Project Preliminary Analysis

Betsy Cowdery

November 21, 2014

Goal

The goal of this project is to develop a multivariate Bayesian meta-analytical model that synthesizes plant trait data from multiple studies while accounting for various sources of uncertainty. Using observed sample mean, sample size, and a sample error statistics for multiple plant traits, we aim to produced well constrained estimates of mean and precision for a single trait. This has be done in PEcAn (the Predictive Ecosystem Analyzer) using a univariate model [1], however, a multivariate model can leverage the fact that many plant traits are highly correlated [2] to constrain our estimates even further. This may be especially useful improving predictions for studies in which observations are missing.

For this project I am using the data set compiled by Wright et. all to develop a "leaf economics spectrum" [2]. This data is from the global plant trait network (Glopnet), a database created to quantify leaf economics across the world's plant species.

From this data set I will be focusing on six plant-traits:

1. Leaf mass per area (LMA)

2. Photosynthetic capacity (Amass) - photosynthetic assimilation rates measured under high light, ample soil moisture and ambient CO2

3. Leaf nitrogen (N)

4. Leaf phosphorus (P)

5. Dark respiration rate (Rmass)

6. Leaf lifespan (LL)

Data

Including studies with missing observations
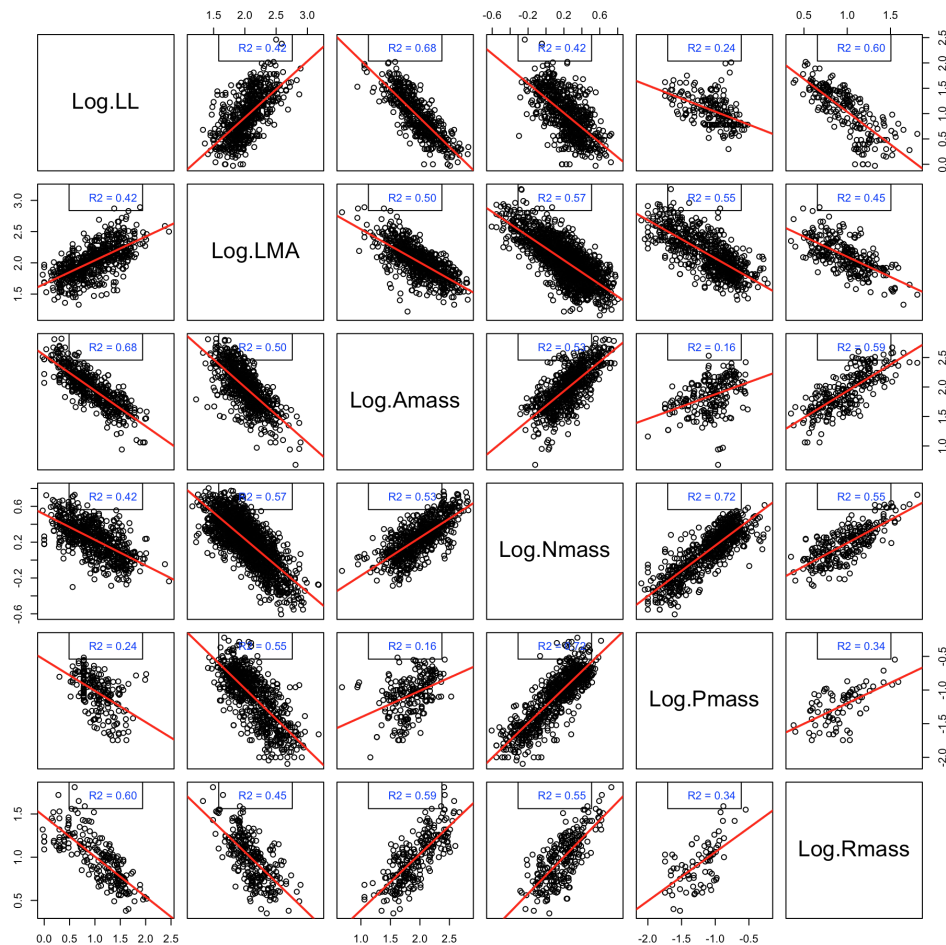
Here we have plotted each the log of each plant trait variable against one another with a regression line in red if the regression is statistically significant. We can see that each plant trait has a statistically significant ($p < .01$) correlation with all the others, which suggests that a multivariate analysis will in fact be informative.

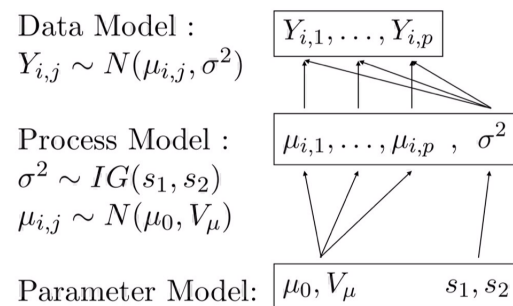**Excluding studies with missing observations**
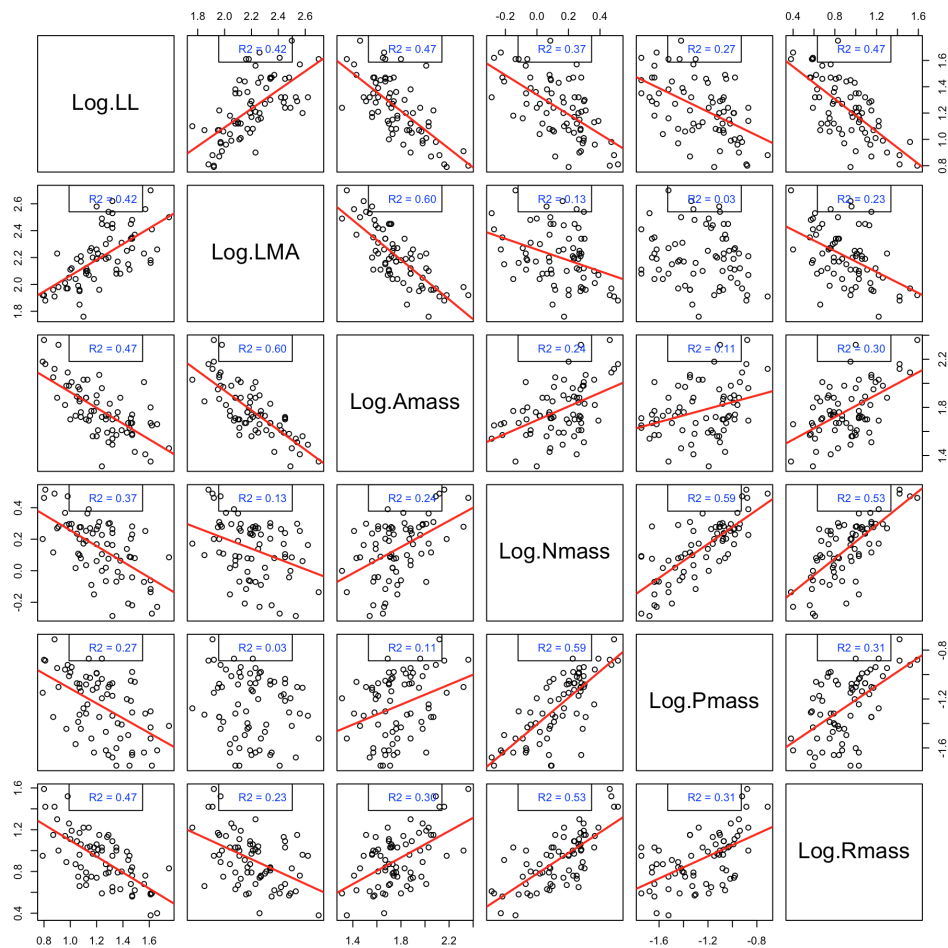
Here we have the exact same plot as above, but only including the studies that do not contain missing observations. Clearly there is less data to work with and thus the R-squared values are lower and one pair of traits no longer have a statistically significant correlation. However, I expect that given the presence of so many statistically significant correlations, the multivariate model will provide improved predictions for the means of each variable.

Univariate Model

Let $Y_{i,j}$ represent the observed value of the jth trait variable in study i.

Model Graph

Data Model :
$$Y_{i,j} \sim N(\mu_{i,j}, \sigma^2)$$
$$\boxed{Y_{i,1}, \ldots, Y_{i,p}}$$

Process Model :
$$\sigma^2 \sim IG(s_1, s_2)$$
$$\mu_{i,j} \sim N(\mu_0, V_\mu)$$
$$\boxed{\mu_{i,1}, \ldots, \mu_{i,p} \;,\; \sigma^2}$$

Parameter Model:
$$\boxed{\mu_0, V_\mu \qquad s_1, s_2}$$

Model in JAGS

```
model{
  prec.sigma~dgamma(.001,.001)
  sigma <- 1/prec.sigma
  for(i in 1:n){mu[i]~dnorm(0,.001)}

  for(i in 1:N){
    for(j in 1:n){
      Y[i,j]~dnorm(mu[j],prec.sigma)
    }
  }
}
```
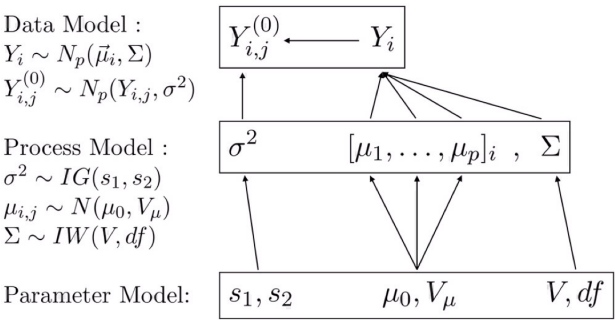
Multivariate Model

Let

Y_i represent the vector of observed value of the j traits variable in study i.

$Y_{i,j}^0$ represent the observed value jth trait variable in study i, taking into account observation error.

Model Graph

Data Model :
$Y_i \sim N_p(\vec{\mu}_i, \Sigma)$
$Y_{i,j}^{(0)} \sim N_p(Y_{i,j}, \sigma^2)$

Process Model :
$\sigma^2 \sim IG(s_1, s_2)$
$\mu_{i,j} \sim N(\mu_0, V_\mu)$
$\Sigma \sim IW(V, df)$

Parameter Model:

Model in JAGS

```
model{
  prec.Sigma~dwish(Vsig[,],n)
  Sigma[1:n,1:n] <- inverse(prec.Sigma[,])


  mu[1:n]~dmnorm(mu0[],Vmu)


  for(i in 1:N){
    Y[i,1:n]~dmnorm(mu[],prec.Sigma[,])
    for(j in 1:n){
      X[i,j]~dnorm(Y[i,j],10000000)
    }
  }
}
```

Comparisons

Overall comparison of computed and estimated means for each of the 6 variables

```
##                            Log.LL Log.LMA Log.Amass Log.Nmass Log.Pmass
## Data                       1.2353   2.203     1.773    0.1389   -1.2570
## Univariate    without NA's 1.2354   2.203     1.773    0.1388   -1.2570
## Multivariate without NA's 1.2353   2.203     1.773    0.1387   -1.2572
## Univariate    with NA's    0.9612   1.991     1.982    0.2273   -1.0990
## Multivariate with NA's     0.9437   1.991     1.982    0.2404   -0.9258
##                            Log.Rmass
## Data                          0.9157
## Univariate    without NA's    0.9155
## Multivariate without NA's    0.9158
## Univariate    with NA's       0.9850
## Multivariate with NA's        1.0470
```
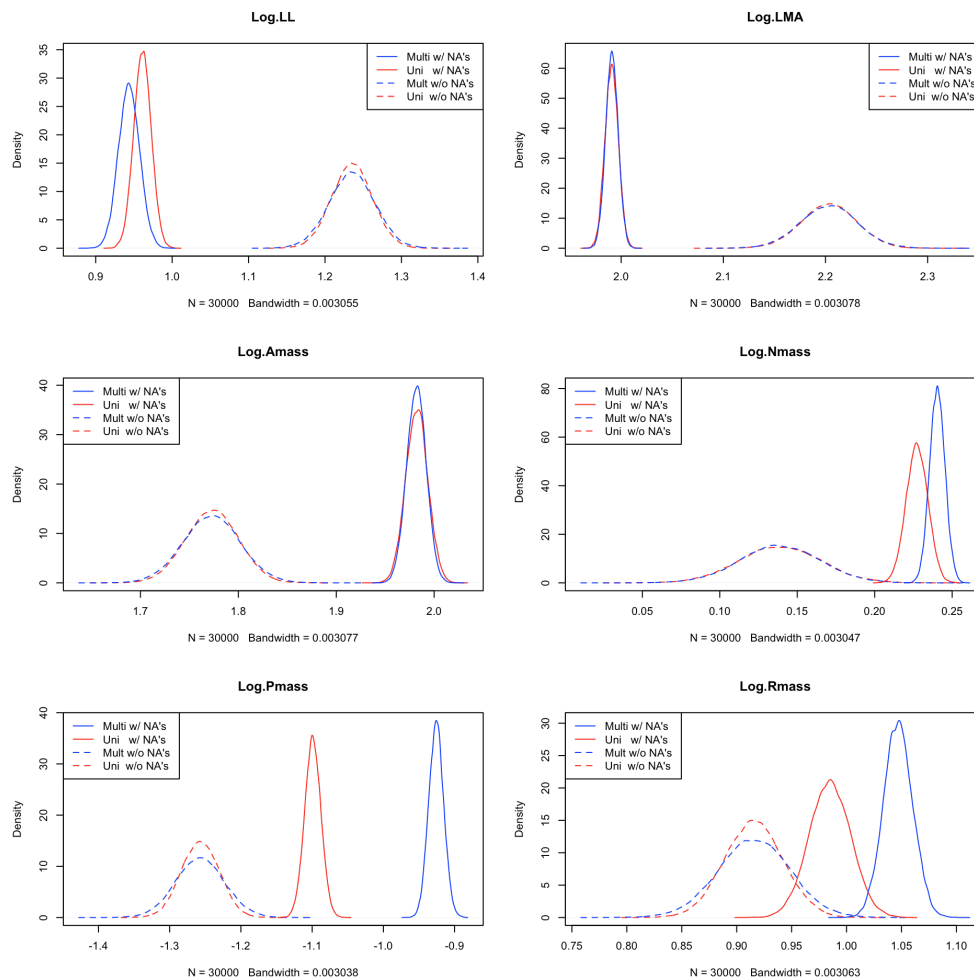
When using data that excludes all studies with missing observations, there is practically no difference between the two model's estimated means for each of the variables. However, for both models, including studies with NA's produces estimated means that are significantly different from those produced with data excluding NA's. For the variables Log.LMA and Log.Amass, the estimated means from the univariate and multivariate models were very close, but for the remaining variables, they were noticeably different, with the estimated mean from the univariate model always closer to the data mean than the estimated mean from the multivariate model.
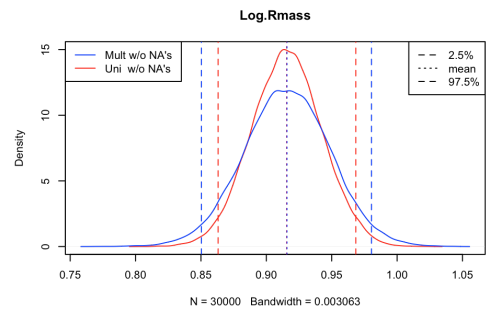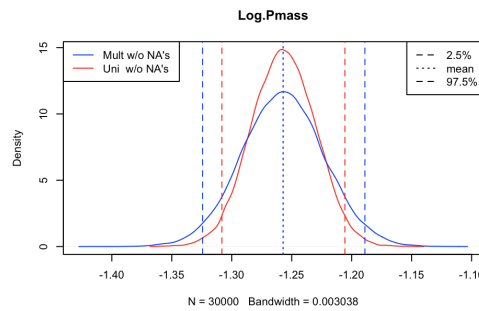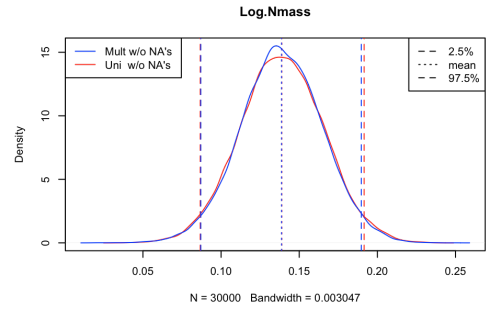
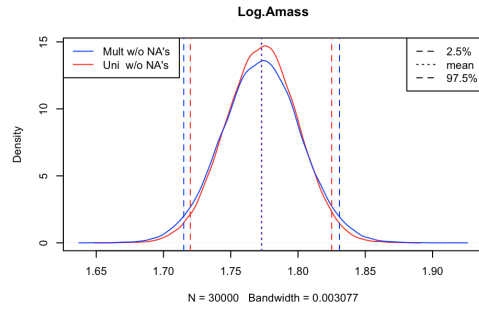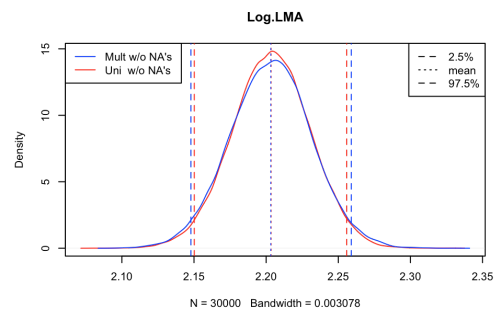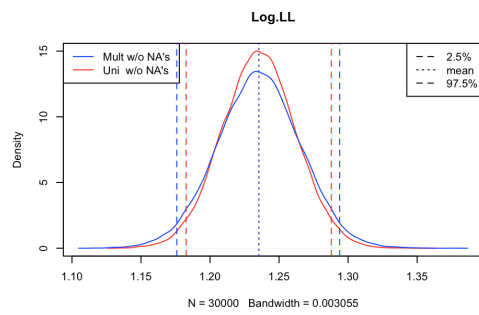### Error around the mean

```
## 
## 
## Log.LL
##                             Mean       SE   2.5%  97.5%
## Data                      1.2353 2.667e-02 0.8078 1.6290
## Univariate   without NA's 1.2354 1.542e-04 1.1827 1.2879
## Multivariate without NA's 1.2353 1.745e-04 1.1759 1.2939
## Univariate   with NA's    0.9612 6.602e-05 0.9388 0.9836
## Multivariate with NA's    0.9437 7.946e-05 0.9167 0.9708
## lowest SE: Univariate   with NA's
## 
## Log.LMA
##                            Mean       SE  2.5% 97.5%
## Data                      2.203 2.422e-02 1.873 2.589
## Univariate   without NA's 2.203 1.556e-04 2.150 2.256
## Multivariate without NA's 2.203 1.628e-04 2.148 2.259
## Univariate   with NA's    1.991 3.756e-05 1.978 2.003
## Multivariate with NA's    1.991 3.523e-05 1.979 2.003
## lowest SE: Multivariate with NA's
## 
## Log.Amass
##                            Mean       SE  2.5% 97.5%
## Data                      1.773 2.550e-02 1.396 2.211
## Univariate   without NA's 1.773 1.552e-04 1.720 1.825
## Multivariate without NA's 1.773 1.696e-04 1.715 1.831
## Univariate   with NA's    1.982 6.478e-05 1.960 2.004
## Multivariate with NA's    1.982 5.735e-05 1.962 2.001
## lowest SE: Multivariate with NA's
## 
## Log.Nmass
##                             Mean       SE     2.5%  97.5%
## Data                      0.1389 2.179e-02 -0.23867 0.4758
## Univariate   without NA's 0.1388 1.536e-04  0.08663 0.1915
## Multivariate without NA's 0.1387 1.503e-04  0.08702 0.1897
## Univariate   with NA's    0.2273 4.023e-05  0.21365 0.2410
## Multivariate with NA's    0.2404 2.860e-05  0.23069 0.2502
## lowest SE: Multivariate with NA's
## 
## Log.Pmass
##                              Mean       SE    2.5%  97.5%
## Data                      -1.2570 3.096e-02 -1.7450 -0.870
## Univariate   without NA's -1.2570 1.532e-04 -1.3082 -1.206
## Multivariate without NA's -1.2572 1.982e-04 -1.3243 -1.189
## Univariate   with NA's    -1.0990 6.607e-05 -1.1215 -1.077
## Multivariate with NA's    -0.9258 6.062e-05 -0.9461 -0.905
## lowest SE: Multivariate with NA's
## 
## Log.Rmass
##                             Mean       SE   2.5%  97.5%
## Data                      0.9157 2.975e-02 0.5240 1.4425
## Univariate   without NA's 0.9155 1.545e-04 0.8631 0.9684
## Multivariate without NA's 0.9158 1.911e-04 0.8503 0.9804
## Univariate   with NA's    0.9850 1.093e-04 0.9478 1.0222
## Multivariate with NA's    1.0470 7.642e-05 1.0211 1.0732
## lowest SE: Multivariate with NA's
```
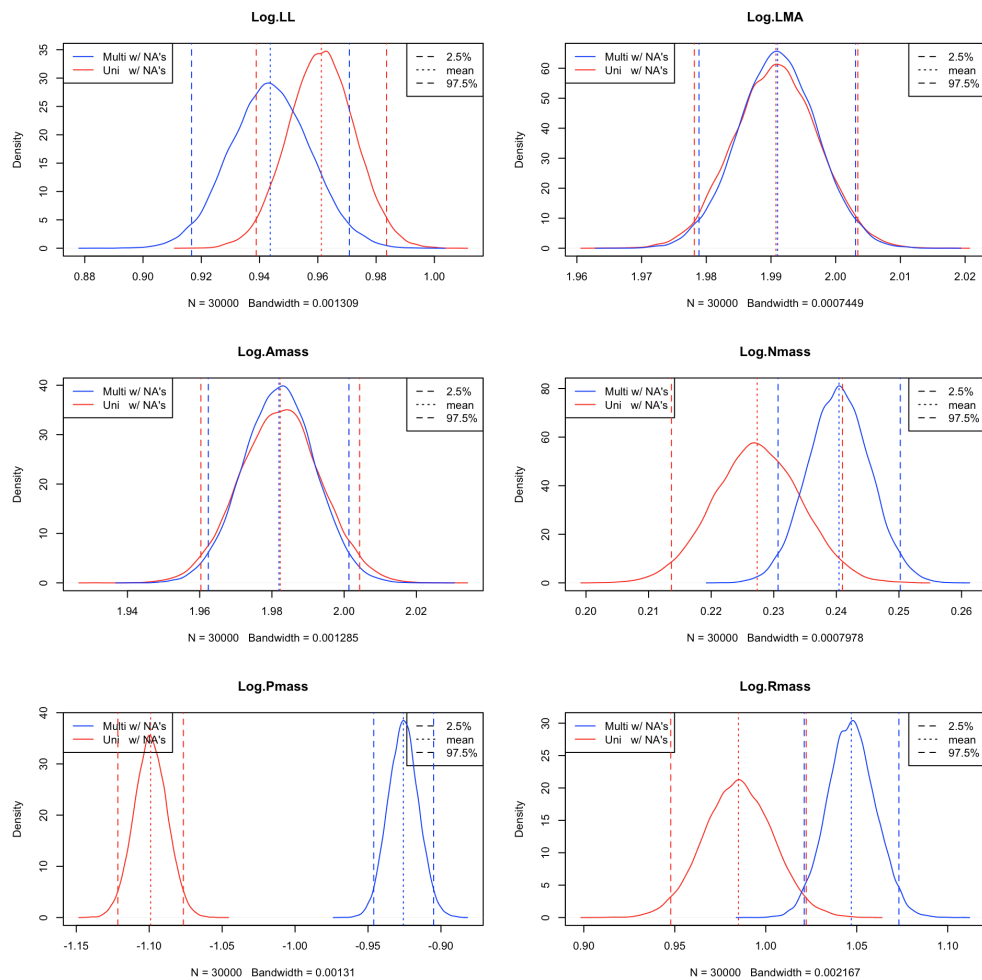
Excluding Studies with Missing Observations

Contrary to what I expected, the variation in the multivariate model showed little to no improvement over the univariate. The multivariate model produced posterior distributions with larger variance around the mean (except for Log.Nmass where the SE for the univariate model was larger by 3.3e-06, a small amount relative to the size of the mean.)

Including Studies with Missing Observations

When studies with missing observations are included, the standard errors begin to behave more like one might expect. The posterior distributions from the multivariate model have smaller variances for all the variables except Log.Pmass. However, it is difficult to see the difference since the expected means of the variables are no longer similar.

Continued Analysis

- Examine the posterior distributions of the estimated variances
- Account for the fact that the data is summary data
- Add random effects into the model

References

[1] LeBauer, D.S., Wang, D., Richter, K.T., Davidson, C.C. & Dietze, M.C. Facilitating feedbacks between field measurements and ecosystem models. Ecological Monographs 83, 133-154 (2013).

[2] Wright, I.J. et al. The worldwide leaf economics spectrum. Nature 428, 821-827 (2004).