# Model selection for Mixed Effects Models: Effects of fire on reproduction of a rare plant

In a prior demo, we demonstrated the advantages of recognizing the nature of our sampling schemes and evaluating the effects of random factors in our models. Here, we discuss how to implement model selection for models with mixed effects including correlated random intercepts. We follow the procedure described in McElreath 2016 (p: 411-419). We already provided evidence that number of reproductive structures of *Hypericum cumulicola* is clearly associated with plant standing height (Quintana-Ascencio *et al.* 2003, 2018, 2019). We also established that there was variation on number of fruits among populations. Now, we want to evaluate the relevance of one more fixed variable to explain variation in fecundity for this species. Time-since-fire (TSF) affects, among other things, nutrient and water availability, and abundance of predators and competitors, potentially influencing plant attributes (age & size) and resources available for reproduction. We also evaluate explicitly the covariance among populations. We obtain a matrix of distance among populations since we expect that populations that are closer to each other are more similar environmentally. A map of the distribution of *Hypericum cumulicola* populations in Archbold is in Figure 2. We use a model selection approach to assess the relative importance of fire and population similarity to explain variation in fruit production of *Hypericum cumulicola.*

**Figure 1**. Fire in the FL scrub!

For this demo you will need
The script `Mixed LMM 2019.R`,
The data files: `hypericum_data_94_07.txt` and `Bald_mat_dis.txt`, `Bald_coordinates.txt`
A `STAN` version that is compatible with your R and the `rstan` and `rethinking` packages.

We prepare the data in the same way as before but add one new variable. We read the two data files. The demographic data and the matrix of population distances.

```
orig_data <- read.table("hypericum_data_94_07.txt", header=T)
dist_data <- read.table("Bald_mat_dis.txt", header=T)
```

We prepare the matrix of population distances to be sent to `Stan`.

```
dist_data <- as.matrix(dist_data/1000,14,14)
dist <- as.data.frame(dist_data)
round(dist_data,2)
site <- unique(orig_data$bald)
ran_sites <-sample(site,6)
ran_sites <- ran_sites[order(ran_sites)]
```

For this example, we work on a subset of populations since the RAM of our personal computers cannot deal with the amount of data. You could run the complete data set in a more powerful computer. We only use 6 of the 14 available populations. Your randomly selected populations set may be different. We will compare our results in class. Remember to identify your set calling `ran_sites`. In this example the populations selected were:

```
> ran_sites
[1] 29 42 59 87 88 91
```

We retrieve the distances among the chosen populations

```
dis_pop_ran <- array(0,c(6,6))
for(i in 1:6){
  for(j in 1:6){
     dis_pop_ran[i,j] <- dist[which(ran_sites[i]==site),
                              which(ran_sites[j]==site)]
  }
}
colnames(dis_pop_ran) <- ran_sites
rownames(dis_pop_ran) <- ran_sites
dis_pop_ran
```

We subset the data to only include the selected populations and the data for the year 1995

```
dt <- subset(orig_data, !is.na(ht_init) & !is.na(st_init) & rp_init > 0 &
               year==1995 & bald == ran_sites )
yr <- unique(dt$year)
```

We obtain the logarithms of height and number of fruits.

```
dt$lgh <- log(dt$ht_init)
dt$lfr <- log(dt$rp_init)
```

We calculate time-since-fire (`tsf`) as the difference between sampling year and the year of the last fire for each population.

```
table(dt$bald,dt$fire_year)
dt$tsf <- dt$year-dt$fire_year
table(dt$bald,dt$tsf)
```

We scale the data to facilitate calculations and minimize the correlation of the interaction coefficients.

```
dt$lgh_s <- scale(dt$lgh)
dt$tsf_s <- scale(dt$tsf)
```

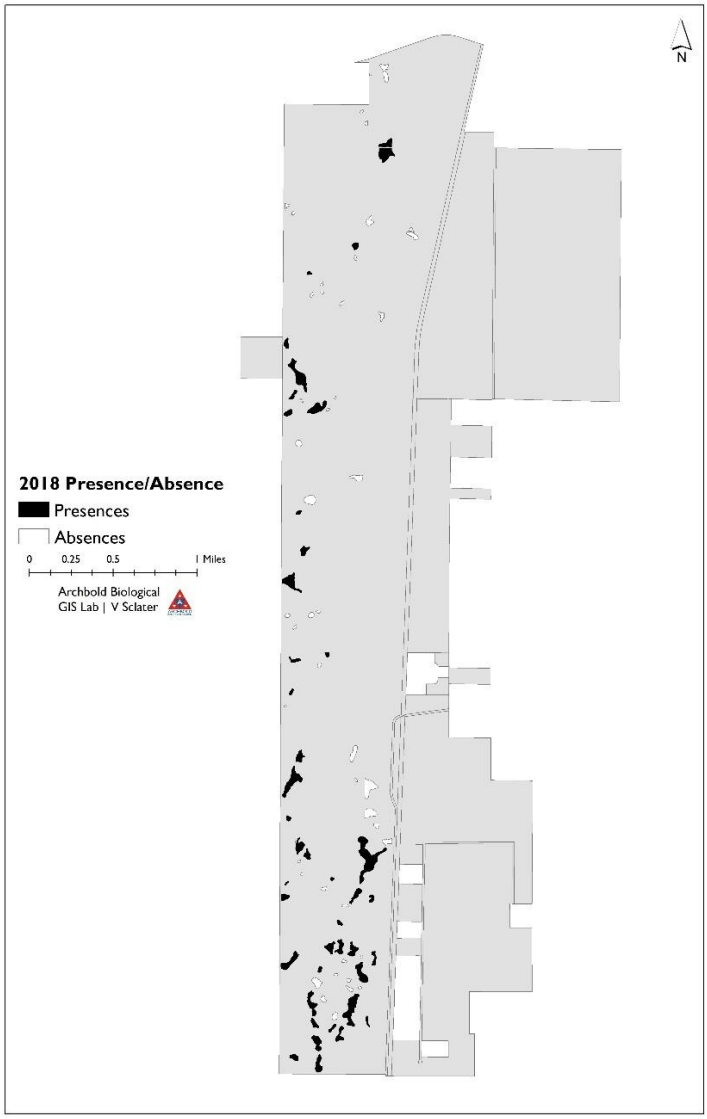


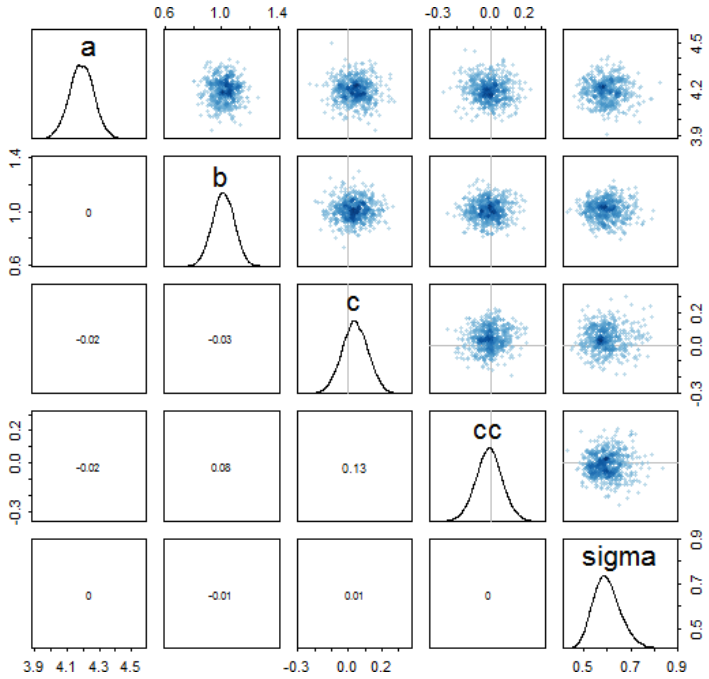

**Figure 2**. Florida rosemary scrub patches and *Hypericum cumulicola* occurrence in Archbold Biological Station. This species only occurs in a fraction of the possible suitable patches in the region (Quintana-Ascencio & Menges 1996; Quintana-Ascencio, Dolan & Menges 1998)

Zuur et al. (2009) caution about the need to start with a model that includes all possible fixed effects to evaluate the best configuration for the random factors. For our data this model includes two single factors, height and time-since-fire, and the two-way interaction among height and time-since-fire. We propose three options for the random configuration: (i) no random effects, (ii) random effects on the intercept given the population, and (iii) correlated random effects among populations. It is your responsibility to inspect that the models are generated properly. After the previous demo the first two models should be familiar.

```
## model with no random effects

m_no <- ulam( alist(
    lfr ~ dnorm(mu,sigma),
    mu <- a + b*lgh_s + c*tsf_s + cc*tsf_s*lgh_s,
    a ~ dnorm(0,50),
    b ~ dnorm(0,1),
    c ~ dnorm(0,1),
    cc ~ dnorm(0,1),
    sigma ~ dunif(0,1)
  ),
  data = dt,chains =3,
  iter=6000,warmup=2000
)
precis(m_no,digits=3)
```

| | Mean | StdDev | lower 0.89 | upper 0.89 | n_eff | Rhat |
|---|---|---|---|---|---|---|
| a | 4.190 | 0.077 | 4.068 | 4.313 | 14549 | 1 |
| b | 1.012 | 0.079 | 0.884 | 1.133 | 14095 | 1 |
| c | 0.041 | 0.079 | -0.090 | 0.162 | 14635 | 1 |
| cc | -0.010 | 0.078 | -0.135 | 0.113 | 14230 | 1 |
| sigma | 0.599 | 0.059 | 0.509 | 0.693 | 11176 | 1 |




```
# Assign populations indices

dt$pj <- 1
for (i in 2: length(ran_sites)){
  dt$pj[dt$bald==ran_sites[i]] <-i}
```

```
# model with random intercepts

m_rinter <- ulam( alist(
    lfr ~ dnorm(mu,sigma),
    mu <- a + p[pj]+ b*lgh_s + c*tsf_s + cc*tsf_s*lgh_s,
    a ~ dnorm(0,50),
    b ~ dnorm(0,1),
    c ~ dnorm(0,1),
    cc ~ dnorm(0,1),
    p[pj] ~ dnorm(0,sigmap),
    sigmap ~ dcauchy(0,1),
    sigma ~ dcauchy(0,1)
  ),
  data = dt,chains =3
)
> precis(m_rinter,depth=2,digits=3)
```

| | Mean | StdDev | lower 0.89 | upper 0.89 | n_eff | Rhat |
|---|---|---|---|---|---|---|
| a | 4.214 | 0.211 | 3.878 | 4.527 | 1189 | 1.001 |
| b | 1.078 | 0.071 | 0.969 | 1.190 | 2639 | 1.000 |
| c | 0.023 | 0.240 | -0.321 | 0.406 | 937 | 1.002 |
| cc | -0.031 | 0.069 | -0.143 | 0.074 | 2204 | 1.000 |
| p[1] | 0.006 | 0.252 | -0.420 | 0.372 | 1322 | 1.002 |
| p[2] | 0.094 | 0.333 | -0.471 | 0.568 | 916 | 1.004 |
| p[3] | 0.041 | 0.497 | -0.710 | 0.816 | 1142 | 1.002 |
| p[4] | -0.332 | 0.257 | -0.755 | 0.048 | 1265 | 1.000 |
| p[5] | 0.513 | 0.275 | 0.074 | 0.939 | 1348 | 1.002 |
| p[6] | -0.284 | 0.240 | -0.635 | 0.094 | 1152 | 1.000 |
| sigmap | 0.493 | 0.261 | 0.127 | 0.828 | 828 | 1.006 |
| sigma | 0.522 | 0.055 | 0.434 | 0.603 | 1821 | 1.000 |

For the third model we use the distances (in km) between each population (after: McElreath 2016) The geographic matrix is displayed below:

```
> dis_pop_ran
    29    42    59    87    88    91
29 0.000 1.658 2.834 4.134 4.057 4.441
42 1.658 0.000 0.931 2.259 2.131 2.471
59 2.834 0.931 0.000 1.283 1.180 1.597
87 4.134 2.259 1.283 0.000 0.278 0.726
88 4.057 2.131 1.180 0.278 0.000 0.490
91 4.441 2.471 1.597 0.726 0.490 0.000
```

For example, population 29 is approximately 1.6 km from population 42. Notice that the diagonal is all zeros and that the matrix is symmetric around the diagonal. This will allow us to estimate varying intercepts for each population that account for non-independence in fruit production as a function of distance. The first lines of the model are familiar. The `p[pj]` will be the varying intercepts in this case. We also include ordinary coefficients for the other variables. The highlight in this model is the multivariate prior for the intercepts

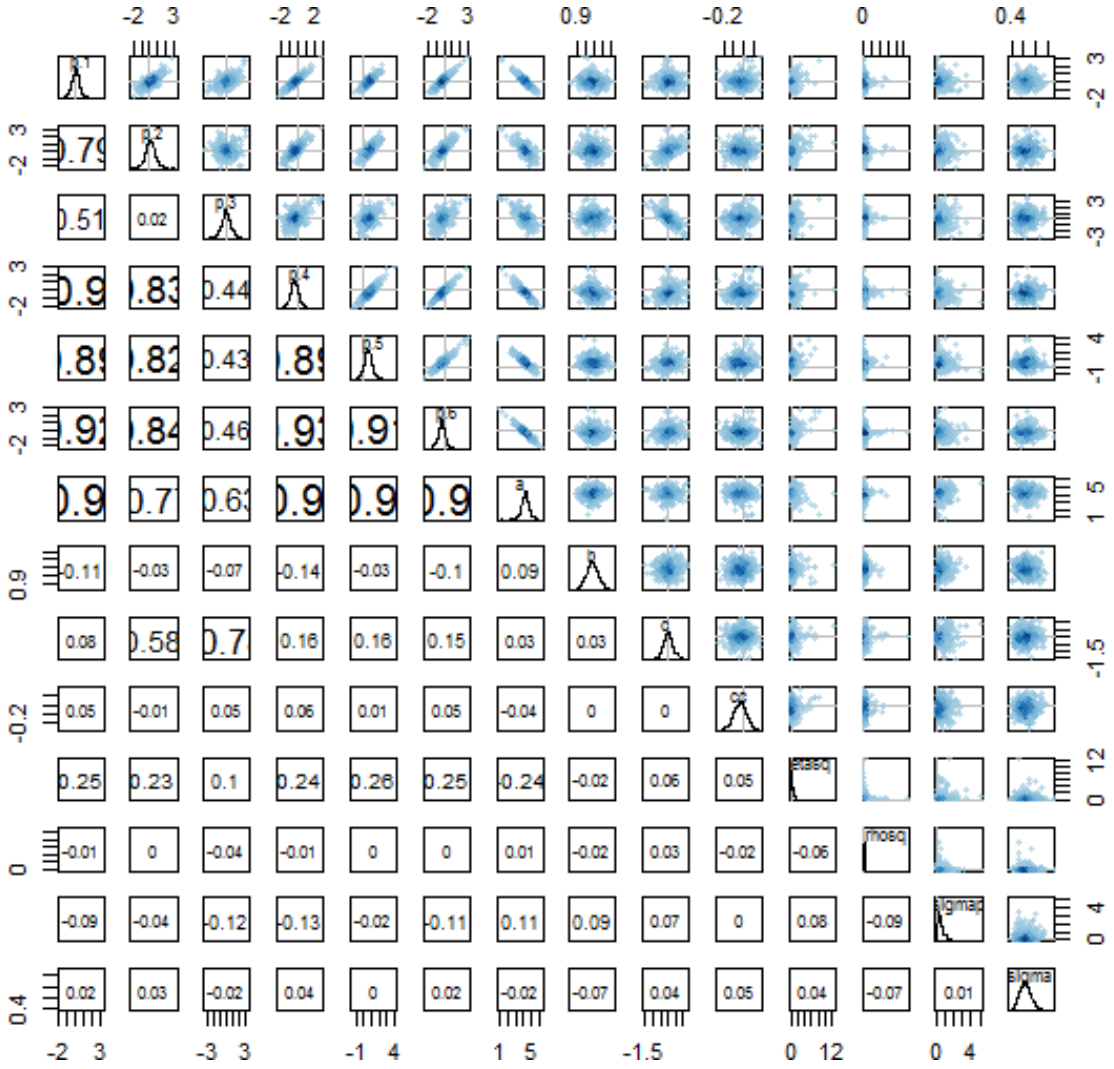
$$\gamma\gamma \sim MVNormal([0, \ldots .0], K) \quad \text{\# prior for intercepts}$$

$$K_{ij} = \eta^2 \exp(-\rho^2 D^2_{ij}) + \delta_{ij}\sigma^2 \quad \text{\# covariance matrix}$$

The first line is a 6-dimensional Gaussian prior for the intercepts. The vector of means is all zeros because the grand mean is in "a" in the model, which make the intercepts deviations from the expectation. The covariance matrix is **K**. This covariance is defined by the formula on the second line above. This function uses three parameters η, ρ, σ to model how the covariance among populations change with distance. The part $\exp(-\rho^2 D^2_{ij})$, where $D$ is distance, indicates that covariance between populations decline exponentially with the square of the distance. The parameter ρ determines the rate of decline. The last two pieces are $\eta^2$, the maximum covariance between two populations i and j, and $\delta_{ij}\sigma^2$, the extra covariance beyond $\eta^2$, when i = j. The model computes the posterior distribution of η, ρ, σ, and needs priors for them.

```
m_rspac <- ulam(
  alist(
    lfr ~ dnorm(mu,sigma),
    mu <- a + p[pj]+ b*lgh_s+ c*tsf_s  + cc*tsf_s*lgh_s,
    p[pj] ~ GPL2(Dmat,etasq,rhosq,sigmap),
    a ~ dnorm(0,50),
    b ~ dnorm(0,1),
    c ~ dnorm(0,1),
    cc ~ dnorm(0,1),
    etasq ~ dcauchy(0,1),
    rhosq ~ dcauchy(0,1),
    sigmap ~ exponential(1),
    sigma ~ dcauchy(0,1)
  ),
```

```
data = list(lfr = dt$lfr,
            lgh_s = dt$lgh_s,
            tsf_s = dt$tsf_s,
            pj = dt$pj,
            Dmat = dis_pop_ran,
            warmup = 2000, iter = 1e4,
            chains =1))
```




```
> precis(m_rspac,depth=2,digits=3)

         Mean StdDev lower 0.89 upper 0.89 n_eff  Rhat
p[1]    0.042  0.579     -0.900      0.853   232 1.005
p[2]    0.137  0.701     -0.926      1.307   269 1.001
p[3]    0.120  0.927     -1.506      1.381   295 1.005
p[4]   -0.351  0.606     -1.323      0.543   242 1.003
p[5]    0.607  0.594     -0.417      1.412   216 1.004
p[6]   -0.276  0.583     -1.199      0.559   218 1.005
```

```
a        4.180   0.568      3.462      5.203   212 1.006
b        1.085   0.071      0.981      1.210   878 0.999
c        0.008   0.386     -0.530      0.676   463 1.001
cc      -0.034   0.067     -0.130      0.074   919 1.000
etasq    0.599   0.920      0.002      1.236   427 0.999
rhosq    4.177  15.396      0.001      5.738   623 1.002
sigmap   0.523   0.496      0.004      1.106   757 0.999
sigma    0.518   0.052      0.435      0.596   758 1.003
```

The WAICs of these models indicate that the ones with random intercepts are most plausible. In this case, adding the spatial relationships among populations does not appear to provide much additional information. What do you think? This approach warrantees that we explore the whole variation associated with the fixed and random factors before deciding the inference from our models.

```
> compare(m_no,m_rinter,m_rspac)
          WAIC pWAIC dWAIC weight    SE   dSE
m_rspac   98.6   7.4   0.0   0.57 10.94    NA
m_rinter 99.2   7.3   0.6   0.43 10.90  0.83
m_no     111.3   3.8  12.6   0.00 10.68  6.92
```

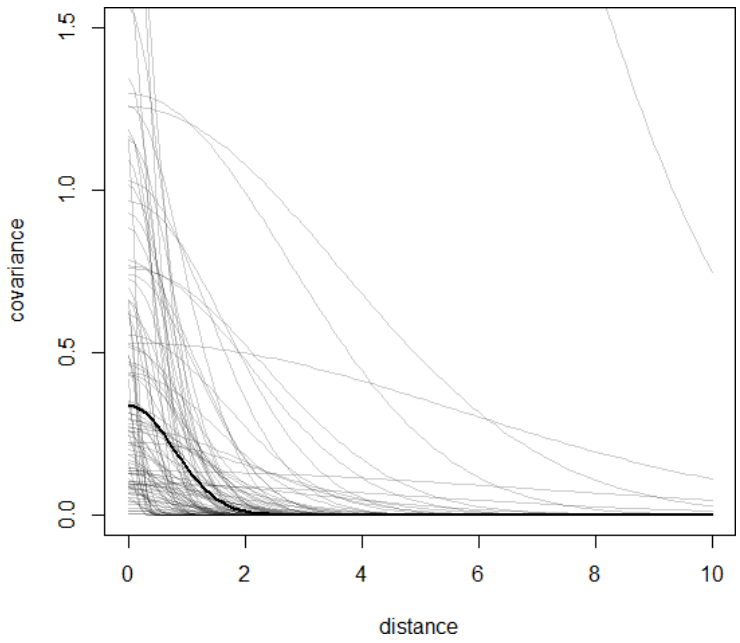


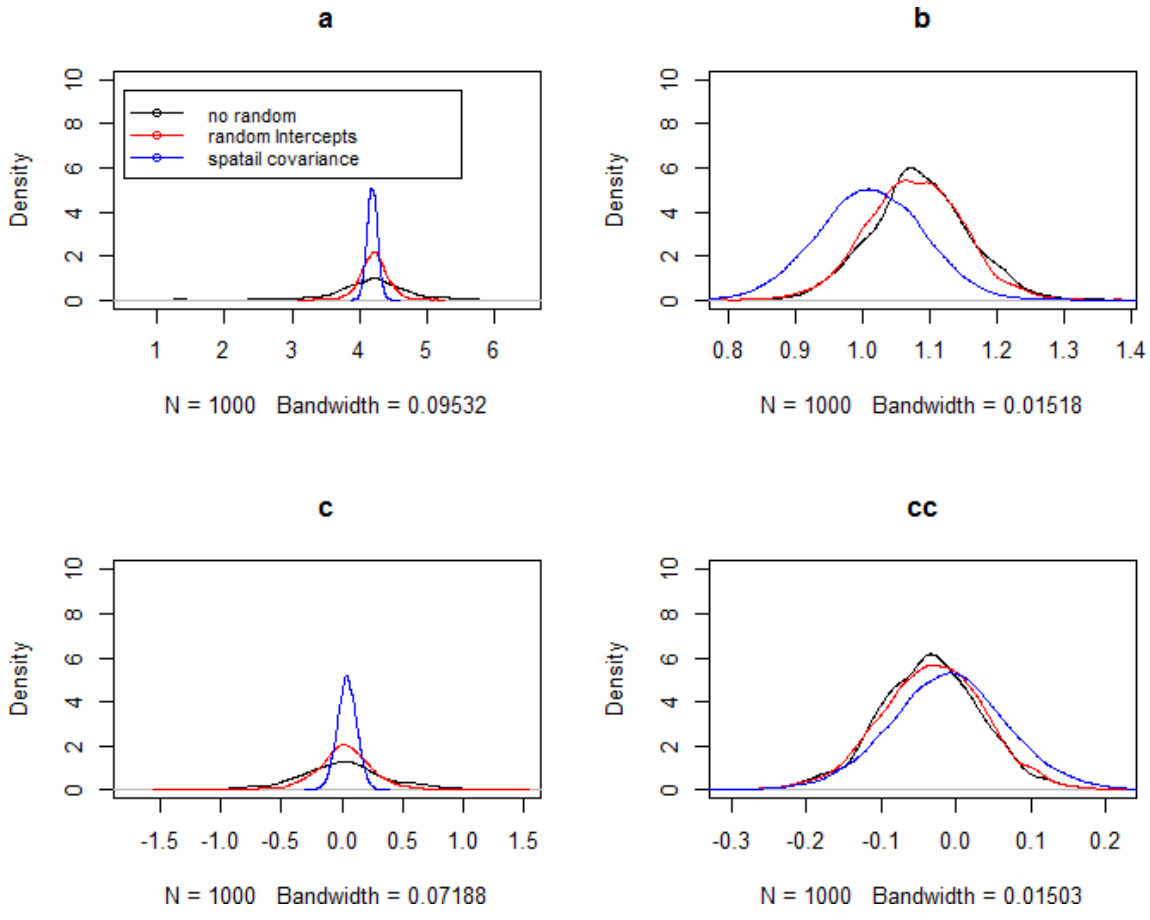

**Figure 6.** Posterior distribution of the spatial covariance between pair of populations. The dark curve displays the posterior media. The thin curves show the 100 realizations sampled from the join posterior distribution of $\eta, \rho$. Notice that, given the distances between the populations, many are virtually independent of each other.

We evaluate the correlations in fruit production among the populations evaluated

```
> Rho
     29   42   59   87   88   91
29 1.0 0.10 0.00 0.00 0.00 0.00
42 0.1 1.00 0.47 0.01 0.02 0.01
59 0.0 0.47 1.00 0.24 0.30 0.11
87 0.0 0.01 0.24 1.00 0.91 0.62
88 0.0 0.02 0.30 0.91 1.00 0.79
91 0.0 0.01 0.11 0.62 0.79 1.00
```

Correlations in fruit production are high among populations 88, 89 and 91 which are near each other in the south portion of Archbold Biological Station and share same time-since-fire.

In the figure panel below observe that the structure of the random effects changes the inference on the fixed factors.



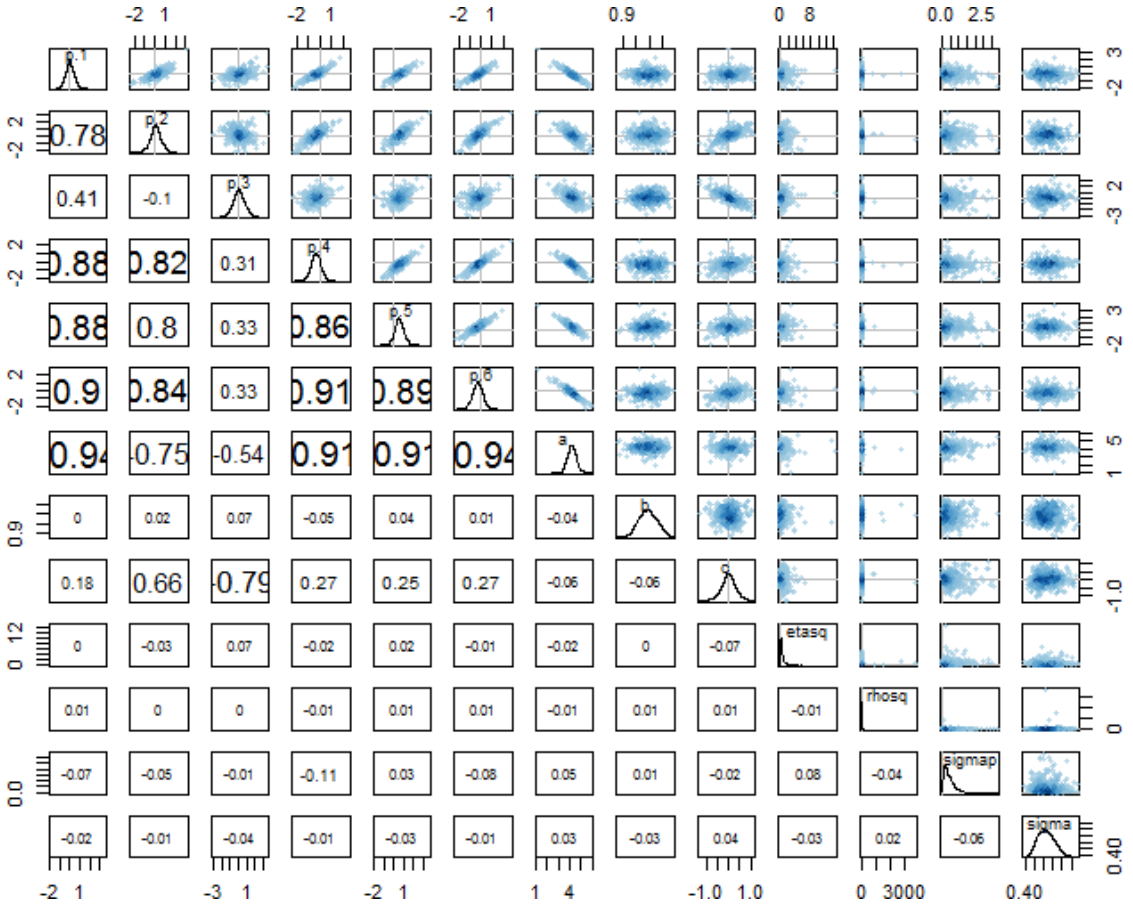

Arguably, fixed factors are the ones in which we are more interested. We now evaluate the information in models with different structure for the fixed factors. We have already evaluated the

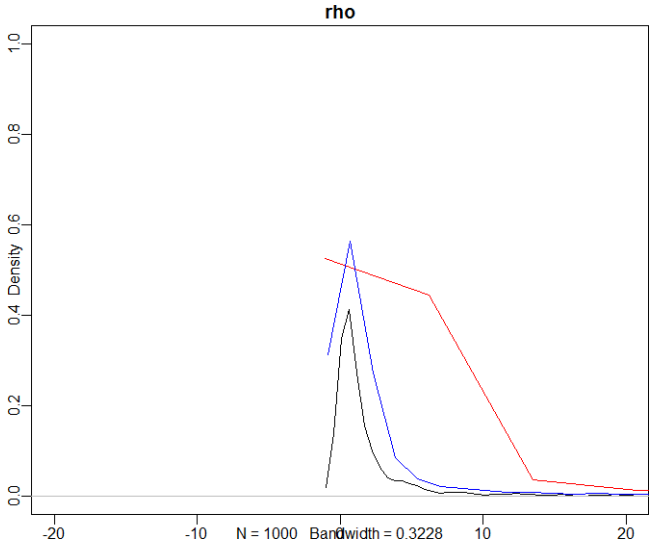
saturated model and now we generate the model without interactions using the configuration with spatial information for random effects by population and compare them.

```
> compare(m_rinter,m_rspac,m_rspac_noint)
              WAIC pWAIC dWAIC weight    SE  dSE
m_rspac_noint 96.6   6.6   0.0   0.61 10.89   NA
m_rspac       98.6   7.4   2.0   0.22 10.94 0.98
m_rinterc     99.2   7.3   2.6   0.16 10.90 1.13
```

It provides weak evidence for variation in number of fruits due to the interaction of height and time-since fire. It also changed (making it more uncertain), the distribution of ρ, the rate of decline

```
> precis(m_rspac_noint,depth=2,digits=3)
         Mean   StdDev lower 0.89 upper 0.89 n_eff  Rhat
p[1]    0.013    0.502     -0.718      0.733   228 1.004
p[2]    0.093    0.642     -0.873      1.127   263 1.001
p[3]    0.108    0.795     -1.080      1.371   371 0.999
p[4]   -0.359    0.508     -1.095      0.429   226 1.000
p[5]    0.582    0.521     -0.297      1.252   254 1.001
p[6]   -0.297    0.510     -1.078      0.469   251 1.000
a       4.199    0.474      3.493      4.841   226 1.001
b       1.088    0.070      0.974      1.195   756 1.000
c       0.007    0.361     -0.568      0.573   405 1.000
etasq   0.597    0.974      0.001      1.314   686 0.999
rhosq  14.182  161.159      0.000      5.383   241 1.004
sigmap  0.492    0.478      0.014      1.014   615 1.001
sigma   0.515    0.050      0.438      0.591   569 0.999
```


p.1
p.2
p.3
p.4
p.5
p.6
a
b
c
etasq
rhosq
sigmap
sigma
0.78
0.41
-0.1
0.88
0.82
0.31
0.88
0.8
0.33
0.86
0.9
0.84
0.33
0.91
0.89
0.94
-0.75
-0.54
0.91
0.91
0.94
0
0.02
0.07
-0.05
0.04
0.01
-0.04
0.18
0.66
-0.79
0.27
0.25
0.27
-0.06
-0.06
0
-0.03
0.07
-0.02
0.02
-0.01
-0.02
0
-0.07
0.01
0
0
-0.01
0.01
0.01
-0.01
0.01
0.01
-0.01
-0.07
-0.05
-0.01
-0.11
0.03
-0.08
0.05
0.01
-0.02
0.08
-0.04
-0.02
-0.01
-0.04
-0.01
-0.03
-0.01
0.03
-0.03
0.04
-0.03
0.02
-0.06
-2 1
-2 1
-2 1
0.9
0 8
0.0 2.5
-2 3
-3 2
-2 3
1 5
-1.0
0
0.40
-2 1
-3 1
-2 1
1 4
-1.0 1.0
0 3000
0.40


rho
Density
N = 1000 Bandwidth = 0.3228
0.0
0.2
0.4
0.6
0.8
1.0
-20
-10
0
10
20

We compare these models to the one without the effect of fire

```
> compare(m_rint,m_rspac,m_rspac_noint,m_rspac_noint_nofire)
                       WAIC pWAIC dWAIC weight    SE   dSE
m_rspac_noint          96.6   6.6   0.0   0.41 10.89    NA
m_rspac_noint_nofire   97.0   6.5   0.4   0.34 10.64  0.35
m_rspac                98.6   7.4   2.0   0.15 10.94  0.98
m_rint                 99.2   7.3   2.6   0.11 10.90  1.13
```

The information that we obtain about the posterior distribution of the coefficient of plant height from the three models with spatial population random effects is consistent. How you interpret this information?

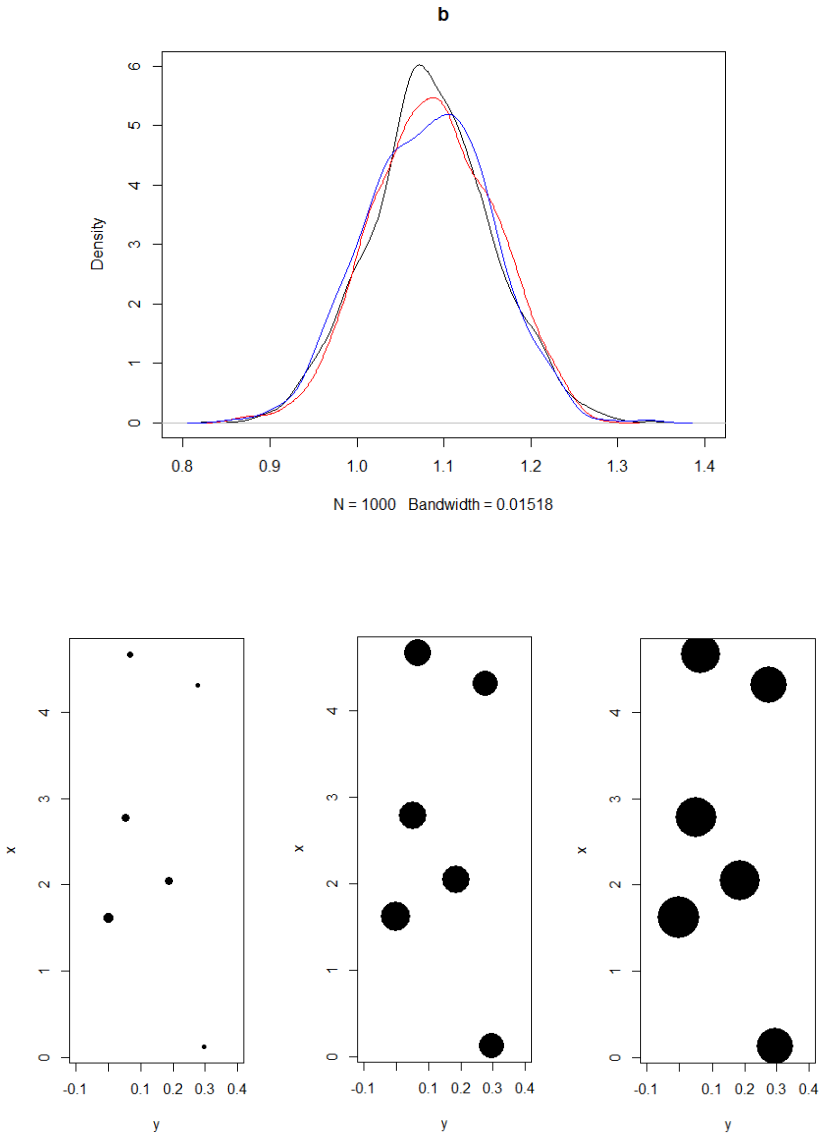



**Figure.** Predicted average relative number of fruits for the smallest, average and largest reproductive plant per site in the longest time-since-fire.

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

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

## References

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

Quintana-Ascencio, P. F and E. S. Menges. 1996. Inferring metapopulation dynamics from patch-level incidence of Florida scrub plants. Conservation Biology, 10: 1210-1219.

Quintana-Ascencio, P. F., R. W. Dolan and E. S. Menges. 1998. *Hypericum cumulicola* demography in unoccupied and occupied Florida scrub patches with different time-since-fire. Journal of Ecology, 86: 640-651.

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. and 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. and 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.

Zuur, A, J.M. Hilbe and E N. Leno. 2015. A beginner's guide to GLM and GLMM with R. Highland Statistics, Ltd.