Hierarchical Bayesian species distribution models with the **hSDM** R Package

June 4, 2014

Adansonia grandidieri Baill. next to Andavadoaka village (southwest Madagascar).

Ghislain Vieilledent[⋆,1] Cory Merow[2] Jérôme Guélat[3]

Andrew M. Latimer[4] Marc Kéry[3]

Alan E. Gelfand[5] Adam M. Wilson[6] Frédéric Mortier[1]

and John A. Silander Jr.[2]

[⋆] **Corresponding author:** \E-mail: ghislain.vieilledent@cirad.fr \Phone: +33.(0)4.67.59.37.51 \Fax: +33.(0)4.67.59.39.09

[1] **Cirad** – UPR BSEF, F–34398 Montpellier, France

[2] **University of Connecticut** – Department of Ecology and Evolutionary Biology, Storrs, CT 06269, USA

[3] **Swiss Ornithological Institute** – 6204 Sempach, Switzerland

[4] **University of California** – Department of Plant Sciences, Davis, CA 95616, USA

[5] **Duke University** – Department of Statistical Science, Durham, NC 27708, USA

[6] **Yale University** – Department of Ecology and Evolutionary Biology, New Haven, CT 06520, USA

Abstract

Species distribution models (SDM) are useful tools to explain or predict species range from various environmental factors. SDM are thus widely used in conservation biology. Based on the observations of the species in the field (occurence or abundance data), SDM face two major problems which lead to bias in models' results: imperfect detection and spatial correlation of the observations.

At the present time, there is a lack of statistical tools to analyse large occurence or abundance data-sets (typically with tens of hundreds observation points) taking into account both imperfect detection and spatial correlation.

Here, we present the **hSDM** R package wich aims at providing user-friendly statistical functions to fill this gap. Functions were developped through a hierarchical Bayesian approach. They call a Metropolis-within-Gibbs algorithm coded in C to estimate model's parameters. Using compiled C code for the Gibbs sampler reduce drastically the computation time.

By making these new statistical tools available to the scientific community, we hope to democratize the use of more complex, but more realistic, statistical models for increasing knowledge in ecology and conserving biodiversity.

Keywords: R, C code, site-occupancy models, CAR process, spatial autocorrelation, biodiversity, SDM, niche modelling, detection probability, counts data, presence-absence, false absence, uncertainty, hierachical Bayesian models, Metropolis, MCMC, Gibbs sampler

1 Introduction

1.1 Species distribution models

Biogeography is the study of the distribution of species over space and time and biogeographers try to understand the factors determining a species distribution (Smith, 1868; Wallace, 1876). A species distribution is often represented with a map (Wallace, 1876). This knowledge on the ecology of the species can be used for several applications such as conservation biology (Thuiller *et al.*, 2014).

Species distribution modelling (alternatively known as "environmental niche modelling", "ecological niche modelling", "predictive habitat distribution modelling", and "climate envelope modelling") refers to the process of using computer algorithms to predict the distribution of species in geographic space on the basis of a mathematical representation of their known distribution in environmental space (i.e. the realized ecological niche). The environment is in most cases represented by climate data (such as temperature, and precipitation), but other variables such as soil type and land cover can also be used. Species distribution models (SDM) allow estimating the probability of presence or abundance of a species on a large geographical range using a limited number of species observations (Elith & Leathwick, 2009; Guisan & Zimmermann, 2000). Species observations can be occurence data (presence-absence data or presence only data) or abundance data (also known as count data).

1.2 Imperfect detection and spatial correlation of the observations

When considering presence-absence or abundance data for species distribution modelling, strong assumptions are usually made (Araujo & Guisan, 2006; Guisan & Thuiller, 2005; Sinclair *et al.*, 2010). Among these assumptions, two can lead to biased estimates of species distribution. The first one deals with imperfect detection and the second one with spatial correlation of the observations.

Regarding imperfect detection, occurrence of a species is typically not observed perfectly. Species traits, survey-specific conditions and site-specific characteristics may influence species detection probability which is often < 1 (Chen *et al.*, 2013). Thus, observations might include false absences. For example, the habitat can be suitable and the species is present but individuals have not been seen during the census. Or the habitat can be suitable but the species has not dispersed yet to the site (typical example for plant species, see Latimer *et al.* (2006)) or was not present on the site at the moment of the observation (typical example for animal species such as birds, see Kéry *et al.* (2005)). Treating observed occurrence and species distributions as the true occurrence and distribution, failing to make amendments for imperfect detection, may lead to problems in species distribution studies, habitat models and biodiversity management (Kéry & Schmidt, 2008; Lahoz-Monfort *et al.*, 2014; Latimer *et al.*, 2006).

Regarding spatial correlation, most species present geographical patchiness (positive

spatial autocorrelation). This pattern is often driven by multiple causes that may be associated to exogenous environmental factors such as climate or soil (which might be partly taken into account in species distribution models), but also to endogeneous biotic processes, called contagious processes, such as dispersal, migration, conspecific attraction or mortality which are rarely considered (Dormann *et al.*, 2007; Legendre, 1993; Lichstein *et al.*, 2002; Sokal & Oden, 1978). Due to the contagious biotic processes, the presence or abundance of a species at one site is influenced by the presence or abundance of the species at surrounding sites. A species might be present at a site where the environment is less suitable because of the presence of the species at neighbouring sites where the environment is higly suitable. Thus, ignoring spatial correlation may lead to biased conclusions about ecological relationships (Lichstein *et al.*, 2002) and even invert the slope of relationships from non-spatial analysis in some particular cases (Kühn *et al.*, 2006). In addition to its ecological significance, spatial autocorrelation is problematic for classical species distribution models which assume independently distributed errors (Dormann *et al.*, 2007; Legendre, 1993; Lichstein *et al.*, 2002).

1.3 Methods and software to account for imperfect detection and spatial correlation

New classes of models, called site-occupancy models (MacKenzie *et al.*, 2002) or zero inflated binomial (ZIB) models (Latimer *et al.*, 2006) for presence-absence data and N-mixture models (Royle, 2004) or zero inflated Poisson (ZIP) models for abundance data (Flores *et al.*, 2009), were developed to solve the problems created by imperfect detection. These models combine two processes, an ecological process which describes habitat suitability and an observation process which takes into account imperfect detection. Because they mix probability distributions to represent the suitability and observation processes, these models have also been called mixture models. Mixture models use information from repeated observations at several sites to estimate detectability. Detectability may vary with site characteristics (e.g., habitat variables) or survey characteristics (e.g., weather conditions), whereas suitability relates only to site characteristics.

One additional point regarding site-occupancy models is that they form a unifying framework for a very large array of capture-recapture models to estimate population size in animal ecology (Nichols, 1992): using parameter-expanded data augmentation (Royle *et al.*, 2007), most models for population size, survival, recruitment and similar demographic quantities (presented in detail in standard references such as Williams *et al.* (2002), Royle & Dorazio (2008) and Kéry & Schaub (2012)) can be cast into the framework of an occupancy model and this makes their fitting much easier.

Several studies have demonstrated the advantages of site-occupancy and N-mixture models over classical models which do not consider imperfect detection. These studies have focused on the distribution of various plant or animal species in marine and terrestrial ecosystems (see Chen *et al.* (2013); Latimer *et al.* (2006) for plants, Dorazio *et al.* (2006); Kéry *et al.* (2005); Rota *et al.* (2011); Royle (2004) for birds, Kéry *et al.* (2010) for insects,

Bailey *et al.* (2004); Chelgren *et al.* (2011); MacKenzie *et al.* (2002) for amphibians, Monk (2014) for fishes, and Gray (2012); Poley *et al.* (2014) for mammals).

Several softwares can be used to fit site-occupancy and N-mixture models (Table 2). Some are based on the maximum likelihood approach (such as the widely used free Windows programs **MARK** and **PRESENCE** and the R package **unmarked**) while other are based on the hierarchical Bayesian approach (such as **WinBUGS** and **OpenBUGS** programs).

Softwares	Socc	Nmix	Sp	Approach	OS	Reference	URL
PRESENCE	1	1	0	ML	MS-W	MacKenzie (2006)	PRESENCE
MARK	1	1	0	ML	MS-W	White & Burnham (1999)	MARK
E-SURGE	1	0	0	ML	MS-W	Choquet *et al.* (2009)	E-SURGE
unmarked	1	1	0	ML	cross-platform	Fiske & Chandler (2011)	unmarked
stocc	1	0	1	Bayesian	cross-platform	Johnson *et al.* (2013)	stocc
JAGS	1	1	0	Bayesian	cross-platform		JAGS
Stan	1	1	0	Bayesian	cross-platform	Stan Development Team (2014)	Stan
WinBUGS	1	1	1	Bayesian	MS-W	Lunn *et al.* (2009)	WinBUGS
OpenBUGS	1	1	1	Bayesian	cross-platform	Lunn *et al.* (2009)	OpenBUGS
hSDM	1	1	1	Bayesian	cross-platform		hSDM

Table 2: **Softwares available for modeling species distribution including imperfect detection.**

A variety of methods have been developed to correct for the effects of spatial autocorrelation in species distribution models based on occurence or abundance data (Cressie & Cassie, 1993; Dormann *et al.*, 2007; Keitt *et al.*, 2002; Miller *et al.*, 2007). In their review article, Dormann *et al.* (2007) described six different statistical approaches to account for spatial autocorrelation: autocovariate regression; spatial eigenvector mapping; generalised least squares; autoregressive models and generalised estimating equations.

Several studies have demonstrated the advantages of these mehods focusing on a variety of plant or animal species (see Gelfand *et al.* (2005); Kühn *et al.* (2006); Latimer *et al.* (2006) for plants, Lichstein *et al.* (2002) for birds, and Johnson *et al.* (2013); Poley *et al.* (2014) for mammals).

Among the methods available to account for spatial autocorrelation, conditional autoregressive (CAR) models, which incorporate spatial autocorrelation through a neighbourhood structure, are commonly implemented in statistical softwares (Dormann *et al.*, 2007). The most commonly used softwares to implement CAR models are **OpenBUGS** and **WinBUGS** softwares (Lunn *et al.*, 2009) which have in-built functions (`car.normal` and `car.proper`) to describe the CAR process. CAR models can also be implemented in **BayesX** (Brezger *et al.*, 2005) and in the following R packages: **R-INLA** (Rue *et al.*, 2009), **CARBayes** (Lee, 2013), **stocc** (for binary data only), **spatcounts** (for count data only), **CARramps** (for Gaussian data only), and **spdep** (for Gaussian data only) (Table 4).

Softwares	Type of data	Approach	OS	Reference	URL
OpenBUGS	all	Bayesian	cross-platform	Lunn *et al.* (2009)	OpenBUGS
WinBUGS	all	Bayesian	MS-W	Lunn *et al.* (2009)	WinBUGS
BayesX	all	Bayesian	cross-platform	Brezger *et al.* (2005)	BayesX
R-INLA	all	Bayesian	cross-platform	Rue *et al.* (2009)	R-INLA
CARBayes	all	Bayesian	cross-platform	Lee (2013)	CARBayes
stocc	all	Bayesian	cross-platform	Johnson *et al.* (2013)	stocc
spatcounts	count	Bayesian	cross-platform		spatcounts
CARramps	Gaussian	Bayesian	cross-platform		CARramps
spdep	Gaussian	ML	cross-platform		spdep
hSDM	binomial and count	Bayesian	cross-platform		hSDM

Table 4: **Softwares available for modeling species distribution including spatial autocorrelation.**

1.4 Objectives of the hSDM R package

Among the available statistical programs, only **OpenBUGS** can be used on any operating system to fit both site-occupancy or N-mixture models including also a spatial autocorrelation process (Table 2 and Table 4). One problem is that **OpenBUGS**, for such models, cannot handle large data-sets (typically, data-sets with tens of thousands sites). Moreover, for smaller data-sets, models can be fitted but computation time can be long due to the fact that the **OpenBUGS** code is interpreted and not compiled. For this reason, we decided to develop the **hSDM** (for hierarchical Bayesian species distribution models) R package. The **stocc** R package (Johnson *et al.*, 2013; Poley *et al.*, 2014), which can handle binary data only, has been developed for the same reasons. The **hSDM** package allows the user to fit mixture models which take into account imperfect detection (site-occupancy, N-mixture, ZIB and ZIP models) and account for spatial autocorrelation. Spatial autocorrelation is represented through an intrinsic CAR process (Besag *et al.*, 1991). Functions in the **hSDM** R package use a Metropolis algorithm (Robert & Casella, 2004) in a Gibbs sampler (Casella & George, 1992; Gelfand & Smith, 1990) to obtain the posterior distribution of model's parameters. The Gibbs sampler is written in C code and compiled to optimize computation efficiency. Thus, the **hSDM** package can be used for very large data-sets while reducing drastically the computation time.

In this vignette, we present examples to illustrate the use of the **hSDM** package in the R statistical environment (R Core Team, 2014). Examples use virtual or real data-sets. Results obtained with functions in the **hSDM** package are compared with the results obtained with other softwares and models.

2 Species distribution models

2.1 Binomial model

2.1.1 Mathematical formulation

Let's consider a random variable y_i representing the total number of presences of a species after several visits v_i at a particular site i. Random variable y_i can take values from 0 to v_i and can be assumed to follow a Binomial distribution having parameters v_i and θ_i (Eq. 1). Parameter θ_i can be interpreted as the probability of presence of the species at site i . Using a logit link function, θ_i can be expressed as a linear model combining explicative variables X_i and parameters β (Eq. 1).

$$y_i \sim \mathcal{B}inomial(v_i, \theta_i)$$
$$\text{logit}(\theta_i) = X_i\beta \tag{1}$$

Using this statistical model, we aim at representing a "suitability process". Given environmental variables X_i, how much is habitat at site i suitable for the species under consideration? Parameters β indicate how much each environmental variable contributes to the suitability process. Like every other function in the **hSDM** R package, function `hSDM.binomial()` estimates the parameters β of such a model in a Bayesian framework. Parameter inference is done using a Gibbs sampler including a Metropolis algorithm. The Gibbs sampler is coded in the C language to optimize computation efficiency.

2.1.2 Data generation

To explore the characteristics of the `hSDM.binomial()` function, we can generate a virtual data-set on the basis of the Binomial model described above (Eq. 1). In the most general case (presence/absence data or presence/background data), a site i is visited once and $v_i = 1$. Thus, the random variable y_i follows a Bernoulli distribution having parameters θ_i and habitat characteristics X_i are fixed for site i. We will generate a virtual data-set in this particular case. For data generation, we can import virtual altitudinal data in R. Altitude will be used as an explicative variable to determine habitat suitability, i.e. the probability of presence of a virtual species. Altitudinal data are loaded at the same time as the **hSDM** R package (data frame `altitude` in the working directory).

These data can be transformed into a raster object using the function `rasterFromXYZ()` from the **raster** package. The raster has 2500 cells (50 columns and 50 rows) and the altitude ranges roughly between 100 and 600 m (Fig. 1). For linear models, explicative variables are usually centered and scaled to facilitate inference and interpretation of model parameters.

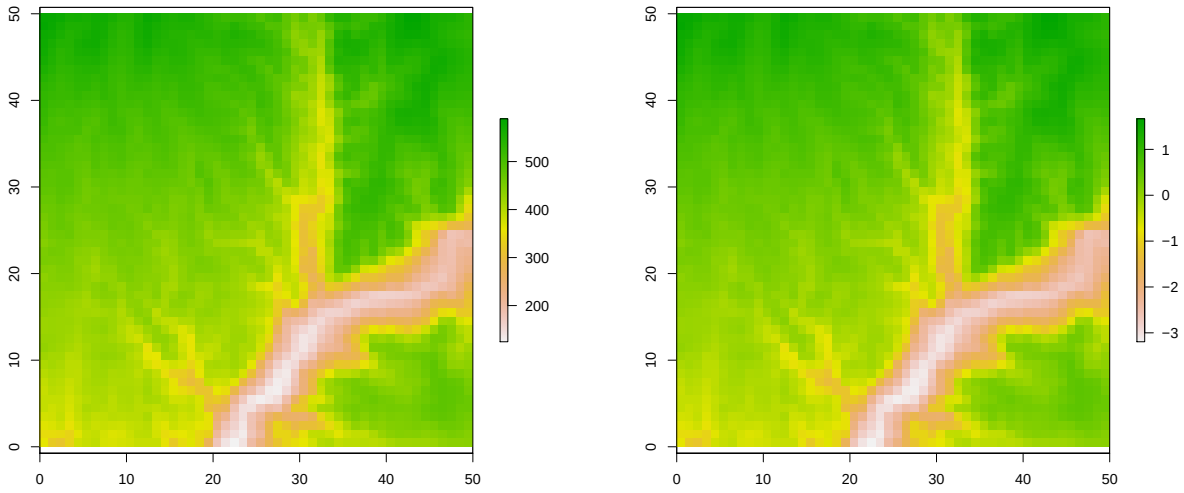

Figure 1: **Altitudinal data**. Original values (in m) on the left. Centered and scaled values on the right.

```
# Load altitudinal data and create raster
library(raster)
data(altitude,package="hSDM")
alt.orig <- rasterFromXYZ(altitude)
extent(alt.orig) <- c(0,50,0,50)
plot(alt.orig)
# Center and scale altitudinal data
alt <- scale(alt.orig,center=TRUE,scale=TRUE)
plot(alt)
```

A linear model including altitude (variable denoted A) is used to compute the probability of presence of the species (Eq. 2).

$$
\begin{aligned}
y_i &\sim \mathcal{B}ernoulli(\theta_i) \\
\text{logit}(\theta_i) &= \beta_0 + \beta_1 A_i
\end{aligned}
\tag{2}
$$

We fix the parameters to $\beta_0 = -1$ and $\beta_1 = 1$. The species has a higher probability of presence at higher altitudes (Fig. 2).

```
# Load hSDM library
library(hSDM)
# Target parameters
```

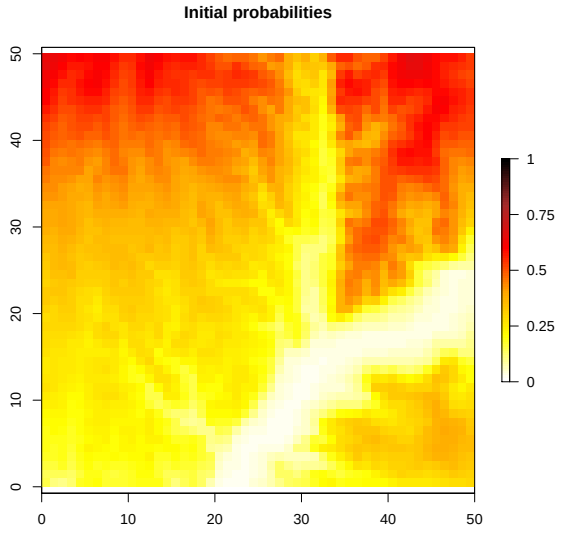

Figure 2: **Probability of presence**.

```
beta.target <- matrix(c(-1,1),ncol=1)
# Matrix of covariates (including the intercept)
ncells <- ncell(alt)
X <- cbind(rep(1,ncells),values(alt))
# Probability of presence as a quadratic function of altitude
logit.theta <- X %*% beta.target
theta <- inv.logit(logit.theta)
# Coordinates of raster cells
coords <- coordinates(alt)
# Transform the probability of presence into a raster
theta <- rasterFromXYZ(cbind(coords,theta))
# Color palette for probability plots
colRP <- colorRampPalette(c("white","yellow","orange",
                            "red","brown","black"))
# Plot the probability of presence
brks <- seq(0,1,length.out=100)
arg <- list(at=seq(0,1,length.out=5), labels=c("0","0.25","0.5","0.75","1"))
nb <- length(brks)-1
plot(theta,main="Initial probabilities",col=colRP(nb),
     breaks=brks,axis.args=arg,zlim=c(0,1))
```

We can assume a number n of sites in the landscape where we have been able to observe or not the presence of the species. We can simulate the presence or absence of the species at these n sites given our model (Fig. 3).

```
# Number of observation sites
nsite <- 200
# Set seed for repeatability
seed <- 1234
# Sample the observations in the landscape
set.seed(seed)
x.coord <- runif(nsite,0,50)
set.seed(2*seed)
y.coord <- runif(nsite,0,50)
library(sp)
sites.sp <- SpatialPoints(coords=cbind(x.coord,y.coord))
# Extract altitude data for sites
alt.sites <- extract(alt,sites.sp)
# Compute theta for these observations
X.sites <- cbind(rep(1,nsite),alt.sites)
logit.theta.site <- X.sites %*% beta.target
theta.site <- inv.logit(logit.theta.site)
# Simulate observations
visits <- rep(1,nsite) # One visit per site for the moment
set.seed(seed)
Y <- rbinom(nsite,visits,theta.site)
# Group explicative and response variables in a data-frame
data.obs.df <- data.frame(Y,visits,alt=X.sites[,2])
# Transform observations in a spatial object
data.obs <- SpatialPointsDataFrame(coords=coordinates(sites.sp),
                                   data=data.obs.df)
# Plot observations
plot(alt.orig)
points(data.obs[data.obs$Y==1,],pch=16)
points(data.obs[data.obs$Y==0,],pch=1)
```

2.1.3 Parameter inference using the `hSDM.binomial()` function

The `hSDM.binomial()` function performs a Binomial logistic regression in a Bayesian framework. Before using this function we need to prepare a bit the data for predictions. We want to have predictions on the whole landscape, not only at observation points. To directly obtain these predictions, we can create a data frame including altitudinal data on the whole landscape. This data frame will be used for the `suitability.pred` argument. The data frame for predictions must include the same column names as those used in the formula for the `suitability` argument (i.e. "alt" our example).

```
data.pred <- data.frame(alt=values(alt))
```

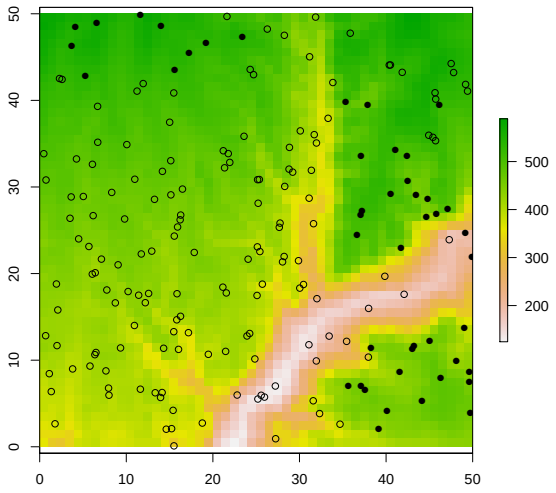

Figure 3: **Observation points**. Presences (full circles) and absences (empty circles) are localized on the altitude map (in m).

We can now call the `hSDM.binomial()` function. Setting parameter `save.p` to 1, we can save in memory the MCMC values for predictions. These values can be used to compute several statistics for each predictions (mean, median, 95% quantiles). For example, mean and 95% quantiles are useful to estimate the uncertainty around the mean predictions.

```
mod.hSDM.binomial <- hSDM.binomial(presences=data.obs$Y,
                                   trials=data.obs$visits,
                                   suitability=~alt,
                                   data=data.obs,
                                   suitability.pred=data.pred,
                                   burnin=1000, mcmc=1000, thin=1,
                                   beta.start=0,
                                   mubeta=0, Vbeta=1.0E6,
                                   seed=1234, verbose=1, save.p=1)
```

2.1.4 Analysis of the results

The `hSDM.binomial()` function returns an MCMC (Markov chain Monte Carlo) for each parameter of the model and also for the model deviance. To obtain parameter estimates, MCMC values can be summarized through a call to the `summary()` function from the **coda** package. We can check that the values of the target parameters, $\beta_0 = -1$ and $\beta_1 = 1$, are within the 95% confidence interval of the parameter estimates.

```
summary(mod.hSDM.binomial$mcmc)

##
## Iterations = 1001:2000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 1000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                     Mean    SD Naive SE Time-series SE
## beta.(Intercept)  -1.413 0.226  0.00713         0.0223
## beta.alt           0.984 0.296  0.00936         0.0328
## Deviance         202.166 2.285  0.07225         0.1661
##
## 2. Quantiles for each variable:
##
##                     2.5%     25%     50%    75%   97.5%
## beta.(Intercept)  -1.843  -1.557  -1.413  -1.27  -0.958
## beta.alt           0.451   0.783   0.969   1.18   1.681
## Deviance         199.895 200.490 201.328 203.19 207.664
```

Parameters estimates can be compared to results obtained with the `glm()` function.

```
#== glm results for comparison
mod.glm <- glm(cbind(Y,visits-Y)~alt,family="binomial",data=data.obs)
summary(mod.glm)

##
## Call:
## glm(formula = cbind(Y, visits - Y) ~ alt, family = "binomial",
##     data = data.obs)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.129  -0.751  -0.604  -0.175   2.728
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.382      0.197   -7.03    2e-12 ***
## alt            0.952      0.276    3.44  0.00057 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 215.71  on 199  degrees of freedom
## Residual deviance: 199.79  on 198  degrees of freedom
## AIC: 203.8
##
## Number of Fisher Scoring iterations: 5
```

MCMC can also be graphically summarized with a call to the `plot.mcmc()` function, also in the **coda** package. MCMC are plotted with a trace of the sampled output and a density estimate for each variable in the chain (Fig. 4). This plot can be used to visually check that the chains have converged.

```
plot(mod.hSDM.binomial$mcmc)
```

The `hSDM.binomial()` function also returns two other objects. The first one, `theta.latent`, is the predictive posterior mean of the latent variable θ (the probability of presence) for each observation.

```
str(mod.hSDM.binomial$theta.latent)

##  num [1:200] 0.2191 0.0992 0.1038 0.1878 0.221 ...

summary(mod.hSDM.binomial$theta.latent)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0171  0.1540  0.2180  0.2300  0.2970  0.4970
```

The second one, `theta.pred` is the set of sampled values from the predictive posterior (if parameter `save.p` is set to 1) or the predictive posterior mean (if `save.p` is set to 0) for each prediction. In our example, `save.p` is set to 1 and `theta.pred` is an `mcmc` object. Values in `theta.pred` can be used to plot the predicted probability of presence on the whole landscape and the uncertainty associated to predictions (Fig 5).

```
# Create a raster for predictions
theta.pred.mean <- raster(theta)
# Create rasters for uncertainty
theta.pred.2.5 <- theta.pred.97.5 <- raster(theta)
# Attribute predicted values to raster cells
theta.pred.mean[] <- apply(mod.hSDM.binomial$theta.pred,2,mean)
theta.pred.2.5[] <- apply(mod.hSDM.binomial$theta.pred,2,quantile,0.025)
theta.pred.97.5[] <- apply(mod.hSDM.binomial$theta.pred,2,quantile,0.975)
# Plot the predicted probability of presence and uncertainty
plot(theta.pred.mean,main="Mean",col=colRP(nb),breaks=brks,
```

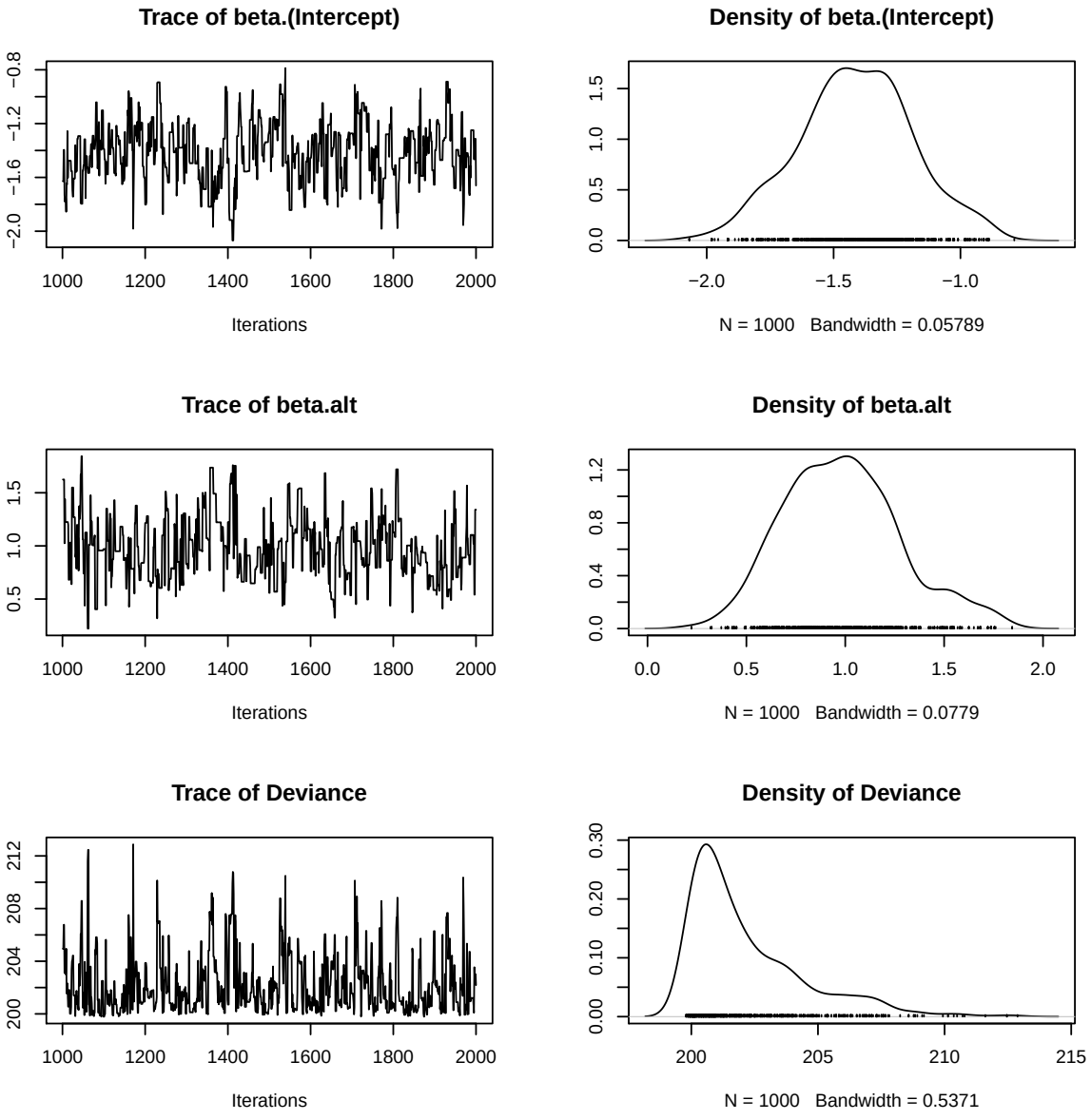

Figure 4: **Trace and density estimate for each variable of the MCMC.**

```
     axis.args=arg,zlim=c(0,1))
plot(theta.pred.2.5,main="Quantile 2.5 %",col=colRP(nb),breaks=brks,
     axis.args=arg,zlim=c(0,1))
plot(theta.pred.97.5,main="Quantile 97.5 %",col=colRP(nb),breaks=brks,
     axis.args=arg,zlim=c(0,1))
```

In our example, we can compare the predictions to the initial probability of presence computed from our model to check that our predictions are correct (Fig. 6).

```
# Comparing predictions to initial values
plot(theta[],theta.pred.mean[],cex.lab=1.4,xlim=c(0,1),ylim=c(0,1))
points(theta[],theta.pred.2.5[],cex.lab=1.4,col=grey(0.5))
points(theta[],theta.pred.97.5[],cex.lab=1.4,col=grey(0.5))
abline(a=0,b=1,col="red",lwd=2)
```

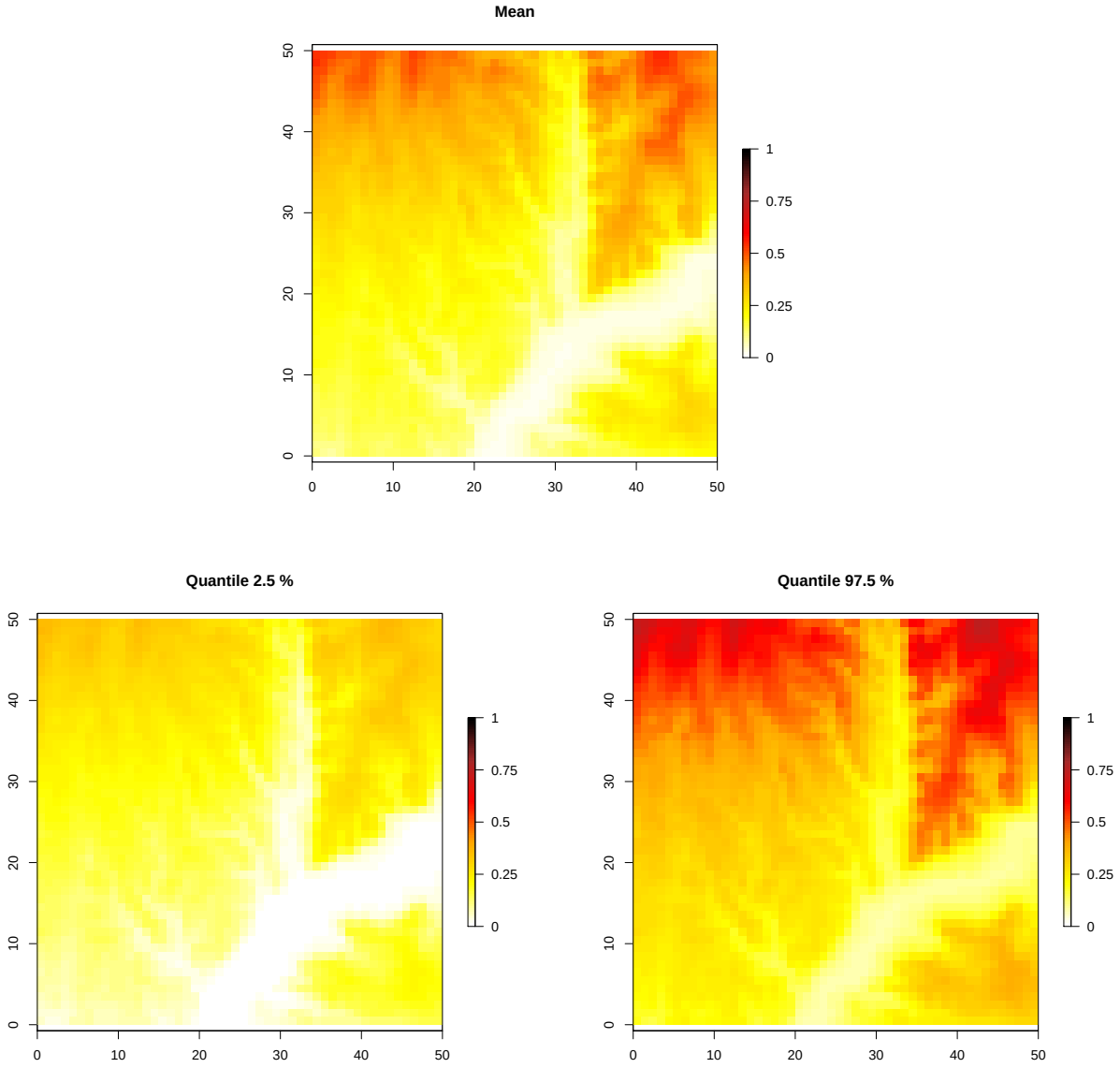

Figure 5: **Predicted probability of presence and uncertainty of predictions**. Mean probability of presence (top), predictions at 2.5% quantile (bottom left) and 97.5% quantile (bottom right) can be plotted from the `mcmc` object `plot.p.pred` returned by function `hSDM.binomial()`.

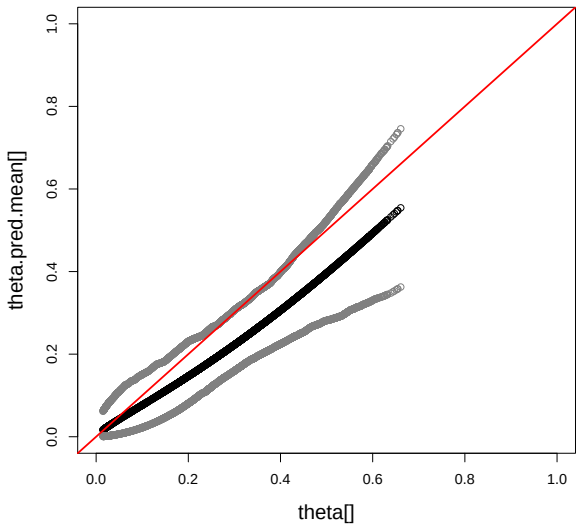

Figure 6: **Predicted vs. initial probabilities of presence**. Initial probabilities of presence are computed from the Binomial logistic regression model with target parameters.

2.2 Site-occupancy model

2.2.1 Mathematical formulation

Let's consider the random variable z_i describing habitat suitability at site i. The random variable z_i can take value 1 or 0 depending on the fact that the habitat is suitable ($z_i = 1$) or not ($z_i = 0$). Habitat at site i is described by environmental variables X_i. Random variable z_i can be assumed to follow a Bernoulli distribution of parameter θ_i (Eq. 3). In this case, θ_i is the probability that the habitat is suitable. Several visits at time t_1, t_2, etc., can occur at site i. Let's consider the random variable y_{it} representing the presence of the species at site i and time t. The species is observed at site i ($\sum_t y_{it} \geq 1$) only if the habitat is suitable ($z_i = 1$). The species is unobserved at site i ($\sum_t y_{it} = 0$) if the habitat is not suitable ($z_i = 0$) or if the habitat is suitable ($z_i = 1$) but the probability δ_{it} of detecting the species at site i and time t is inferior to 1. Thus, y_{it} is assumed to follow a Bernoulli distribution of parameter $z_i\delta_{it}$. Using a logit link function, δ_{it} can be expressed as a linear model combining explicative variables W_{it} and parameters γ (Eq. 3). Typically, explicative variables W_{it} are site characteristics (e.g., habitat variables) or survey characteristics (e.g., weather conditions). The function `hSDM.siteocc()` estimates the parameters β and γ of such a model.

$$
\begin{aligned}
&\textbf{Ecological process:} \\
&z_i \sim \mathcal{B}ernoulli(\theta_i) \\
&\text{logit}(\theta_i) = X_i\beta \\
&\\
&\textbf{Observation process:} \\
&y_{it} \sim \mathcal{B}ernoulli(z_i\delta_{it}) \\
&\text{logit}(\delta_{it}) = W_{it}\gamma
\end{aligned}
\tag{3}
$$

2.2.2 Data generation

To explore the characteristics of the `hSDM.siteocc()` function, we can generate a new virtual data-set on the basis of the site-occupancy model described above (Eq. 3). In the most general case, the observation protocol includes severals visits with varying survey conditions (e.g. weather conditions) to several sites with fixed sites characteristics (e.g. habitat variables). We will generate a virtual data-set following this protocole using the altitudinal data in the previous example for the Binomial model (Sec. 2.1).

We draw at random the number of visits at each site of the previous example (see Fig. 3 of Sec. 2.1).

```
# Number of visits associated to each observation point
set.seed(seed)
visits <- rpois(nsite,lambda=3) # Mean number of visits ~3
# NB: Setting a too low mean number of visits per site (lambda < 3)
# leads to inaccurate parameter estimates
visits[visits==0] <- 1 # Number of visits must be > 0
# Vector of observation sites
sites <- vector()
for (i in 1:nsite) {
    sites <- c(sites,rep(i,visits[i]))
}
```

The survey conditions for each visit are determined by two explicative variables, w_1 and the altitude (variable denoted A). These two variables explain the observability of the species (Eq. 4).

$$
\begin{aligned}
&y_{it} \sim \mathcal{B}ernoulli(z_i\delta_{it}) \\
&\text{logit}(\delta_{it}) = \gamma_0 + \gamma_1 w_{1it} + \gamma_2 A_{it}
\end{aligned}
\tag{4}
$$

We fix the intercept and the effects of these two variables: $\gamma_0 = -1$, $\gamma_1 = 1$ and $\gamma_2 = -1$ for determining the detection probability. In our case, the detection probability decreases with altitude ($\gamma_2 < 0$).

```r
# Explicative variables for observation process
nobs <- sum(visits)
set.seed(seed)
w1 <- rnorm(n=nobs,0,1)
W <- cbind(rep(1,nobs),w1,X.sites[sites,2])
# Target parameters for observation process
gamma.target <- matrix(c(-1,1,-1),ncol=1)
```

Using covariates and parameters for the two processes, we compute the probability that the habitat is suitable (θ_i) and the species detection probability (δ_i). We also draw the random variables z_i and y_i and construct the observation data-set.

```r
# Ecological process (suitability)
logit.theta.site <- X.sites %*% beta.target
theta.site <- inv.logit(logit.theta.site)
set.seed(seed)
Z <- rbinom(nsite,1,theta.site)

# Observation process (detectability)
logit.delta.obs <- W %*% gamma.target
delta.obs <- inv.logit(logit.delta.obs)
set.seed(seed)
Y <- rbinom(nobs,1,delta.obs*Z[sites])

# Data-sets
data.obs <- data.frame(Y,w1,alt=X.sites[sites,2],site=sites)
data.suit <- data.frame(alt=X.sites[,2])
```

2.2.3 Parameter inference using the `hSDM.siteocc()` function

The `hSDM.siteocc()` function estimates the parameter of a site-occupancy model in a Bayesian framework.

```r
mod.hSDM.siteocc <- hSDM.siteocc(# Observations
                                 presence=data.obs$Y,
                                 observability=~w1+alt,
                                 site=data.obs$site,
                                 data.observability=data.obs,
                                 # Habitat
                                 suitability=~alt,
                                 data.suitability=data.suit,
                                 # Predictions
                                 suitability.pred=data.pred,
                                 # Chains
```

```
                                burnin=1000, mcmc=1000, thin=1,
                                # Starting values
                                beta.start=0,
                                gamma.start=0,
                                # Priors
                                mubeta=0, Vbeta=1.0E6,
                                mugamma=0, Vgamma=1.0E6,
                                # Various
                                seed=1234, verbose=1, save.p=1)
```

2.2.4 Analysis of the results

```
summary(mod.hSDM.siteocc$mcmc)

##
## Iterations = 1001:2000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 1000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                      Mean    SD Naive SE Time-series SE
## beta.(Intercept)   -0.826 0.327  0.01035         0.0282
## beta.alt            1.110 0.445  0.01406         0.0341
## gamma.(Intercept)  -1.287 0.221  0.00698         0.0229
## gamma.w1            0.941 0.208  0.00658         0.0205
## gamma.alt          -0.930 0.220  0.00695         0.0200
## Deviance          295.730 3.092  0.09779         0.2765
##
## 2. Quantiles for each variable:
##
##                      2.5%     25%     50%     75%   97.5%
## beta.(Intercept)   -1.500  -1.041  -0.850  -0.586  -0.191
## beta.alt            0.417   0.738   1.031   1.445   2.077
## gamma.(Intercept)  -1.775  -1.438  -1.264  -1.142  -0.864
## gamma.w1            0.517   0.781   0.925   1.066   1.392
## gamma.alt          -1.398  -1.067  -0.924  -0.806  -0.504
## Deviance          291.769 293.442 295.091 297.205 304.031
```

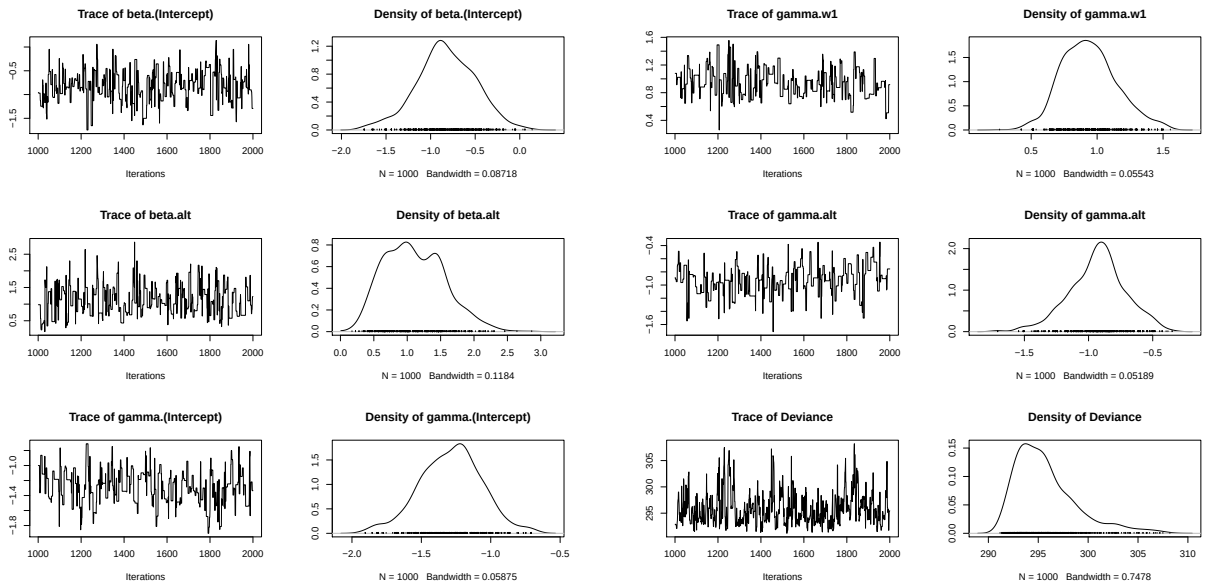

Figure 7: **Trace and density estimate for each variable of the MCMC.**

```
plot(mod.hSDM.siteocc$mcmc)
```

```
# Create a raster for predictions
theta.pred.mean <- raster(theta)
# Computing mean and quantiles for uncertainty
theta.pred.mean[] <- apply(mod.hSDM.siteocc$theta.pred,2,mean)
theta.pred.2.5 <- apply(mod.hSDM.siteocc$theta.pred,2,quantile,0.025)
theta.pred.97.5 <- apply(mod.hSDM.siteocc$theta.pred,2,quantile,0.975)
# Plot the predicted probability of presence
plot(theta.pred.mean,main="hSDM.siteocc",col=colRP(nb),breaks=brks,
     axis.args=arg,zlim=c(0,1))
```

```
# Comparing predictions to initial values
plot(theta[],theta.pred.mean[],xlim=c(0,1),ylim=c(0,1),cex.lab=1.4)
points(theta[],theta.pred.2.5[],cex.lab=1.4,col=grey(0.5))
points(theta[],theta.pred.97.5[],cex.lab=1.4,col=grey(0.5))
abline(a=0,b=1,col="red",lwd=2)
```

Parameters estimates can be compared to results obtained with the `glm()` function assuming a perfect detection.

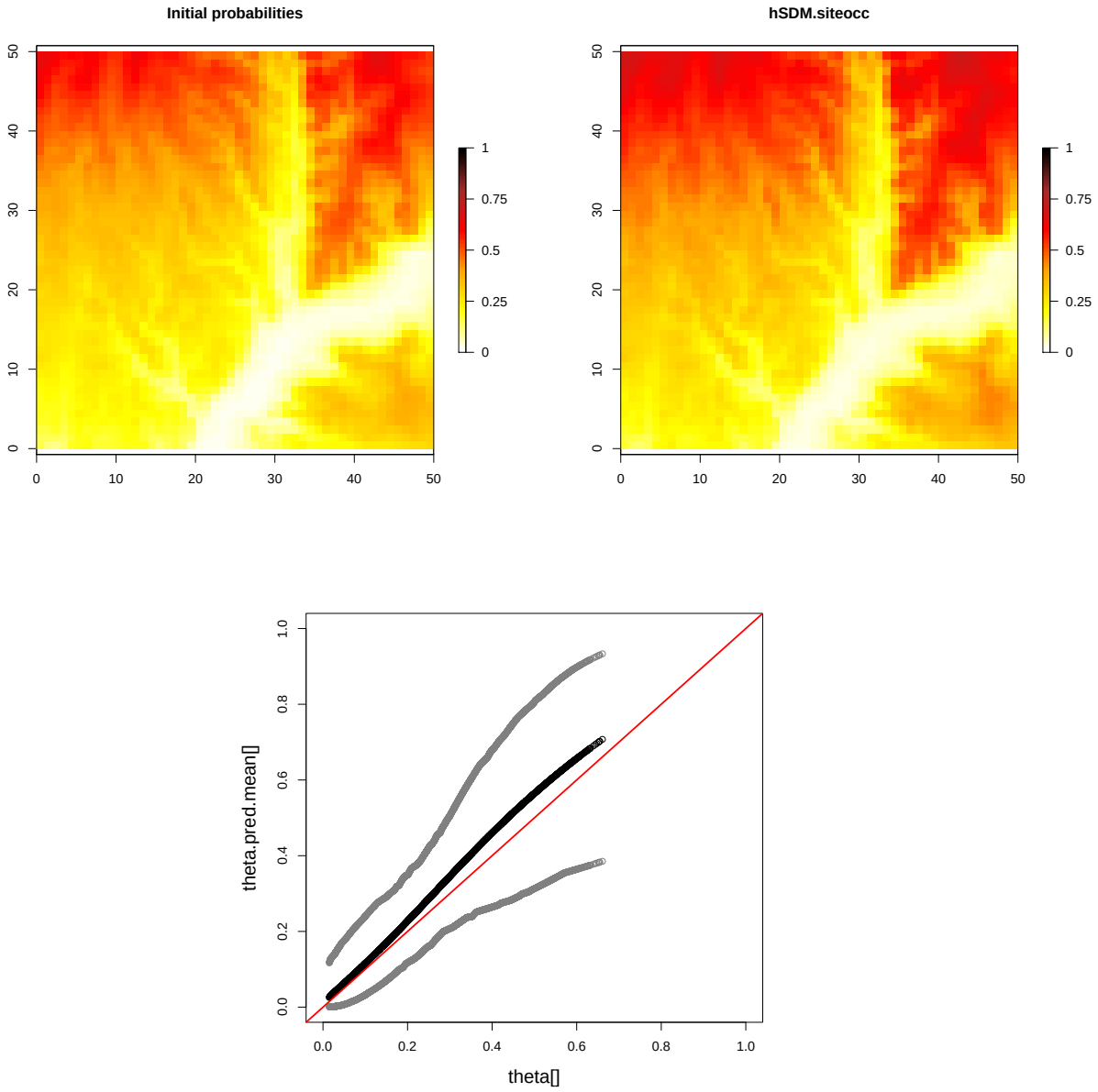

Figure 8: **Comparing predicted probability of presence with initial probabilities.**

```r
#== glm results for comparison
mod.glm <- glm(Y~alt,family="binomial",data=data.obs)
summary(mod.glm)
```

```
##
## Call:
## glm(formula = Y ~ alt, family = "binomial", data = data.obs)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.528  -0.442  -0.427  -0.408   2.265
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.318      0.144  -16.06   <2e-16 ***
## alt           -0.133      0.129   -1.03      0.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 362.10  on 594  degrees of freedom
## Residual deviance: 361.08  on 593  degrees of freedom
## AIC: 365.1
##
## Number of Fisher Scoring iterations: 5
```

```r
# Create a raster for predictions
theta.pred.glm <- raster(theta)
# Attribute predicted values to raster cells
theta.pred.glm[] <- predict.glm(mod.glm,newdata=data.pred,type="response")
# Plot the predicted probability of presence
plot(theta.pred.glm,main="GLM",col=colRP(nb),breaks=brks,
     axis.args=arg,zlim=c(0,1))
```

```r
# Comparing predictions to initial values
plot(theta[],theta.pred.glm[],
     xlim=c(0,1),ylim=c(0,1),cex.lab=1.4)
points(theta[],theta.pred.mean[],col=grey(0.5))
abline(a=0,b=1,col="red",lwd=2)
```

On Figure 14, we can see that using a GLM in the case of imperfect detection can lead to very inaccurate parameter estimates and predictions for the probability of presence of

the species. This is particularly true when detection probability is negatively correlated to presence probability (through an explicative variable such as the altitude in our example). This has been clearly demonstrated in an article by Lahoz-Monfort *et al.* (2014).

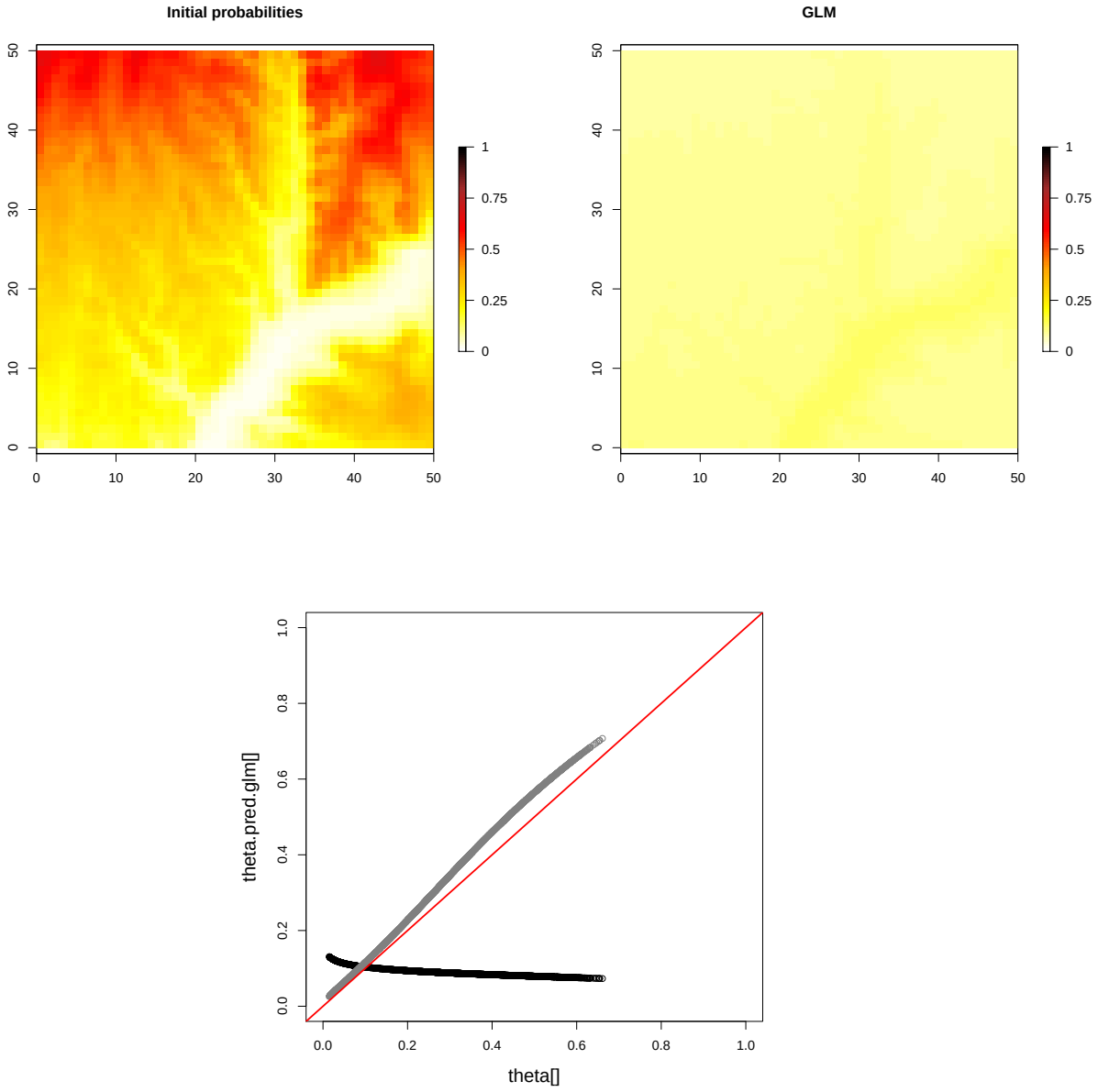

Figure 9: **Comparing predicted probability of presence using GLM with initial probabilities.** Grey dots figure the predictions with the `hSDM.siteocc()` function whereas black dots figure the prediction using the `glm()` function.

2.3 Binomial iCAR model

2.3.1 Mathematical formulation

2.3.2 Data generation with iCAR

```
# Rasters must be projected to correctly compute the neighborhood
crs(alt) <- '+proj=utm +zone=1'
# Neighborhood matrix
neighbors.mat <- adjacent(alt, cells=c(1:ncells), directions=8,
                          pairs=TRUE, sorted=TRUE)
# Number of neighbors by cell
n.neighbors <- as.data.frame(table(as.factor(neighbors.mat[,1])))[,2]
# Adjacent cells
adj <- neighbors.mat[,2]
# Generate symmetric adjacency matrix, A
A <- matrix(0,ncells,ncells)
index.start <- 1
for (i in 1:ncells) {
    index.end <- index.start+n.neighbors[i]-1
    A[i,adj[c(index.start:index.end)]] <- 1
    index.start <- index.end+1
}
```

```
# Function to draw in a multivariate normal
rmvn <- function(n, mu=0, V=matrix(1), seed=1234) {
    p <- length(mu)
    if (any(is.na(match(dim(V), p)))) {
        stop("Dimension problem!")
    }
    D <- chol(V)
    set.seed(seed)
    t(matrix(rnorm(n*p),ncol=p)%*%D+rep(mu,rep(n,p)))
}
```

```
# Generate spatial random effects
Vrho.target <- 1 # Variance of spatial random effects
d <- 1  # Spatial dependence parameter = 1 for intrinsic CAR
Q <- diag(n.neighbors)-d*A + diag(.0001,ncells) # Add small constant to
                                                # make Q non-singular
covrho <- Vrho.target*solve(Q) # Covariance of rhos
rho <- c(rmvn(1,mu=rep(0,ncells),V=covrho,seed=seed)) # Spatial Random Effects
rho <- rho-mean(rho) # Centering rhos on zero
```

```
rho.rast <- rasterFromXYZ(xyz=cbind(coords,rho))
# Probability of presence
theta.cells <- inv.logit(X %*% beta.target + rho)
theta <- rasterFromXYZ(cbind(coords,theta.cells))
```

```
# Ecological process (suitability)
cells <- extract(alt,sites.sp,cell=TRUE)[,1]
set.seed(seed)
logit.theta.site <- X.sites %*% beta.target + rho[cells]
theta.site <- inv.logit(logit.theta.site)
set.seed(seed)
Y <- rbinom(nsite,visits,theta.site)
# Data-sets
data.suit <- data.frame(Y,visits,alt=X.sites[,2],cells)
data.pred <- data.frame(alt=values(alt),cells=c(1:ncells))
# Transform observations into a spatial object
data.suit <- SpatialPointsDataFrame(coords=coordinates(sites.sp),
                                    data=data.suit)
```

```
# Plot spatial random effects
plot(rho.rast,main="Spatial random effects")
# Plot initial probabilities and observations
plot(theta,main="Initial probabilities (iCAR model)",col=colRP(nb),breaks=brks,
     axis.args=arg,zlim=c(0,1))
points(data.suit[data.suit$Y>0,],pch=16)
points(data.suit[data.suit$Y==0,],pch=1)
```

2.3.3 Parameter inference using the `hSDM.binomial.iCAR()` function

```
Start <- Sys.time() # Start the clock
mod.hSDM.binomial.iCAR <- hSDM.binomial.iCAR(presences=data.suit$Y,
                                             trials=data.suit$visits,
                                             suitability=~alt,
                                             spatial.entity=data.suit$cells,
                                             data=data.suit,
                                             n.neighbors=n.neighbors,
                                             neighbors=adj,
                                             suitability.pred=data.pred,
                                             spatial.entity.pred=data.pred$cells,
                                             burnin=5000, mcmc=5000, thin=5,
                                             beta.start=0,
```

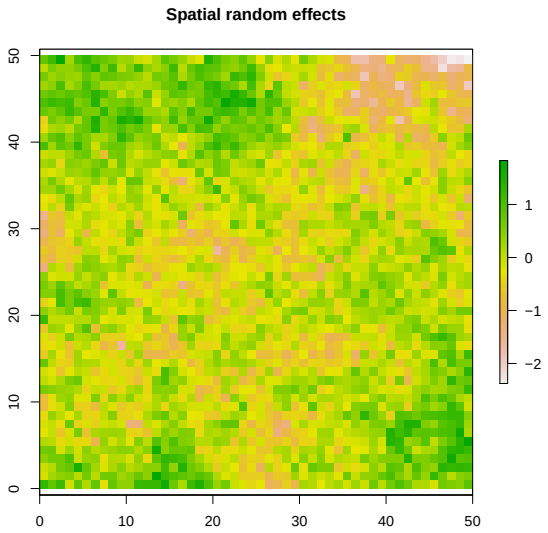

Figure 10: **Spatial random effects.**

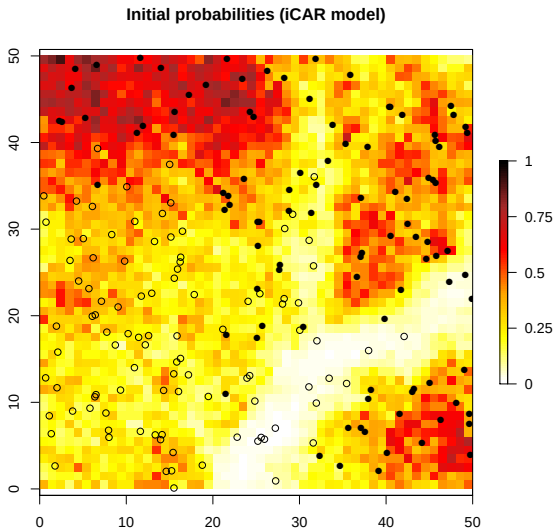

Figure 11: **Initial probability of presence and observations.** Presences (full circles) and absences (empty circles).

```
                                                Vrho.start=1,
                                                priorVrho="1/Gamma",
                                                #priorVrho="Uniform",
                                                #priorVrho=1,
                                                mubeta=0, Vbeta=1.0E6,
                                                shape=0.5, rate=0.0005,
                                                Vrho.max=10,
                                                seed=1234, verbose=1,
                                                save.rho=1, save.p=0)
Time.hSDM <- difftime(Sys.time(),Start,units="sec") # Time difference
```

2.3.4 Analysis of the results with iCAR

```
summary(mod.hSDM.binomial.iCAR$mcmc)

##
## Iterations = 5001:9996
## Thinning interval = 5
## Number of chains = 1
## Sample size per chain = 1000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                     Mean     SD Naive SE Time-series SE
## beta.(Intercept)  -1.42  0.171  0.00542         0.0179
## beta.alt           1.14  0.221  0.00700         0.0318
## Vrho               3.24  0.932  0.02946         0.2401
## Deviance         259.97 10.747  0.33984         0.9438
##
## 2. Quantiles for each variable:
##
##                      2.5%    25%    50%    75%  97.5%
## beta.(Intercept)  -1.754  -1.53  -1.42  -1.30  -1.10
## beta.alt           0.734   1.00   1.14   1.28   1.63
## Vrho               1.930   2.56   3.05   3.78   5.55
## Deviance         241.580 252.20 259.24 267.29 281.26
```

```
# Predictions for spatial random effects
rho.pred <- apply(mod.hSDM.binomial.iCAR$rho.pred,2,mean)
rho.pred.rast <- rasterFromXYZ(cbind(coords,rho.pred))
plot(rho.pred.rast,main="Predictions rho")
```

```r
# Predictions for probability of presence
theta.pred <- mod.hSDM.binomial.iCAR$theta.pred
theta.pred.rast <- rasterFromXYZ(cbind(coords,theta.pred))
plot(theta.pred.rast,main="Predictions theta",col=colRP(nb),breaks=brks,
     axis.args=arg,zlim=c(0,1))
# Predictions vs. initial spatial random effects
plot(rho[-cells],rho.pred[-cells],xlab="rho target",ylab="Predictions rho")
points(rho[cells],rho.pred[cells],col="blue",pch=16)
abline(a=0,b=1,col="red")
# Predictions vs. initial probabilities
plot(values(theta)[-cells],theta.pred[-cells],xlab="theta target",
     ylab="Predictions theta")
points(values(theta)[cells],theta.pred[cells],col="blue",pch=16)
abline(a=0,b=1,col="red")
```

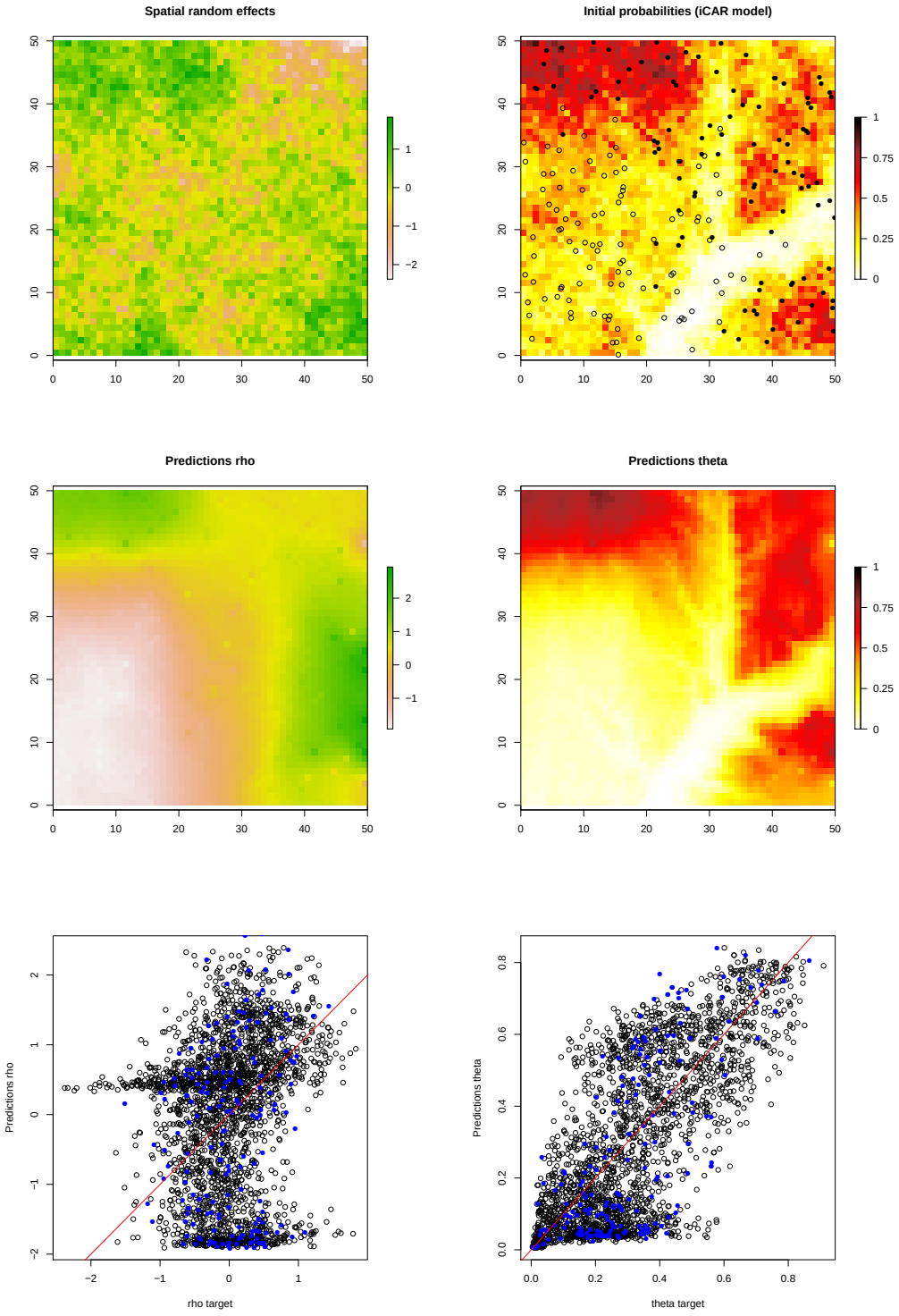

Figure 12: **Predictions vs. initial values**

2.3.5 With OpenBUGS

```
# BUGS model
model.txt <-
"model {

# likelihood
for (n in 1:nobs) {
  y[n] ~ dbin(theta[n], visits[n])
  logit(theta[n]) <- Xbeta[n] + rho[IdCell[n]]
  Xbeta[n] <- beta[1] + beta[2]*x1[n]
}

# CAR prior distribution for spatial random effects:
rho[1:ncells] ~ car.normal(adj[], weights[], num[], tau)
for(k in 1:sumNumNeigh) {
  weights[k] <- 1 # set equal weights for all neighbors
}

# Other priors
for (i in 1:2) {
  beta[i] ~ dnorm(0, 1.0E-6)
}
Vrho <- 1/tau
tau ~ dgamma(0.5,0.0005)

}"

# Create model.txt file in the working directory
system(paste("echo \"",model.txt,"\" > model.txt",sep=""))

# Data for OpenBUGS
y <- data.suit$Y
visits <- data.suit$visits
IdCell <- data.suit$cells
x1 <- data.suit$alt
num <- n.neighbors
adj <- adj
nobs <- length(y)
ncells <- length(n.neighbors)
sumNumNeigh <- length(adj)
data <- list("y","visits","IdCell","x1","num",
             "adj","nobs","ncells","sumNumNeigh")

# Inits
```

Value	OpenBUGS	hSDM
β_0	-1.43	-1.42
β_1	1.19	1.14
V_ρ	3.27	3.24
Deviance	260.26	259.97
Time (secs)	98	7

Table 5: **Comparison between hSDM and OpenBUGS outputs.**

```
inits <- list(list(beta=rep(0,2),rho=rep(0,ncells),tau=1))

# OpenBUGS call
library(R2OpenBUGS)
Start <- Sys.time() # Start the clock
Open <- bugs(data,inits,
             model.file="model.txt",
             parameters=c("beta","Vrho","rho"),
             n.chains=1,
             OpenBUGS.pgm="/usr/local/bin/OpenBUGS",
             n.iter=2000,
             n.burnin=1000,
             n.thin=5,
             DIC=TRUE,
             debug=FALSE,
             clearWD=FALSE)
Time.OpenBUGS <- difftime(Sys.time(),Start,units="sec") # Time difference

# Time difference
ratio.time <- as.numeric(Time.OpenBUGS)/as.numeric(Time.hSDM)
ratio.time # For this example, hSDM is X times faster

#== Outputs
attach.bugs(Open,overwrite=TRUE)
Open$DIC
Open$pD
beta.pred.Open <- apply(Open$sims.list$beta,2,mean)
Vrho.pred.Open <- mean(Open$sims.list$Vrho)
deviance.Open <- mean(Open$sims.list$deviance)
rho.OpenBUGS <- apply(Open$sims.list$rho,2,mean)
plot(rho.pred,rho.OpenBUGS)
abline(a=0,b=1,col="red")
```

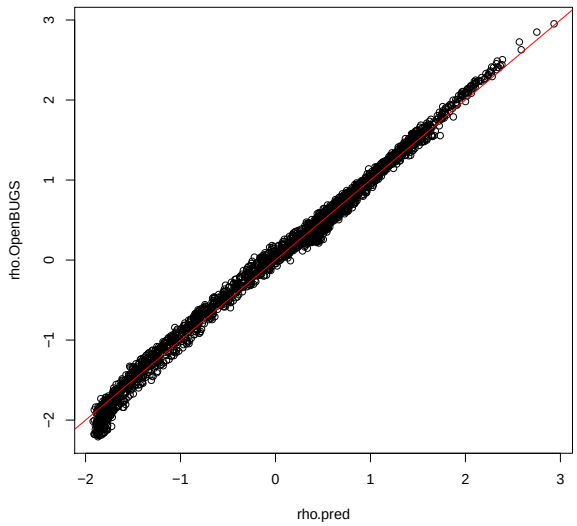

Figure 13: **Comparison between hSDM and OpenBUGS for spatial random effect estimates.**

2.3.6 With a GLM

```
#== glm results for comparison
mod.glm <- glm(cbind(Y,visits)~alt,family="binomial",data=data.suit)
summary(mod.glm)

##
## Call:
## glm(formula = cbind(Y, visits) ~ alt, family = "binomial", data = data.suit)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.738  -0.883  -0.511   0.291   2.257
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.2228     0.0924   -13.2  < 2e-16 ***
## alt           0.6302     0.1126     5.6  2.2e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.80  on 199  degrees of freedom
```

```
## Residual deviance: 127.05  on 198  degrees of freedom
## AIC: 339.9
##
## Number of Fisher Scoring iterations: 4
```

```
# Create a raster for predictions
theta.pred.glm <- raster(theta)
# Attribute predicted values to raster cells
theta.pred.glm[] <- predict.glm(mod.glm,newdata=data.pred,type="response")
# Plot the predicted probability of presence
plot(theta.pred.glm,main="GLM for iCAR",col=colRP(nb),breaks=brks,
     axis.args=arg,zlim=c(0,1))
```

```
# Comparing predictions to initial values
plot(theta[],theta.pred.glm[],
     xlim=c(0,1),ylim=c(0,1),cex.lab=1.4)
abline(a=0,b=1,col="red",lwd=2)
```

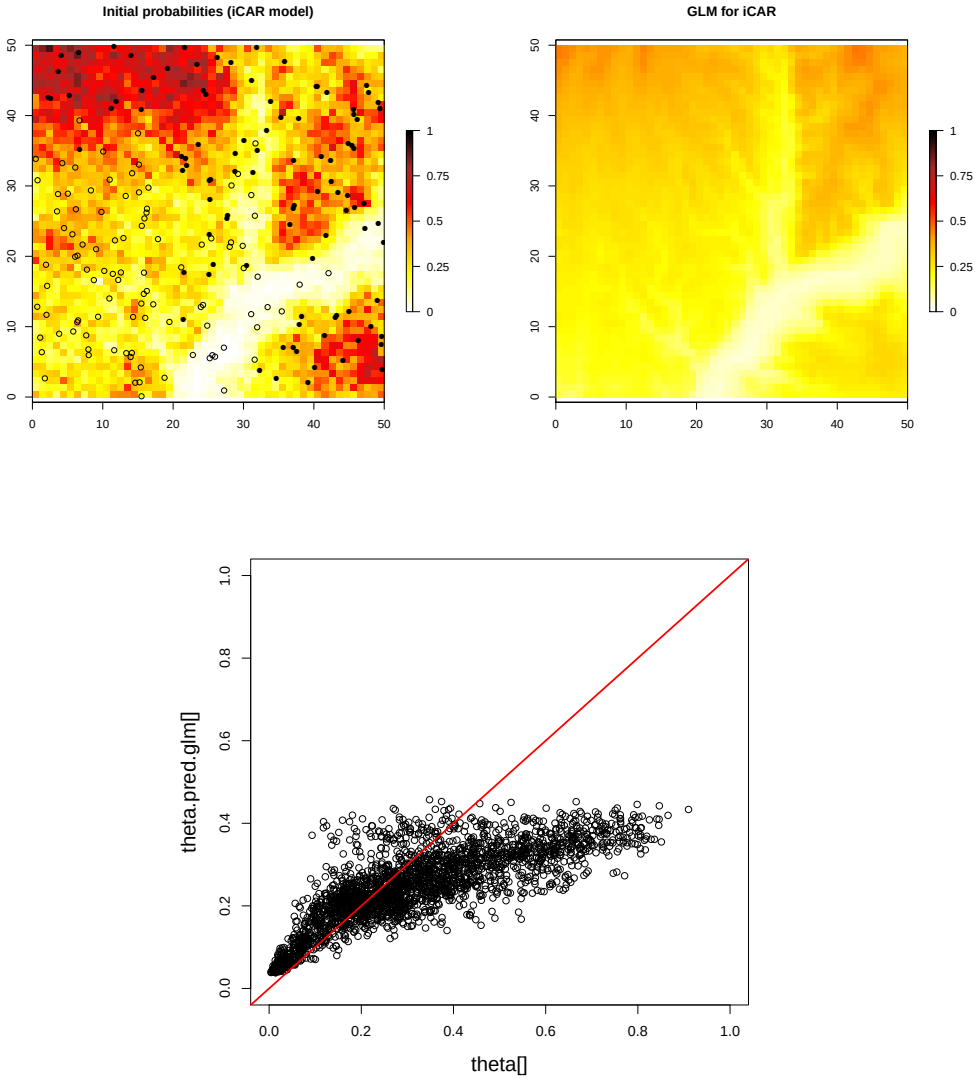

Figure 14: **Comparing predicted probability of presence using GLM with initial probabilities for a binomial model with iCAR process.**

2.4 Site-occupancy iCAR model

3 Acknowledgements

Support was provided by Cirad and FRB (Fondation pour la Recherche sur la Biodiversité) through the BioSceneMada project (project agreement AAP-SCEN-2013 I).

References

Araujo MB, Guisan A (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, **33**, 1677–1688.

Bailey LL, Simons TR, Pollock KH (2004) Estimating site occupancy and species detection probability parameters for terrestrial salamanders. *Ecological Applications*, **14**, 692–702.

Besag J, York J, Mollié A (1991) Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**, 1–20.

Brezger A, Kneib T, Lang S (2005) Bayesx: Analyzing bayesian structural additive regression models. *Journal of Statistical Software*, **14**, 1–22. URL http://www.jstatsoft.org/v14/i11.

Casella G, George EI (1992) Explaining the Gibbs Sampler. *American Statistician*, **46**, 167–174.

Chelgren ND, Adams MJ, Bailey LL, Bury RB (2011) Using multilevel spatial models to understand salamander site occupancy patterns after wildfire. *Ecology*, **92**, 408–421.

Chen G, Kéry M, Plattner M, Ma K, Gardner B (2013) Imperfect detection is the rule rather than the exception in plant distribution studies. *Journal of Ecology*, **101**, 183–191.

Choquet R, Rouan L, Pradel R (2009) Program e-surge: a software application for fitting multievent models. In: *Modeling demographic processes in marked populations*, pp. 845–865. Springer.

Cressie NA, Cassie NA (1993) *Statistics for spatial data*, vol. 900. Wiley New York.

Dorazio RM, Royle JA, Soderstrom B, Glimskar A (2006) Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, **87**, 842–854.

Dormann CF, McPherson JM, Araujo M, *et al.* (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30**, 609–628. URL http://dx.doi.org/10.1111/j.2007.0906-7590.05171.x.

Elith J, Leathwick JR (2009) Species distribution models: Ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.*, **40**, 677–697. URL http://dx.doi.org/10.1146/annurev.ecolsys.110308.120159.

Fiske I, Chandler R (2011) unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, **43**, 1–23. URL http://www.jstatsoft.org/v43/i10/.

Flores O, Rossi V, Mortier F (2009) Autocorrelation offsets zero-inflation in models of tropical saplings density. *Ecological Modelling*, **220**, 1797–1809.

Gelfand AE, Schmidt AM, Wu S, Silander JA, Latimer A, Rebelo AG (2005) Modelling species diversity through species level hierarchical modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 1–20.

Gelfand AE, Smith AFM (1990) Sampling-Based Approaches to Calculating Marginal Densities. *Journal of American Statistical Association*, **85**, 398–409.

Gray TN (2012) Studying large mammals with imperfect detection: Status and habitat preferences of wild cattle and large carnivores in eastern cambodia. *Biotropica*, **44**, 531–536.

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Johnson DS, Conn PB, Hooten MB, Ray JC, Pond BA (2013) Spatial occupancy models for large data sets. *Ecology*, **94**, 801–808.

Keitt TH, Bjørnstad ON, Dixon PM, Citron-Pousty S (2002) Accounting for spatial pattern when modeling organism-environment interactions. *Ecography*, **25**, 616–625.

Kühn I, Bierman SM, Durka W, Klotz S (2006) Relating geographical variation in pollination types to environmental and spatial factors using novel statistical methods. *New Phytologist*, **172**, 127–139.

Kéry M, Gardner B, Monnerat C (2010) Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, **37**, 1851–1862.

Kéry M, Royle JA, Schmid H (2005) Modeling avian abundance from replicated counts using binomial mixture models. *Ecological applications*, **15**, 1450–1461.

Kéry M, Schaub M (2012) *Bayesian population analysis using WinBUGS: a hierarchical perspective*. Academic Press.

Kéry M, Schmidt BR (2008) Imperfect detection and its consequences for monitoring for conservation. *Community Ecology*, **9**, 207–216.

Lahoz-Monfort JJ, Guillera-Arroita G, Wintle BA (2014) Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, **23**, 504–515. doi:10.1111/geb.12138. URL http://dx.doi.org/10.1111/geb.12138.

Latimer AM, Wu SS, Gelfand AE, Silander JA (2006) Building statistical models to analyze species distributions. *Ecological Applications*, **16**, 33–50.

Lee D (2013) Carbayes: An r package for bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, **55**. URL http://www.jstatsoft.org/v55/i13.

Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, 1659–1673.

Lichstein JW, Simons TR, Shriner SA, Franzreb KE (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, **72**, 445–463.

Lunn D, Spiegelhalter D, Thomas A, Best N (2009) The bugs project: Evolution, critique and future directions. *Statistics in medicine*, **28**, 3049–3067.

MacKenzie DI (2006) *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Academic Press.

MacKenzie DI, Nichols JD, Lachman GB, Droege S, Andrew Royle J, Langtimm CA (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

Miller J, Franklin J, Aspinall R (2007) Incorporating spatial dependence in predictive vegetation models. *Ecological Modelling*, **202**, 225–242.

Monk J (2014) How long should we ignore imperfect detection of species in the marine environment when modelling their distribution? *Fish and Fisheries*, **15**, 352–358.

Nichols JD (1992) Capture-recapture models. *BioScience*, pp. 94–102.

Poley LG, Pond BA, Schaefer JA, Brown GS, Ray JC, Johnson DS (2014) Occupancy patterns of large mammals in the far north of ontario under imperfect detection and spatial autocorrelation. *Journal of Biogeography*, **41**, 122–132.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Robert CP, Casella G (2004) *Monte Carlo statistical methods*, vol. 319. Citeseer.

Rota CT, Fletcher RJ, Evans JM, Hutto RL (2011) Does accounting for imperfect detection improve species distribution models? *Ecography*, **34**, 659–670.

Royle JA (2004) N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, **60**, 108–115.

Royle JA, Dorazio RM (2008) *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Academic Press.

Royle JA, Dorazio RM, Link WA (2007) Analysis of multinomial models with unknown index using data augmentation. *Journal of Computational and Graphical Statistics*, **16**, 67–85.

Rue H, Martino S, Chopin N (2009) Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, **71**, 319–392.

Sinclair SJ, White MD, Newell GR (2010) How useful are species distribution models for managing biodiversity under future climates? *Ecology and Society*, **15**, 8.

Smith SI (1868) The geographical distribution of animals. *The American Naturalist*, **2**, pp. 124–131. URL http://www.jstor.org/stable/2447129.

Sokal RR, Oden NL (1978) Spatial autocorrelation in biology: 2. some biological implications and four ap- plications of evolutionary and ecological interest. *Biological Journal of the Linnean Society*, **10**, 229–249.

Stan Development Team (2014) *Stan Modeling Language Users Guide and Reference Manual, Version 2.2*. URL http://mc-stan.org/.

Thuiller W, Guéguen M, Georges D, *et al.* (2014) Are different facets of plant diversity well protected against climate and land cover changes? a test study in the french alps. *Ecography*.

Wallace AR (1876) *The geographical distribution of animals: with a study of the relations of living and extinct faunas as elucidating the past changes of the earth's surface*. Macmillan & Co., London.

White GC, Burnham KP (1999) Program mark: survival estimation from populations of marked animals. *Bird study*, **46**, S120–S139.

Williams BK, Nichols JD, Conroy MJ (2002) *Analysis and management of animal populations: modeling, estimation, and decision making*. Academic Press.