MIIN Part 4: Meta-dataset overview

Marissa Lee

June 1, 2015

Filename: MIIN_4_datasetOverview.Rmd
This markdown file does the following tasks: 1. Article selection statistics

2. Number of papers and observations

3. Types of observations

4. Plant species statistics

5. Cover data statistics

6. Trait data statistics

7. Soil measurement statistics

8. Effect size statistics

9. CWM trait value statistics

```
knitr::opts_chunk$set(cache=TRUE)

citation()
```

```
##
## To cite R in publications use:
##
##   R Core Team (2015). R: A language and environment for
##   statistical computing. R Foundation for Statistical Computing,
##   Vienna, Austria. URL http://www.R-project.org/.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {R: A Language and Environment for Statistical Computing},
##     author = {{R Core Team}},
##     organization = {R Foundation for Statistical Computing},
##     address = {Vienna, Austria},
##     year = {2015},
##     url = {http://www.R-project.org/},
##   }
##
## We have invested a lot of time and effort in creating R, please
## cite it when using it for data analysis. See also
## 'citation("pkgname")' for citing R packages.
```

```
library(plyr)
if(nchar(system.file(package="plyr"))) citation("plyr")
```

```
##
## To cite plyr in publications use:
##
##   Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data
##   Analysis. Journal of Statistical Software, 40(1), 1-29. URL
##   http://www.jstatsoft.org/v40/i01/.
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {The Split-Apply-Combine Strategy for Data Analysis},
##     author = {Hadley Wickham},
##     journal = {Journal of Statistical Software},
##     year = {2011},
##     volume = {40},
##     number = {1},
##     pages = {1--29},
##     url = {http://www.jstatsoft.org/v40/i01/},
##   }
```

```r
library(doBy)
```

```
## Loading required package: survival
```

```r
library(ggplot2)
if(nchar(system.file(package="ggplot2"))) citation("ggplot2")
```

```
##
## To cite ggplot2 in publications, please use:
##
##   H. Wickham. ggplot2: elegant graphics for data analysis.
##   Springer New York, 2009.
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     author = {Hadley Wickham},
##     title = {ggplot2: elegant graphics for data analysis},
##     publisher = {Springer New York},
##     year = {2009},
##     isbn = {978-0-387-98140-6},
##     url = {http://had.co.nz/ggplot2/book},
##   }
```

```r
library(reshape2)
library(gridExtra)
library(metafor)
```

```
## Loading required package: Matrix
## Loading 'metafor' package (version 1.9-7). For an overview
## and introduction to the package please type: help(metafor).
```

```r
if(nchar(system.file(package="metafor"))) citation("metafor")
```

```
##
## To cite the metafor package in publications, please use:
##
##   Wolfgang Viechtbauer (2010). Conducting meta-analyses in R with
##   the metafor package. Journal of Statistical Software, 36(3),
##   1-48. URL http://www.jstatsoft.org/v36/i03/.
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {Conducting meta-analyses in {R} with the {metafor} package},
##     author = {Wolfgang Viechtbauer},
##     journal = {Journal of Statistical Software},
##     year = {2010},
##     volume = {36},
##     number = {3},
##     pages = {1--48},
##     url = {http://www.jstatsoft.org/v36/i03/},
##   }
```

```r
source('CODE/mytheme.R')
```

```
## Loading required package: grid
```

```r
figuresPath<-file.path(getwd()[1], "FIGURES_TABLES", "overview") #where to put the saved plots
fig.height<-2.5 #inches
fig.width<- 2.5 #inches
fig.res<-300

#from MIIN_3_calcEffectSizes.Rmd
papers<-read.table("DATA/DATA_SYNTHESIZED/calcES/papers.txt", sep="\t")
observations<-read.table("DATA/DATA_SYNTHESIZED/calcES/observations.txt", header=TRUE, sep="\t")
cover<-read.table("DATA/DATA_SYNTHESIZED/calcES/cover.txt", header=TRUE, sep="\t")
species<-read.table("DATA/DATA_SYNTHESIZED/calcES/species.txt", header=TRUE, sep="\t")
traits<-read.table("DATA/DATA_SYNTHESIZED/calcES/traits.txt", header=TRUE, sep="\t")
measures<-read.table("DATA/DATA_SYNTHESIZED/calcES/measures.txt", header=TRUE, sep="\t")
cwm<-read.table("DATA/DATA_SYNTHESIZED/calcES/cwm.txt", header=TRUE, sep="\t")
spIDcover<-read.table("DATA/DATA_SYNTHESIZED/calcES/spIDcover.txt", header=TRUE, sep="\t")
spIDtraits<-read.table("DATA/DATA_SYNTHESIZED/calcES/spIDtraits.txt", header=TRUE, sep="\t")
metaDataset<-read.table("DATA/DATA_SYNTHESIZED/calcES/metaDataset.txt", header=TRUE, sep="\t")
```

---

1. Article selection statistics

```
### Number of papers detected by source ###
summ.papers <- ddply(papers,~source,summarise,
                     numPapers=length(read),
                     numAcceptedPapers=sum(reject=='No'))
summ.papers<-orderBy(~-numPapers, summ.papers)
summ.papers
```

```
##                                 source numPapers numAcceptedPapers
## 19                      search2_111714       388                36
## 18                      search1_111714       219                46
## 12                            Liao2007        94                47
## 11 independent search for plant traits         3                 0
## 3                         cited by 249         2                 2
## 8                         cited by 368         2                 2
## 10                        cited by 626         2                 1
## 1                         cited by 155         1                 1
## 2                         cited by 229         1                 0
## 4                          cited by 25         1                 1
## 5                         cited by 256         1                 0
## 6                          cited by 29         1                 1
## 7                         cited by 317         1                 1
## 9                         cited by 455         1                 1
## 13                        ReferencedBy         1                 1
## 14                  related record 181         1                 1
## 15                  related record 188         1                 1
## 16                    related record 4         1                 0
## 17                  related record 570         1                 1
```

```
### Number of unique number of papers detected ###
summ.papers2 <- ddply(papers,~source+rejectRationale,summarise,
                     numPapers=length(read),
                     numAcceptedPapers=sum(reject=='No'))
summ.papers2<-orderBy(~-numPapers, summ.papers2)
totalNumReturned<-sum(summ.papers$numPapers) #total number of papers detected
numAlreadyFound<-sum(summ.papers2[summ.papers2$rejectRationale == 'alreadyFound' & !is.na(summ.papers2$
numUnique<-totalNumReturned - numAlreadyFound #total number of unique papers detected
paste(numUnique, 'unique papers identified by search criteria and their references')
```

```
## [1] "483 unique papers identified by search criteria and their references"
```

```
paste(sum(summ.papers$numAcceptedPapers), 'papers were accepted')
```

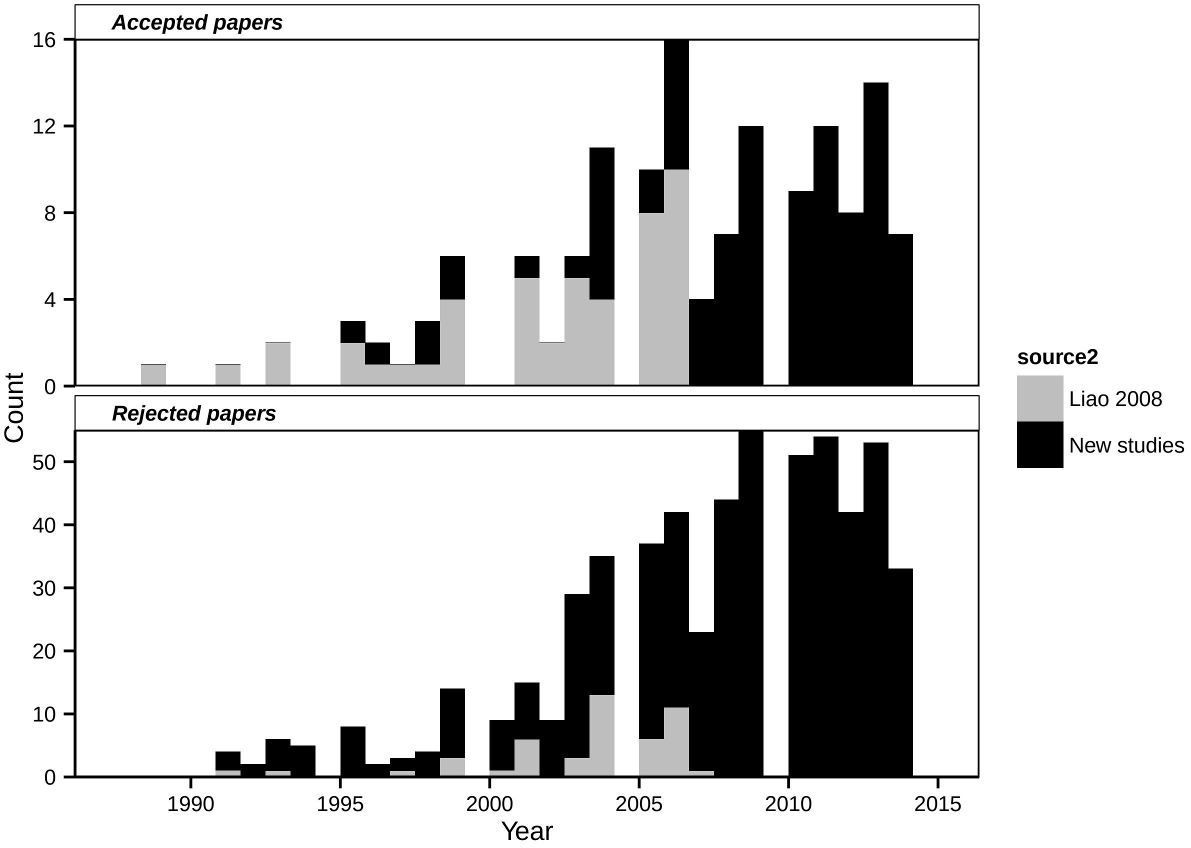
```
## [1] "143 papers were accepted"
```

```
length(unique(metaDataset$paperID)) #this should be the same number
```

```
## [1] 143
```

```
### Subset papers detected by the previous meta-analysis, Liao2007 ###
papers$source2<-'New studies'
```

```
papers[papers$source=='Liao2007','source2']<-'Liao 2008'
papers$reject<-revalue(papers$reject, c("No"="Accepted papers", "Yes"="Rejected papers"))

#plot number of papers that were accepted/rejected from Liao 2008 and this search
pHist_papers<-ggplot(papers, aes(x=year, fill=source2)) + mytheme +
  facet_wrap(~reject, scales="free_y", ncol=1) +
  geom_histogram() +  scale_y_continuous(expand = c(0,0)) +
  ylab('Count') + xlab('Year') +
  scale_fill_manual(values=c('gray','black'))
pHist_papers
```



```
newfilename<-"pHist_papers.png"
png(paste(figuresPath,newfilename, sep='/'),
    units='in', width = fig.width*1.5, height = fig.height*2, res=fig.res)
pHist_papers
dev.off()
```

```
## pdf
##   2
```

```
#what was the year of the most recent data included in Liao 2008?
maxLiaoyr<-max(papers[papers$source == 'Liao2007','year'])
paste(maxLiaoyr, 'was the most recent year that data was included in the Liao 2008 meta-analysis')
```

```
## [1] "2007 was the most recent year that data was included in the Liao 2008 meta-analysis"
```

```
#how many accepted papers were published after the most recent Liao 2008 reference?
accepted.after<-subset(papers, source != 'Liao2007' & reject == 'Accepted papers' & year > maxLiaoyr)
paste(dim(accepted.after)[1], 'papers were accepted after the most recent reference included in Liao 20
```

```
## [1] "69 papers were accepted after the most recent reference included in Liao 2008"
```

```
#how many papers were rejected that were referenced in Liao 2008? Remember that Liao 2008 also addresse
rejected.Liao<-subset(papers, source == 'Liao2007' & reject == 'Rejected papers')
numLiaoRej<-dim(rejected.Liao)[1]
all.Liao<-subset(papers, source == 'Liao2007')
numLiaoAll<-dim(all.Liao)[1]
paste(dim(rejected.Liao)[1], 'papers that were used in Liao 2008 were rejected from this study, or', rou
```

```
## [1] "47 papers that were used in Liao 2008 were rejected from this study, or 50 % of Liao references
```

```
#How many papers were accepted that were published before the most recent Liao 2008 reference and were
accepted.before<-subset(papers, source != 'Liao2007' & reject == 'Accepted papers' & year < maxLiaoyr)
numAccBef<-dim(accepted.before)[1]
paste(dim(accepted.before)[1], 'papers that were published before the most recent Liao 2008 reference we
```

```
## [1] "23 papers that were published before the most recent Liao 2008 reference were included in this
```

```
#attached the 'source2' column to the metaDataset
temp_indx<-papers[,c('paperID','source2')]
metaDataset<-merge(metaDataset, temp_indx, by='paperID')

#re-write the metaDataset file so that it has the source2 column
newfilename<-'metaDataset.txt'
synthdataPath<-file.path(getwd()[1], "DATA", "DATA_SYNTHESIZED", "calcES")
write.table(metaDataset, file=paste(synthdataPath,newfilename, sep='/'), sep='\t')
```

---

2. Number of papers and observations

```
#how many observations?
paste(length(unique(observations$obsID)), 'observations in the full dataset')
```

```
## [1] "404 observations in the full dataset"
```

```
length(unique(metaDataset$obsID)) #these should be the same
```

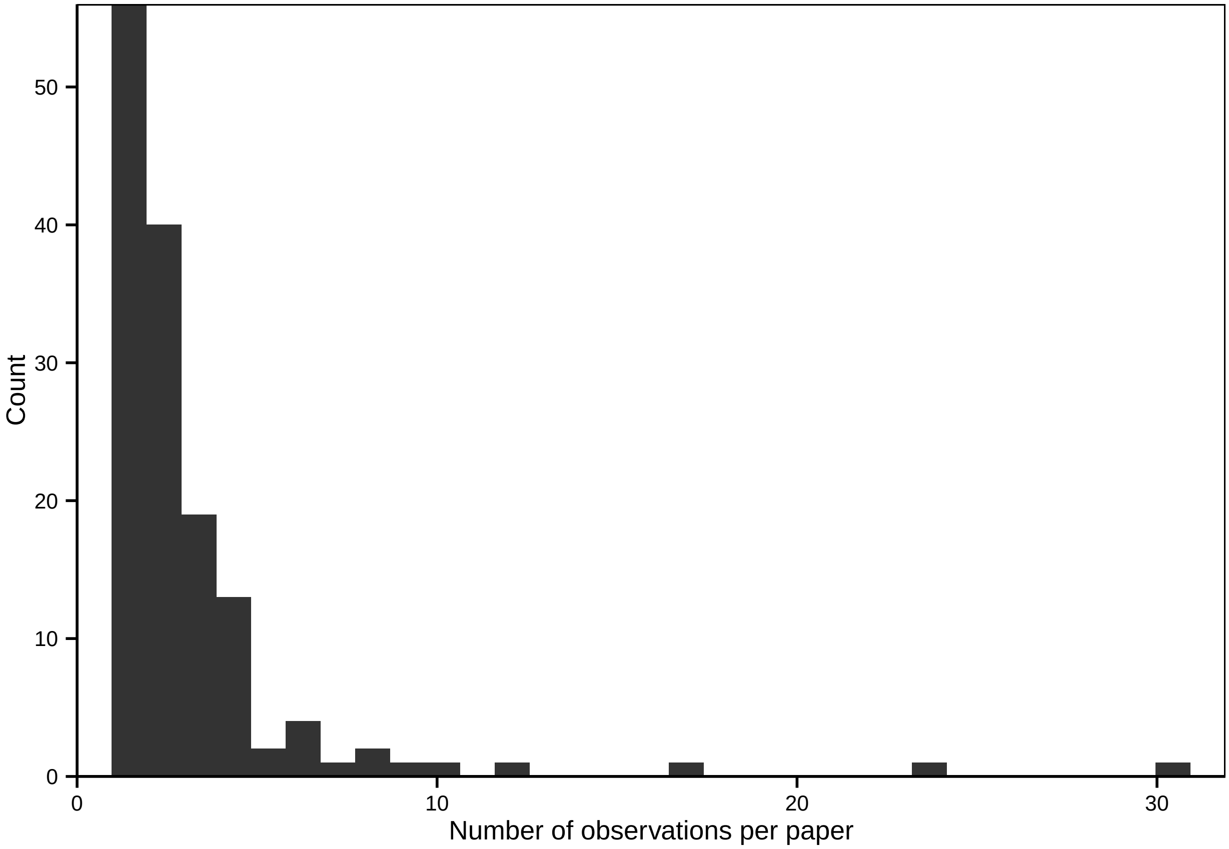
```
## [1] 404
```

```
#how many observations per paper?
summ.obs <- ddply(observations,~paperID,summarise, numObs=length(paperID))
median(summ.obs$numObs); range(summ.obs$numObs)
```

```
## [1] 2
```

```
## [1]  1 30
```

```
pHist_obs<-ggplot(summ.obs, aes(x=numObs)) +
  scale_y_continuous(expand=c(0,0)) + scale_x_continuous(expand=c(0,0)) +
  geom_histogram() + mytheme +
  ylab('Count') + xlab('Number of observations per paper')
pHist_obs
```



3. Types of observations

```
summ.obs.eco <- ddply(observations,~ecosystCat,summarise, numObs=length(paperID))
summ.obs.st <- ddply(observations,~studyType,summarise, numObs=length(paperID))
summ.obs.nfix <- ddply(observations,~Nfix,summarise, numObs=length(paperID))
factorlist<-list(summ.obs.eco, summ.obs.st, summ.obs.nfix)
factortab<-ldply(factorlist)
factortab$factor<-c(rep('ecosystem',5), rep('studyType', 4), rep('Nfix',4))
factortab$level<-NA
factortab[!is.na(factortab$ecosystCat),'level']<-as.character(factortab[!is.na(factortab$ecosystCat),'ec
factortab[!is.na(factortab$studyType),'level']<-as.character(factortab[!is.na(factortab$studyType),'stu
factortab[!is.na(factortab$Nfix),'level']<-as.character(factortab[!is.na(factortab$Nfix),'Nfix'])
factortab1<-factortab[,c('factor','level','numObs')]
factortab1
```

```
##       factor                          level numObs
## 1  ecosystem                         forest    123
## 2  ecosystem                      grassland    176
## 3  ecosystem                          other      4
## 4  ecosystem                      shrubland     73
## 5  ecosystem                        wetland     28
## 6  studyType            field expt addition     45
## 7  studyType             field expt removal     25
## 8  studyType                    field study    273
## 9  studyType                greenhouse expt     61
## 10      Nfix Invasive and resident N-fixers     20
## 11      Nfix         Invasive N-fixers only     51
## 12      Nfix                    No N-fixers    295
## 13      Nfix         Resident N-fixers only     38
```

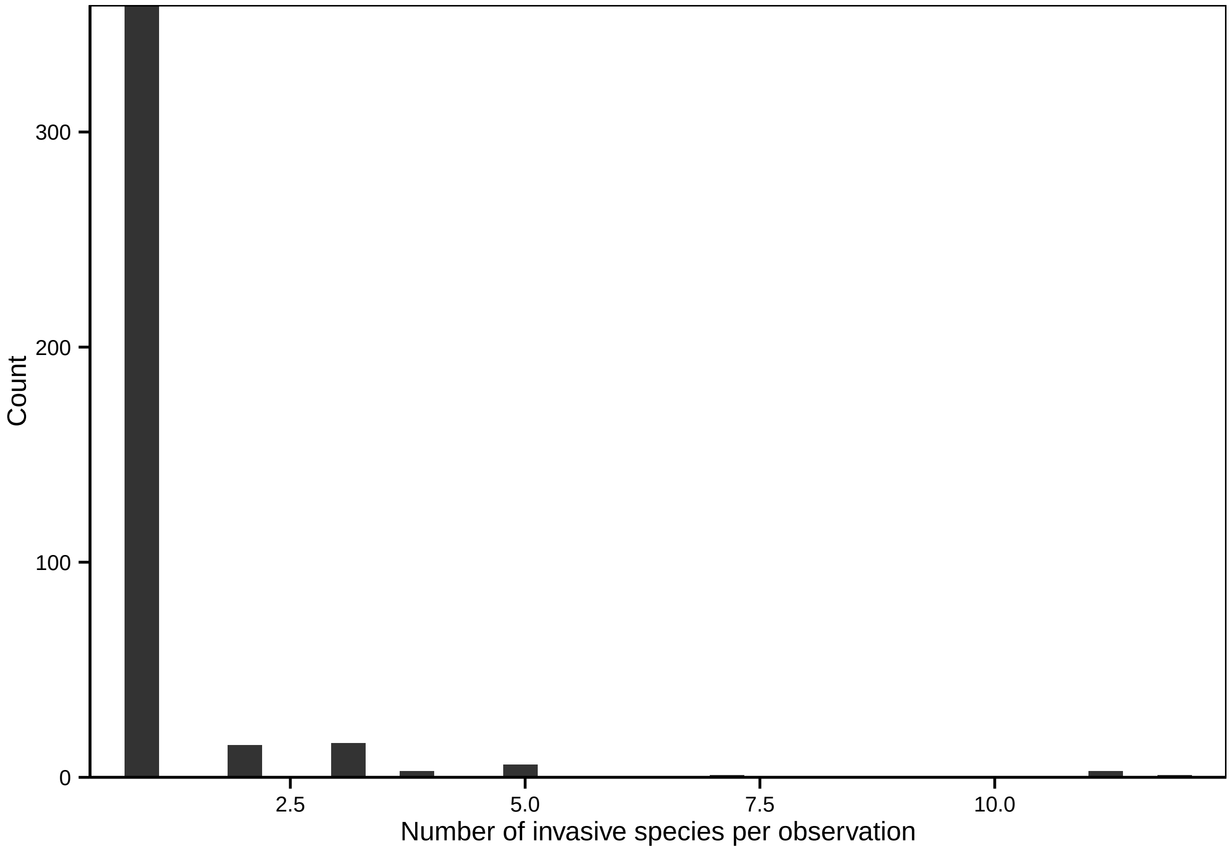
```
newfilename<-'numObsTable.txt'
write.table(factortab1, file=paste(figuresPath,newfilename, sep='/'), sep='\t')
```

---

4. Plant species statistics

What is the distribution of invasive species per observation? Native species? Are certain invasive species over-represented?

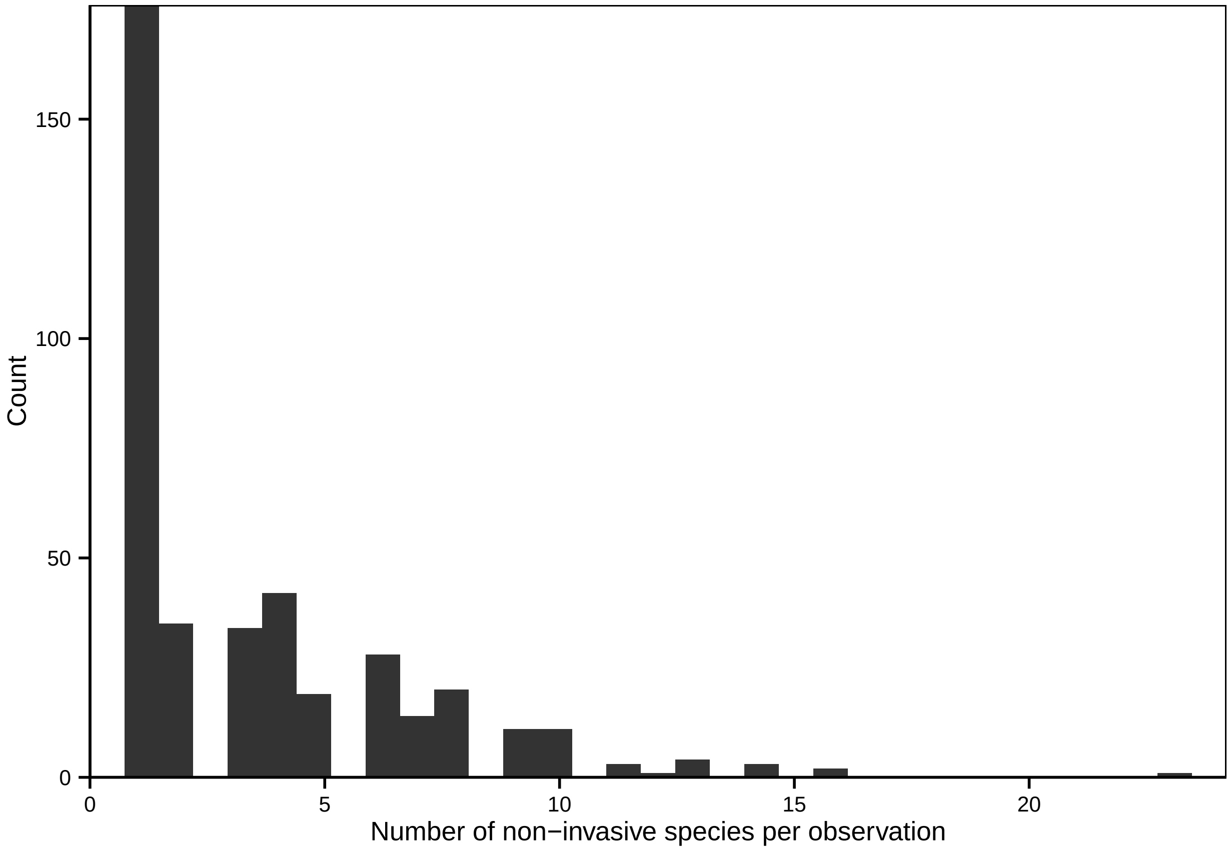
```
#what is the distribution of invasive species per observation?
summ.spp <- ddply(species,~obsID,summarise,
                  numTotalspp=length(obsID),
                  numInvspp=sum(spInvasive=='invasive' & spExotic=='exotic' & spFocal=='focal'),
                  numNonInvspp=sum(spInvasive=='not invasive'),
                  numOthers=numTotalspp-(numInvspp + numNonInvspp))
hist_Inv<-ggplot(summ.spp, aes(x=numInvspp)) + geom_histogram() +
  scale_y_continuous(expand=c(0,0)) + scale_x_continuous(expand=c(0,0)) +
  mytheme +
  ylab('Count') + xlab('Number of invasive species per observation')
hist_Inv; median(summ.spp$numInvspp); range(summ.spp$numInvspp)
```



```
## [1] 1
```

```
## [1]  1 12
```

```
hist_Nat<-ggplot(summ.spp, aes(x=numNonInvspp)) + geom_histogram() +
  scale_y_continuous(expand=c(0,0)) + scale_x_continuous(expand=c(0,0)) +
  mytheme +
  ylab('Count') + xlab('Number of non-invasive species per observation')
hist_Nat; median(summ.spp$numNonInvspp); range(summ.spp$numNonInvspp)
```



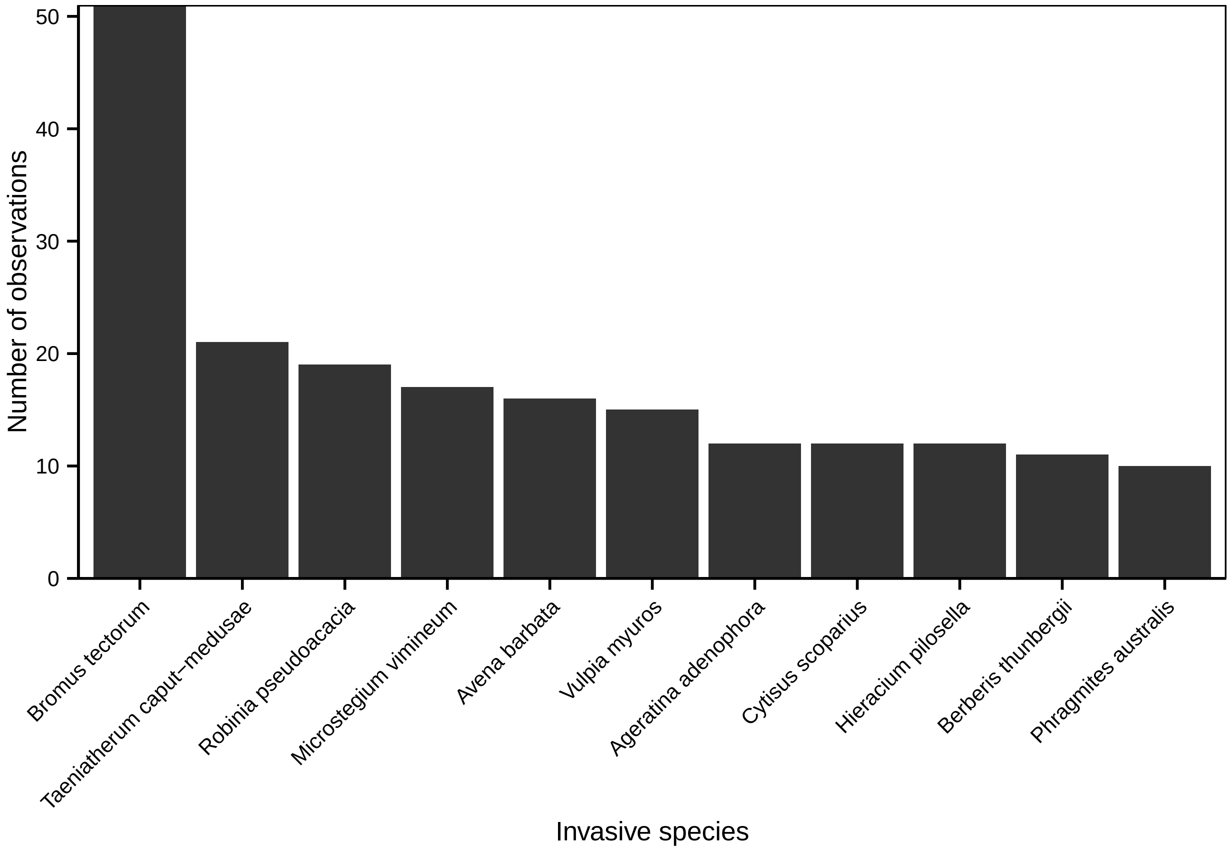
```
## [1] 2
```

```
## [1]  1 23
```

```
#number of observations per species
summ.spp.nam <- ddply(species,~spName+spFocal+spExotic,summarise,
                  numObs=length(obsID),
                  numPapers=length(unique(paperID)))
spp.many<-summ.spp.nam[which(summ.spp.nam$numObs > 9 & summ.spp.nam$spFocal == 'focal'),] #more than 9
spp.many.o<-orderBy(~-numObs, spp.many)
spp.many.o
```

```
##                         spName spFocal spExotic numObs numPapers
## 130            Bromus tectorum   focal   exotic     51        19
## 616 Taeniatherum caput-medusae   focal   exotic     21         3
## 551       Robinia pseudoacacia   focal   exotic     19         7
## 405      Microstegium vimineum   focal   exotic     17         7
## 92               Avena barbata   focal   exotic     16         5
## 660              Vulpia myuros   focal   exotic     15         4
## 29        Ageratina adenophora   focal   exotic     12         1
## 216          Cytisus scoparius   focal   exotic     12         6
## 314        Hieracium pilosella   focal   exotic     12         5
## 103        Berberis thunbergii   focal   exotic     11         4
## 463       Phragmites australis   focal   exotic     10         5
```

```
positions<-spp.many.o$spName
pHist_spp<-ggplot(spp.many.o, aes(x=spName, y=numObs)) + geom_bar(stat='identity') +
  scale_y_continuous(expand=c(0,0)) + scale_x_discrete(limits = positions) +
  mytheme +  theme(axis.text.x=element_text(angle=45, hjust=1)) +
  ylab('Number of observations') + xlab('Invasive species')
pHist_spp
```



```
newfilename<-'pHist_spp.png'
png(paste(figuresPath,newfilename, sep='/'),
    units='in', width = fig.width*2, height = fig.height*2, res=fig.res)
pHist_spp
dev.off()
```

```
## pdf
##   2
```

```
#which species appear both as exotic and native species in the dataset?
summ.spp <- ddply(species,~spName+spExotic, summarise,
                  numObs=length(obsID),
                  numPapers=length(unique(paperID)))
summ.spp.nam2 <- ddply(summ.spp,~spName,summarise,
                  numInvNat=length(spExotic))
summ.spp.nam2[summ.spp.nam2$numInvNat==2,] # if the length of spInvasive col==2, then there is native a
```

```
##                       spName numInvNat
## 5       Acacia longifolia         2
## 10           Acer negundo         2
## 61  Anthoxanthum odoratum         2
## 115          Briza maxima         2
## 118     Bromus hordeaceus         2
## 208     Cytisus scoparius         2
## 209    Dactylis glomerata         2
## 269   Festuca arundinacea         2
## 305        Holcus lanatus         2
## 451  Phragmites australis         2
## 477         Poa pratensis         2
## 498       Prunus serotina         2
## 553   Schedonorus phoenix         2
## 615          Trifolium sp         2
```

---

5. Cover data statistics

What percent of observations have measured cover data?

```
summ.cov.obs <- ddply(cover,~obsID,summarise,
                  numMeasured= sum(covQuality=='measured'))
numMeasured.obs<-sum(summ.cov.obs$numMeasured > 0) #number of observations with cover measurement value
numtotal.obs<-length(summ.cov.obs$numMeasured > 0) #total number of observations
cov.obs.perc<-round((numMeasured.obs / (numtotal.obs) ) *100, digits=2)
paste(cov.obs.perc, '% of observations have any cover data at all that was measured in the original pape
```

```
## [1] "27.97 % of observations have any cover data at all that was measured in the original paper"
```

```
#What is the frequency of cover observations for each cover measure type?
summ.cov <- ddply(cover,~covCat,summarise,
                  numMeas = length(obsID),
                  numObs=length(unique(obsID)),
                  numPapers=length(unique(paperID)))
orderBy(~-numMeas, summ.cov)
```

```
##        covCat numMeas numObs numPapers
## 3 sp_plantcov    1141    404       143
## 1  sp_biomass      81     32        13
## 2      sp_ind      80     22         6
```

```
#What units are commonly reported for each cover measure type?
summ.covUnit <- ddply(cover,~covCat+covUnit,summarise,
                      numMeas = length(obsID),
                      numObs=length(unique(obsID)),
                      numPapers=length(unique(paperID)))
COVCAT<-unique(summ.covUnit$covCat)
covUnitList<-list()
i<-0
for(i in 1:length(COVCAT)){
  subdf<-summ.covUnit[summ.covUnit$covCat==COVCAT[i],]
  covUnitList[[as.character(COVCAT[i])]]<-orderBy(~-numMeas, subdf)
}
covUnitList
```

```
## $sp_biomass
##       covCat covUnit numMeas numObs numPapers
## 2 sp_biomass    g/m2      38     13         6
## 5 sp_biomass  ind/ha      12      6         1
## 7 sp_biomass   m2/ha      11      3         1
## 4 sp_biomass   g/pot       6      3         1
## 6 sp_biomass   kg/m2       6      3         2
## 1 sp_biomass       %       4      2         1
## 3 sp_biomass  g/m2*y       4      2         1
##
## $sp_ind
##    covCat     covUnit numMeas numObs numPapers
## 11 sp_ind notReported      34      8         1
## 9  sp_ind    ind/30m2      19      1         1
## 13 sp_ind    stems/m2      11      5         1
## 12 sp_ind   plants/m2      10      5         1
## 8  sp_ind    ind/10m2       4      1         1
## 10 sp_ind      ind/ha       2      2         1
##
## $sp_plantcov
##         covCat covUnit numMeas numObs numPapers
## 14 sp_plantcov       %    1141    404       143
```

```
#A more detailed look at cover data quality as it contributes to CWM values...
cwm.calc<-subset(cwm, qualityCWMcalc == 'calculated')
summ.cwm <- ddply(cwm.calc,~traitCat+invType,summarise,
                  numObs=length(unique(obsID)),
                  num1spAll_1=sum(qualityCover=='Measured=All, 1sp=All'),
                  num1spAll_2=sum(qualityCover=='Measured=None, 1sp=All'),
                  num1spAll_3=sum(qualityCover=='Measured=NA, 1sp=NA'),
                  num1spAll_4=sum(qualityCover=='Measured=Mid, 1sp=All'),
                  totalspAll=sum(num1spAll_1, num1spAll_2, num1spAll_3, num1spAll_4),
                  perc1spAll=(totalspAll/numObs) *100,
                  percEqual=100-perc1spAll)
summ.cwm #cover data quality by traitCat and invType
```

```
##       traitCat      invType numObs num1spAll_1 num1spAll_2 num1spAll_3
## 1           cn      InvArea    198          15         138           0
## 2           cn InvSpInvArea    198          18         154           0
## 3           cn      NatArea    198          15          78          57
## 4     littercn      InvArea     40           5          27           0
## 5     littercn InvSpInvArea     40           8          31           0
## 6     littercn      NatArea     40           5          14           6
## 7  litterpercN      InvArea     42           5          24           0
## 8  litterpercN InvSpInvArea     42           6          29           0
## 9  litterpercN      NatArea     42           5          17           7
## 10       percN      InvArea    318          32         189           2
## 11       percN InvSpInvArea    318          52         232           2
## 12       percN      NatArea    318          36         144           23
##    num1spAll_4 totalspAll perc1spAll percEqual
## 1            2        155   78.28283  21.71717
## 2            0        172   86.86869  13.13131
## 3            0        150   75.75758  24.24242
## 4            1         33   82.50000  17.50000
## 5            0         39   97.50000   2.50000
## 6            1         26   65.00000  35.00000
## 7            0         29   69.04762  30.95238
## 8            0         35   83.33333  16.66667
## 9            0         29   69.04762  30.95238
## 10           6        229   72.01258  27.98742
## 11           0        286   89.93711  10.06289
## 12           1        204   64.15094  35.84906
```

```
summ.cwm2 <- ddply(summ.cwm,~invType,summarise,
                  mean1sp=mean(perc1spAll),
                  meanEqual=mean(percEqual),
                  seEqual=sd(percEqual)/sqrt(length(percEqual)))
summ.cwm2 #aggregated across traitCat
```

```
##        invType  mean1sp meanEqual  seEqual
## 1      InvArea 75.46076  24.53924 3.034981
## 2 InvSpInvArea 89.40978  10.59022 3.015379
## 3      NatArea 68.48903  31.51097 2.647894
```

```
summ.cwm3 <- ddply(cwm,~invType+traitCat+obsID,summarise,
                  numReported=sum(qualityCWMcalc=='reported'))
summ.cwm4 <- ddply(summ.cwm3,~traitCat+invType,summarise,
                  count=sum(numReported != 0),
                  total=length(obsID),
                  percCWMReported=(count/total)*100)
summ.cwm4 #percent of CWM data that was reported in the original paper (rather than calculated based on
```

```
##       traitCat      invType count total percCWMReported
## 1           cn      InvArea    14   212        6.603774
## 2           cn InvSpInvArea    14   212        6.603774
## 3           cn      NatArea    14   212        6.603774
## 4     littercn      InvArea    21    61       34.426230
## 5     littercn InvSpInvArea    21    61       34.426230
## 6     littercn      NatArea    21    61       34.426230
## 7  litterpercN      InvArea    26    68       38.235294
## 8  litterpercN InvSpInvArea    26    68       38.235294
## 9  litterpercN      NatArea    26    68       38.235294
## 10       percN      InvArea    53   371       14.285714
## 11       percN InvSpInvArea    53   371       14.285714
## 12       percN      NatArea    53   371       14.285714
```

6. Trait data statistics

```
# what percent of observations had trait data reported within the original article?
n.ot<-length(unique(traits$obsID)) # number of observations with trait data
n.o<-length(unique(observations$obsID)) # total number of observations
tr.obs.perc<-round((n.ot/n.o) *100, digits=2) # percent of observations with trait data
paste(tr.obs.perc, '% of observations with species-level trait data from the original paper',collapse='
```

```
## [1] "34.41 % of observations with species-level trait data from the original paper"
```

```
summ.tr <- ddply(traits,~traitCat,summarise,
                 numObs = length(unique(obsID)),
                 numPapers = length(unique(paperID)))
summ.tr.o<-orderBy(~-numObs, summ.tr)
summ.tr.o
```

```
##         traitCat numObs numPapers
## 4        sp_percN    106        40
## 1           sp_cn     54        21
## 3 sp_litterpercN     48        22
## 2    sp_littercn     32        18
```

```
positions<-summ.tr.o$traitCat
pBar.tr<-ggplot(summ.tr.o, aes(x=traitCat, y=numObs)) + geom_bar(stat='identity') +
  scale_y_continuous(expand=c(0,0)) +
  scale_x_discrete(limits = positions,
                   labels = c("sp_percN" = "Leaf %N",
                              "sp_cn" = "Leaf C:N",
                              "sp_litterpercN" = "Litter %N",
                              "sp_littercn" = "Litter C:N")) +
  mytheme +  theme(axis.text.x=element_text(angle=45, hjust=1)) +
  ylab('Number of observations') + xlab('Trait type (species-level)')

#What units and methods are commonly reported for each measurement?
summ.traitUnit <- ddply(traits,~traitCat+traitUnit,summarise,
                   numMeas = length(obsID),
                   numObs=length(unique(obsID)),
                   numPapers=length(unique(paperID)))
summ.traitUnit
```

```
##          traitCat traitUnit numMeas numObs numPapers
## 1           sp_cn         %      16      4         2
## 2           sp_cn     %C/%N      11      5         5
## 3           sp_cn molC/molN     128     45        14
## 4     sp_littercn     %C/%N       6      2         2
## 5     sp_littercn molC/molN      81     30        16
## 6  sp_litterpercN         %      85     29        14
## 7  sp_litterpercN      g/kg       6      4         2
## 8  sp_litterpercN      mg/g      38     15         6
## 9        sp_percN         %     160     58        23
## 10       sp_percN      g/kg      50     21         3
## 11       sp_percN      mg/g      66     23         11
## 12       sp_percN   mmol/kg       4      1          1
## 13       sp_percN      ug/g      14      2          1
## 14       sp_percN     ug/mg       2      1          1
```

```
TRAITCAT<-unique(summ.traitUnit$traitCat)
traitUnitList<-list()
i<-0
for(i in 1:length(TRAITCAT)){
  subdf<-summ.traitUnit[summ.traitUnit$traitCat==TRAITCAT[i],]
  traitUnitList[[as.character(TRAITCAT[i])]]<-orderBy(~-numMeas, subdf)
}
traitUnitList
```

```
## $sp_cn
##   traitCat traitUnit numMeas numObs numPapers
## 3    sp_cn molC/molN     128     45        14
## 1    sp_cn         %      16      4         2
## 2    sp_cn     %C/%N      11      5         5
##
## $sp_littercn
##      traitCat traitUnit numMeas numObs numPapers
## 5 sp_littercn molC/molN      81     30        16
## 4 sp_littercn     %C/%N       6      2         2
##
## $sp_litterpercN
##         traitCat traitUnit numMeas numObs numPapers
## 6 sp_litterpercN         %      85     29        14
## 8 sp_litterpercN      mg/g      38     15         6
## 7 sp_litterpercN      g/kg       6      4         2
##
## $sp_percN
##    traitCat traitUnit numMeas numObs numPapers
## 9  sp_percN         %     160     58        23
## 11 sp_percN      mg/g      66     23        11
## 10 sp_percN      g/kg      50     21         3
## 13 sp_percN      ug/g      14      2         1
## 12 sp_percN   mmol/kg       4      1         1
## 14 sp_percN     ug/mg       2      1         1
```

---

7. Soil measurement statistics

```
summ.meas <- ddply(measures,~measCat,summarise, numObs=length(unique(obsID)))
summ.meas.o<-orderBy(~-numObs, summ.meas)
summ.meas.o
```

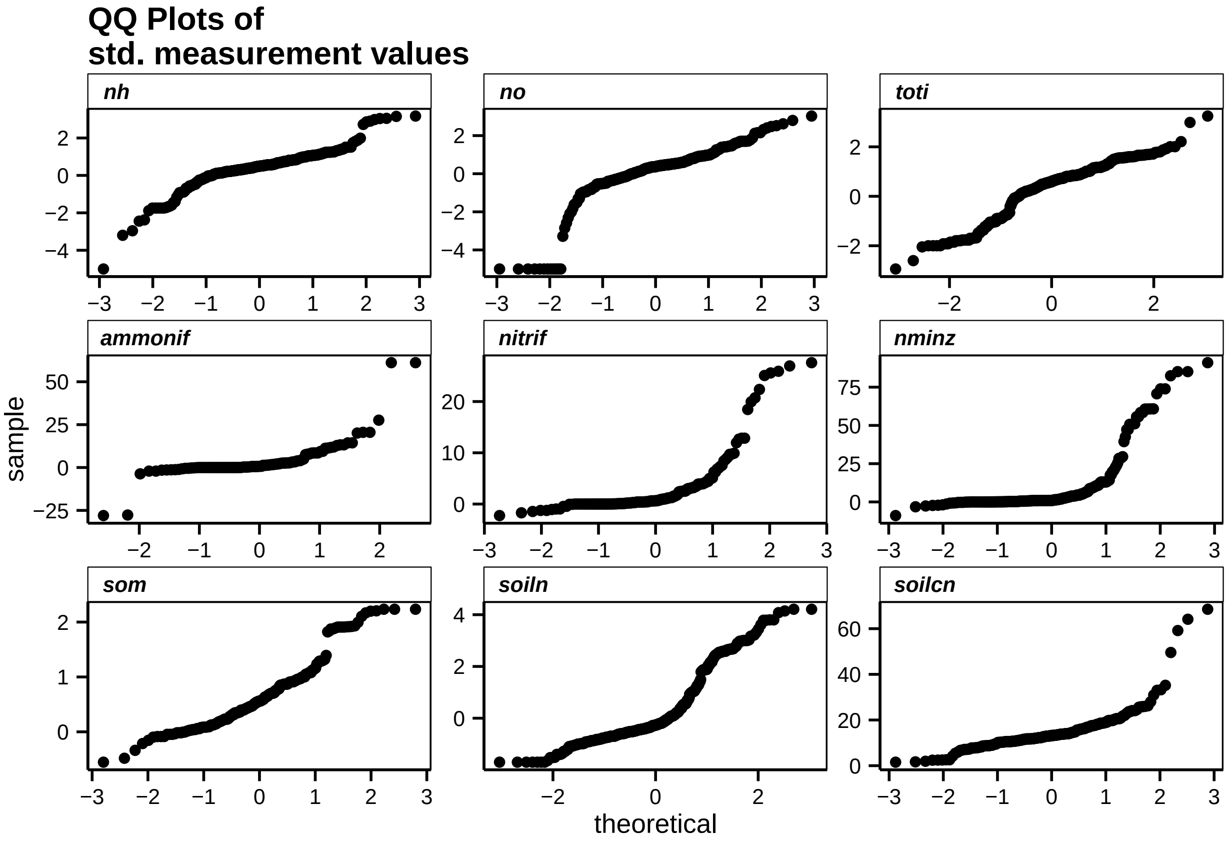
```
##   measCat numObs
## 9    toti    225
## 7   soiln    212
## 5      no    177
## 2      nh    162
## 4   nminz    128
## 6  soilcn    126
## 8     som     97
## 3  nitrif     85
## 1 ammonif     54
```

```
#What units and methods are commonly reported for each measurement?
summ.measUnit <- ddply(measures,~measCat+unit,summarise,
                   numMeas = length(obsID),
                   numObs=length(unique(obsID)))
MEASCAT<-unique(summ.measUnit$measCat)
measUnitList<-list()
i<-0
for(i in 1:length(MEASCAT)){
  subdf<-summ.measUnit[summ.measUnit$measCat==MEASCAT[i],]
  measUnitList[[as.character(MEASCAT[i])]]<-orderBy(~-numMeas, subdf)
}
measUnitList
```

```
## $ammonif
##    measCat              unit numMeas numObs
## 4  ammonif           mg/kg*d      14     14
## 14 ammonif            ug/g*d       7      7
## 12 ammonif         ug/g*2wks       5      5
## 17 ammonif           ug/g*mo       5      5
## 15 ammonif           ug/g*hr       4      4
## 2  ammonif         mg/kg*10d       2      2
## 8  ammonif       notReported       2      2
## 9  ammonif               ppm       2      2
## 10 ammonif              ug/g       2      2
## 11 ammonif          ug/g*2wk       2      2
## 18 ammonif          umol/g*d       2      2
## 1  ammonif meq per 100g soil       1      1
## 3  ammonif         mg/kg*28d       1      1
## 5  ammonif          mg/kg*mo       1      1
## 6  ammonif           mg/m2*d       1      1
## 7  ammonif           mg/m2*y       1      1
## 13 ammonif          ug/g*30d       1      1
## 16 ammonif          ug/g*IER       1      1
##
## $nh
##    measCat              unit numMeas numObs
## 38      nh              ug/g      59     59
## 28      nh             mg/kg      44     44
## 32      nh           mmol/kg      17     17
## 29      nh              mg/L       7      7
## 35      nh               ppm       5      5
## 26      nh            mg/bag       4      4
## 27      nh              mg/g       4      4
## 21      nh              g/m2       3      3
## 34      nh       notReported       3      3
## 37      nh        ug/capsule       3      3
## 22      nh              g/m3       2      2
## 31      nh           mg/m2*y       2      2
## 19      nh                 %       1      1
## 20      nh     cmol/kg resin       1      1
## 23      nh              g/mg       1      1
## 24      nh             kg/ha       1      1
## 25      nh meq per 100g soil       1      1
## 30      nh             mg/m2       1      1
## 33      nh            ng/g*d       1      1
## 36      nh      ug/10cm2*35d       1      1
## 39      nh             ug/kg       1      1
##
## $nitrif
##    measCat              unit numMeas numObs
## 56  nitrif            ug/g*d      23     23
## 44  nitrif           mg/kg*d      19     19
## 54  nitrif         ug/g*2wks       5      5
## 57  nitrif           ug/g*hr       5      5
## 59  nitrif           ug/g*mo       5      5
## 40  nitrif          g/m2*6mo       4      4
## 52  nitrif          ug/g*14d       3      3
## 42  nitrif         mg/kg*10d       2      2
## 47  nitrif          mg/m2*mo       2      2
## 49  nitrif       notReported       2      2
## 50  nitrif           ppm/30d       2      2
## 51  nitrif              ug/g       2      2
## 53  nitrif          ug/g*2wk       2      2
## 60  nitrif          umol/g*d       2      2
## 41  nitrif meq per 100g soil       1      1
## 43  nitrif         mg/kg*28d       1      1
## 45  nitrif          mg/kg*mo       1      1
## 46  nitrif           mg/m2*d       1      1
## 48  nitrif           mg/m2*y       1      1
## 55  nitrif          ug/g*30d       1      1
## 58  nitrif          ug/g*IER       1      1
##
## $nminz
##    measCat        unit numMeas numObs
## 80   nminz      ug/g*d      24     24
## 74   nminz mmol/kg*30d      17     17
## 70   nminz     mg/kg*d      16     16
## 68   nminz       mg/kg      11     11
## 64   nminz      g/m2*s       6      6
## 66   nminz   mg/g*382d       6      6
## 67   nminz     mg/g*wk       6      6
## 78   nminz   ug/g*2wks       5      5
## 62   nminz    g/m2*6mo       4      4
## 81   nminz     ug/g*hr       4      4
## 65   nminz      g/m2*y       3      3
## 75   nminz        ug/g       3      3
## 76   nminz    ug/g*14d       3      3
## 83   nminz     ug/g*mo       3      3
## 61   nminz      g/ha*d       2      2
## 69   nminz   mg/kg*60d       2      2
## 72   nminz    mg/m2*mo       2      2
## 77   nminz    ug/g*2wk       2      2
## 84   nminz      ug/g*y       2      2
## 85   nminz    umol/g*d       2      2
## 63   nminz      g/m2*d       1      1
## 71   nminz     mg/m2*d       1      1
## 73   nminz     mg/m2*y       1      1
## 79   nminz    ug/g*30d       1      1
## 82   nminz    ug/g*IER       1      1
##
## $no
##     measCat              unit numMeas numObs
## 104      no              ug/g      59     59
## 95       no             mg/kg      57     57
## 98       no           mmol/kg      17     17
## 100      no       notReported       7      7
## 101      no               ppm       5      5
## 93       no            mg/bag       4      4
## 94       no              mg/g       4      4
## 96       no              mg/L       4      4
## 88       no              g/m2       3      3
## 103      no        ug/capsule       3      3
## 89       no              g/m3       2      2
## 97       no           mg/m2*y       2      2
## 106      no    umol/capsule*d       2      2
## 86       no                 %       1      1
## 87       no     cmol/kg resin       1      1
## 90       no              g/mg       1      1
## 91       no             kg/ha       1      1
## 92       no meq per 100g soil       1      1
## 99       no            ng/g*d       1      1
## 102      no      ug/10cm2*35d       1      1
## 105      no             ug/kg       1      1
##
## $soilcn
##     measCat      unit numMeas numObs
## 109  soilcn molC/molN      79     79
## 107  soilcn     %C/%N      39     39
## 108  soilcn     gC/gN       8      8
##
## $soiln
##     measCat    unit numMeas numObs
## 110   soiln       %     106    106
## 111   soiln    g/kg      33     33
## 115   soiln    mg/g      29     29
## 112   soiln    g/m2      18     18
## 116   soiln   mg/kg      12     12
## 119   soiln    ug/g       6      6
## 114   soiln   kg/m3       2      2
## 118   soiln     ppm       2      2
## 120   soiln   ug/kg       2      2
## 113   soiln   kg/ha       1      1
## 117   soiln mmol/kg       1      1
```

```
##
## $som
##     measCat unit numMeas numObs
## 121     som    %      91     91
## 123     som g/kg       4      4
## 122     som   cm       1      1
## 124     som mg/g       1      1
##
## $toti
##     measCat           unit numMeas numObs
## 140    toti           ug/g      79     79
## 130    toti          mg/kg      52     52
## 133    toti         mg/pot      24     24
## 134    toti        mmol/kg      17     17
## 128    toti         mg/bag      10     10
## 141    toti      ug/gIER*d       8      8
## 129    toti           mg/g       7      7
## 131    toti           mg/L       6      6
## 136    toti            ppm       5      5
## 139    toti       ug/bag*d       4      4
## 126    toti           g/m2       3      3
## 135    toti    notReported       3      3
## 142    toti umol/capsule*d       2      2
## 125    toti              %       1      1
## 127    toti          index       1      1
## 132    toti          mg/m2       1      1
## 137    toti    ug/10cm2*35d       1      1
## 138    toti         ug/bag       1      1
```

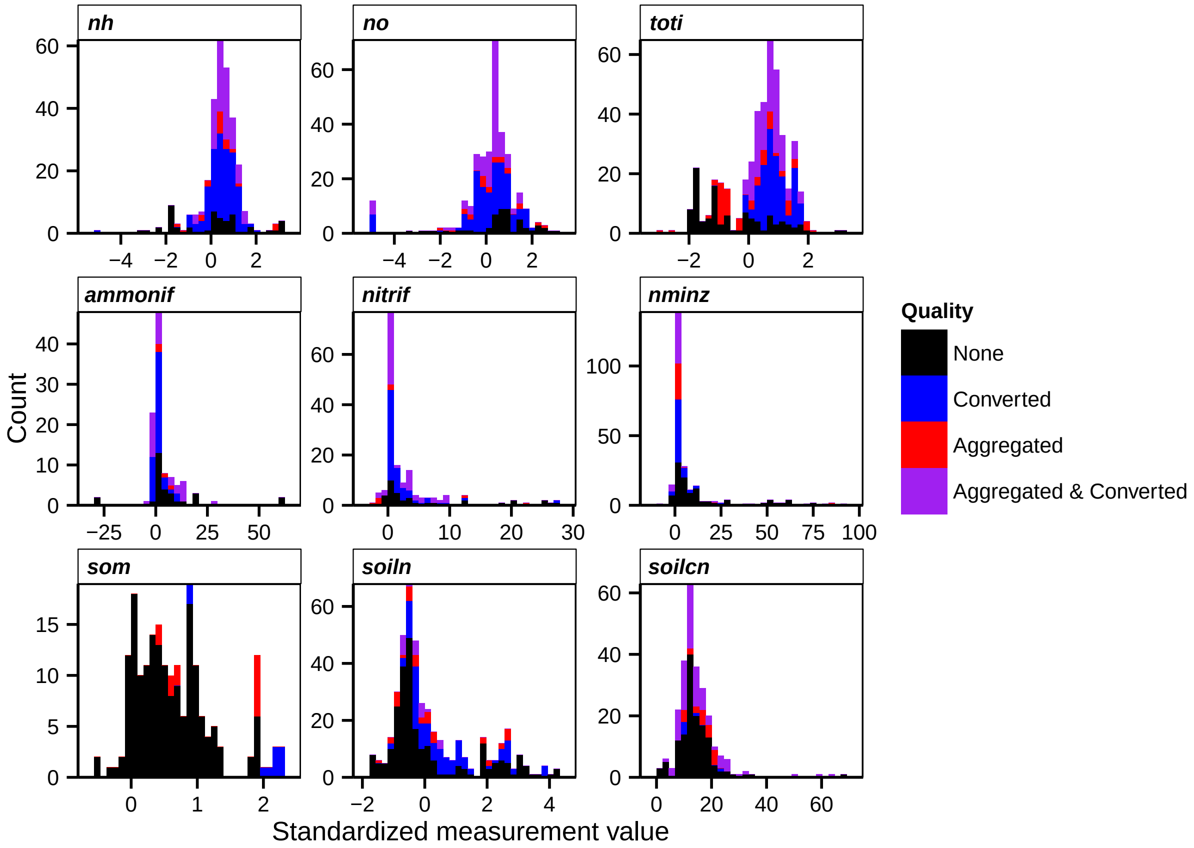
```r
#re-order measCat levels
metaDataset$measCat <- factor(metaDataset$measCat, levels = measCat_order)

#re-shape measures so that inv and nat are in the same column temporarily
tmp<-ddply(metaDataset, ~obsID+measCat, summarize,
      m1i_logt = unique(m1i_logt),
      m2i_logt = unique(m2i_logt),
      measQuality = unique(measQuality))
tmp$obsID<-as.factor(tmp$obsID)
m.tmp<-melt(tmp, idcols=c('obsID','measCat','measQuality'))
m.tmp$invType<-rep(NA,length(dim(m.tmp)[1]))
m.tmp[m.tmp$variable == 'm1i_logt','invType']<-'inv'
m.tmp[m.tmp$variable == 'm2i_logt','invType']<-'ref'

#Shapiro Test
# ddply(measures, ~measCat, summarise,
#       shapTest=shapiro.test(value)$p.value,
#       shapTest.Ln=shapiro.test(log(value+1))$p.value)
#none are normal according to Shapiro test

# Q-Q plots
qq<-ggplot(m.tmp, aes(sample=value)) +
  facet_wrap(~measCat, scales='free', ncol=3) +
  stat_qq() + mytheme + ggtitle('QQ Plots of \nstd. measurement values')
qq
```



```
newfilename<-"qq_meas.png"
png(paste(figuresPath,newfilename, sep='/'),
    units='in', width = fig.width*3, height = fig.height*6, res=fig.res)
qq
dev.off()
```

```
## pdf
##   2
```

```
# Plot Quality Histograms
#re-order measQuality levels
m.tmp$measQuality <- factor(m.tmp$measQuality, levels = c('NoAgg.NoConv','NoAgg.Conv','Agg.NoConv','Agg

pHist_measQual<-ggplot(data=m.tmp, aes(x=value,fill=measQuality)) + mytheme +
  facet_wrap(~measCat, scales='free', ncol=3) + geom_histogram() +
  scale_y_continuous(expand = c(0,0)) +
  scale_fill_manual(name = "Quality",
                    labels = c("Agg.Conv"="Aggregated & Converted",
                               "Agg.NoConv"="Aggregated",
                               "NoAgg.Conv"="Converted",
                               "NoAgg.NoConv"="None"),
                    values=c("Agg.Conv" = "purple",
                             "Agg.NoConv" = "red",
                             "NoAgg.Conv" = "blue",
                             "NoAgg.NoConv" = "black")) +
  ylab('Count') + xlab('Standardized measurement value')
pHist_measQual
```



```
newfilename<-"pHist_measQual.png"
png(paste(figuresPath,newfilename, sep='/'),
    units='in', width = fig.width*3.5, height = fig.height*6, res=fig.res)
pHist_measQual
dev.off()
```

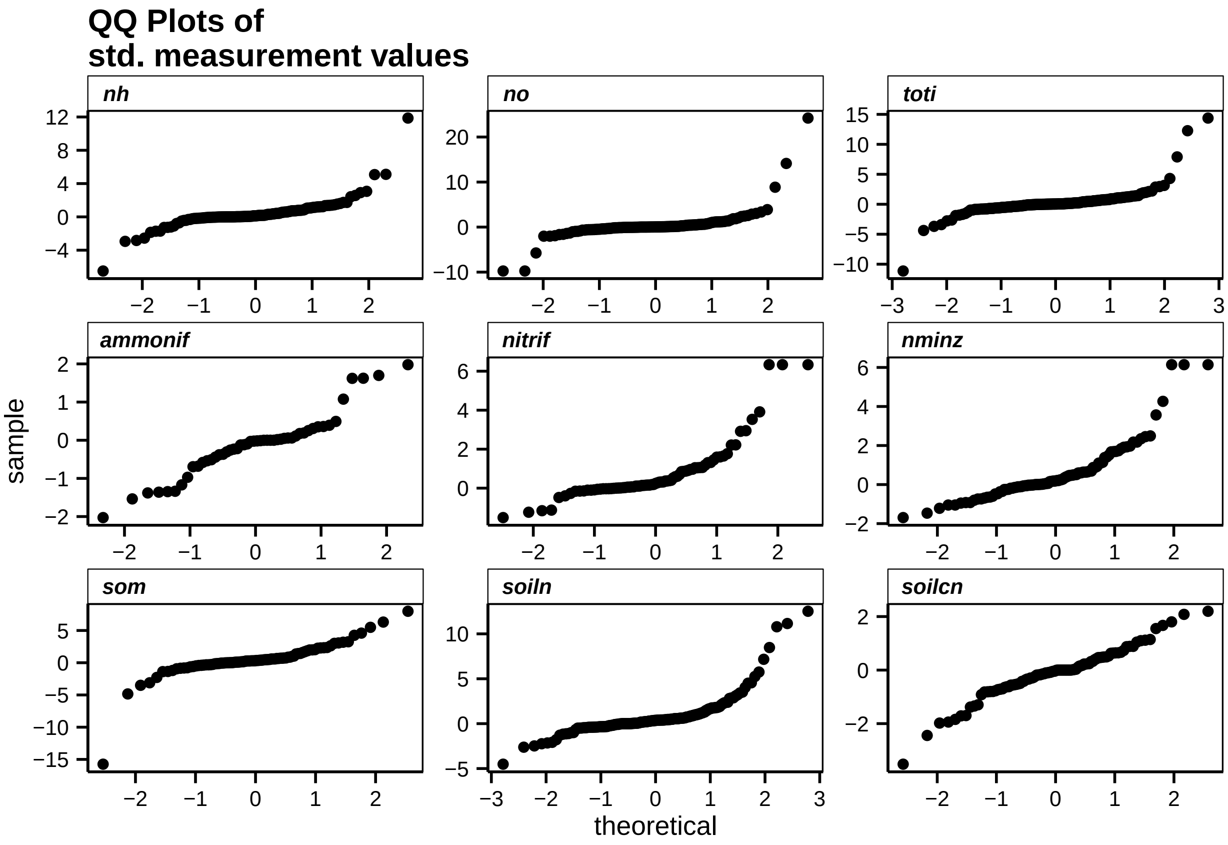
```
## pdf
##   2
```

---

8. Effect size statistics

```
#re-shape measures so that inv and nat are in the same column temporarily
tmp<-ddply(metaDataset, ~obsID+measCat, summarize,
      yi = unique(yi),
      measQuality = unique(measQuality))
tmp$obsID<-as.factor(tmp$obsID)

#Shapiro Test
# ddply(measures, ~measCat, summarise,
#        shapTest=shapiro.test(value)$p.value,
#        shapTest.Ln=shapiro.test(log(value+1))$p.value)
```

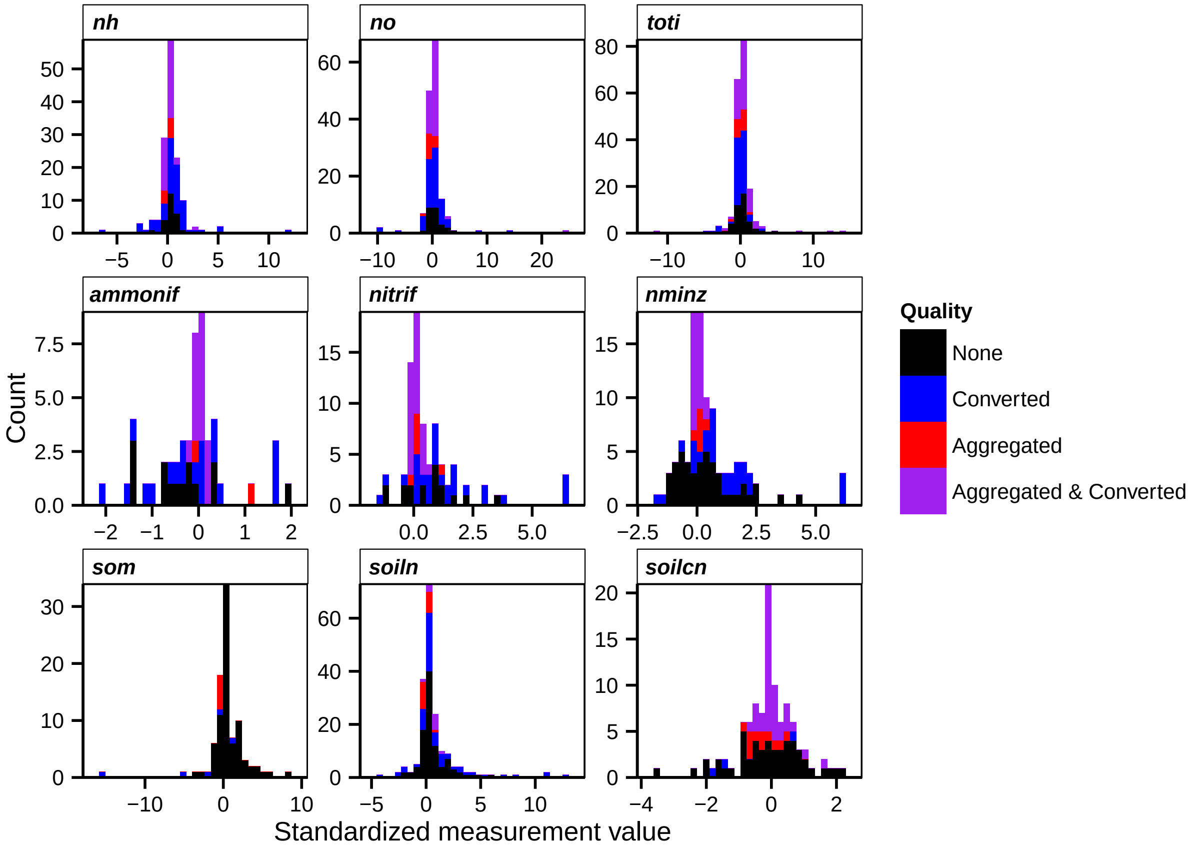
```
#none are normal according to Shapiro test

# Q-Q plots
qq<-ggplot(tmp, aes(sample=yi)) +
  facet_wrap(~measCat, scales='free', ncol=3) +
  stat_qq() + mytheme + ggtitle('QQ Plots of \nstd. measurement values')
qq
```



```
newfilename<-"qq_ESmeas.png"
png(paste(figuresPath,newfilename, sep='/'),
    units='in', width = fig.width*3, height = fig.height*6, res=fig.res)
qq
dev.off()
```

```
## pdf
##   2
```

```
#Plot Quality Histograms
#re-order measQuality levels
tmp$measQuality <- factor(tmp$measQuality, levels = c('NoAgg.NoConv','NoAgg.Conv','Agg.NoConv','Agg.Con

pHist_ESmeasQual<-ggplot(data=tmp, aes(x=yi,fill=measQuality)) + mytheme +
  facet_wrap(~measCat, scales='free', ncol=3) + geom_histogram() +
  scale_y_continuous(expand = c(0,0)) +
  scale_fill_manual(name = "Quality",
                    labels = c("Agg.Conv"="Aggregated & Converted",
                               "Agg.NoConv"="Aggregated",
                               "NoAgg.Conv"="Converted",
                               "NoAgg.NoConv"="None"),
                    values=c("Agg.Conv" = "purple",
                             "Agg.NoConv" = "red",
                             "NoAgg.Conv" = "blue",
                             "NoAgg.NoConv" = "black")) +
  ylab('Count') + xlab('Standardized measurement value')
pHist_ESmeasQual
```

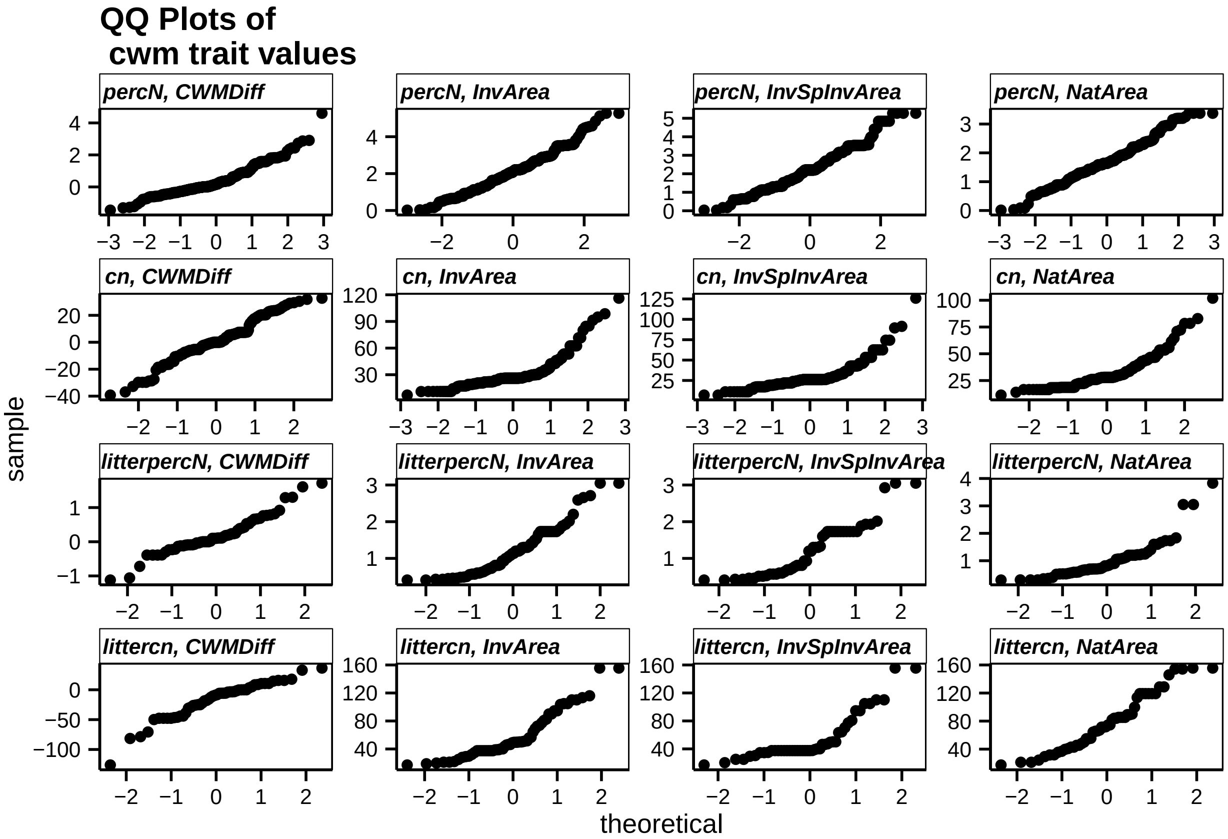


```
newfilename<-"pHist_ESmeasQual.png"
png(paste(figuresPath,newfilename, sep='/'),
    units='in', width = fig.width*3.5, height = fig.height*6, res=fig.res)
pHist_measQual
dev.off()
```

```
## pdf
##   2
```

9. CWM trait value statistics

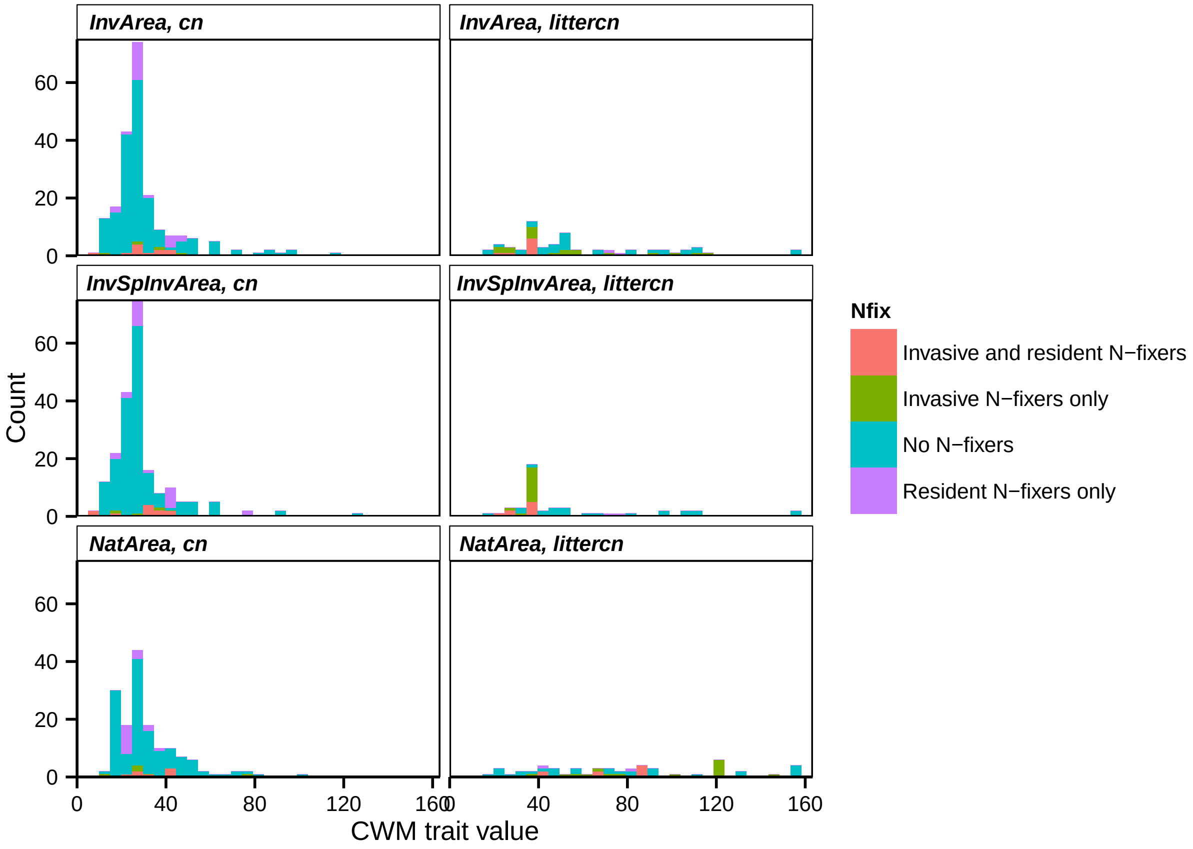
```
#re-order measCat levels
metaDataset$traitCat <- factor(metaDataset$traitCat, levels = traitCat_order)
#re-shape measures so that inv and nat are in the same column temporarily
tmp<-ddply(metaDataset, ~obsID+traitCat, summarize,
      InvArea_cwm = unique(InvArea_cwm),
      InvSpInvArea_cwm = unique(InvSpInvArea_cwm),
      NatArea_cwm = unique(NatArea_cwm),
      CWMDiff_cwm = unique(CWMDiff_cwm),
      InvArea_qualRank = unique(InvArea_qualRank),
      InvSpInvArea_qualRank = unique(InvSpInvArea_qualRank),
      NatArea_qualRank = unique(NatArea_qualRank),
      CWMDiff_qualRank = unique(InvArea_qualRank)+ unique(NatArea_qualRank))

tmp$obsID<-as.factor(tmp$obsID)
m.tmp<-melt(tmp, idcols=c('obsID','traitCat'))
m.tmp$dataType<-rep(NA,length(dim(m.tmp)[1])) #dataType
m.tmp[grepl('_qualRank', m.tmp$variable),'dataType']<-'qualRank'
m.tmp[grepl('_cwm', m.tmp$variable),'dataType']<-'cwm'
m.tmp$invType<-rep(NA,length(dim(m.tmp)[1])) #invType
m.tmp[grepl('InvArea', m.tmp$variable),'invType']<-'InvArea'
m.tmp[grepl('InvSpInvArea', m.tmp$variable),'invType']<-'InvSpInvArea'
m.tmp[grepl('NatArea', m.tmp$variable),'invType']<-'NatArea'
m.tmp[grepl('CWMDiff', m.tmp$variable),'invType']<-'CWMDiff'
c.tmp<-dcast(m.tmp, obsID+traitCat+invType~dataType)
c.tmp<-c.tmp[!is.na(c.tmp$cwm),]

# #Shapiro Test
# ddply(cwm, ~traitCat, summarise,
#       shapTest=shapiro.test(cwm)$p.value,
#       shapTestLn=shapiro.test(log10(cwm))$p.value)
# #none are normal according to Shapiro test

# Q-Q plots
qq<-ggplot(c.tmp, aes(sample=cwm)) +
  facet_wrap(~traitCat+invType, scales='free', ncol=4) +
  stat_qq() + mytheme + ggtitle('QQ Plots of \n cwm trait values')
qq
```



```
newfilename<-"qq_cwm.png"
png(paste(figuresPath,newfilename, sep='/'),
    units='in', width = fig.width*4, height = fig.height*3, res=fig.res)
qq
dev.off()
```

```
## pdf
##   2
```

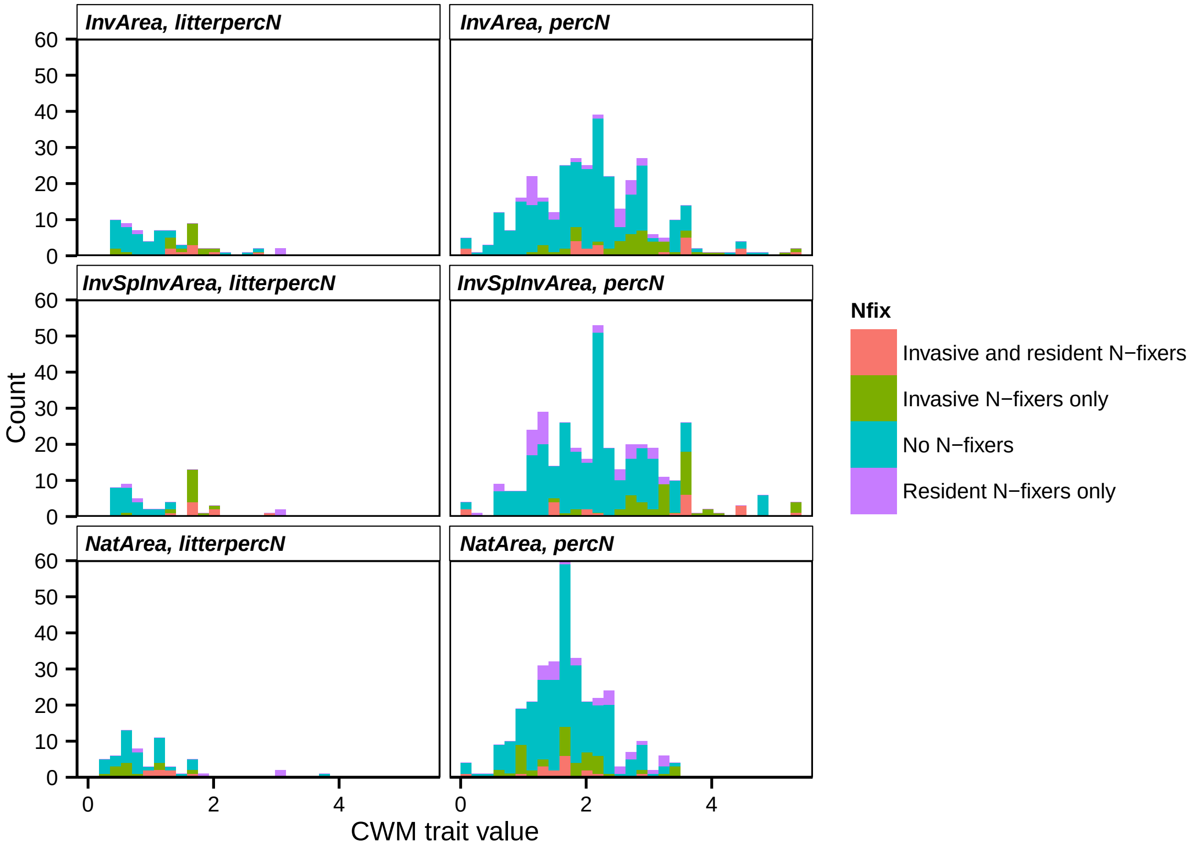
```
#Plot Factor Histograms
#InvType
cwm$obsID<-as.factor(cwm$obsID)
cwm$n_invSp_invArea<-as.factor(cwm$n_invSp_invArea)
cwm$n_invSp_natArea<-as.factor(cwm$n_invSp_natArea)
cwm$n_natSp_invArea<-as.factor(cwm$n_natSp_invArea)
cwm$n_natSp_natArea<-as.factor(cwm$n_natSp_natArea)

#Nfix
cwm.tmp<-merge(cwm, observations, by='obsID')

#plot
cwm.tmp.cn<-subset(cwm.tmp, traitCat %in% c('cn','littercn'))
pHist_cwm_cn<-ggplot(data=cwm.tmp.cn, aes(x=cwm,fill=Nfix)) +
  facet_wrap(~invType+traitCat, scales='fixed',ncol=2) +
  scale_y_continuous(expand=c(0,0)) + scale_x_continuous(expand=c(0,0)) +
  geom_histogram() + mytheme +
  ylab('Count') + xlab('CWM trait value')
pHist_cwm_cn
```



```
newfilename<-"pHist_cwm_cn.png"
png(paste(figuresPath,newfilename, sep='/'),
    units='in', width = fig.width*2.5, height = fig.height*3, res=fig.res)
pHist_cwm_cn
dev.off()
```

```
## pdf
##   2
```

```
cwm.tmp.percn<-subset(cwm.tmp, traitCat %in% c('percN','litterpercN'))
pHist_cwm_percn<-ggplot(data=cwm.tmp.percn, aes(x=cwm,fill=Nfix)) +
  facet_wrap(~invType+traitCat, scales='fixed',ncol=2) +
  scale_y_continuous(expand=c(0,0)) + scale_x_continuous(expand=c(0,0)) +
  geom_histogram() + mytheme +
  ylab('Count') + xlab('CWM trait value')
pHist_cwm_percn
```



```
newfilename<-"pHist_cwm_percn.png"
png(paste(figuresPath,newfilename, sep='/'),
    units='in', width = fig.width*2.5, height = fig.height*3, res=fig.res)
pHist_cwm_percn
dev.off()
```

```
## pdf
##   2
```