ExaminingCovariatesVII

PetraGuy

27 March 2018

From previous analysis it became obvious (maybe should have been in the first place) that most of the variables I had "created" in order to express heterogeneity were already contained within the existing variates. Eg, the standard deviation of pH and number of major soil groups. In addition, I wasn't sure what standardizing a standard deviation in the PCA would mean, there were so many variables the analysis was becoming harder rather than simpler, none of the additional variables had an obvious correlation with richness and since many were computationally related, they were correlated. So I decided to strip the variables back to those that were originally measured and recorded and start again.

Here I will go through these and remove firstly, those that are not correlated with species richness and/or are correlated with each other - eg. we don't want all of area, perimeter and area_ratio, but which is most important for richness? It has been suggested that the geometry of the wood can be related to richness - eg, an undisturbed core could contain poorly dispersing woodland species, or a long narrow wood would more easily allow recolonization. So area ratio might be the most useful variable.

There are also a few outliers that may affect the analysis, one site(74) has area > 300ha, the others have areas less that 150ha. Including this site gives a stronger correlation with richness and area than when it is removed. Two sites have positive heterogeneity indices above 78, the rest are below 40. Again, the correlation with richness is reduced if they are removed. Plus, the positive heterogeneity index may need tweaking (do some things need to be removed? Are there other woodland management codes which should be included?)

Bearing in mind that the purpose of this analysis is to look for the factors that effect richness, and might therefore be related to z and c. It is not to create a predictive model using richness as a function of the variables. Therefore looking at R2 and AIC etc may not be best way to proceed. I tried looking at PCA, R2 and p values for various subsets, and the amount of results and graphs produced didn't add a lot of information. I have therefore decided to simplify the exploration to first looking at the multiple linear regression plots using visref because these show which variables, in the presence of the others, correlate with richness. Secondly I will look at correlations between variables using a standardized correlation matrix. These correlations will be explored to see whether the correlated variables might be confounding.

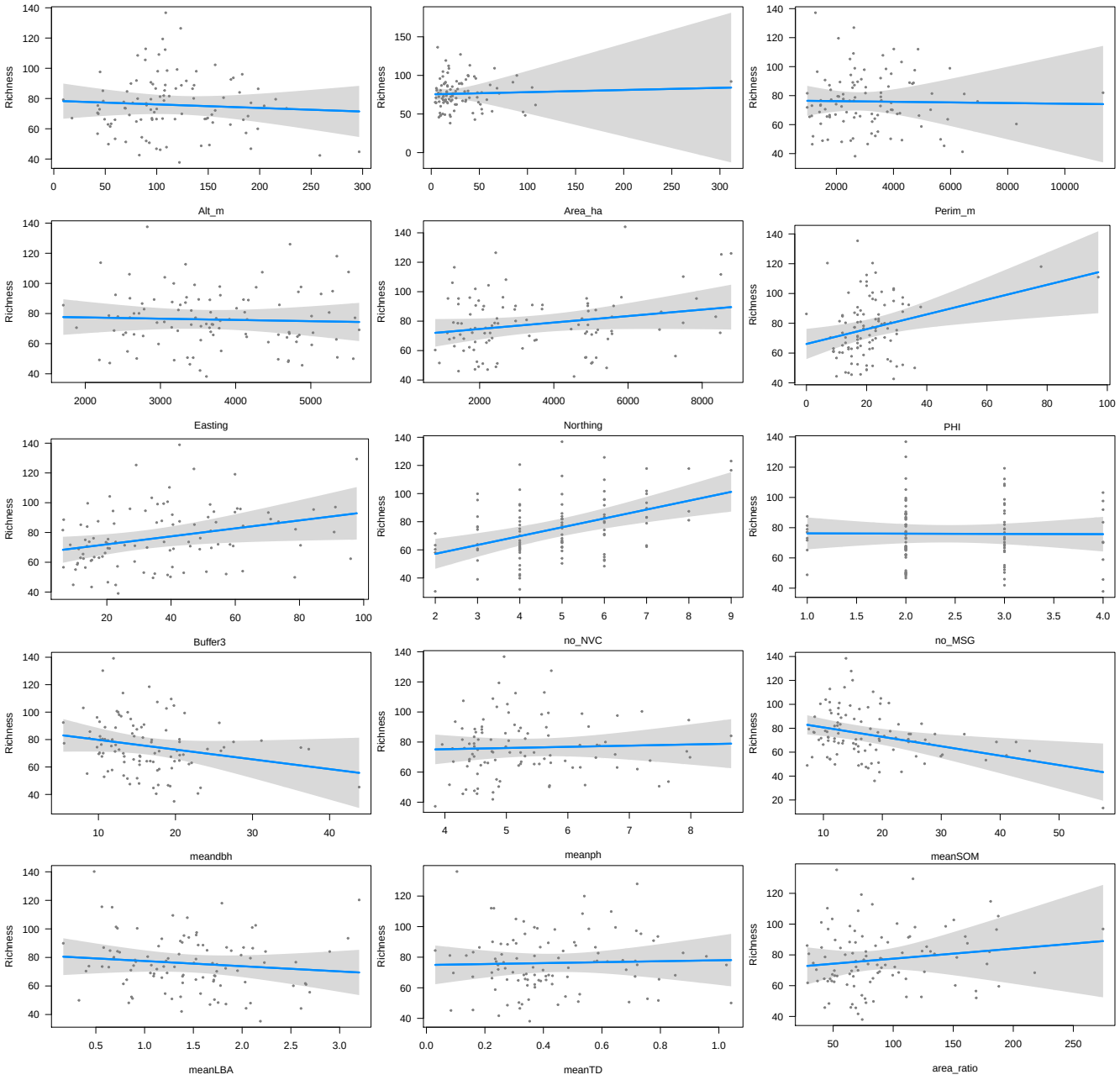
Initial set of variables.

```
##  [1] "Alt_m"      "Area_ha"    "Perim_m"    "Easting"    "Northing"
##  [6] "PHI"        "Buffer3"    "Richness"   "no_NVC"     "no_MSG"
## [11] "meandbh"    "meanph"     "meanSOM"    "meanLBA"    "meanTD"
## [16] "area_ratio"
```

(PS PHI = positive heterogeneity index, meanTD = mean tree density, no_trees = number of sites with no trees, no_NVC = number of NVC codes, no_MSG = number of major soil groups).

An initial correlation matrix (not shown because huge) shows the expected strong correlations between area, perimeter and area ratio and meanLBA, tree density, and meandbh. Since area_ratio is potentially the more interesting, this will be retained, and since mean dbh has the strongest correlation with richness,(see visref plots below) that will also be retained along with tree density, which I think might possibly express richness in a slightly different way to meandbh. But aren't meanLBA meandbh are basically the same thing? so I will drop meanLBA because the correlation with richness is weaker.

Similarly, area, perimeter and area ratio are not all required, area_ratio shows a slight correlation with richness, whereas area and perimeter do not, (see visref plots), and is potentially the more interesting, so this will be retained.

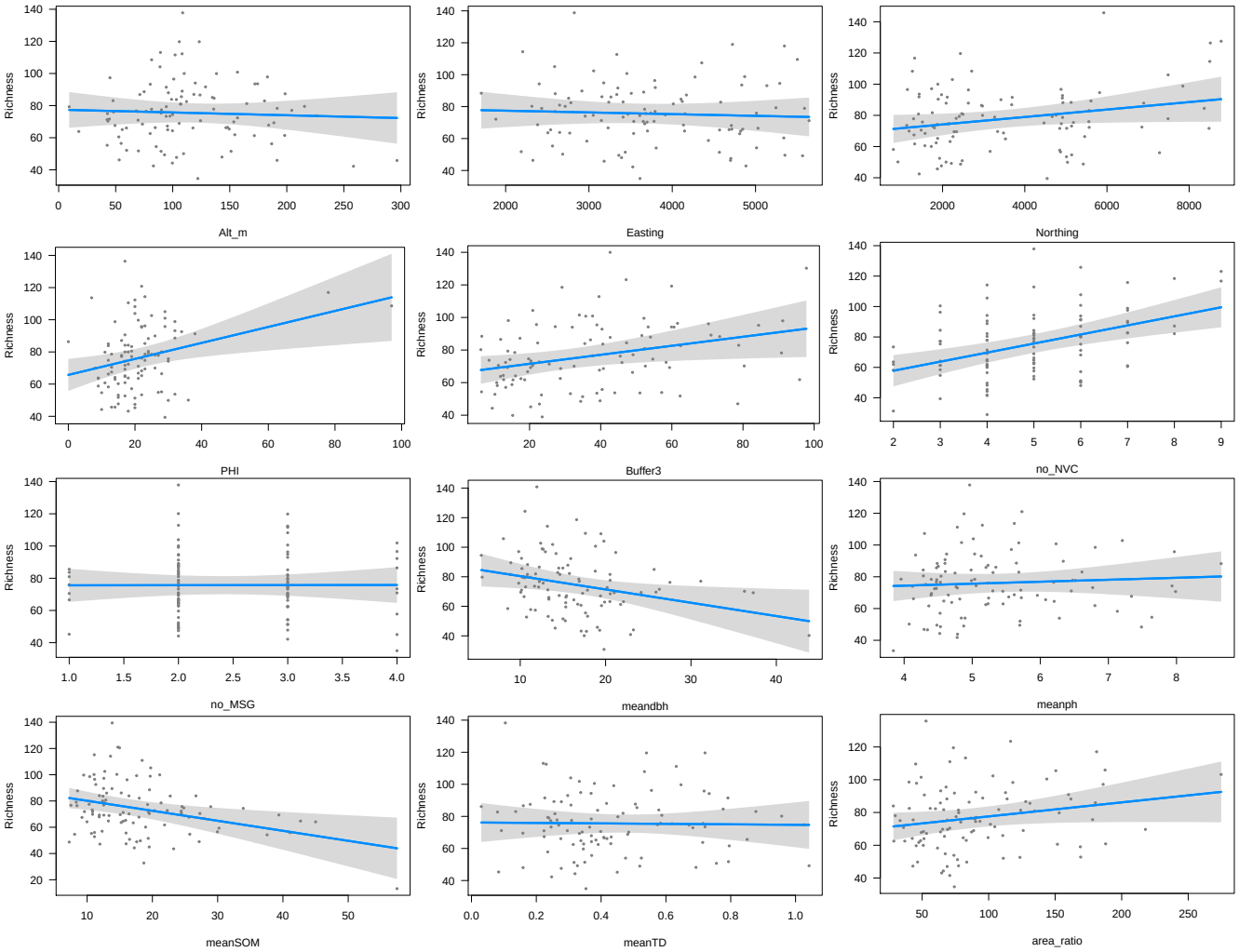


The multiple linear regression using all variables shows the following variables are correlated with richness : Buffer,Northing, no_NVC, meandbh, meanSOM, meanTD, area ratio, meanLBA (slightly)

The new reduced sest of variables will be....

```
##  [1] "Alt_m"      "Easting"    "Northing"   "PHI"        "Buffer3"
##  [6] "Richness"   "no_NVC"     "no_MSG"     "meandbh"    "meanph"
## [11] "meanSOM"    "meanTD"     "area_ratio"
```

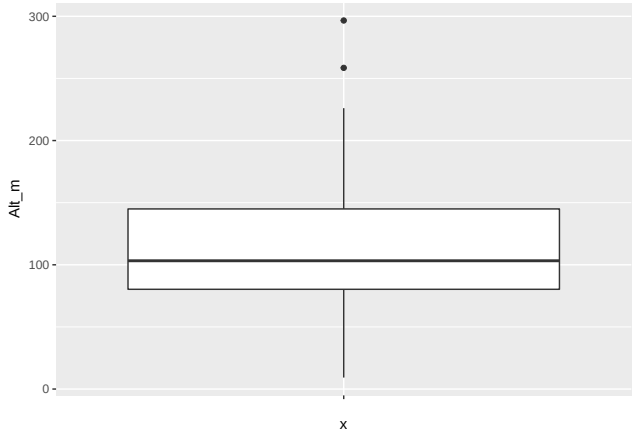
This is a more workable and realistic set of variables - in the sense that none are clearly correlated because they are computationally related, and they are all based on the original measured data with little manipulation.



The visref plots show no correlation with richness for altitude, Easting, no_MSG , meanpH or mean tree density.

Altitude

The range in altitude of the sites is not great, none are particularly high, half the woods are between about 75m and 150m, the highest 300m. It is unlikely that these altitudes would affect richness, so I'll drop this before we go on.



The new set of reduced variales is now...

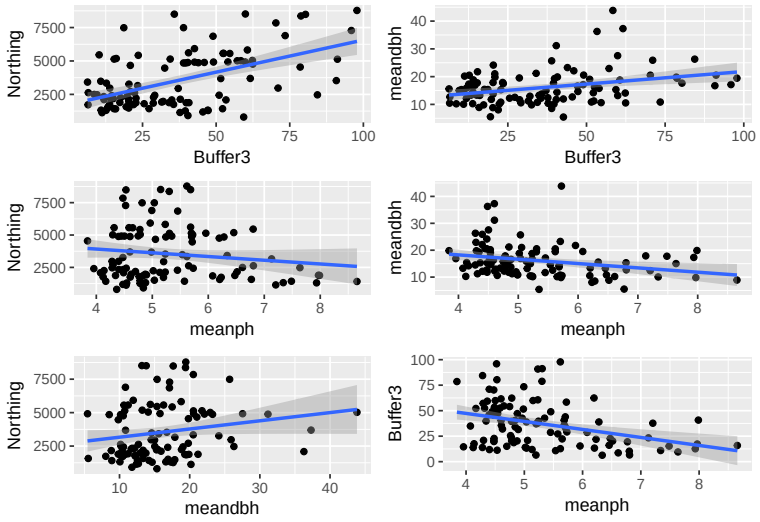
```
##  [1] "Easting"    "Northing"   "PHI"        "Buffer3"    "Richness"
##  [6] "no_NVC"     "no_MSG"     "meandbh"    "meanph"     "meanSOM"
## [11] "meanTD"     "area_ratio"
```

```
##            Easting Northing   PHI Buffer3 no_NVC no_MSG meandbh meanph
## Easting       1.00    -0.27 -0.07   -0.29  -0.05  -0.10   -0.11   0.27
## Northing     -0.27     1.00  0.08    0.52   0.15   0.26    0.18  -0.14
## PHI          -0.07     0.08  1.00    0.08  -0.05   0.00    0.04  -0.01
## Buffer3      -0.29     0.52  0.08    1.00  -0.05   0.26    0.32  -0.36
## no_NVC       -0.05     0.15 -0.05   -0.05   1.00   0.08   -0.24   0.08
## no_MSG       -0.10     0.26  0.00    0.26   0.08   1.00    0.05  -0.37
## meandbh      -0.11     0.18  0.04    0.32  -0.24   0.05    1.00  -0.26
## meanph        0.27    -0.14 -0.01   -0.36   0.08  -0.37   -0.26   1.00
## meanSOM      -0.42     0.19  0.13    0.29  -0.04   0.25    0.06  -0.18
## meanTD        0.37    -0.42 -0.18   -0.38  -0.07  -0.19   -0.44   0.26
## area_ratio    0.09    -0.10  0.01    0.17  -0.10   0.20   -0.02  -0.10
##            meanSOM meanTD area_ratio
## Easting      -0.42   0.37       0.09
## Northing      0.19  -0.42      -0.10
## PHI           0.13  -0.18       0.01
## Buffer3       0.29  -0.38       0.17
## no_NVC       -0.04  -0.07      -0.10
## no_MSG        0.25  -0.19       0.20
## meandbh       0.06  -0.44      -0.02
## meanph       -0.18   0.26      -0.10
## meanSOM       1.00  -0.17       0.20
## meanTD       -0.17   1.00      -0.04
## area_ratio    0.20  -0.04       1.00
```

Correlations over 0.3: Buffer and Northing 0.52 meandbh and buffer 0.32 meanph and buffer -0.36 meanph and no_MSG -0.37 meanTD and Easting 0.37 meanTD and Northing -0.42 meanTD and meandbh -0.44 meanSOM and Easting -0.42

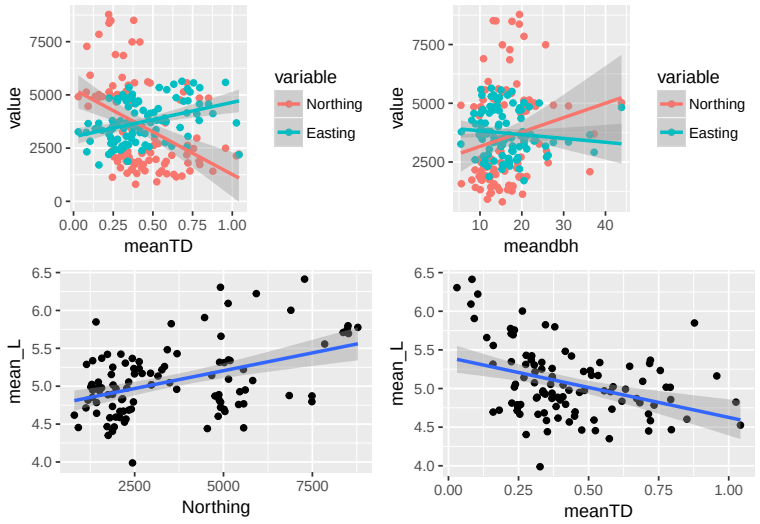
Buffer, Northing, meanpH, meandbh and meanTD

First consider the set of correlations with Northing to buffer to meanph to meandbh. I think these are potentially confounded because why should meandbh be related to meanpH, or meanpH to buffer?



The northern woods tend to have larger buffers, a lower ph and are more mature. There is no reason for meandbh , buffer or meanpH to correlate with each other, so it seems likely that the correlation is related through the latitude to the distribution and age class of the trees and the pH of the woodland. I.e. Buffer is correlated with meanpH because the Northern woods have lower pH and larger buffers.

There is also the correlation of richness to buffer to consider, which may have ecological reasons, does the buffer provide a seed bank for recolonization or protect the woodland from nitrification. But I think if you look at the density of the woods,then the relationship is explained.



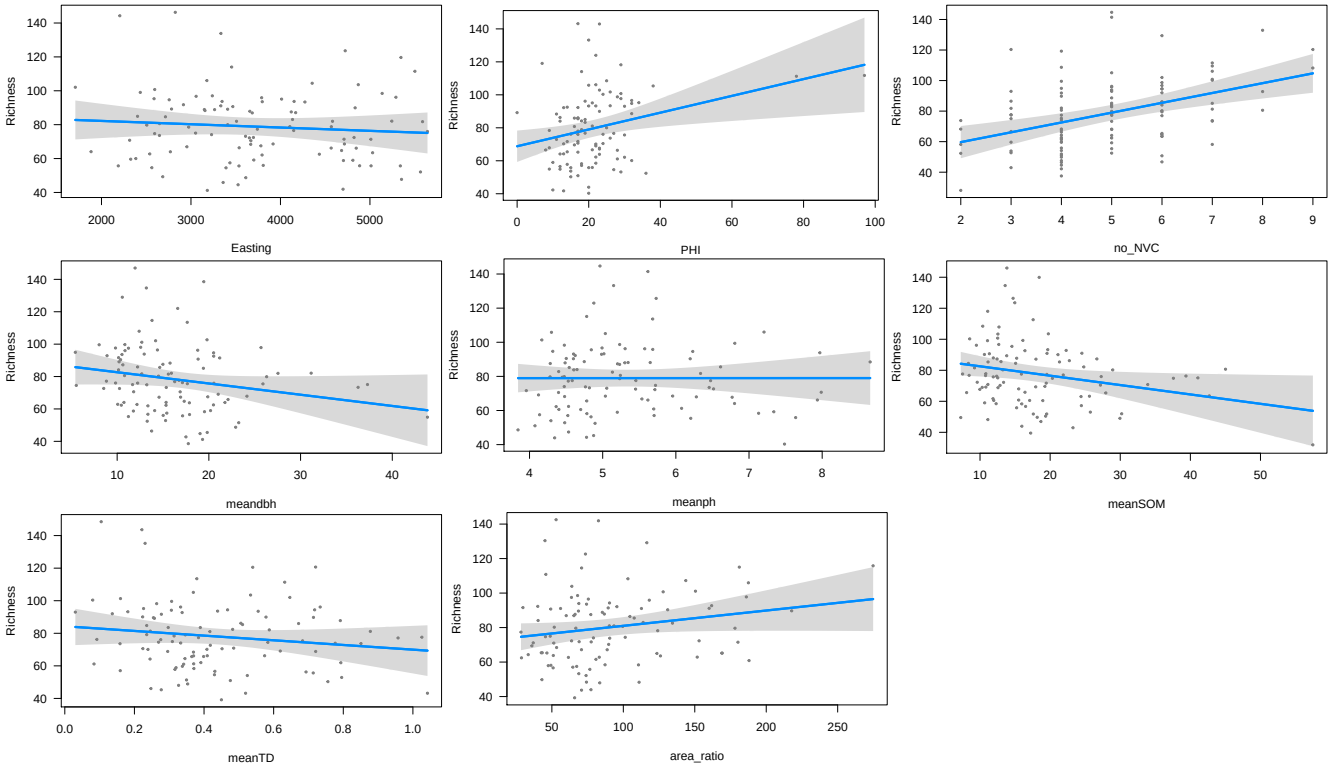
The northern woods are both less dense, and have trees with greater dbh and more light loving plants. This suggests that there is an increase in richness due to the reduced tree density which allows the addition of light loving species. It is probably not the buffer or the latitude that is contributing to the richness, but the openness of the woodlands. Therefore buffer and Northing are probably not required to express the richness, rather the openness of the woodland via the meanTD is what we want.

pH and number of major soil groups

pH is correlated with number of major soil groups, which is not surprising. Number of major soil groups does not correlate with richness and neither does meanpH but I would prefer to retain meanpH instead of number of major soil groups because: Number of major soil groups is likely to be represented by the meanpH, the variable has small range and is not correlated with any other variables, or richness and we know that plant richness correlates with pH even if we arent seeing it here.

The variables are now reduced to...

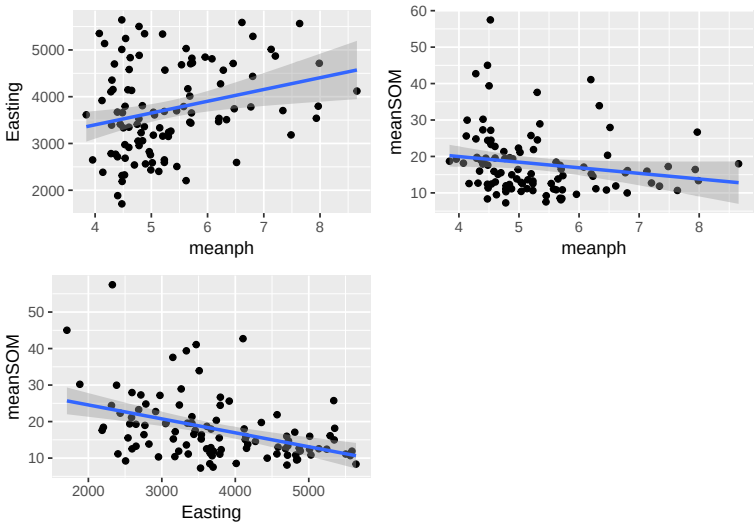
```
## [1] "Easting"    "PHI"        "Richness"   "no_NVC"     "meandbh"
## [6] "meanph"     "meanSOM"    "meanTD"     "area_ratio"
```




```
##             Easting   PHI no_NVC meandbh meanph meanSOM meanTD area_ratio
## Easting        1.00 -0.07  -0.05   -0.11   0.27   -0.42   0.37       0.09
## PHI           -0.07  1.00  -0.05    0.04  -0.01    0.13  -0.18       0.01
## no_NVC        -0.05 -0.05   1.00   -0.24   0.08   -0.04  -0.07      -0.10
## meandbh       -0.11  0.04  -0.24    1.00  -0.26    0.06  -0.44      -0.02
## meanph         0.27 -0.01   0.08   -0.26   1.00   -0.18   0.26      -0.10
## meanSOM       -0.42  0.13  -0.04    0.06  -0.18    1.00  -0.17       0.20
## meanTD         0.37 -0.18  -0.07   -0.44   0.26   -0.17   1.00      -0.04
## area_ratio     0.09  0.01  -0.10   -0.02  -0.10    0.20  -0.04       1.00
```

meanSOM, meanpH and Easting.

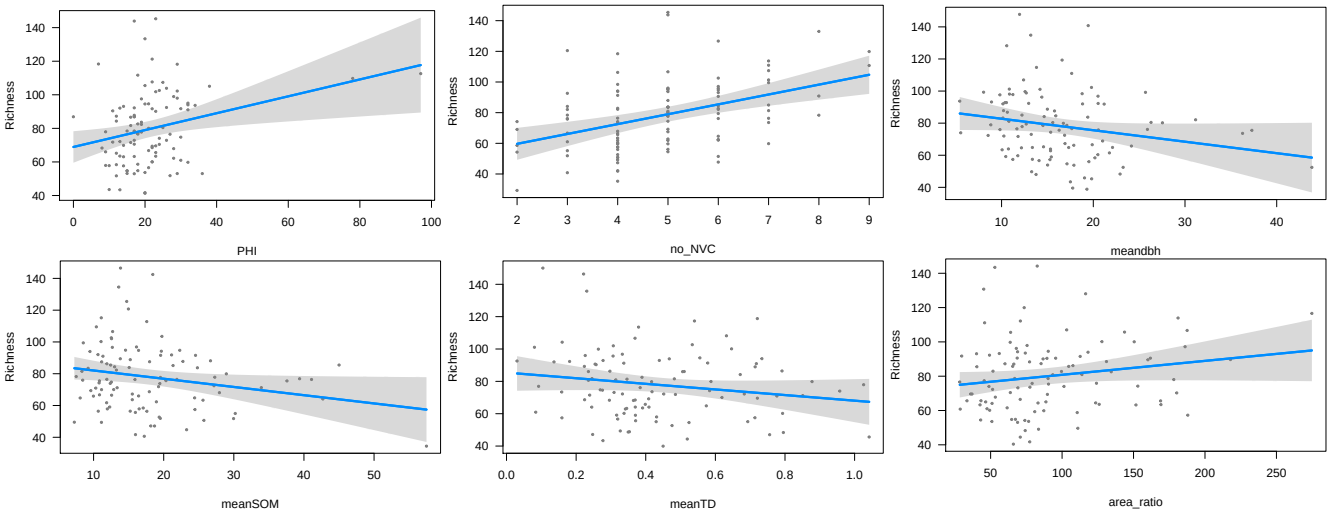
There is a correlation with meanSOM and Easting which we saw previously.



As expected, mean pH decreases with increasing mean SOM. The mean SOM in the Eastern woods is less hence the pH is less. The relationship between Easting and meanSOM is due to relationship between meanSOM and meanpH, and the fact that the meanSOM is reduced in the eastern woods. Since meanSOM is strongly correlated with richness, it is the meanSOM we require, not the Easting. Likewise, since meanSOM explains meanpH, we can probably drop mean pH as well

The variables are now reduced to...

```
## [1] "PHI"        "Richness"   "no_NVC"     "meandbh"    "meanSOM"
## [6] "meanTD"     "area_ratio"
```



All the variables are correlated with richness.

```
##              PHI no_NVC meandbh meanSOM meanTD area_ratio
## PHI         1.00  -0.05    0.04    0.13  -0.18       0.01
## no_NVC     -0.05   1.00   -0.24   -0.04  -0.07      -0.10
## meandbh     0.04  -0.24    1.00    0.06  -0.44      -0.02
## meanSOM     0.13  -0.04    0.06    1.00  -0.17       0.20
## meanTD     -0.18  -0.07   -0.44   -0.17   1.00      -0.04
## area_ratio  0.01  -0.10   -0.02    0.20  -0.04       1.00
```

There are no correlations between variables (greater than0.24), but meandbh is correlated with meanTD(-0.44). Are both required?

```
##                p   R2
## (Intercept)    0 0.31
## PHI         0.01    -
## no_NVC         0    -
## meandbh     0.08    -
## meanSOM     0.04    -
## meanTD      0.14    -
## area_ratio  0.09    -
```

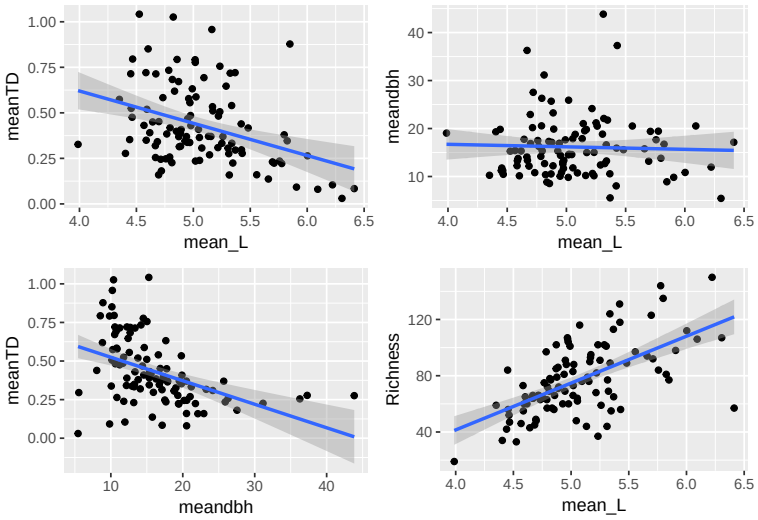
meandbh is more significant than meanTD.

```
##                p   R2
## (Intercept)    0 0.29
## PHI         0.01    -
## no_NVC         0    -
## meanSOM     0.05    -
## meanTD      0.47    -
## area_ratio  0.07    -
```

Removing meandbh does not make meanTD more significant.

```
##                p    r
## (Intercept)    0 0.29
## PHI            0    -
## no_NVC         0    -
## meandbh     0.23    -
## meanSOM     0.06    -
## area_ratio  0.07    -
```

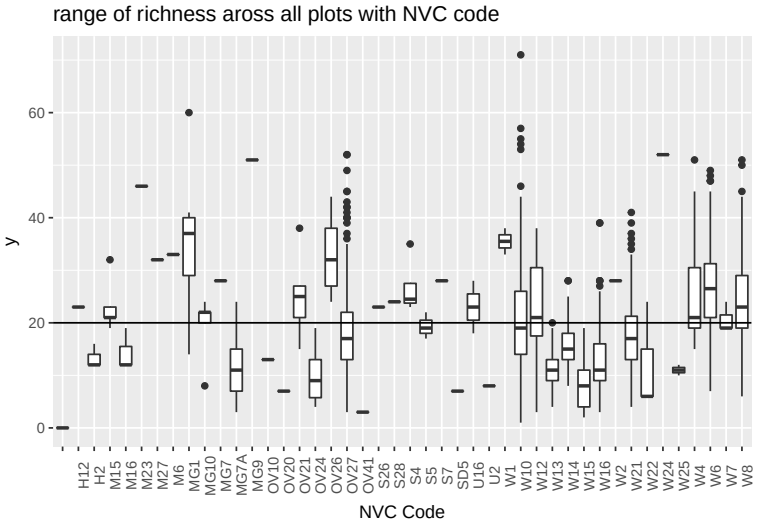
But removing meanTD makes meandbh less significant.



Is a correlation of -0.44 OK?

The graphs above suggest that meanTD is important. As the meanTD reduces, the mean Ellenberg L increases. This is not the case for mean dbh. And as mean Ellenberg L increases, richness increases. We can also see that as meanTD decreases, meandbh increases - as you would expect, you are getting high density for smaller trees, low density for larger trees. Looking at these plots I would be tempted to drop meandbh - but the reduction in significance when they are both included is a bit confusing.

I think you weren't convinced about using number of NVC codes Simon, because it might be too subjective. The following plot made me want to include it.



There is an obvious reduction in richness for the acid oak,beech woods - W14 - W16, and the shady W13 Taxxus baccata. W12, being richer than the others due to its location on more calcareous soils. The damper W4 - W8 woodlands also being richer - due to the dampness. W10, being more neutral, falling in the intermediate richness range - perhaps because they can sometimes be dominated by bramble and bracken, and sometimes not.

MG7A - is the the enriched, species poor Lolium perenne leys, and OV24 is likely to be overrun with nettle. So it looks like the richness for the plots does correspond with their classification - but the question is, does their classification express something about the richness that is NOT already expressed by the less subjective variables I have already chosen - meanSOM, tree density, DBH, PHI (positive heterogeneity index), area ratio.

Looking at the above boxplot, I get the impression that the NVC classification is expressing richness in the way you would expect. But I am using number of NVC codes, not the code itself - because I was thinking that the more codes you have, the more different habitats, the more different species. But if you have W13, W14,W15 and W16, you are getting more codes, but proably not more species? So its not a perfect concept. I would like to do some models on richness and the NVC code, or pull out the occurrences of beech, bramble etc and see how they relate to richness - but I'm also thinking that I need to move on to relating these variables to z and c, and then looking at zeta - then come back to this question as an extra later??

You know whether you think this would be variable that could be published, and how reliable the NVC codes are, so I leave it up to you whether it's included or not.

PS - these are the positive heterogeneity codes I selected, shall I remove some??

```
##      X SD_code     Description
## 1    1       8      Cop. stool
## 2    2       9    Singled cop.
## 3    3      10    Rec.cut cop.
## 4    4      12  Stump hard.new
## 5    5      13  Stump hard.old
## 6    6      14   Stump con.new
## 7    7      15   Stump con.old
## 8    8      16    Stump ovgwn.
## 9    9      17   Brash/pruning
## 10  10      18     Brash heaps
## 11  11      24      Fire sites
## 12  12      54    Log.v rotten
## 13  13      55  Fall bnb.>10cm
```

```
## 14 14      56    Hollow trees
## 15 15      57       Rot holes
## 16 16      58     Stump <10cm
## 17 17      59     Stump >10cm
## 18 18      61       Bryo.base
## 19 19      62      Bryo.trunk
## 20 20      63     Bryo.branch
## 21 21      64    Lichen trunk
## 22 22      65   Lichen branch
## 23 23      86     Pond 1-20m2
## 24 24      87  Pond/lake>20m2
## 25 25      88   Strm.slow <1m
## 26 26      89   Strm.fast <1m
## 27 27      90   Riv.slow 1-5m
## 28 28      91   Riv.fast 1-5m
## 29 29      92    Riv.slow >5m
## 30 30      93    Riv.fast >5m
## 31 31     105   Gld.5-12m grs
## 32 32     106    Gld.>12m grs
## 33 33     107  Gld.3-12m mxd.
## 34 34     108        "Gld.>12
## 35 35     109   Gld.5-12m bgy
## 36 36     110    Gld.>12m bgy
## 37 37     111  Rky.knoll <12m
## 38 38     112  Rky.knoll >12m
## 39 39     113           Field
## 40 40     114       Path 1-5m
## 41 41     115        Ride >5m
## 42 42     116 Track non-prep.
## 43 43     150       Moss bank
## 44 44     151       Fern bank
## 45 45     152      Grass bank
## 46 46     153      Leaf drift
## 47 47     157 Macrofungi soil
## 48 48     158 Macrofungi wood
## 49 49     208   Rough grazing
## 50 50     209  Heath/moorland
## 51 51     210   Marsh/fen/bog
## 52 52     211           River
## 53 53     212            Lake
```