Analyses v2

Cape vs SWA

Ruan van Mazijk

2019-06-13

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:

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

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), 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.

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 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 2). Notably, the GCFR is more pronouncedly heterogeneous at broad scales in terms of clay (Figure 2). In general (i.e. regarding PC1; Figure 2), the GCFR is more environmentally heterogeneous than the SWAFR, and particularly so at fine spatial 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	*

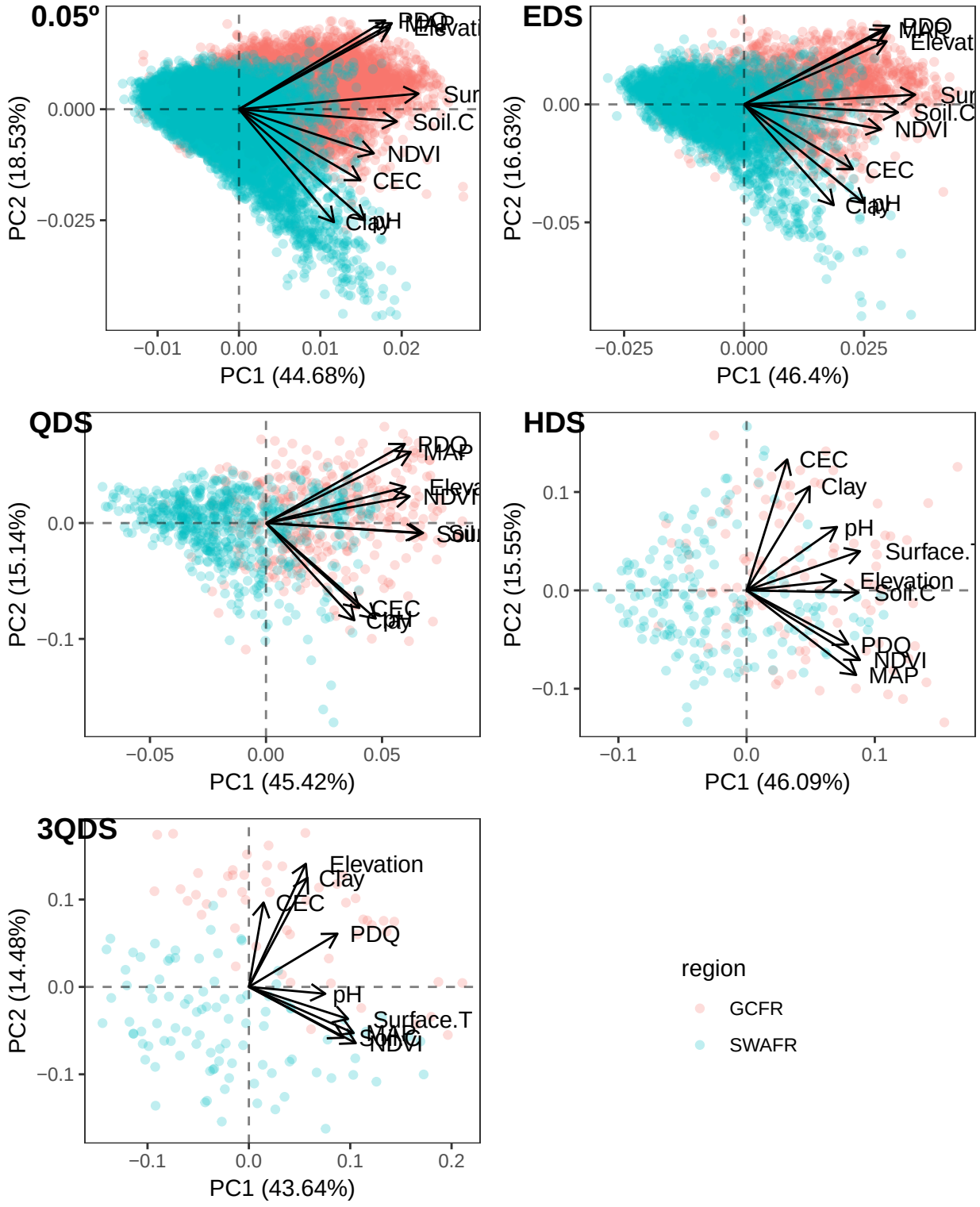


Figure 1: Scatter plots of the first and second principal components (PC1, PC2) of environmental heterogeneity following principal components analyses (PCAs) of the various forms of environmental heterogeneity, repeated at the five spatial scales. The proportion of variation accounted for by each axis 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.

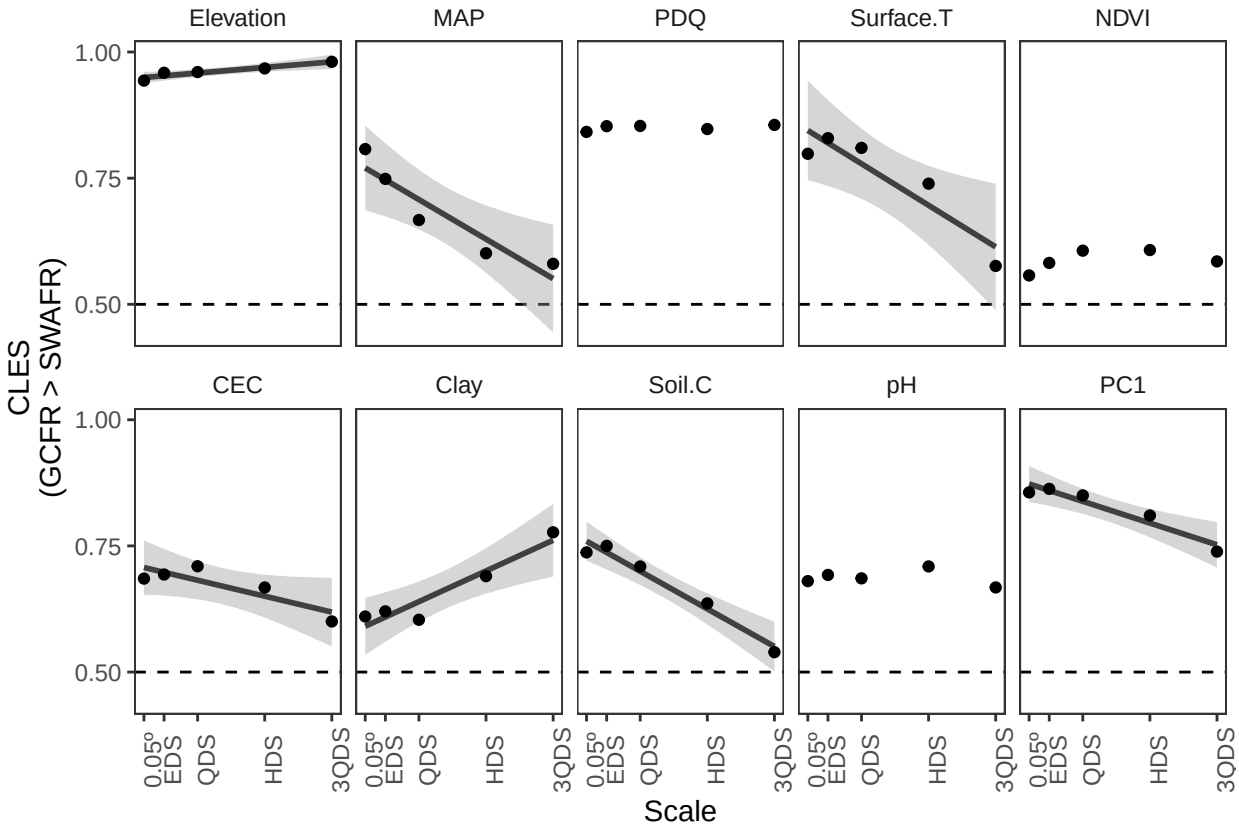


Figure 2: Regressions of the common language effect size ($CLES$) of various forms of environmental heterogeneity, and the first principal component thereof (PC1, see Figure 1). Only significant or marginally significant fits are plotted (Table 2). Across spatial scales, all $CLES$ values differed significantly from zero following a two-side t-test ($P < 0.001$), demonstrating the difference in heterogeneity between the regions.

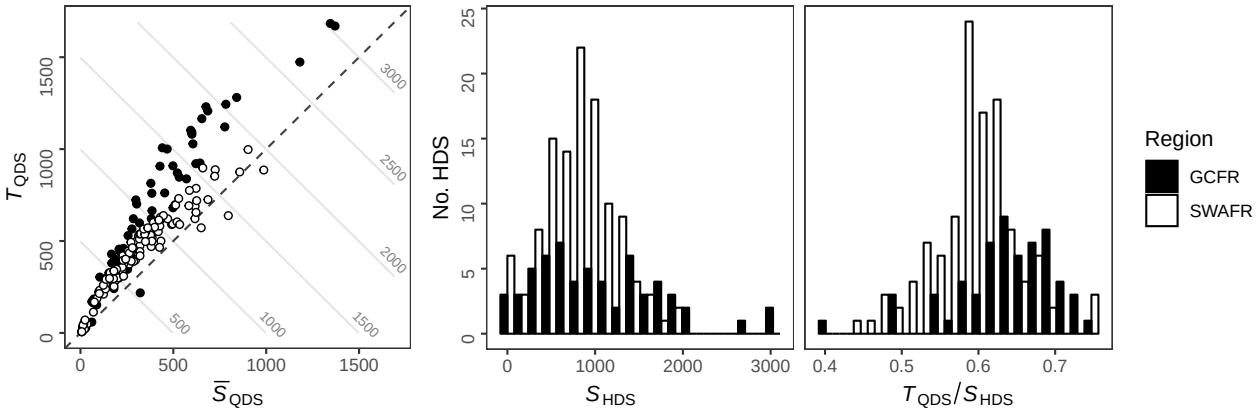


Figure 3: Scatter plot of $\overline{S}_{\text{QDS}}$ and T_{QDS} with contour lines denoting the S_{HDS} that would arise as their sum (i.e. increasing from lower-left to upper-right).

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 3. 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 4.

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.

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}	0.584	0.066
S_{QDS}	0.576	0.002
$T_{\text{QDS}}/S_{\text{HDS}}$	0.738	< 0.001

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

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.

The richness and turnover measures used are the same as in the previous section, while ...

$$
\begin{aligned}
X &= \{x_1, x_2, x_3, x_4\} \\
R(X) &= \frac{1}{n} \sum_{i=1}^{n} D(x_i) \\
D(x_i) &= \frac{1}{m} \sum_{j=1}^{m} |x_i - y_j| \\
Y &= X \setminus x_i \\
n &= 4 \\
m &= n - 1
\end{aligned}
$$

[Interpretations to follow after meeting.]

3.1. Separate-regions models with combinations of variables

Table 4: Results of bi-directional stepwise multiple linear regressions of three richness and turnover responses in the against additive combinations of environmental heterogeneity variables. The stepwise regression procedure started with all variables included. (See Figure 5 for a graphical representation.)

Response	Predictor	Slope	P_{slope}	
GCFR S_{HDS}	Clay	180.429	0.032	*
GCFR S_{HDS}	Elevation	145.884	0.138	
GCFR S_{HDS}	MAP	657.587	0.000	*
GCFR S_{HDS}	pH	-305.324	0.018	*
GCFR $\overline{S}_{\text{QDS}}$	MAP	167.976	0.000	*
GCFR $\overline{S}_{\text{QDS}}$	NDVI	121.452	0.000	*
GCFR $\overline{S}_{\text{QDS}}$	PDQ	-48.642	0.116	
GCFR $\overline{S}_{\text{QDS}}$	pH	-137.680	0.000	*
GCFR $\overline{S}_{\text{QDS}}$	Soil.C	81.429	0.022	*
GCFR T_{QDS}	CEC	0.021	0.016	*
GCFR T_{QDS}	Clay	-0.017	0.047	*
GCFR T_{QDS}	Elevation	0.028	0.003	*
GCFR T_{QDS}	MAP	-0.016	0.127	
SWAFR S_{HDS}	CEC	-147.946	0.000	*
SWAFR S_{HDS}	Clay	60.151	0.031	*
SWAFR S_{HDS}	Elevation	189.173	0.000	*
SWAFR S_{HDS}	MAP	120.403	0.000	*
SWAFR S_{HDS}	NDVI	54.642	0.119	
SWAFR S_{HDS}	PDQ	161.140	0.004	*
SWAFR S_{HDS}	Surface.T	83.543	0.083	.
SWAFR $\overline{S}_{\text{QDS}}$	CEC	-29.634	0.009	*

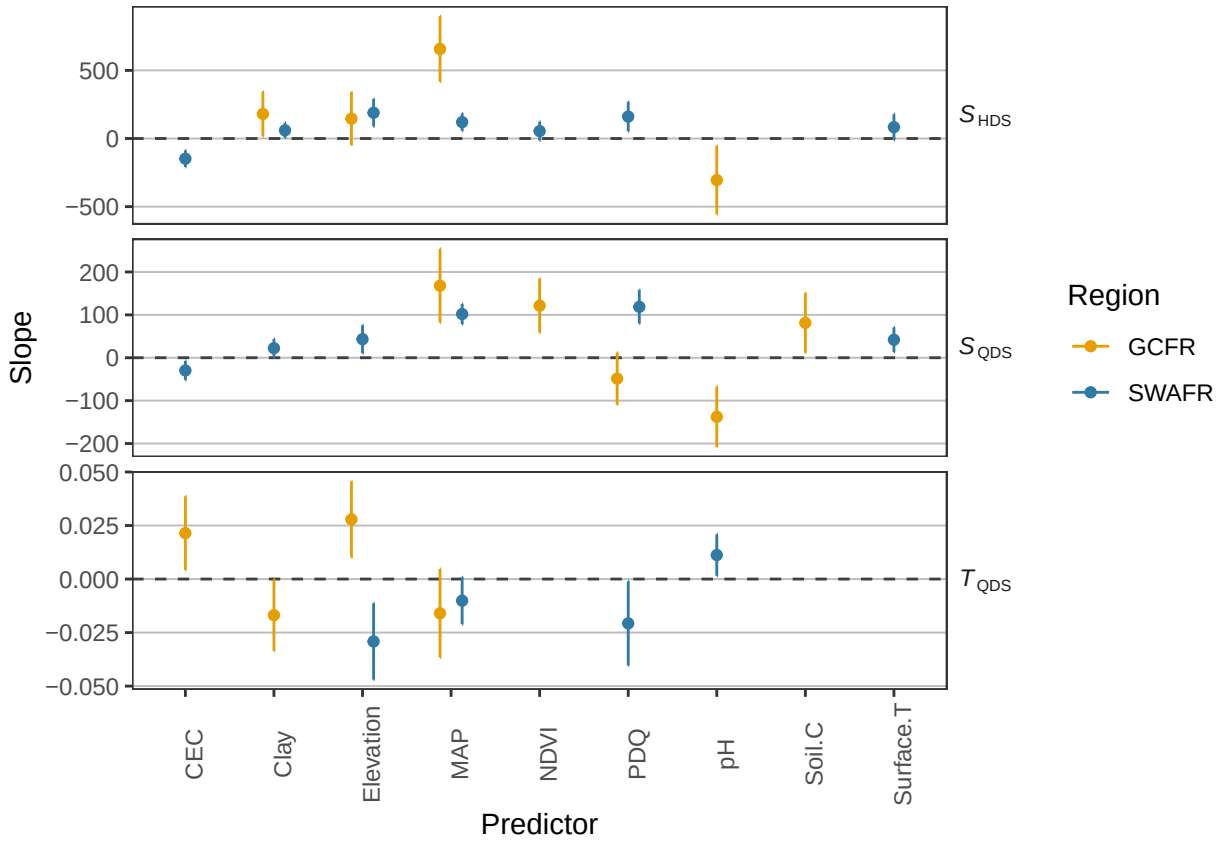


Figure 4: Slopes from Table 5, with error bars denoting 95% confidence intervals about each slope estimate.

Response	Predictor	Slope	P_{slope}	
SWAFR $\overline{S}_{\text{QDS}}$	Clay	22.302	0.043	*
SWAFR $\overline{S}_{\text{QDS}}$	Elevation	43.273	0.010	*
SWAFR $\overline{S}_{\text{QDS}}$	MAP	101.879	0.000	*
SWAFR $\overline{S}_{\text{QDS}}$	PDQ	118.732	0.000	*
SWAFR $\overline{S}_{\text{QDS}}$	Surface.T	41.950	0.005	*
SWAFR T_{QDS}	Elevation	-0.029	0.002	*
SWAFR T_{QDS}	MAP	-0.010	0.070	.
SWAFR T_{QDS}	PDQ	-0.021	0.039	*
SWAFR T_{QDS}	pH	0.011	0.024	*

Table 5: Adjusted R^2-values of the models in Table 5.

Response	${R^2}_{adj.}$
GCFR S_{HDS}	0.365
GCFR $\overline{S}_{\text{QDS}}$	0.265
GCFR T_{QDS}	0.247
SWAFR S_{HDS}	0.541
SWAFR $\overline{S}_{\text{QDS}}$	0.334
SWAFR T_{QDS}	0.208

3.2. Combined-regions models with individual variables

3.2.1. Environmental heterogeneity variables

[Plots of the models referred to in Tables 6–8 can be discussed in a meeting.]

Table 6: Results of separate simple linear regressions of S_{HDS} against environmental heterogeneity variables with no region-term.

Predictor	R^2	P_{slope}	
CEC_roughness	0.012	0.157	
Clay_roughness	0.053	0.002	*
Elevation_roughness	0.160	0.000	*
MAP_roughness	0.301	0.000	*
NDVI_roughness	0.160	0.000	*
PDQ_roughness	0.143	0.000	*
pH_roughness	0.039	0.009	*
Soil.C_roughness	0.151	0.000	*
Surface.T_roughness	0.131	0.000	*

Table 7: Results of separate simple linear regressions of S_{HDS} against environmental heterogeneity variables with an additive region-term.

Predictor	R^2	P_{slope}		P_{region}	
CEC_roughness	0.049	0.796		0.010	*
Clay_roughness	0.089	0.006	*	0.010	*
Elevation_roughness	0.165	0.000	*	0.325	
MAP_roughness	0.302	0.000	*	0.685	
NDVI_roughness	0.170	0.000	*	0.140	
PDQ_roughness	0.143	0.000	*	0.855	
pH_roughness	0.066	0.074		0.026	*
Soil.C_roughness	0.154	0.000	*	0.463	
Surface.T_roughness	0.131	0.000	*	0.880	

Table 8: Results of separate simple linear regressions of S_{HDS} against environmental heterogeneity variables with an interaction-region-term.

Predictor	R^2	P_{slope}		P_{region}		$P_{slope:region}$	
CEC_roughness	0.077	0.051		0.042	*	0.023	*
Clay_roughness	0.091	0.252		0.009	*	0.558	
Elevation_roughness	0.173	0.002	*	0.354		0.218	
MAP_roughness	0.329	0.000	*	0.196		0.008	*
NDVI_roughness	0.173	0.001	*	0.197		0.460	
PDQ_roughness	0.151	0.002	*	0.761		0.219	
pH_roughness	0.078	0.027	*	0.071		0.136	
Soil.C_roughness	0.155	0.002	*	0.565		0.641	
Surface.T_roughness	0.135	0.016	*	0.805		0.350	

3.2.2. PC1 models

Here, I present my findings with raw R-code, because I don't have the time to format it neatly.

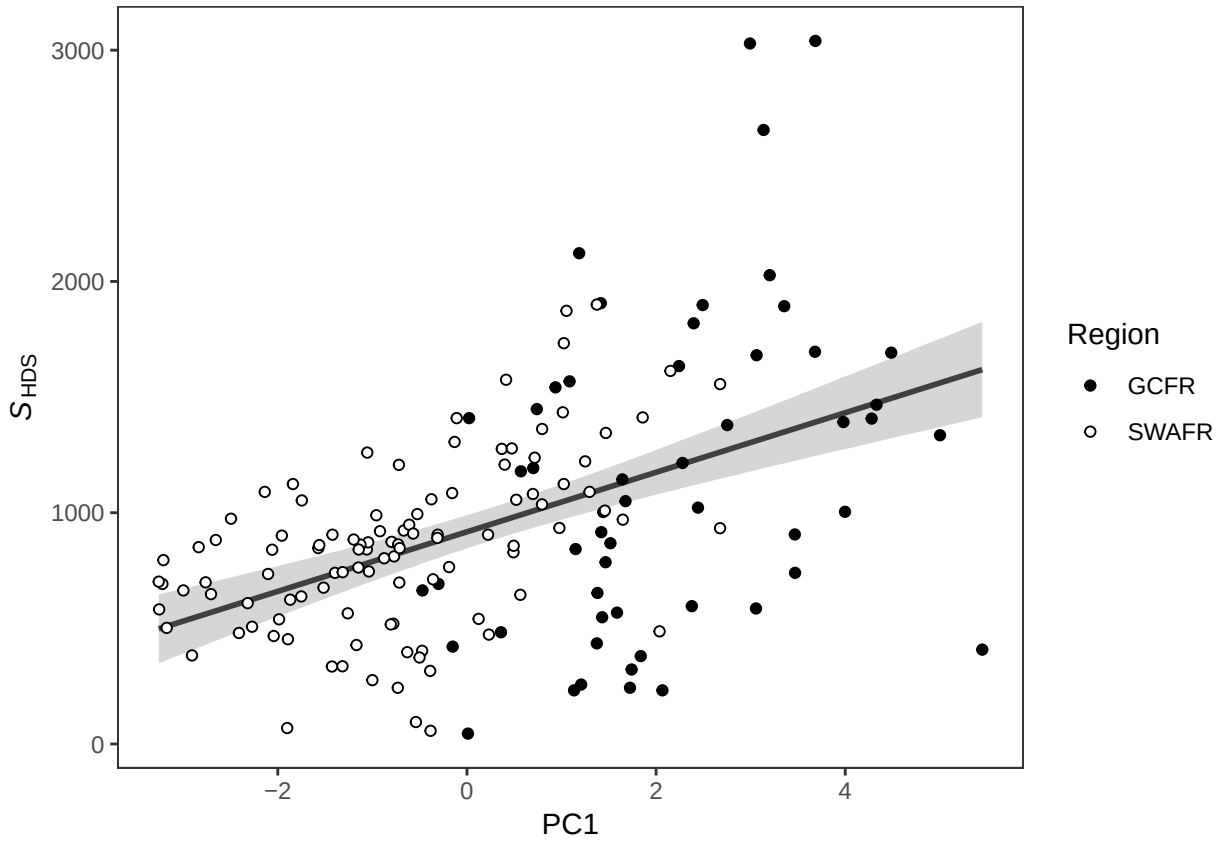
```
# Richness (HDS)

m1 <- lm(HDS_richness ~ PC1,          filter(HDS, n_QDS == 4))
m2 <- lm(HDS_richness ~ PC1 + region, filter(HDS, n_QDS == 4))
m3 <- lm(HDS_richness ~ PC1 : region, filter(HDS, n_QDS == 4))
m4 <- lm(HDS_richness ~ PC1 * region, filter(HDS, n_QDS == 4))
AIC(m1, m2, m3, m4)
```

```
##    df      AIC
## m1  3 2433.144
## m2  4 2433.558
## m3  4 2435.144
## m4  5 2434.958
```

```
summary(m1)
```

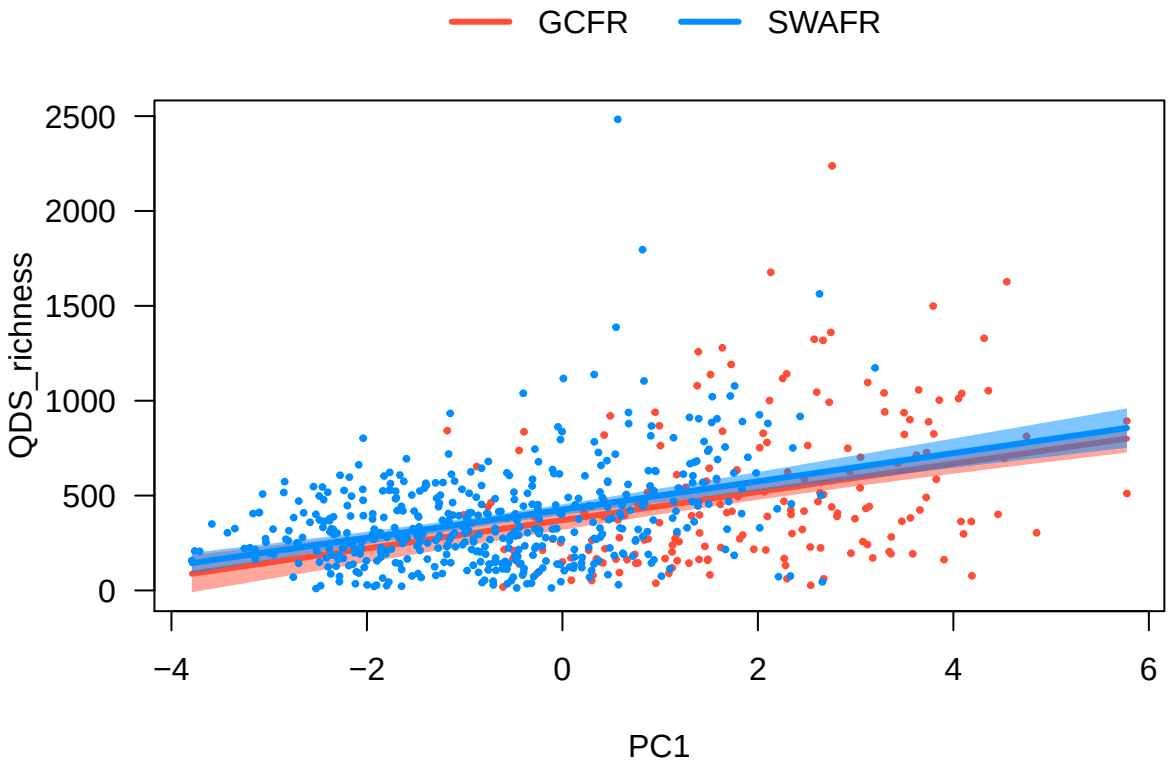
```
##
## Call:
## lm(formula = HDS_richness ~ PC1, data = filter(HDS, n_QDS ==
##     4))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1210.75  -226.28    13.83   235.50  1726.38
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   917.42      36.32  25.256  < 2e-16 ***
## PC1           128.72      18.82   6.838 1.63e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 457.2 on 159 degrees of freedom
## Multiple R-squared:  0.2273, Adjusted R-squared:  0.2224
## F-statistic: 46.76 on 1 and 159 DF,  p-value: 1.631e-10
```

```
ggplot(filter(HDS, n_QDS == 4), aes(PC1, HDS_richness)) +
  geom_smooth(method = lm, colour = "grey25") +
  geom_point(aes(fill = region), shape = 21, colour = "black") +
  ylab(bquote(italic("S")["HDS"])) +
  scale_fill_manual(name = "Region", values = c("black", "white"))
```



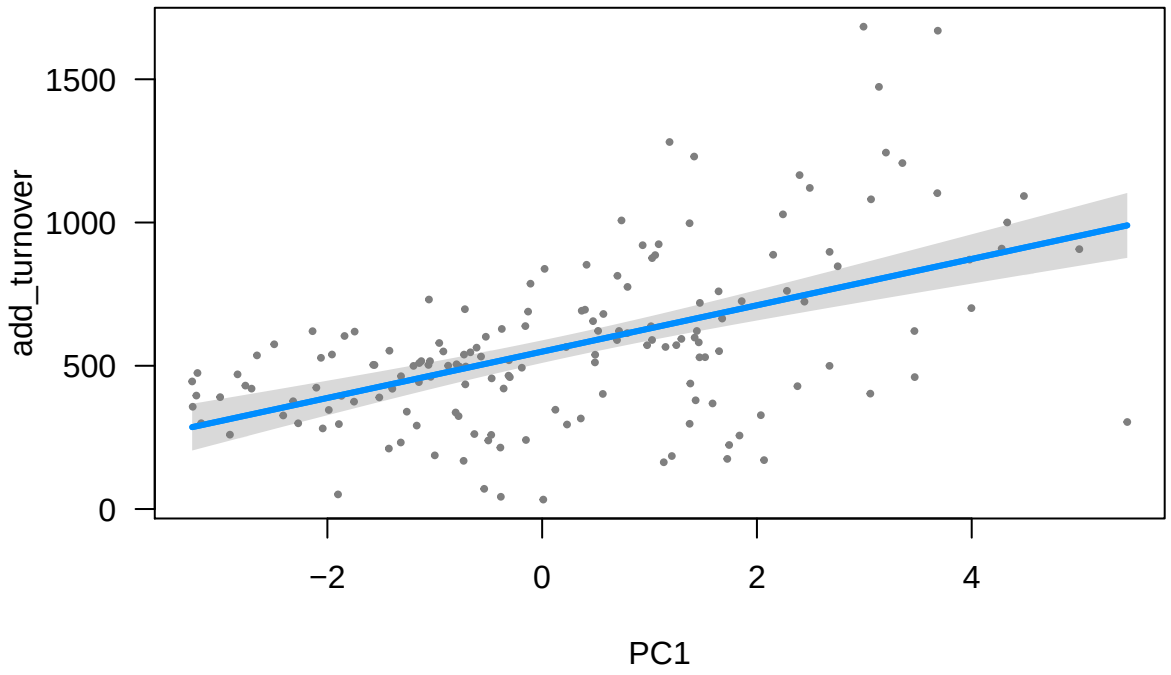

```
##    df      AIC
## m1  3 9205.960
## m2  4 9204.961
## m3  4 9207.951
## m4  5 9205.652
```

```
##
## Call:
## lm(formula = QDS_richness ~ PC1 + region, data = filter(QDS,
##     n_EDS == 4))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -604.9 -192.3  -32.8  133.8 2014.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  370.081     25.926  14.275   <2e-16 ***
## PC1           74.426      7.885   9.439   <2e-16 ***
## regionSWAFR   55.887     32.309   1.730   0.0842 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 283.5 on 648 degrees of freedom
## Multiple R-squared:  0.1571, Adjusted R-squared:  0.1545
## F-statistic: 60.39 on 2 and 648 DF,  p-value: < 2.2e-16
```



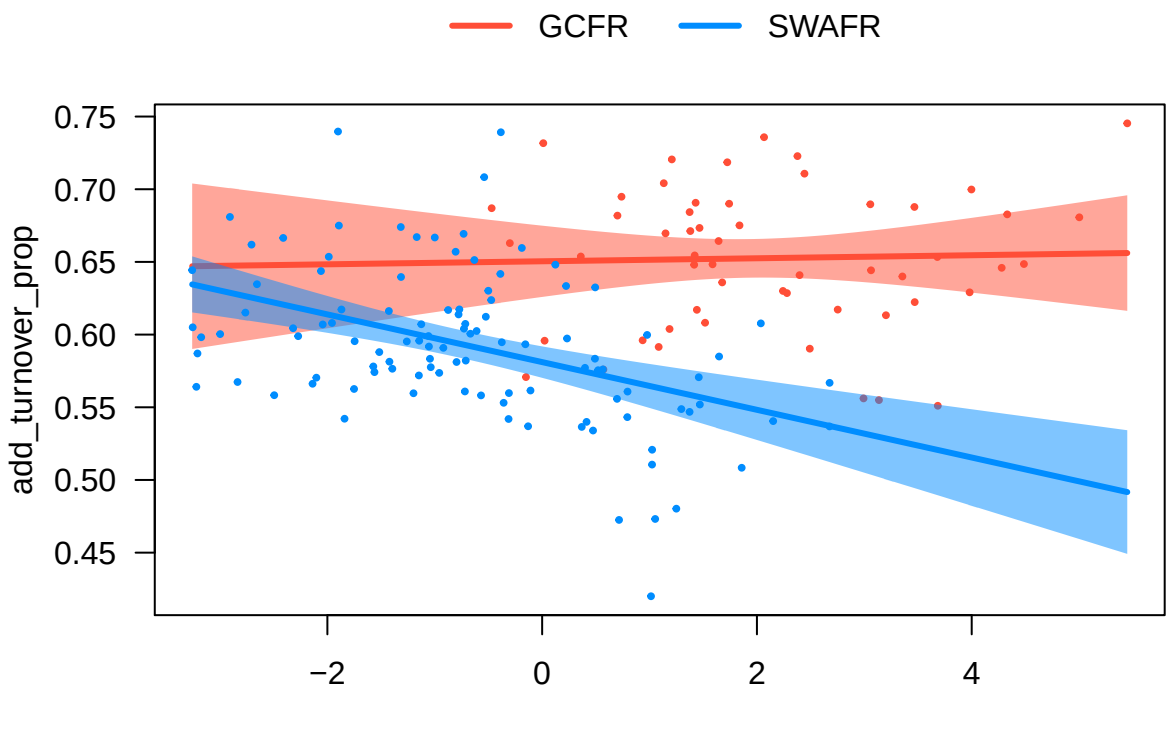

```
##    df      AIC
## m1  3 2240.186
## m2  4 2242.178
## m3  5 2241.587
```

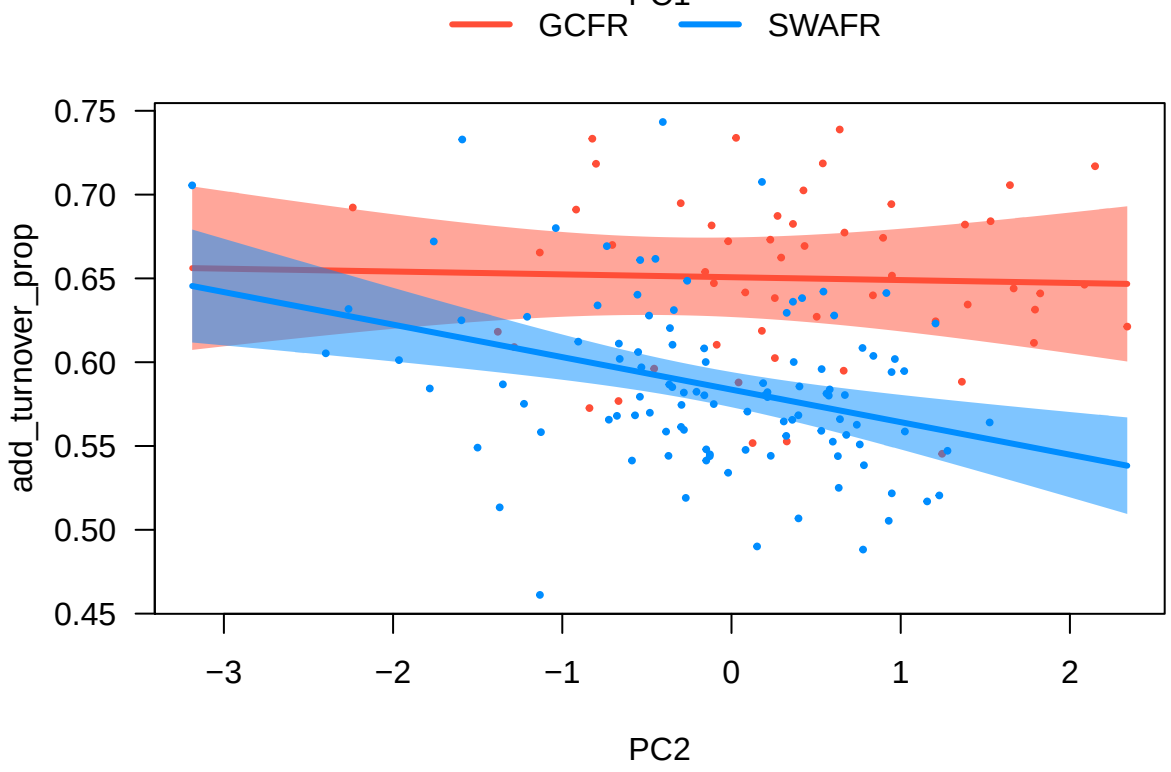
```
##
## Call:
## lm(formula = add_turnover ~ PC1, data = filter(HDS, n_QDS ==
##     4))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -686.17 -131.58    6.89  114.15  892.17
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   549.34      19.95  27.536  < 2e-16 ***
## PC1            80.86      10.34   7.821 6.81e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 251.1 on 159 degrees of freedom
## Multiple R-squared:  0.2778, Adjusted R-squared:  0.2733
## F-statistic: 61.18 on 1 and 159 DF,  p-value: 6.806e-13
```

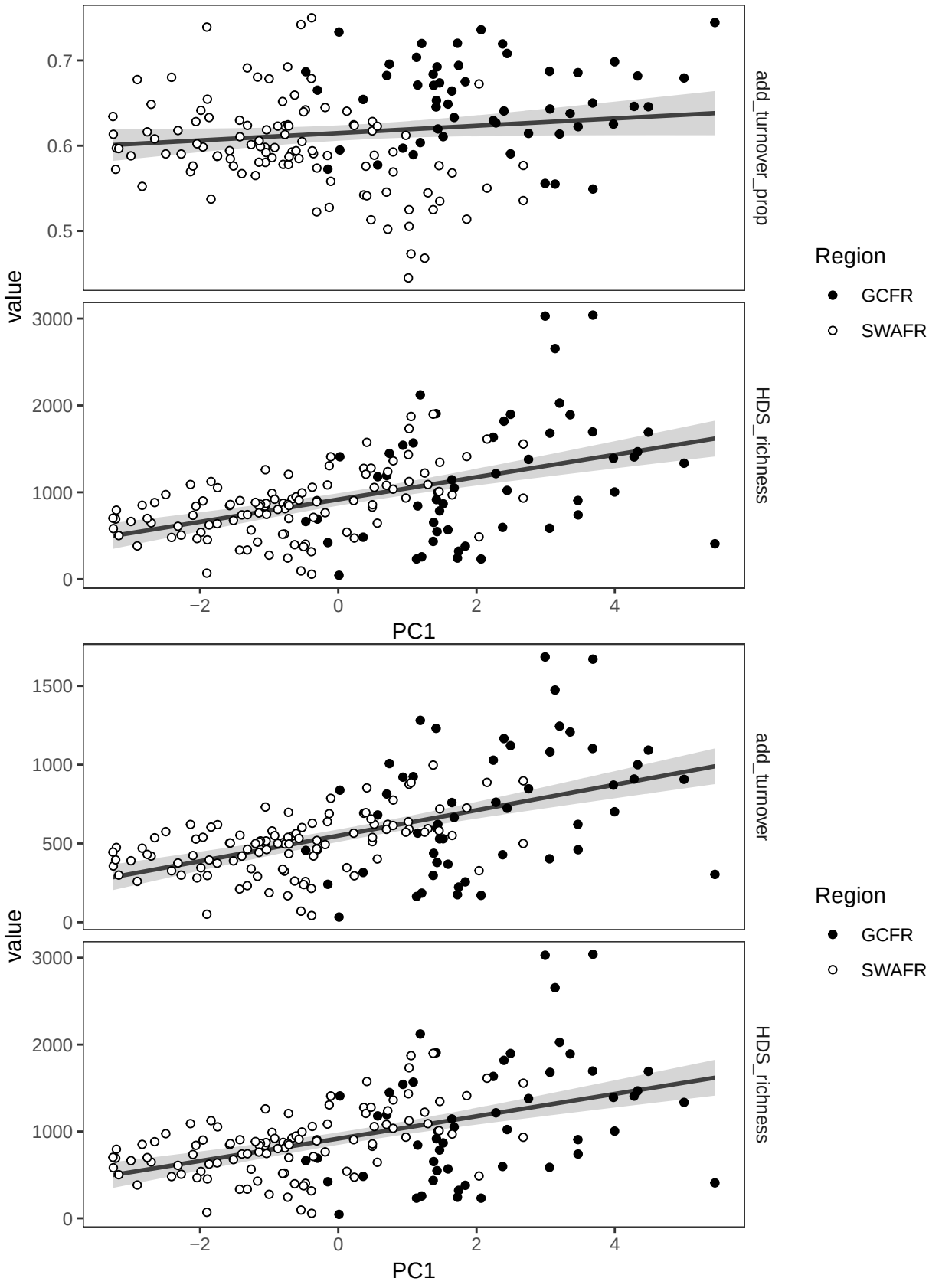



```
##    df       AIC
## m1  4 -457.3396
## m2  5 -501.9416
## m3  7 -513.2655
```

```
##
## Call:
## lm(formula = add_turnover_prop ~ PC1 * region + PC2 * region,
##     data = filter(HDS, n_QDS == 4))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.144429 -0.031686 -0.004089  0.030877  0.151809
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.650679   0.012117  53.702  < 2e-16 ***
## PC1              0.001040   0.005396   0.193  0.84737
## regionSWAFR     -0.066673   0.013213  -5.046 1.25e-06 ***
## PC2             -0.001701   0.007602  -0.224  0.82325
## PC1:regionSWAFR -0.017447   0.006369  -2.739  0.00688 **
## regionSWAFR:PC2 -0.017699   0.009327  -1.898  0.05960 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04796 on 155 degrees of freedom
## Multiple R-squared:  0.3363, Adjusted R-squared:  0.3149
## F-statistic: 15.71 on 5 and 155 DF,  p-value: 1.698e-12
```


GCFR
SWAFR
add_turnover_prop
0.75
0.70
0.65
0.60
0.55
0.50
0.45
−2
0
2
4
PC1


GCFR
SWAFR
add_turnover_prop
0.75
0.70
0.65
0.60
0.55
0.50
0.45
−3
−2
−1
0
1
2
PC2


add_turnover_prop
HDS_richness
add_turnover
HDS_richness
value
PC1
Region
GCFR
SWAFR

3.3. Combined-regions models with combinations of variables

[To be discussed in meeting.]