This material has been published in CAB Reviews, issue 7, No. 016, 2012, doi: 10.1079/PAVSNNR20127018, the only accredited archive of the content that has been certified and accepted after peer review.

Potential of GLMM in modelling invasive spread: Supplement B

Modelling invasion probability of giant hogweed (*Heracleum mantegazzianum*) with logistic GLMM

J. Thiele1* and B. Markussen2

Address: [1] Institute of Landscape Ecology, University of Muenster, Robert-Koch-Str. 28, 48149 Muenster, Germany. [2] Department of Basic Sciences and Environment, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark.
*Correspondence: J. Thiele. Email: jan.thiele@uni-muenster.de

Introduction

In this worked example we model the invasion probability of an invasive plant species, giant hogweed (*Heracleum mantegazzianum*), based on field surveys of 20 study areas of 1 km² which were situated in areas of Germany most invaded by this species (see Thiele et al. 2008 for more details). The statistical analyses will be conducted in R (R Development Core Team 2011) using the packages *lme4* (version 0.999375-41; Bates et al. 2011), *gof* (0.7-6; Holst 2011), *glmmML* (0.81-8; Broström and Holmberg 2011), *lattice* (0.19-17; Sarkar 2008), and *MuMIn* (1.0.0; Barton 2011).

Field study and dataset

Data collection

The presence or absence of giant hogweed was recorded for a total of 343 patches of suitable habitat nested in the 20 study areas. However, the patches were not homogenous regarding habitat suitability and history. A patch could comprise both open, herbaceous habitat (optimal habitat) and tree-dominated habitat (suboptimal habitat). Further, an analysis of historical aerial images showed that habitat history, e.g. trajectories from agricultural grassland to fallow land, could vary within patches. Hence, the patches were divided into subpatches of homogenous habitat suitability and history. These subpatches were the entities of the data analysis, i.e. the rows in the data table.

While the division of patches led to a nicely large sample size of 1559 subpatches, we have to consider that observations made in subpatches that are grouped together in one contiguous patch will not be statistically independent. Further, the observations made in one study area may be more similar to each other than to observations made in other study areas. For instance, we could imagine that the level of giant hogweed invasion varied among study areas, e.g. due to longer or shorter residence times, so that patches or subpatches in one study area would have a higher probability of being invaded regardless of the environmental variables that we will use to predict invasion probability. Thus, we have two levels of (potential) spatial dependence: study areas and patches nested within study areas. This means that we should conduct a mixed effects model analysis that includes random effects of study areas and patches.

We will use the presence or absence of giant hogweed in subpatches as the dependent variable. Further, there are 9 potential fixed predictor variables in our dataset (3 categorical, 6 continuous) and, finally, the random effect variables study area and patch (Table 1).

Table 1 Variables used in data analysis.

Variable name	Description
Dependent variable	
Hogweed	Presence-absence of giant hogweed in subpatch
Categorical predictors	
Habitat	Habitat suitability (suboptimal, sub; optimal, opt)
Landuse	Land use of subpatch, either fallow ('fallow') or maintenance mowing, ca. once a year ('mowing')
Terrain	Type of terrain, 'valley', 'slope', 'hilltop' or 'plateau'
Continuous predictors	
Proximity	Proximity index of patch (McGarigal & Marks 1995)
Roaddist	Minimum distance between subpatch edge and the closest road
Riverdist	Minimum distance between subpatch edge and the closest brook or river
Shapei	Shape index of subpatch
Neighbor	Average cover of giant hogweed in adjacent subpatches
Parea	Subpatch area
Random variables	
Starea	Study area (n = 20)
Patch	Main habitat patch (n = 343), often comprising several subpatches

Checking the dataset

Before we start with the analysis, we load the data table into the R workspace and take a look at its structure

```
> str(Heracleum)

'data.frame':   1559 obs. of  12 variables:
 $ starea   : Factor w/ 20 levels "att","aus","bre",..: 11 11 11 11 11
11
 $ patch    : Factor w/ 343 levels "attp1","attp10",..: 184 177 179
193 193
 $ hogweed  : int  0 0 1 1 0 0 1 0 0 1 ...
 $ habitat  : Factor w/ 2 levels "opt","sub": 2 2 1 1 1 1 1 2 1 1 ...
 $ landuse  : Factor w/ 2 levels "fallow","mowing": 1 1 1 1 1 1 1 1 1
1 ...
 $ terrain  : Factor w/ 4 levels "hilltop","plateau",..: 3 3 4 4 4 4 3
3 3
 $ proximity: int  24 0 2 175 175 175 175 24 586 586 ...
 $ roaddist : int  44 207 86 92 173 181 173 105 138 2 ...
 $ riverdist: int  9 186 6 17 12 10 12 33 28 18 ...
 $ shapei   : num  1.44 1.48 1.51 1.56 1.19 1.33 1.13 1.26 1.21 1.26
...
 $ neighbor : num  0 0 0.73 0.8 0.67 0.21 1 0 0.2 0.06 ...
 $ parea    : int  656 3427 5623 2678 1699 3115 1244 920 93 1540 ...
```

The variable `hogweed` is coded as 0 for absence and 1 for presence of giant hogweed. The categorical predictor variables `habitat` and `landuse` have two levels,

while `terrain` has four. Some of the continuous predictor variables are marked as integers ('int') because their values were rounded to zero decimal places.

Categorical predictors

We arrange the sequences of levels of the categorical variables so that the first level, which will not be included in the model as a dummy variable, would make a reasonable baseline that facilitates interpretation. Here, we make those levels the baseline which might be associated with lower invasion probability (suboptimal habitat, landuse mowing) or which are 'endpoints of a gradient' (hilltop in comparison to plateau, slope and valley).

```
> Heracleum$habitat<- factor(Heracleum$habitat, levels=c("sub", "opt")
)
> Heracleum$landuse<- factor(Heracleum$landuse, levels=c("mowing",
"fallow") )
> Heracleum$terrain<- factor(Heracleum$terrain, levels=c("hilltop",
"plateau", "slope", "valley") )
```

Continuous predictors

The continuous variables are non-negative and have a right-skew distribution (Fig. 1).

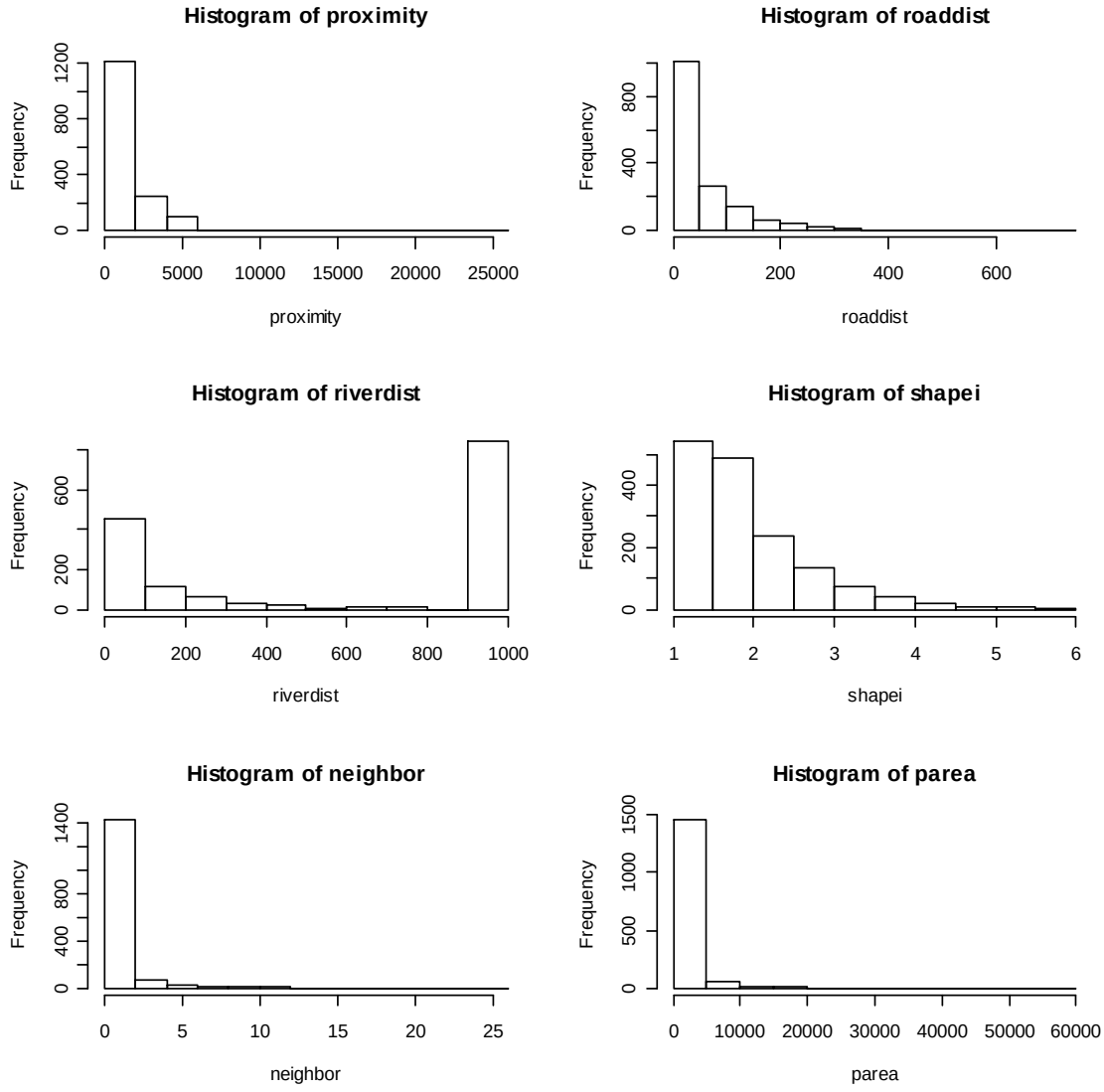


Figure 1 Histograms of fixed continuous predictor variables used in GLMM analysis.

The distribution of the variable `riverdist` has got two peaks, one at zero and one at 1000. A value of zero means that the patch is adjacent to a river or brook, while 1000 was assigned to patches in study areas without any rivers or brooks.

In principle, skewed and bimodal distributions of predictor variables are not problematic for linear modelling. But sometimes linearity and convergence of the model may be improved by transformations that make the distributions of the predictor variables more symmetric and that remove bimodalities. This turns out to be the case here. If we do an analysis with the original predictors, then a z-transformation is needed to make the GLMM converge. However, a plot of the cumulative residuals against `neighbor` indicates that this variable presumably should be log-transformed to be used in the linear model. In order to simplify the following interpretation we prefer to log-transform all the continuous predictor variables.

The predictor variables are non-negative, but since the variables `proximity`, `roaddist`, `riverdist` and `neighbor` contain many zeros, the log-transformation can't be applied directly on these variables. A solution to this problem is to add a positive number to the variables before taking the logarithm. We will use this for `proximity`, `roaddist` and `riverdist`, i.e. we will use the variables `log(1+proximity)` etc. in the analysis. Since the choice of the positive constant, here 1, is arbitrary, this solution is somewhat ad hoc.

A zero for the variable `neighbor` means that giant hogweed hasn't been found in any of the adjacent subpatches, so it makes sense to give these observations special attention. One way to do this is to use a separate parameter for the observations with `neighbor`=0 and to use a linear slope against `log(neighbor)` for the observations with `neighbor`>0. Mathematically this may be done using two variables `zero.neighbor` and `log.neighbor` defined such that `zero.neighbor`=0 and `log.neighbor`=`log(neighbor)` if `neighbor`>0, and `zero.neighbor`=1 and `log.neighbor`=0 if `neighbor`=0. In this way the slope against `zero.neighbor` quantifies the effect when there is no giant hogweed in the adjacent patches, and the slope against `log.neighbor` quantifies the effect when there is giant hogweed in the adjacent patches. The same technique may be used to take care of the value '1000' for the variable `riverdist`. This value isn't a distance, but means that there are no rivers or brooks in the study area. Thus, we define the following variables in R,

```
> attach(Heracleum)
> log.proximity   <- log(1+proximity)
> log.roaddist    <- log(1+roaddist)
> log.riverdist   <- log(1+riverdist)*(riverdist<1000)
> large.riverdist <- as.numeric(riverdist==1000)
> log.shapei      <- log(shapei)
> zero.neighbor   <- as.numeric(neighbor==0)
> log.neighbor    <- log((neighbor==0)+neighbor)
> log.parea       <- log(parea)
> detach(Heracleum)
```

Pre-analyses

Next we would like to get a first impression of relationships between the predictor variables and `hogweed` using box-and-whisker plots, in case of continuous variables, or contingency tables, in case of categorical predictors. For instance, let us look at the relationship between presence-absence of giant hogweed and the distance from the closest road by drawing a `bwplot`

```
> library(lattice)
```

```
> bwplot(hogweed ~ log10(1+roaddist)|starea, Heracleum, scales=
list(y=list(labels=c("absence", "presence"))) )
```

which is shown in Fig. 2. We have plotted one panel for each study area to get an impression of variation of this relationship among study areas. Invaded subpatches seem to have a tendency to be closer to roads than uninvaded ones, but the opposite pattern can be observed in a few study areas (e.g. 'eng').

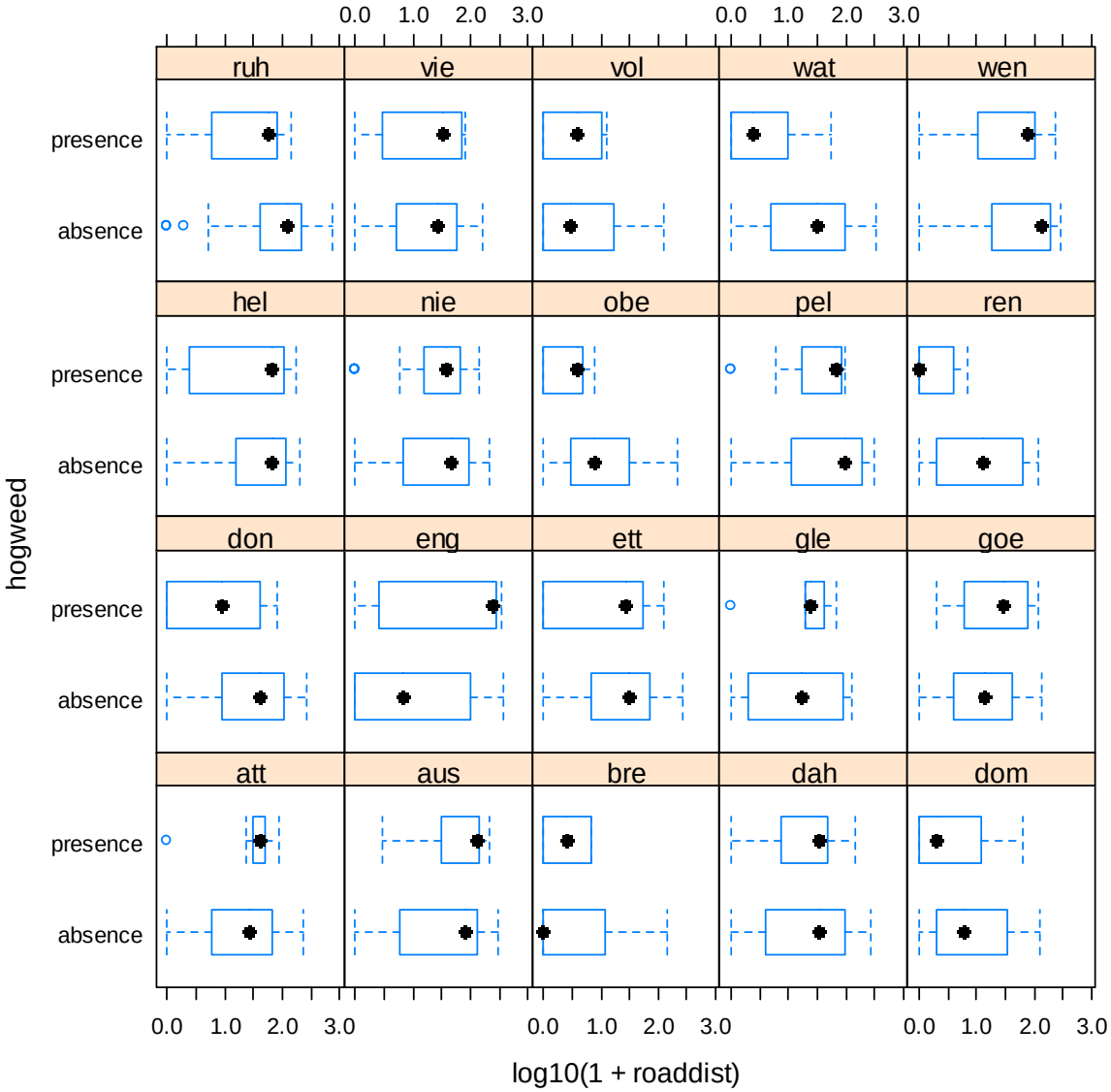


Figure 2 Lattice box-and-whisker plot of giant hogweed presence-absence against distance of subpatches from the closest road. Note that distances were log10-transformed.

To investigate the relationship of habitat (or other categorical predictors) with hogweed presence-absence we can cross-tabulate the two variables

```
> with(Heracleum, table(hogweed, habitat) )
       habitat
hogweed sub opt
      0 915 311
      1 194 139
```

and see that optimal habitats are invaded at a higher rate than suboptimal ones. A chi-square test

```
> chisq.test(with(Heracleum, table(hogweed, habitat) ) )

        Pearson's Chi-squared test with Yates' continuity correction

data:  with(Heracleum, table(hogweed, habitat))
X-squared = 33.4038, df = 1, p-value = 7.488e-09
```

indicates that this simple relationship is significant.

Choosing the model setup

As our dependent variable is binary, we will use the binomial distribution for modelling. We will first use the logit link, but we will also try other link functions.

Now we need to find a suitable method for estimating the model. Penalized Quasi-Likelihood (PQL) is fast, but not suitable for binary data. So we have to choose a somewhat more robust technique, at the cost of longer computation time, and decide to use Laplace approximation which is the fasted valid estimation method here.

The significance of fixed effects can be tested with Wald chi² or with likelihood ratio (LR) chi² tests. Random effects should be tested with LR tests (although significance testing of random effects is not our main goal here).

Model building

Strategy

As we have many predictor variables, we would like to see if we really need all of them for modelling hogweed invasion. A more parsimonious subset of variables would be desirable regarding both computation and interpretation of the model. Thus, we need to decide on a criterion for comparing different models. For our dataset – binary data (i.e. no over-/underdispersion), and large sample size – Akaike's Information Criterion (AIC) is a good measure for comparing models that are fit to the same dataset. Our strategy for finding the final model will be 'best subset' judged by AIC.

Maximum model

Finding the best model is a challenge with this dataset because it is not feasible to fit a full model that contains all fixed effects, interactions of fixed effects, random intercepts and random slopes. The algorithm would not converge. Thus, we need to define a 'maximum model' which is a bit slimmer than the complete model, as starting point of the model building process.

We decide to use a fairly simple structure of random effects: (1|starea/ patch), i.e. a random intercept for study areas and a random intercept for patches nested in study areas. Regarding fixed effects, we first include all main effects into the model and then add interactions of fixed effects, one at a time, and assess their significance using Wald tests reported in the 'summary' table provided by `lmer`. In this way, we find a maximum model that contains the 11 main effects (incl. terms modelling the '1000s' in `riverdist` and the zeros in `neighbor`), 5 interactions of the fixed predictors, and the random-effect structure mentioned above.

We calculate the maximum model

```
> library(lme4)
```

```
> Her.max<- lmer(hogweed ~ habitat + landuse + terrain + log.proximity
+ log.roaddist + large.riverdist + log.riverdist + log.shapei +
zero.neighbor + log.neighbor + log.parea + habitat:terrain +
landuse:terrain + log.roaddist:terrain + log.shapei:landuse +
zero.neighbor:landuse + (1|starea/patch), Heracleum, family=binomial )
```

and get the following results:

```
> print(Her.max)

Generalized linear mixed model fit by the Laplace approximation
Formula: hogweed ~ habitat + landuse + terrain + log.proximity +
log.roaddist +      large.riverdist + log.riverdist + log.shapei +
+zero.neighbor +      log.neighbor + log.parea + habitat:terrain +
landuse:terrain +      log.roaddist:terrain + log.shapei:landuse +
zero.neighbor:landuse +      (1 | starea/patch)
   Data: Heracleum
 AIC  BIC logLik deviance
 904 1049   -425      850
Random effects:
 Groups       Name        Variance Std.Dev.
 patch:starea (Intercept) 0.016520 0.12853
 starea       (Intercept) 0.060452 0.24587
Number of obs: 1559, groups: patch:starea, 343; starea, 20

Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  0.85739    1.13875   0.753 0.451495
habitatopt                   0.19117    0.56539   0.338 0.735271
landusefallow               -0.63136    0.82697  -0.763 0.445185
terrainplateau              -2.24366    1.39434  -1.609 0.107589
terrainslope                -2.14160    0.80846  -2.649 0.008073 **
terrainvalley               -2.27919    0.80422  -2.834 0.004596 **
log.proximity                0.10050    0.05936   1.693 0.090424 .
log.roaddist                -0.80990    0.23455  -3.453 0.000554 ***
large.riverdist             -2.55479    0.48845  -5.230 1.69e-07 ***
log.riverdist               -0.46783    0.10893  -4.295 1.75e-05 ***
log.shapei                  -0.05208    0.62560  -0.083 0.933656
zero.neighbor               -3.12774    0.53830  -5.810 6.23e-09 ***
log.neighbor                 0.56928    0.06057   9.399  < 2e-16 ***
log.parea                    0.38186    0.08281   4.611 4.00e-06 ***
habitatopt:terrainplateau   -0.18805    0.81440  -0.231 0.817388
habitatopt:terrainslope      1.76891    0.69117   2.559 0.010488 *
habitatopt:terrainvalley     1.03474    0.63225   1.637 0.101716
landusefallow:terrainplateau 1.91356    1.39040   1.376 0.168739
landusefallow:terrainslope   0.31916    0.77037   0.414 0.678659
landusefallow:terrainvalley  1.07691    0.72000   1.496 0.134731
terrainplateau:log.roaddist  0.61624    0.29001   2.125 0.033598 *
terrainslope:log.roaddist    0.60987    0.26005   2.345 0.019015 *
terrainvalley:log.roaddist   0.63277    0.24391   2.594 0.009479 **
landusefallow:log.shapei     0.61963    0.68751   0.901 0.367446
landusefallow:zero.neighbor -1.20376    0.62371  -1.930 0.053606 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

[correlation matrix of fixed effects omitted]

The logit link is a common standard for binary models, but other link functions might be more suitable for our dataset. So we calculate the maximum model with probit and complementary log-log link and check the residual deviance:

```
> Her.max.probit<- update(Her.max, family=binomial(link = "probit") )
> Her.max.cloglog<- update(Her.max, family=binomial(link = "cloglog")
)
```

We find that the residual deviance is 849 (AIC 903) with probit and 862 (AIC 916) with complementary log-log-link as compared to 850 (AIC 904) with logit link. Thus, the probit model performs slightly better than the logit, but as the difference is small we will keep the logit link for further analysis.

Diagnostics of the maximum model

We check linearity of the relationship between hogweed and fixed predictor variables using a GLM that includes all fixed effects of the maximum model and the `cumres` function of the `gof` package which calculates cumulative residuals ordered after values of the predictors:

```
> Her.glm <- glm(hogweed ~ habitat + landuse + terrain + log.proximity
+ log.roaddist + large.riverdist + log.riverdist + log.shapei +
zero.neighbor + log.neighbor + log.parea + habitat:terrain +
landuse:terrain + log.roaddist:terrain + log.shapei:landuse +
zero.neighbor:landuse, Heracleum, family=binomial )
>
> library(gof)
> g0 <- cumres(Her.glm)

> x11(); par(mfrow=c(2,2)); plot(g0,idx=1:4)
[..]
```

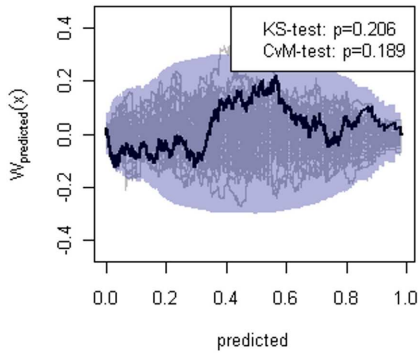
The resulting Fig. 3 shows that the observed cumulative residuals in general are within the typical range of the theoretical cumulative residuals, i.e. that the model is valid. The only exceptions are the parameters log.riverdist and log.neighbor, where the Cramer-von-Mises goodness-of-fit tests indicate some problems. Comparing the cumulative residuals with Figure 2a and 2c in Lin et al. (2002), we see that a possible solution might be to apply an additional log-transformation on `log.riverdist` and to add a cubic term in `log.neighbor`. But since the interpretation of such terms is difficult, we will not extend the model any further.

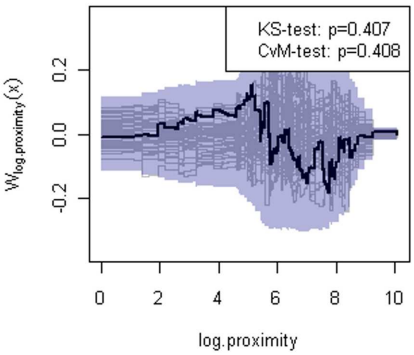
Significance tests of random effects

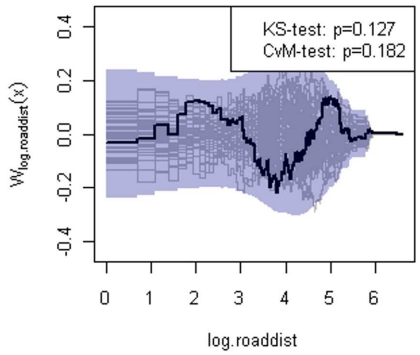
Next, we will test the significance of the random effects. This is not the most important issue here, because the random effects are nuisance variables that we include into the model to account for spatial independence in order to get valid estimates and p-values for the fixed effects. However, in case the random effects were far from being significant, we might consider dropping them from the analysis and conducting a GLM.

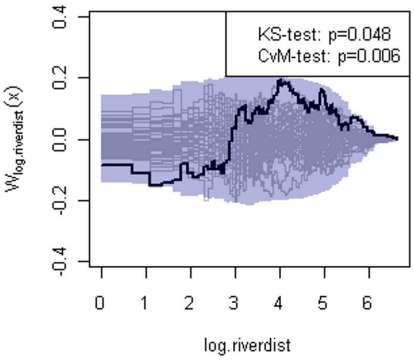
We use a LR test to assess the significance of the random intercept of patch nested in study area. That is, we compare a reduced model without random intercept for patch

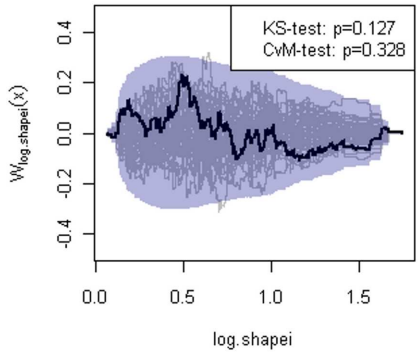
```
> Her.red.patch<- lmer(hogweed ~ habitat + landuse + terrain +
log.proximity + log.roaddist + large.riverdist + log.riverdist +
log.shapei + zero.neighbor + log.neighbor + log.parea +
habitat:terrain + landuse:terrain + log.roaddist:terrain +
log.shapei:landuse + zero.neighbor:landuse + (1|starea), Heracleum,
family=binomial )
```

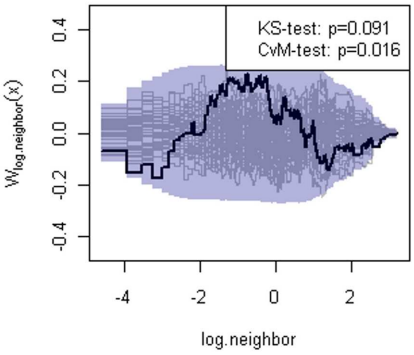

KS-test: p=0.206
CvM-test: p=0.189
$W_{predicted}(x)$
predicted

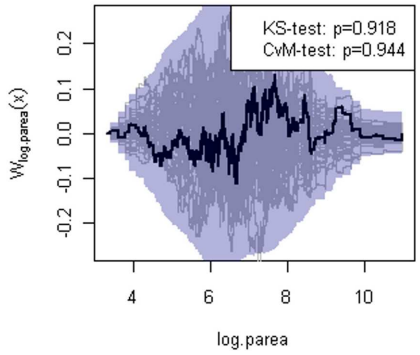

KS-test: p=0.407
CvM-test: p=0.408
$W_{log.proximity}(x)$
log.proximity

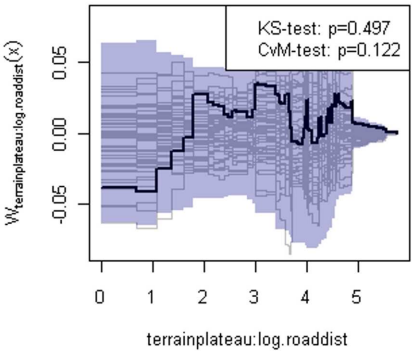

KS-test: p=0.127
CvM-test: p=0.182
$W_{log.roaddist}(x)$
log.roaddist


KS-test: p=0.048
CvM-test: p=0.006
$W_{log.riverdist}(x)$
log.riverdist


KS-test: p=0.127
CvM-test: p=0.328
$W_{log.shapei}(x)$
log.shapei


KS-test: p=0.091
CvM-test: p=0.016
$W_{log.neighbor}(x)$
log.neighbor


KS-test: p=0.918
CvM-test: p=0.944
$W_{log.parea}(x)$
log.parea


KS-test: p=0.497
CvM-test: p=0.122
$W_{terrainplateau:log.roaddist}(x)$
terrainplateau:log.roaddist

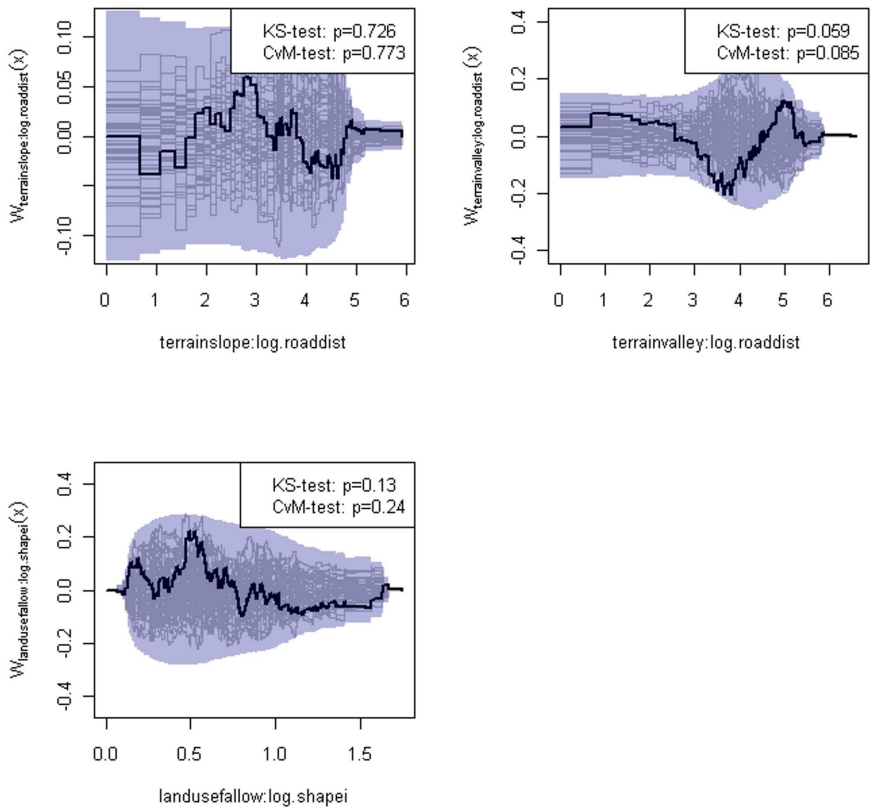


Figure 3 Cumulative residuals ordered after continuous predictor variables with `cumres` in package *gof*. Calculations are based on a GLM with same fixed effects as the maximum GLMM.

with the maximum model using the `anova` function.

```
> anova(Her.max, Her.red.patch)
```

and the resulting analysis of deviance table

```
[..]
             Df    AIC    BIC  logLik Chisq Chi Df Pr(>Chisq)
Her.red.patch 26 902.04 1041.2 -425.02
Her.max       27 904.01 1048.5 -425.01 0.028      1     0.8672
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

shows that the random intercept of patch does not improve the model fit. Thus, we will continue the analysis without this random effect.

The random intercept of study area cannot be tested with `lmer` because the reduced model would not contain any random effects, i.e. it would be a GLM, which is not provided there. But we can use `glmmML` to calculate LR and parametric bootstrap test for `starea`:

```
> library(glmmML)

> Her.glmmML<- glmmML(hogweed ~ habitat + landuse + terrain +
log.proximity + log.roaddist + large.riverdist + log.riverdist +
log.shapei + zero.neighbor + log.neighbor + log.parea +
habitat:terrain + landuse:terrain + log.roaddist:terrain +
```

```
log.shapei:landuse + zero.neighbor:landuse, family=binomial,
Heracleum, cluster=starea, boot=2000 )
```

Computing this model, particularly the bootstrap, takes a while although we only use the recommended minimum of 2000 bootstrap samples. The result

```
> summary(Her.glmmML)

[..]
        LR p-value for H_0: sigma = 0:  0.1118

 Bootstrap p-value for H_0: sigma = 0:  0.715 ( 2000 )
```

suggests that the random variation of intercepts among study areas is not significantly different from zero. Nevertheless, we will keep `starea` in our model because its estimated standard deviation is not too low and it is part of our survey design.

We recalculate the maximum model with simplified random effects structure

```
> Her.max.2<- lmer(hogweed ~ habitat + landuse + terrain +
log.proximity + log.roaddist + large.riverdist + log.riverdist +
log.shapei + zero.neighbor + log.neighbor + log.parea +
habitat:terrain + landuse:terrain + log.roaddist:terrain +
log.shapei:landuse + zero.neighbor:landuse
+ (1|starea), Heracleum, family=binomial )
```

before model selection of fixed effects.

Best subset of fixed effects

For finding the best subset of fixed effects from our maximum model, we can use the `dredge` function in the *MuMIn* package, which automatically fits all different combinations of fixed predictor variables to the data and calculates AIC values. However, there is one problem: given our 16 fixed effects (main effects and interactions) in the maximum model there would be 2^16 = 65,536 possible combinations and it would take approximately two weeks (!) to calculate all of them on an ordinary 5-year-old pc (AMD Athlon XP 2200+, 1800 MHz, 32 bits, 1 MB RAM, Kubuntu Linux (Ubuntu 10.04.3 LTS)). Thus, we have to reduce the number of candidate models. This can be done by 'fixing' some of the fixed effects and only allowing the remaining ones to be permuted. Here, we 'fix' all main effects of variables that had p-values < 0.05 in the summary table of the maximum model (Wald tests).

Now we run `dredge` with 7 effects being fixed, so that the number of candidate models is 2^9 = 512.

```
> library(MuMIn)

> best.subsets<- dredge(Her.max.2, rank="AIC", trace=TRUE, fixed= ~
terrain + log.roaddist + large.riverdist + log.riverdist +
zero.neighbor + log.neighbor + log.parea )

> print(best.subsets, abbrev.names=FALSE)
```

and get a list of candidate models ranked by AIC (Table 2).

Table 2 Ranking of candidate models by AIC calculated with the `dredge` function of the *MuMIn* package (R output modified).

	(Intercept)	habitat	landuse	log.proximity	log.shapei	habitat:terrain	landuse:log.shapei	landuse:terrain	landuse:zero.neighbor	log.roaddist:terrain	large.riverdist	log.neighbor	log.parea	log.riverdist	log.roaddist	terrain	zero.neighbor	k	Dev.	AIC	delta	weight
416	0.0824	+	+	0.1001	0.4537	+			+	+	-2.590	0.5708	0.3622	-0.4837	-0.8357	+	-3.128	22	854.9	898.9	0.0000	0.077
408	0.3717	+	+	0.1024		+			+	+	-2.652	0.5641	0.3931	-0.5018	-0.8642	+	-3.083	21	857.5	899.5	0.6005	0.057
278	0.7519	+		0.1022		+				+	-2.678	0.5722	0.4028	-0.5076	-0.8827	+	-4.046	19	861.7	899.7	0.8576	0.050
412	1.0250	+	+		0.4683	+			+	+	-2.727	0.5555	0.3442	-0.5237	-0.7939	+	-3.122	21	857.7	899.7	0.8855	0.049
286	0.5924	+		0.0983	0.3735	+				+	-2.644	0.5796	0.3843	-0.4984	-0.8542	+	-4.064	20	859.7	899.7	0.8880	0.049
448	0.4343	+	+	0.1019	-0.0647	+	+		+	+	-2.572	0.5707	0.3697	-0.4793	-0.8440	+	-3.089	23	854.0	900.0	1.1470	0.043
288	0.2906	+	+	0.1036	0.4763	+				+	-2.588	0.5768	0.3617	-0.4800	-0.8625	+	-4.048	21	858.4	900.4	1.5240	0.036
404	1.3480	+	+			+			+	+	-2.797	0.5479	0.3757	-0.5439	-0.8215	+	-3.075	20	860.5	900.5	1.6520	0.034
282	1.4940	+			0.3989	+				+	-2.775	0.5650	0.3635	-0.5360	-0.8125	+	-4.087	19	862.6	900.6	1.7320	0.032
274	1.7060	+				+				+	-2.819	0.5560	0.3824	-0.5479	-0.8405	+	-4.068	18	864.8	900.8	1.9770	0.029
480	0.5396	+	+	0.0986	0.4525	+		+	+	+	-2.578	0.5704	0.3719	-0.4728	-0.7964	+	-3.161	25	850.8	900.8	1.9790	0.029
444	1.3620	+	+		-0.0071	+	+		+	+	-2.713	0.5552	0.3508	-0.5202	-0.7992	+	-3.086	22	857.0	901.0	2.1480	0.026
160	-0.5229	+	+	0.0892	0.5124	+			+		-2.673	0.5710	0.3555	-0.5006	-0.2229	+	-3.060	19	863.3	901.3	2.3970	0.023
280	0.6113	+	+	0.1057		+				+	-2.653	0.5696	0.3930	-0.4987	-0.8920	+	-4.035	20	861.3	901.3	2.3970	0.023
472	0.8100	+	+	0.1009		+		+	+	+	-2.641	0.5640	0.4033	-0.4899	-0.8239	+	-3.118	24	853.4	901.4	2.5480	0.022
156	0.3542	+	+		0.5220	+			+		-2.794	0.5574	0.3403	-0.5356	-0.2157	+	-3.066	18	865.5	901.5	2.6300	0.021
284	1.2700	+	+		0.4907	+				+	-2.730	0.5614	0.3424	-0.5211	-0.8182	+	-4.075	20	861.5	901.5	2.6660	0.020
476	1.4420	+	+		0.4677	+		+	+	+	-2.714	0.5561	0.3540	-0.5126	-0.7579	+	-3.158	24	853.7	901.7	2.8140	0.019
320	0.6078	+	+	0.1051	0.0234	+	+			+	-2.572	0.5758	0.3678	-0.4758	-0.8680	+	-4.044	22	857.8	901.8	2.9600	0.018
512	0.8733	+	+	0.1005	-0.0454	+	+	+	+	+	-2.562	0.5704	0.3795	-0.4684	-0.8057	+	-3.121	26	850.0	902.0	3.1860	0.016

Significance tests of fixed effects

We see that models without the interactions `landuse:log.shapei` and `landuse:terrain` perform better than the maximum model. So we drop these interactions and calculate the final model

```
> Her.final<- lmer(hogweed ~ habitat + landuse + terrain +
log.proximity + log.roaddist + large.riverdist + log.riverdist +
log.shapei + zero.neighbor + log.neighbor + log.parea +
habitat:terrain + log.roaddist:terrain + zero.neighbor:landuse +
(1|starea), Heracleum, family=binomial )

> print(Her.final)

Generalized linear mixed model fit by the Laplace approximation
Formula: hogweed ~ habitat + landuse + terrain + log.proximity +
log.roaddist +      large.riverdist + log.riverdist + log.shapei +
+zero.neighbor +      log.neighbor + log.parea + habitat:terrain +
log.roaddist:terrain +      zero.neighbor:landuse + (1 | starea)
   Data: Heracleum
   AIC  BIC logLik deviance
 898.9 1017 -427.4    854.9
Random effects:
 Groups Name        Variance Std.Dev.
 starea (Intercept) 0.073522 0.27115
Number of obs: 1559, groups: starea, 20

Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  0.08249    1.04545   0.079 0.937111
habitatopt                   0.41069    0.54298   0.756 0.449431
landusefallow                0.57339    0.32318   1.774 0.076032 .
terrainplateau              -0.69046    0.76634  -0.901 0.367597
terrainslope                -1.93154    0.56153  -3.440 0.000582 ***
terrainvalley               -1.41685    0.55976  -2.531 0.011368 *
log.proximity                0.10009    0.05936   1.686 0.091745 .
log.roaddist                -0.83565    0.23968  -3.487 0.000489 ***
```

```
large.riverdist1             -2.58971    0.48817  -5.305 1.13e-07 ***
log.riverdist                -0.48375    0.10858  -4.455 8.38e-06 ***
log.shapei                    0.45366    0.28179   1.610 0.107411
zero.neighbor1               -3.12839    0.52462  -5.963 2.47e-09 ***
log.neighbor                  0.57082    0.06018   9.486  < 2e-16 ***
log.parea                     0.36221    0.08127   4.457 8.32e-06 ***
habitatopt:terrainplateau    -0.40656    0.78636  -0.517 0.605146
habitatopt:terrainslope       1.62257    0.66022   2.458 0.013986 *
habitatopt:terrainvalley      0.69669    0.59407   1.173 0.240900
terrainplateau:log.roaddist   0.66772    0.28910   2.310 0.020907 *
terrainslope:log.roaddist     0.61536    0.26389   2.332 0.019706 *
terrainvalley:log.roaddist    0.66836    0.24836   2.691 0.007123 **
landusefallow:zero.neighbor1 -1.18310    0.60968  -1.941 0.052314 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
[correlation matrix of fixed effects omitted]

which we will use for assessment of significance and interpretation of fixed effects.

We test the significance of fixed effects and their interactions with LR tests using the `anova` function. For this purpose, we calculate reduced models missing the effect that we want to test, e.g. `log.parea`

```
> Her.red.log.parea<- lmer(hogweed ~ habitat + landuse + terrain +
log.proximity + log.roaddist + large.riverdist + log.riverdist +
log.shapei + zero.neighbor + log.neighbor + habitat:terrain +
log.roaddist:terrain + zero.neighbor:landuse + (1|starea), Heracleum,
family=binomial )
```

and compare it to the final model,

```
> anova(Her.red.log.parea, Her.final)

[..]
                  Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
Her.red.log.parea 21 917.06 1029.5 -437.53
Her.final         22 898.86 1016.6 -427.43 20.204      1   6.96e-06
***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

that is, we are conducting a type III LR test.

Interpreting the effects

In binomial models with logit link, the linear combination of predictor variables models the logarithm of the odds of the dependent variable. In this example, the odds are the probability *p* of a subpatch being invaded divided by the probability of not being invaded, i.e. *p*/(1-*p*). The effect of a single predictor variable *x* can be assessed by its odds ratio (OR) which is the odds if *x* = 1 divided by the odds if *x* = 0. The OR is calculated by exponentiating the estimate *b* (regression coefficient) of *x*:

$$OR = \frac{p_{x=1}/(1-p_{x=1})}{p_{x=0}/(1-p_{x=0})} = \frac{e^{b_x}}{e^0} = e^{b_x}$$

For instance, the odds ratio for `habitat` is exp(0.41069) = 1.51. This means the odds of invasion probability in optimal habitat is roughly 1.5 times larger than in suboptimal habitat.

Since the continuous predictors are on a logarithmic scale like the log odds, we get a power relation between OR and the continuous predictors. Suppose for instance that the distance from road is doubled from x to $2x$. Then the OR is given by

$$OR = \frac{e^{b*\log(1+2x)}}{e^{b*\log(1+x)}} = \left(\frac{1+2x}{1+x}\right)^b \approx 2^b$$

The addition of 1 in the above equation comes from our ad hoc approach to handle the zeros in `roaddist`. For distance to roads we have b = -0.83565. Thus, if this distance is doubled from 20 m to 40 m, say, then the odds of invasion probability further from the road is $2^{-0.83565} = 0.5603$ times the odds closer to the road.

References

Barton K (2011). MuMIn: Multi-model inference. R package version 1.0.0. http://CRAN.R-project.org/package=MuMIn.

Bates D, Maechler M, Bolker B (2011). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-41. http://CRAN.R-project.org/package=lme4.

Broström G, Holmberg H (2011). glmmML: Generalized linear models with clustering. R package version 0.81-8. http://CRAN.R-project.org/package=glmmML.

Holst K.K (2011). gof: Model-diagnostics based on cumulative residuals. R package version 0.7-6.

Lin EY, Wei LJ, Ying Z. Model-checking techniques based on cumulative residuals. Biometrics 2002;58:1–12.

McGarigal K, Marks BJ. FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. USDA For Serv Gen Tech Rep PNW-351;1995.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Sarkar D (2008) Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5.

Thiele J, Schuckert U, Otte A (2008) Cultural landscapes are patch-corridor-matrix mosaics for an invasive megaforb. Landscape Ecology 23:453-465.