ST 540 – MID-TERM 2

NI LUH PUTU S P PARAMITA

1. INTRODUCTION

The spatial distribution of the American black bear is studied using a dataset from https://www.inaturalist.org. The objective of this study is i) to examine ecological niche of black bears, and ii) to test for their local adaptation by ecoregion. The dataset contains n = 927 50 km^2 regions, each located in either one of four ecoregions (Marine West Coast Forest, Mediterranean California, North American Deserts, Northwestern Forested Mountains). There are 621 regions that include protected lands and the other 306 do not include protected lands. The covariates used to understand the ecological niche are proportion of the region that is forest, proportion of the region that is grassland, proportion of the region that is cropland, annual average temperature, annual average precipitation, and human population.

2. MODELS AND HYPOTHESES

The response variable is the number of black bear reports in region i, $Y_i \in \{0,1,\dots,N_i\}$ for $i = 1,2,\dots,n$. The likelihood is $Y_i \sim Binomial(N_i, p_i)$, where p_i is the true probability of observing at least one black bear in region i. The covariates X_{ij} for $j = 1,\dots,6$ are standardized to each has a mean of 0 and a standard deviation of 1. The analyses are done separately for regions that include and do not include protected lands since there is a tendency in the data that more black bear reports submitted for surveys in regions that include protected lands. Thus, it is reasonable to assume that the variation of the number of black bear reports and their ecological niche are different for protected and unprotected regions.

It is worth to assess the importance of each covariates to predict the ecological niche of black bear, and we may want to only include the "important" covariates in the models. To do this, we perform stochastic search variable selection (SSVS) using the logistic regression model:

$$logit(p_i) = \beta_0 + \sum_{j=1}^{6} \beta_j X_{ij}$$

with priors $\beta_0, \beta_j \sim Normal(0, 0.01)$. Then, in further analyses, we include covariates with inclusion posterior probability > 0.5 given by SSVS, considering this as a measurement of importance.

Since the SSVS is also done separately for protected and unprotected regions, we may see different covariates are included in the models for both types of region.

The following three logistic regression models are built each for protected and unprotected regions, where *p* is the number of covariates included in the models after SSVS and $e_i \in \{1,2,3,4\}$ is the ecoregion where region *i* is located.

- Model 1: The slopes and the intercept are treated as fixed effects, thus they are the same for all ecoregions. The logistic regression model is the same as the model used for SSVS, except that the number of covariates *p* is different (covariates are selected based on SSVS).
- Model 2: The slopes are still fixed, but the intercept is treated as mixed effect with random effects for the ecoregion. The logistic regression model is

$$logit(p_i) = \beta_0 + \theta_{e_i} + \sum_{j=1}^{p} \beta_j X_{ij}$$

with priors $\beta_0, \beta_j \sim Normal(0,0.01)$ and double-exponential random effects $\theta_{e_i} \sim DE(0,\tau)$, with prior $\tau \sim Gamma(0.1,0.1)$.

- Model 3: Both the intercept and the slopes are now treated as mixed effects allowing them to vary based on the ecoregion. The logistic regression model is

$$logit(p_i) = \beta_0 + \theta_{0,e_i} + \sum_{j=1}^{p} (\beta_j + \theta_{j,e_i}) X_{ij}$$

with priors $\beta_0, \beta_j \sim Normal(0,0.01)$ and double-exponential random effects $\theta_{0,e_i} \sim DE(0,\tau_0)$ and $\theta_{j,e_i} \sim DE(0,\tau_j)$, with priors $\tau_0 \sim Gamma(0.1,0.1)$ and $\tau_j \sim Gamma(0.1,0.1)$.

3. COMPUTATION

All models were fit using JAGS package in R, separately for regions with protected lands and with no protected lands. For SSVS, the first 10,000 samples were discarded as a burn-in then 50,000 samples were drawn for each of 3 chains with a thinning factor of 5. For building Model 1, Model 2, and Model 3, the burn-in samples are the first 30,000 then 100,000 samples were drawn for each of 2 chains. The model convergence is checked by looking at the trace plot, the Gelman-

Rubin statistic (< 1.1 indicates convergency), and the effective sample size (> 1000 indicates convergency). The trace plots show that convergency is attained for all parameters in all models. The Gelman-Rubin statistics are calculated for each parameter in every model, and the multivariate potential scale reduction factors (PSRF) for five models presented in this report are 1, while another model gets 1.04. The effective sample size (ESS) for all parameters in all models are also calculated, in general all models look good having ESS far greater than 1000. However, in Model 2 and Model 3 for regions with protected lands, not all random effects achieve ESS > 1000 (although they are all > 100)[1]. Regardless, the Gelman-Rubin statistics indicate convergency, thus all models still seem reasonable.

4. MODEL COMPARISONS

Based on the results from SSVS (Table 1), we select covariates with inclusion posterior probability > 0.5 to be included in Model 1, Model 2, and Model 3. The covariates included i) in the model for regions with protected lands are proportion of the region that is grassland, proportion of the region that is cropland, annual average precipitation, human population (p = 4), and ii) in the model for regions with no protected lands are proportion of the region that is forest, annual average temperature (p = 2). It is then reasonable to say that the ecological niche of black bear in regions with protected lands is more determined by four out of six covariates, while in regions with no protected lands this is explained rather by the other two covariates.

Table 1 Summary of the posteriors from SSVS

	Regions with protected lands				Regions with no protected lands			
	Incl. Prob.	50%	5%	95%	Incl. Prob.	50%	5%	95%
β_{forest}	0.25	0	-0.19	0	**0.77**	0.35	0	0.68
β_{grass}	**1**	0.40	0.26	0.53	0.50	0	-0.73	0.02
β_{crop}	**0.98**	-0.45	-0.75	-0.19	0.38	0	-0.63	0.09
β_{temp}	0.50	0	-0.25	0	**0.94**	-0.59	-1.01	0
β_{precip}	**1**	0.58	0.47	0.70	0.24	0	-0.23	0
β_{human}	**0.81**	-0.10	-0.18	0	0.35	0	-0.64	0

[1] I tried to increase the number of iterations to 500,000 and use different combinations of random effects (Gaussian/Double-Exponential) and priors (Gamma/Half-Cauchy), the ESS did not improve. Using DE random effects actually gives better ESS (> 100) for the random effects that do not achieve ESS > 1000, using Gaussian random effects only gives ESS around 20 – 30 for these parameters.

The model performance are compared using the deviance information criteria (DIC) and the Watanabe-Akaike information criterion (WAIC). The comparisons of the three models given these criteria are summarized in Table 2 for both regions with protected lands and regions with no protected lands. Although Model 3 gives the lowest DIC and WAIC for both regions with protected lands and with no protected lands, we use Model 2 to examine ecological niche of black bear and to test for their local adaptation. These models are easier to interpret and seem to be enough to address the objectives of this study. It is also harder to compare protected and unprotected regions using Model 3 to test for black bear's local adaption.

Table 2 Model comparisons for regions that include protected lands

	Regions with protected lands			Regions with no protected lands		
	DIC	WAIC	Multi PSRF	DIC	WAIC	Multi PSRF
Model 1	1267	1356	1	180	183	1
Model 2	1162	1278	1	167	171	1
Model 3	1114	1241	1.04	162	168	1

5. RESULTS

Applying Model 2, the posteriors of the parameters are summarized in Table 3 (for regions with protected lands) and Table 4 (for regions with no protected lands). All fixed effects parameters in the protected region's model are significant, while in the unprotected region's model only β_0 and β_f are significant[2]. Based on these results, we can suggest the following regarding the ecological niche of black bear:

- In the regions that include protected lands, with all other covariates held fixed: i) increasing the proportion of the region that is grassland by one multiplies the odds of observing a black bear by $\exp(0.45)$, ii) increasing the proportion of the region that is cropland by one divides the odds by $\exp(0.67)$, iii) increasing the precipitation by one multiplies the odds by $\exp(0.45)$, iv) increasing the human population by one divides the odds by $\exp(0.25)$.

[2] However, using Model 1, all fixed effects parameters in the unprotected region's model are significant.

- In the regions that do not include protected lands, with all other covariates held fixed: i) increasing the proportion of the region that is forest by one multiplies the odds by $\exp(0.64)$, ii) increasing the temperature by one divides the odds by $\exp(0.04)$ but note this is not significant.

Thus, we can conclude that the ecological niche of black bear in the regions that include protected lands is grassland area with higher precipitation. Meanwhile, the ecological nice of black bear in the regions that do not include protected lands is forest area.

To test for black bear's local adaptation, we look at the random effects parameters. In the protected region's model, the significant random effect is θ_{e_3}, and in the unprotected region's model, the significant random effect is θ_{e_4}[3]. This indicates that, with all covariates set to be zero, we can observe the following.

- In the protected regions, the odds of observing black bear is lower in North American Deserts than other ecoregions, which is $\exp(-10.84)$.
- In the unprotected regions, the odds is higher in Northwestern Forested Mountains than other ecoregions, which is $\exp(-3.53)$.

Thus, it seems reasonable to say that there is an indication of local adaption although the rest of random effects parameters are not significant.

Table 3 Summary of the posteriors for regions with protected lands

	Mean	2.5%	97.5%
β_0	**-4.01**	**-7.97**	**-1.63**
θ_{e_1}	0.58	-1.80	4.54
θ_{e_2}	0.80	-1.58	4.76
θ_{e_3}	**-6.83**	**-19.93**	**-1.82**
θ_{e_4}	0.22	-2.16	4.18
β_{grass}	**0.45**	**0.29**	**0.60**
β_{crop}	**-0.67**	**-1.06**	**-0.34**
β_{precip}	**0.43**	**0.29**	**0.58**
β_{human}	**-0.25**	**-0.34**	**-0.16**

Table 4 Summary of the posteriors for regions with no protected lands

	Mean	2.5%	97.5%
β_0	**-5.28**	**-7.02**	**-3.43**
θ_{e_1}	-0.29	-2.27	1.50
θ_{e_2}	-0.74	-3.47	0.94
θ_{e_3}	-0.22	-2.39	1.56
θ_{e_4}	**1.75**	**-0.05**	**3.79**
β_{forest}	**0.64**	**0.27**	**1.00**
β_{temp}	-0.04	-0.68	0.60

[3] The 95% credible interval slightly includes zero, but we consider this as significant.

R Code

SSVS (for regions with protected lands)

```
m <- textConnection("model{
 #Likelihood
 for(i in 1:n){
 Y[i] ~ dbin(p[i],N[i])
 logit(p[i]) <- alpha + inprod(X[i,],beta[])}

 #Prior
 for(j in 1:6){
 beta[j] <- gamma[j]*delta[j]
 gamma[j] ~ dbern(0.5)
 delta[j] ~ dnorm(0,tau)}

 alpha ~ dnorm(0,0.01)
 tau   ~ dgamma(0.1,0.1)
}")

data   <- list(Y=Y,X=X,N=N,n=n)
burn   <- 10000; iters  <- 50000; chains <- 3
model  <- jags.model(m,data = data, n.chains=chains,quiet=TRUE)
update(model, burn, progress.bar="none")
samps <- coda.samples(model, variable.names=c("beta"), thin=5, n.iter=iters,
progress.bar="none")

#summarize posterior of beta
beta    <- NULL
for(l in 1:chains){
  beta <- rbind(beta,samps[[l]])
}

Inc_Prob <- apply(beta!=0,2,mean)
Q        <- t(apply(beta,2,quantile,c(0.5,0.05,0.95)))
out      <- cbind(Inc_Prob,Q)
```

Model 2: random effects on intercept for ecoregion (for regions with protected lands)

```
m <- textConnection("model{
 #Likelihood
  for(i in 1:n){
  Y[i] ~ dbin(p[i],N[i])
  logit(p[i]) <- theta[er[i]] + inprod(X[i,],beta[])}

  #Random effect
   for(j in 1:J){theta[j] ~ ddexp(0,tau)}
   tau ~ dgamma(0.1,0.1)

  #Prior for fixed effects
   for(j in 1:p){beta[j] ~ dnorm(0,0.01)}\

  #WAIC calculations
   for(i in 1:n){like[i] <- dbin(Y[i],p[i],N[i])}
}")

#load the model
dat <- list(Y=Y,N=N,X=X,n=n,er=er)
init <- list(theta=rep(0,J),beta=rep(0,p+1))
model5 <- jags.model(m5, inits=init, data = dat, n.chains=2,quiet=TRUE)

#generate samples
update(model, 30000, progress.bar="none")
samp <- coda.samples(model, variable.names=c("theta","beta"), n.iter=100000,
progress.bar="none")

#compile results
ESS <- effectiveSize(samp)
GEL <- gelman.diag(samp)

stat <- summary(samp)$statistics
quant <- summary(samp)$quantiles

#compute DIC
dic5 <- dic.samples(model,n.iter=100000,progress.bar="none")

#compute WAIC
waic <- coda.samples(model, variable.names=c("like"), n.iter=100000, progress.bar="none")
like <- waic5[[1]]
fbar <- colMeans(like)
P <- sum(apply(log(like),2,var))
WAIC <- -2*sum(log(fbar))+2*P
```