S_{HDS}).

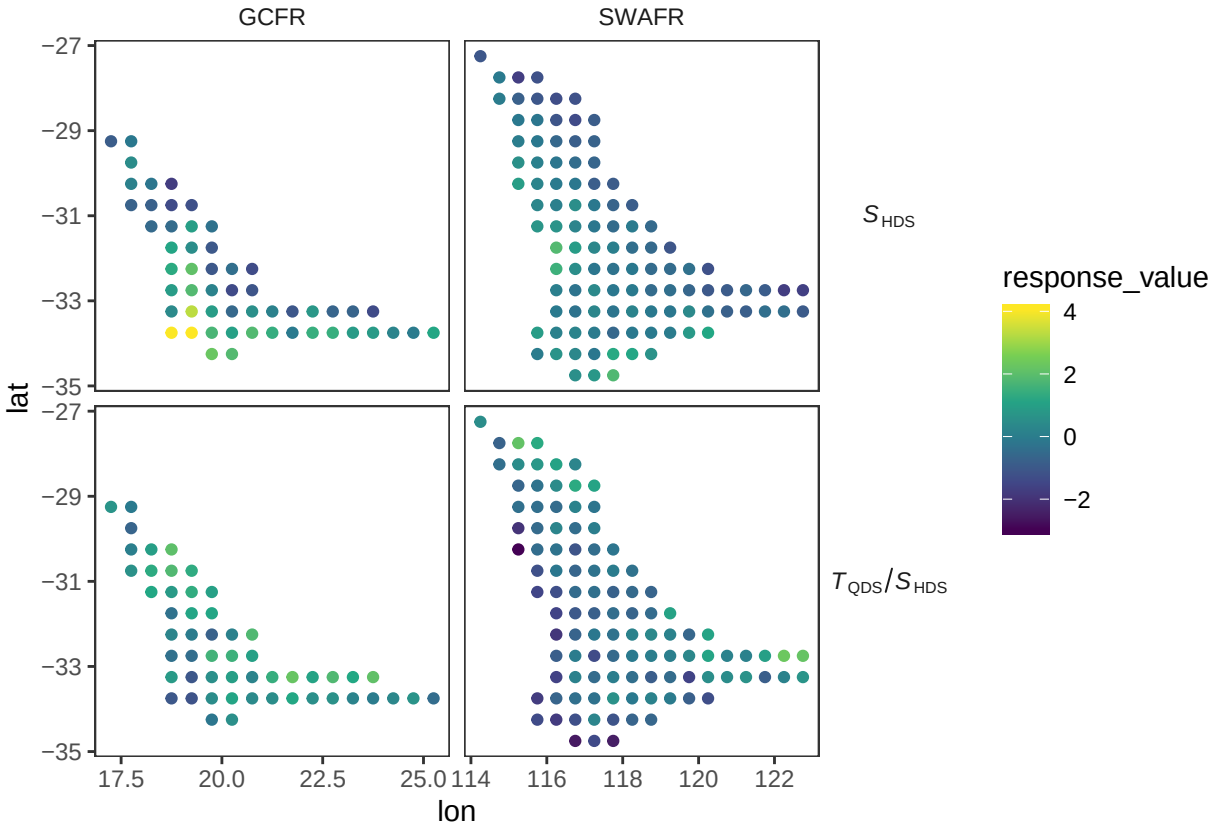


Table 3: Results of Mann-Whitney U-tests and the $CLES$ of GCFR vs SWAFR for various species richness and turnover metrics.

Metric	$CLES$	P_U
S_{HDS}	0.612	0.020
S_{QDS}	0.595	< 0.001
T_{QDS}/S_{HDS}	0.784	< 0.001

3. Relating heterogeneity to species richness & turnover

Does heterogeneity explain differences in richness and turnover between the regions?

Here I fit various linear regressions of richness and turnover as functions of environmental heterogeneity across the two regions. The richness and turnover measures used are the same as in the previous section, while the environmental heterogeneity was recalculated in the same grid-wise fashion as the richness and turnover measures. These analyses were carried out at both the HDS- and QDS-scales, insofar as species occurrence data from GBIF is only accurate to the QDS-scale. These analyses were only carried out on HDS-scale data for HDS-cells that contained four QDS-cells, and similarly for QDS-scale data for QDS-cells that contained four EDS-cells.

Environmental "roughness" here was calculated for each HDS- and QDS-cell in both regions as the mean of each consituent QDS- and EDS-cell's mean absolute difference in environmental conditions from the other three cells within that HDS- or QDS-cell.

In other words, roughness was calculated by first calculating the average absolute-difference in environmental values between each QDS and it's three neighbours in a given HDS. Then, these four values (assuming four QDS in an HDS) are averaged. This roughness index is presented mathematically below. This index allows each of the four values to be similarly independent, and thus more sutiable for our averaging and analyses, as opposed to if it were simly the direct average of pairwise differences [expand?].

$$Roughness_{cellular}(\{x_1, x_2, x_3, x_4\}) = \frac{1}{4}\sum_i f(x_i) = \frac{1}{4}\sum_i \left(\frac{1}{3}\sum_{j \neq i} |x_i - x_j| \right)$$

In R, this is implemented this as follows:

```
roughness_cellular <- function(x) {
  out <- vector(mode = "numeric", length = length(x))
  for (i in seq_along(x)) {
    out[[i]] <- mean(abs(x[i] - x[-i]))
  }
  mean(out)
}
```

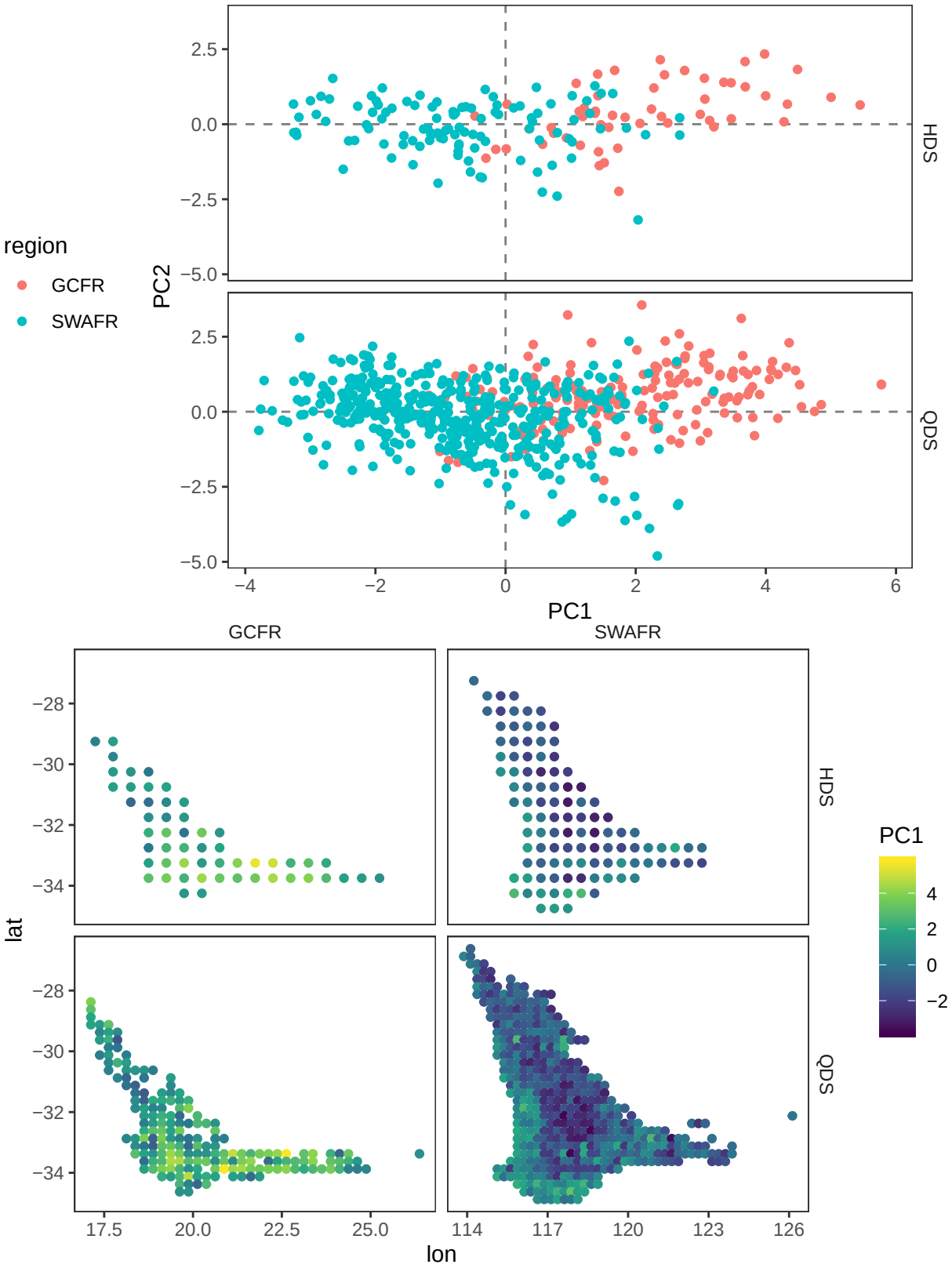
Or:

```
roughness_cells <- function(x) {
  mean(c(
    mean(abs(x[1] - x[-1])),
    mean(abs(x[2] - x[-2])),
    mean(abs(x[3] - x[-3])),
    mean(abs(x[4] - x[-4]))
  ))
}
```

For all analyses in this section, roughness values were $log(x + 1)$-transformed to ensure normality for linear regressions. Indeed, I also repeated the PCA on these transformed data. The PC1 values from this PCA are used in the analyses in this section.

In order to understand the relationships between environmental heterogeneity and species richness and turnover, and how those relationships differ between the GCFR and the SWAFR, I fit both simple and multiple linear regression models of S_{HDS} and S_{QDS} as functions of environmental roughness values. The rationale behind the univariate models is to describe empirical patterns of covariance between roughness axes and species richness. The multivariate models allow us to account for differences in richness across multiple roughness axes.

The PCA used here, done only at the QDS- and HDS-scales, looks almost exactly like the ones in Figure 1 for those scales:



That's encouraging!

3.1. Separate environmental variable models

Tables 5–7 present the results of simple linear regressions of each form of environmental heterogeneity separately as predictors of each of the three response variables (S_{HDS}, S_{QDS} and $T_{\mathrm{QDS}}/S_{\mathrm{HDS}}$ respectively).

In each table, the "best" model (sensu AIC) was select as the simplest model with $\Delta AIC < 2$—i.e. a more complex model was only justified when it had the lowest AIC-score by more than 2 AIC-units.

For S_{HDS} (Table 4), there is evidence for a great difference in the slopes of the GCFR and SWAFR's relationships with heterogeneity in MAP. Heterogneity in NDVI and clay only present evidence for the same slope in each region, but differing intercepts. Heterogeneity in CEC and pH have non-significant slopes and signficiant region-effects—suggesting that these variables values' have weak relationships with S_{HDS}, and that the region-effect explains more of the variance. Other variables (heterogeneity in elevation, PDQ, surface T and soil C) only present evidence for a continuous effect of that heterogeneity, explaining the difference in the regions' S_{HDS} in terms of the roughness values themselves, without the need to invoke a region term. Think of it this way:

- If there is no need for any information concerning the region a cell belongs to, then the environmental roughness "rule" is followed well across the two regions in a similar way.
- If the region-effect is significant, but not the roughness effect, then that roughness axis isn't doing a very good job of explaining anything, and must defer to the region-effect.
- When both the region- and roughness-effect are significant, this represents a softer version of the above, where the roughness axis can explain some variance, but not all.
- When there is a significant interaction between region and roughness, then each region is playing a whole new game with that axes in terms of how richness is being driven.

In Tables 4 and 5, I also regressed against PC1. Like heterogeneity in elevation and surface T, PC1 was the only explanatory variable "needed" in regressions for S_{HDS} (also see Figure 4) and S_{QDS}. Figure 4 shows quite nicely how, in general, the GCFR and SWAFR are following the same "rule" (species richness increases with increasing environmental heterogeneity (PC1)) but occupy different areas along that relationship (the GCFR being more rich and more rough than the SWAFR).

It is worth noting that PC1 is representatve of (ca. 45% of) the multivariate trends of species richness and turnover versus environmental heterogeneity. Some of the individual roughness variables show similar trends, some not, and to varying extents [expand].

Table 4: Results of separate simple linear regressions of S_{HDS} against environmental heterogeneity variables. Asterisks beside each ΔAIC denote the simples model with $\Delta AIC < 2$. Asterisks beside regression coefficients denote P-values < 0.05, while periods (".") denote P-values > 0.05 but < 0.10.

Heterogeneity predictor	Region-term	ΔAIC		Slope		SWAFR effect		Slope:SWAFR	
Elevation	None	0.00	*	194.24	***				
	Additive	1.63		223.32	***	78.46			
	Interaction	0.82		118.12		17.01		202.78	.
MAP	None	6.95		290.95	***				
	Additive	8.87		286.33	***	-21.96			
	Interaction	0.00	*	489.38	***	81.42		-276.80	**
PDQ	None	0.00	*	214.37	***				
	Additive	1.94		207.59	***	-23.49			
	Interaction	1.74		165.19	**	-10.69		156.09	
Surface T	None	0.00	*	182.97	***				
	Additive	1.52		160.90	**	-72.64			
	Interaction	2.02		101.81		-97.31		121.88	
NDVI	None	2.49		209.89	***				
	Additive	0.20	*	185.09	***	-170.47	*		
	Interaction	0.00		267.47	***	-143.64	.	-122.27	
CEC	None	7.61		60.35					
	Additive	0.00	*	-10.02		-300.96	**		
	Interaction	1.64		28.73		-280.70	**	-57.56	
Clay	None	7.97		100.55	*				
	Additive	0.00	*	82.77	*	-265.84	**		
	Interaction	1.66		49.58		-270.82	**	49.14	
Soil C	None	0.00	*	199.51	***				
	Additive	0.51		173.77	***	-111.22			
	Interaction	2.44		192.54	*	-100.71		-25.60	
pH	None	6.32		83.37	*				
	Additive	0.00	*	37.38		-261.18	**		
	Interaction	1.43		94.20		-236.29	*	-74.86	
PC1	None	0.00	*	128.72	***				
	Additive	0.41		151.06	***	131.47			
	Interaction	1.81		179.00	***	180.42		-42.08	

Table 5: Results of separate simple linear regressions of S_{QDS} against environmental heterogeneity variables. Asterisks beside each ΔAIC denote the simples model with $\Delta AIC < 2$. Asterisks beside regression coefficients denote P-values < 0.05, while periods (".") denote P-values > 0.05 but < 0.10.

Heterogeneity predictor	Region-term	ΔAIC		Slope		SWAFR effect		Slope:SWAFR	
Elevation	None	0.00	*	93.52	***				
	Additive	0.81		108.10	***	42.93			
	Interaction	2.13		126.34	***	59.77		-29.74	
MAP	None	0.00	*	147.31	***				
	Additive	1.97		148.14	***	4.28			
	Interaction	2.54		172.38	***	19.06		-32.44	
PDQ	None	12.96		110.34	***				
	Additive	14.52		116.25	***	21.26			
	Interaction	0.00	*	72.77	***	7.68		121.09	***
Surface T	None	0.00	*	100.03	***				
	Additive	0.27		91.47	***	-38.53			
	Interaction	1.53		78.32	***	-46.01		22.93	
NDVI	None	24.27		83.66	***				
	Additive	12.99		72.22	***	-97.45	***		
	Interaction	0.00	*	145.14	***	-71.40	**	-101.83	***
CEC	None	23.79		15.92					
	Additive	0.00	*	-4.26		-142.61	***		
	Interaction	1.91		-10.10		-145.09	***	8.32	
Clay	None	22.89		25.08	*				
	Additive	0.20	*	15.16		-134.03	***		
	Interaction	0.00		-14.38		-140.79	***	40.23	
Soil C	None	9.59		89.73	***				
	Additive	5.07		77.02	***	-71.52	*		
	Interaction	0.00	*	136.51	***	-38.11		-78.03	**
pH	None	24.72		11.04					
	Additive	0.00	*	-4.66		-142.17	***		
	Interaction	1.90		3.15		-139.37	***	-9.73	
PC1	None	1.00	*	65.68	***				
	Additive	0.00		74.43	***	55.89	.		
	Interaction	0.69		88.08	***	77.84	*	-19.58	

3.2. Combined-regions models with combinations of variables

See Figure 6 and Table 10.

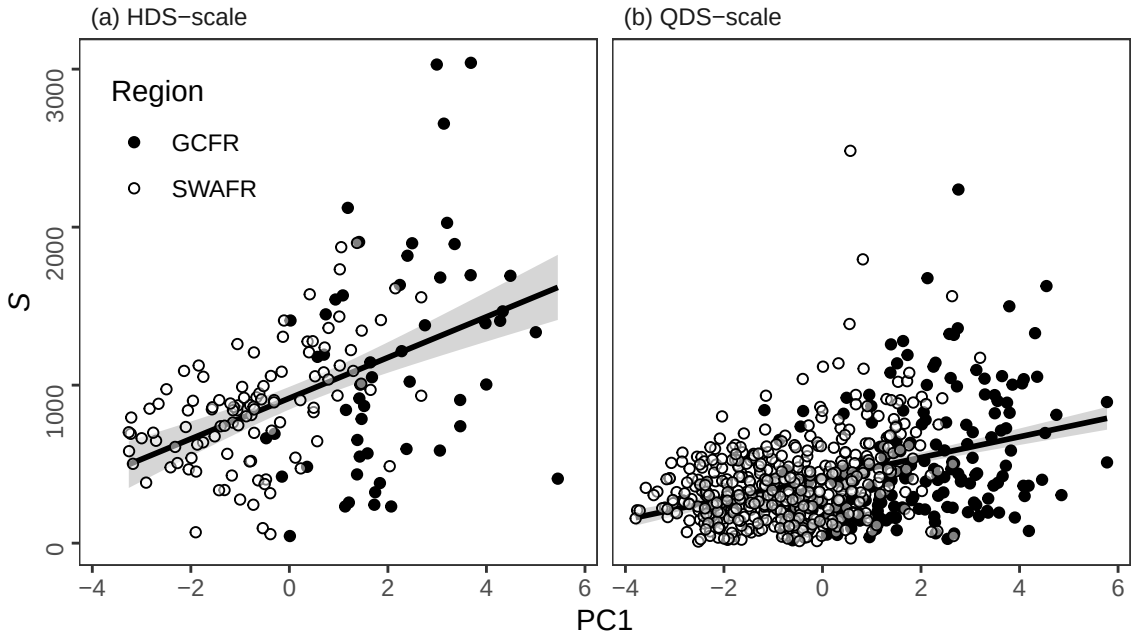


Figure 4: The fits of the common regressions of (a) S_{HDS} and (b) S_{QDS} against each respective scale's PC1 scores (Table 4 and 5). Grey bands denote 95% confidence intervals.

Table 6: See Figure 6. Significance is denoted with asterisks as follows: "***", $P < 0.001$; "**", $0.001 < P < 0.01$; "*", $0.01 < P < 0.05$; ".", $0.05 < P < 0.1$; and non-significant results ($P \geq 0.1$) being left blank.

Response	R^2_{adj}	Heterogeneity predictor	Slope	
HDS_richness	0.49	regionSWAFR	208.14	.
		regionGCFR:Elevation	-2.61	
		regionSWAFR:Elevation	201.36	**
		regionGCFR:MAP	917.41	***
		regionSWAFR:MAP	122.22	**
		regionGCFR:PDQ	-140.12	.
		regionSWAFR:PDQ	166.66	*
		regionGCFR:CEC	42.80	
		regionSWAFR:CEC	-114.76	*
		regionGCFR:Clay	164.20	**
		regionSWAFR:Clay	45.06	
		regionGCFR:Soil.C	-118.02	
		regionSWAFR:Soil.C	50.54	
		regionGCFR:pH	-267.80	**
		regionSWAFR:pH	15.48	
QDS_richness	0.33	Elevation	33.33	*
		MAP	100.38	***
		CEC	-29.15	*
		Clay	17.68	
		regionSWAFR	88.20	*
		regionGCFR:PDQ	-30.85	
		regionSWAFR:PDQ	114.63	***
		regionGCFR:Surface.T	-5.55	
		regionSWAFR:Surface.T	53.45	**
		regionGCFR:NDVI	137.31	***
		regionSWAFR:NDVI	-5.17	
		regionGCFR:Soil.C	111.32	***
		regionSWAFR:Soil.C	4.90	
		regionGCFR:pH	-164.78	***
		regionSWAFR:pH	-7.66	

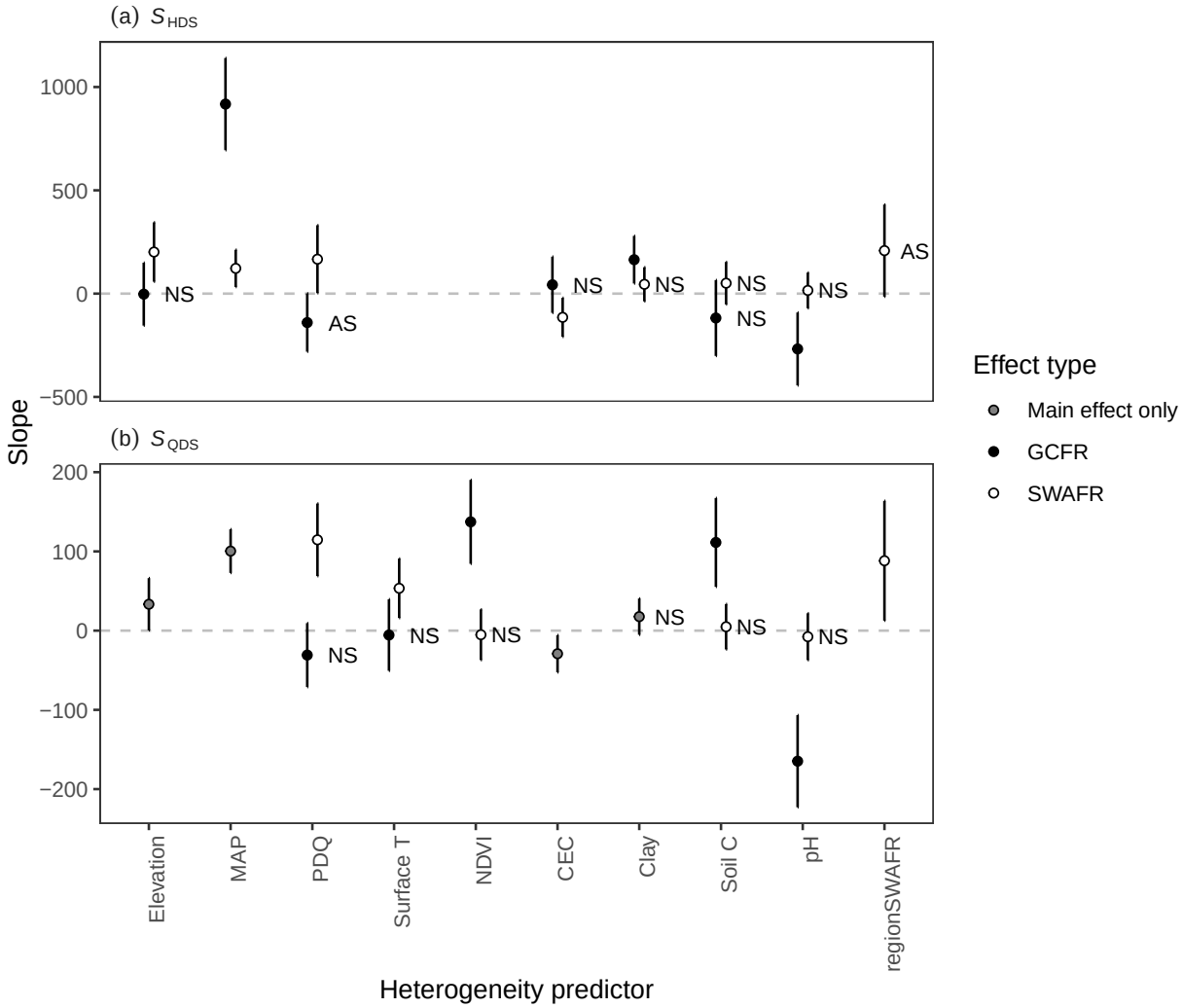


Figure 5: Slope estimates of the three multiple linear regressions of (a) S_{HDS} and (b) S_{QDS} againt environmental heterogeneity variables interacting with region. Each model was simplified using reverse stepwise regression model selection using AIC-scores. Points with error bars denote slope estimates and their 95% confidence intervals. All estimates were significant ($P < 0.05$) unless otherwise stated ($0.10 > P > 0.05$, almost significant, "AS"; $P > 0.05$, not significant, "NS").