Why Bayesian approaches?
The average height of a rare plant

Estimation of averages is an important step in many ecological analyses and demographic models. In this demonstration you will be introduced to *R* and *Rstan* functions which will allow you to reach this goal. You will be confronted with two alternative analysis frameworks, Frequentist and Bayesian, to estimate averages. These two approaches represent different ways to evaluate information. We offer you an example that will help you to decide their relative advantages to obtain inference from data. We use *Hypericum cumulicola* data analyzed in a previous demo but to answer a different question (Quintana-Ascencio et al. 2003, 2018, 2019). We want to characterize the variation in height of *Hypericum cumulicola* and obtain a robust estimate of its mean and uncertainty.

Figure 1. *Hypericum cumulicola*

Part I. Plotting **histograms and calculation using a frequentist approach**

For this part, you will need to download three files from the course website to a folder on your desktop: 1) the text file containing the *Hypericum cumulicola* height data (`hypericum_data_94_07`); 2) the text file containing the *Hypericum cumulicola* population means height data (`popmeanHc.txt`), and 3) the file containing the R script (`Averages 2019.R`). In the program below, the variables and settings of the model are specified in the R script and sent to Stan, so the analysis is run directly in R.

Open the R script `Averages 2019.R`. To see how this script works, let's take a look at the code. For this part we will concentrate in the upper portion of the script. The first line in the program `rm(list=ls())`allows you to clear the memory. This is a good practice when you start a new program. The second and third lines are used to obtain the current directory and set the directory that will be used in the program. You will need to modify this line to adapt it to your directory pathway. The function `read.table("file name.txt", header=T)` obtains the data.

Since 1994, Quintana-Ascencio et al. (2003, 2018, 2019) have collected demographic data of *Hypericum cumulicola* in 14 populations at Archbold Biological Station, Florida USA. The line:

```
Height_data <-Hc_data$ht_init[Hc_data$stage !="sg" & Hc_data$ht_init<90]
```

filters the plant height data to include only plants older than one year (`Hc_data$stage !="sg"`) and eliminates a questionable case with an extreme unrealistic size we detected in a prior demonstration (`Hc_data$ht_init<90`). The same line allocates the filtered data to a new data frame "Height_data".

```
rm(list=ls())
getwd()
# example of a directory
setwd("C:/Users/Pedro/Documents/Classes/Methods in Ecology/2012 fall")
# Before running the script make sure you change the working directory to yours
Hc_data <- read.table("Hcdata.txt", header=T)
Hc_pop  <- read.table("popmeanHc.txt", header=T)

### Analysis using a frequentist approach
Height_data <-Hc_data$ht_init[Hc_data$stage !="sg" & Hc_data$ht_init<90]
hist(Height_data,100,main="Histogram of Hypericum cumulicola height (cm)")
abline(v=Hc_pop$pop_mean, col="blue")
abline(v=mean(Height_data), col="red")

mean(Height_data)
round(sqrt(var(Height_data))/sqrt(length(Height_data)),4)
summary(lm(Height_ ~1))
```

It is always necessary to check the distribution of the data. We will use the function `hist( )` to create a histogram of adult plant heights, and the function `abline()` to add in blue the location of 14 population means and in red the location of the overall mean. Your plot should look like the one below where the data seems to follow a truncated normal distribution:

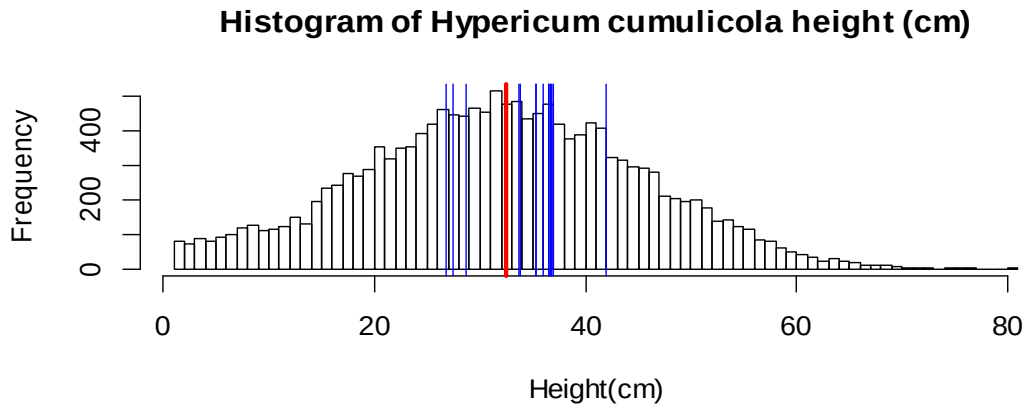


Figure 2. Histogram of individual height of *Hypericum cumulicola* (1994-1997, n =15687). In red is the average height of the whole sample, in blue is the average of the means of 14 populations in Archbold Biological Station.

There are alternative ways to obtain an average in R. Here we first use the function `mean( )` and get an overall plant height mean of 32.4 cm (variance = 170.8). We also obtain the value of the standard error of the plant height mean (SE= 0.10 cm) with a series of R functions. We calculate the square root of the variance and divide it by the square root of the sample size and round the result to four digits.

```
mean(Height_data)
round(sqrt(var(Height_data))/sqrt(length(Height_data)),4)
```

R has functions that can accomplish these calculations in fewer steps.

```
summary(lm(Height_data_95~1))
```

Their output is as follows:

```
Call:
lm(formula = Height_data ~ 1)

Residuals:
   Min     1Q Median     3Q    Max
-31.44  -8.74  -0.44   8.56  48.56

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.4401     0.1043   310.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.07 on 15686 degrees of freedom
```

We also calculate the mean and variance of the population height means and variances. (n=14 populations; Figure 3). The mean and variances overall and at population level are different (32.4 vs 34.4 and 170.8 vs 160.6).

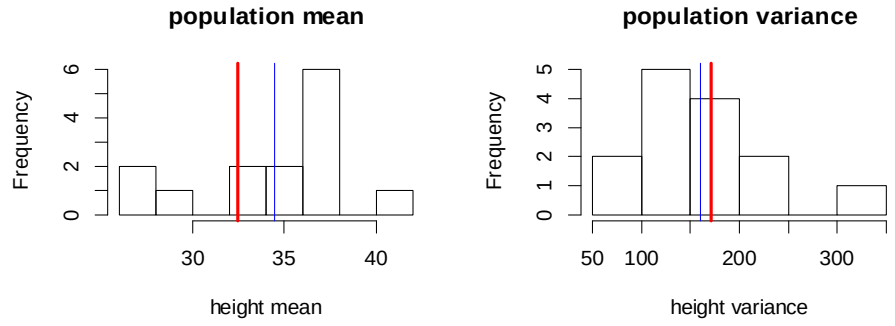


Figure 2. Histograms of population level mean and variance of height among populations of *Hypericum cumulicola* (1994-1997, n =14 populations). In blue are the mean of the

averages (34.4) and the mean of the variances (160.6) among 14 populations. In red is the overall population mean and variance.

Part II. **Calculation of averages using a Bayesian approach**

We can claim that with a sample of 15687 individuals we have a reasonable estimate of the average *Hypericum cumulicola* height and its variance. We will use this information to evaluate the relative efficiency of the frequentist and the Bayesian approaches to estimate means. We will now collect random samples of individuals from our data set and estimate their sample mean and uncertainty. We will compare estimates of this samples obtained using Frequentist and Bayesian approaches with informed and uniformed priors and to the average of the whole data set.

We create several variables that will be used in the model: n is the sample size, x is a vector that contains that random sample of the whole data. At population level, pop.mean.mean is the mean of the means of the population averages, pop.mean.var is the variance of the population means, log.pop.var.mean is the mean of the logarithm of the variances and log.pop.var.var is the variance of the logarithm of the population variances.

```
#### 3. Sampling from the dataset ####
size <-10
n <- length(size)
pop.mean.mean <- mean(Hc_pop$pop_mean)
pop.mean.sd <- sqrt(var(Hc_pop$pop_mean))
log.pop.var.mean <- log(mean(Hc_pop$pop_sd))
log.pop.sd.sd <- log(sqrt(var(Hc_pop$pop_sd)))

## Call  the package

library(rethinking)
library(rstan)
library(ggplot2)
```

We use the function `library(rstan)` to call the program that connects with STAN, and write code as described below. For the **first model**, we define two uninformative priors. The first one is the diffuse prior for the mean `dnorm(0,100)`. The second one is the diffuse prior for the variance `dcauchy(0,1)`. Notice that the first prior describes a normal distribution while the second corresponds to a uniform distribution. The diffuse prior for the mean has a mean of zero and a large variance implying our assumed lack of prior information. Notice that the specification of the distributions is not consistent between R and STAN.

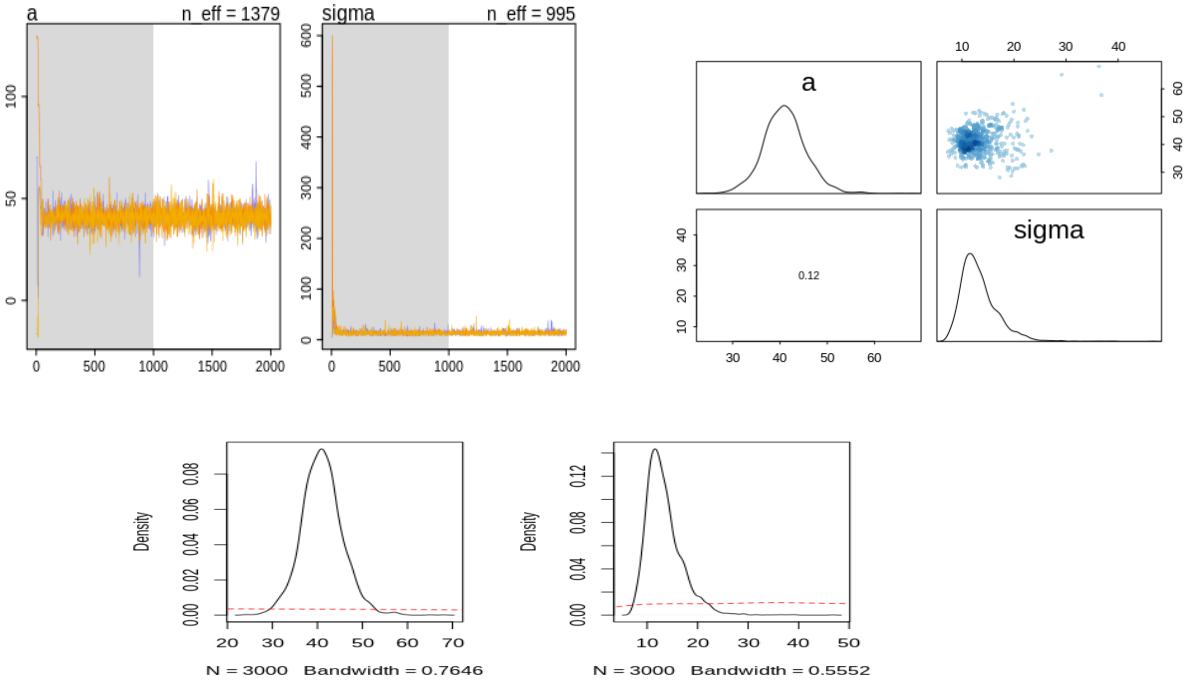
In the model, we specify the calculation of the likelihood `dnorm(mu,sigma)`for each element of the sample. For more information read McCarthy (2007) and McElreath (2016). We use the function `sample( )` to obtain a sample of 10 random plant heights and allocated them to the variable `mu`. We allocate the data to `height`, and enter settings for the Monte Carlo Markov chain (MCMC): `chains =` number of chains

```
##################
# size  10
##################

x <- sample(Height_data,size)
samp_data <- as.data.frame(cbind(1:size,x))
colnames(samp_data) <- c("id","height")

model.difuse <- map2stan(
  alist(
    height ~ dnorm(mu,sigma),
    mu <- a,
    a ~ dnorm(0,100),
    #sigma ~ dunif(0,100)
    sigma ~ dcauchy(0,1)
  ),
  data = samp_data ,chains =2,
)
```

The function `map2stan()` determines the procedure for the model and allocates the results into `model.difuse`. We use the function `precis` to call the results. A realization of the program is presented below.

```
> precis(model.difuse,digits=2)
       Mean StdDev lower 0.89 upper 0.89 n_eff Rhat
a     41.02   4.56      33.98      48.20  1379    1
```

```
sigma 13.37   3.65       8.42      18.36   995    1
```



For this implementation with diffuse priors (in red above), we obtained an estimate (in black above) of the sample mean = 41.02 and SD = 4.56 SD refers to the standard deviation of the parameters estimated of this sampled data) and the variance (mean=13.37, with SD =3.65) and their 89% credible intervals. Remember that the results will vary for each sample. These estimates are different but commensurate to those obtained with the frequentist approach.

For the **second model** we use informed priors based on the data of the populations. In this case we have prior relevant information. We have a mean and a variance from each population. For the overall mean: dnorm (34.47, 4.19) [`pop.mean.mean,pop.mean.var]` and for the overall variance: dlnorm(2.52, 0.82) [`log.pop.var.mean, var(log(Hc_pop$pop_sd^2))]` among populations. Compare the distributions of the means and priors in Figure 4. You can think of a scenario where two different people sample independently *Hypericum cumulicola*. The first person collects a more extensive census of several populations while the second one takes a smaller overall sample of 10 individuals. The second person uses as prior the information from the larger sample. The data available for the second person is the sample of 10 individuals but she incorporates the information from the first person. We need to replace the following lines of code in the R script part to calculate informed parameters of the prior:

```
## Priors
    a ~ dnorm(34.47,4.19),
    sigma ~ dlnorm(2.52,0.82)
```

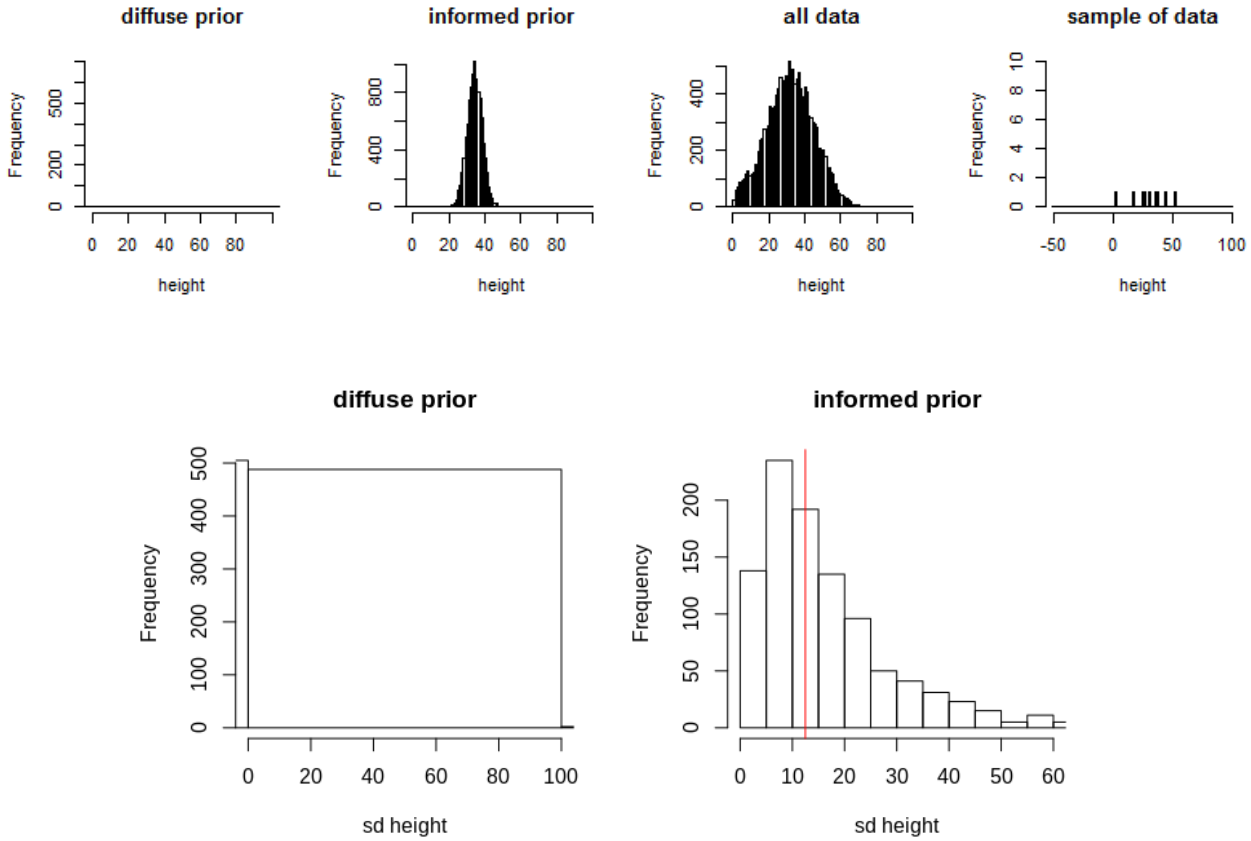


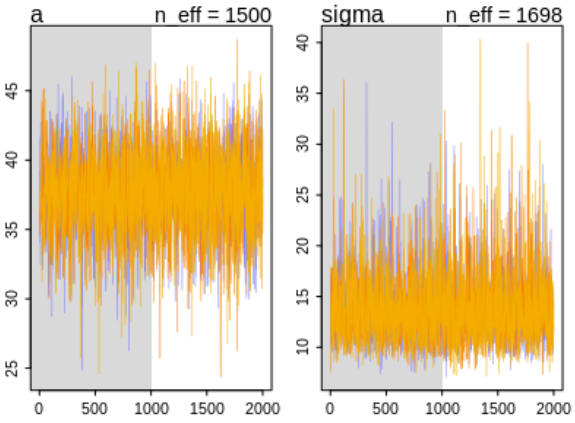
Figure Distributions of priors and the data for the models of mean plant height.

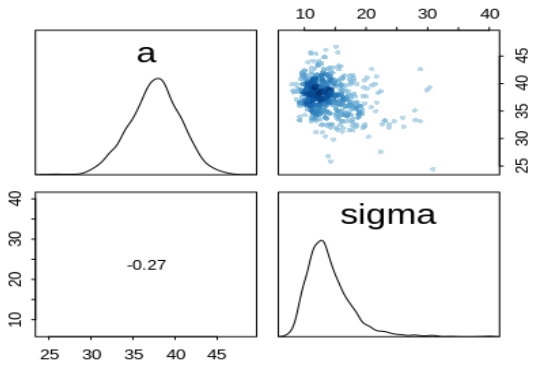
A realization of the program is presented below.

```
> precis(model.informed,digits=2)
       Mean StdDev lower 0.89 upper 0.89 n_eff Rhat
a     37.61   3.08      32.56      42.39  1500    1
sigma 13.95   3.67       8.96      18.68  1698    1
```

For the implementation of the model with informative priors, we obtained an estimate of the sample mean (37.61) and the sigma (13.95) and their 89% credible intervals (remember that the results will vary for each sample):

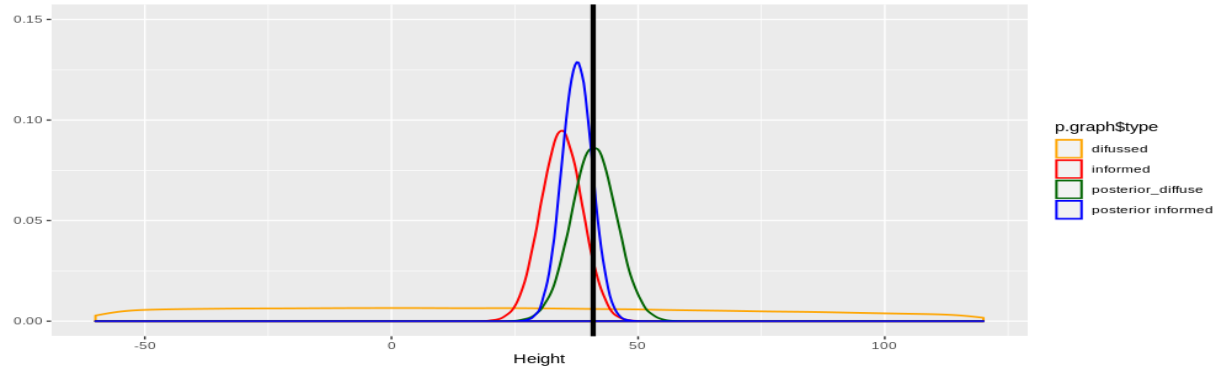
We now implement a frequentist approach estimate for the sample of 10 individuals and allocate the results in a table to compare the results.






```
> round(Results,2)
                Mean StdDev l m 0.89 u m 0.89 Sigma Stdev l sd 0.89 u sd 0.89
Frequentist        40.89   4.06     35.54     46.24 12.83     NA         NA         NA
B diffuse priors   41.02   4.56     33.98     48.20 13.37   3.65       8.42      18.36
B informed priors  37.61   3.08     32.56     42.39 13.95   3.67       8.96      18.68
Population         34.47     NA        NA        NA 12.48     NA       0.00       0.00
Overall            32.44   0.10     32.31     32.57 13.07     NA       0.00       0.00
```

This realization of the Bayesian model with informed priors had closer values to the mean of the populations and much narrower credible intervals (32.56-42.39) than those estimated with the uninformed Bayesian model (33.98-48.20) or the confidence interval of the frequentist approach (35.54-46.24). Repeat the whole process but with a sample of 100 (`size <- 100`) to evaluate how sample size affects the outcomes.



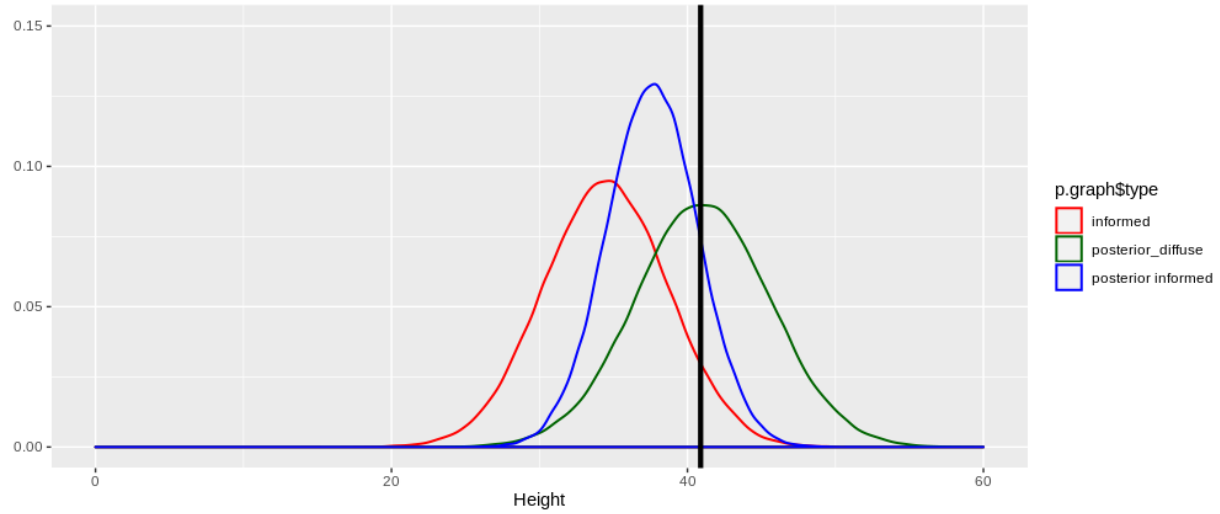


Figure 6. Distributions of priors and posteriors for the diffuse and informed models. The black line indicates the position of the frequentist estimate. The plot on the bottom is the same as the one on the top but with a smaller x scale.

NOTE: all the materials for this demo can be found at:

https://sciences.ucf.edu/biology/d4lab/methods-2/

References.

McCarthy. M.A. 2007. Bayesian Methods for Ecology. Cambridge University Press.

McElreath, R.M. 2016. Statistical Rethinking: a Bayesian course with examples in R and Stan. Chapman and Hall.

Quintana-Ascencio, P. F., E. S. Menges, and C. Weekley. 2003. A fire-explicit population viability analysis of Hypericum cumulicola in Florida rosemary scrub. Conservation Biology 17: 433-449.

Quintana-Ascencio, P.F. Koontz, S., Smith, V., David, A., Sclater, V. L. & E. S Menges. 2018. Predicting landscape-level distribution and abundance: Integrating demography, fire, elevation, and landscape habitat configuration. Journal of Ecology, 106: 2395-2408

Quintana-Ascencio, P.F. Koontz, S.M., Ochocki, B., Sclater, V. L., López-Borghesi, F., Li, H. & E. S Menges. 2019. Assessing the roles of seed bank, seed dispersal and historical disturbances for metapopulation persistence of a pyrogenic herb. Journal of Ecology, 107: 2760-2771.