Report on the statistical analysis investigating the relationship between the terrain slope and the observed counts of beaver dams along a slope range

1) Regression analysis

In ecology we often wish to model a response variable that is a count variable. Counts are usually Poisson distributed, which implies that the model errors are not approximately Gaussian distributed, making the use of linear models impossible (Kabacoff 2015, Quinn and Keough 2002, Wilkinson 2017).

Generalized linear models (GLMs) extend the linear modelling theory to cover response variables of the exponential family (in addition to the Gaussian) (Kabacoff 2015, Wilkinson 2017). Common applications of GLMs are the logistic regression (where the response variable is binary),the Poisson regression (where the response variable are counts) and the log-linear models (where the response variable is categorical) (Quinn and Keough 2002). For the dataset on the observed counts along a slope range the statistical analysis started with a Poisson regression using the glm function of the stats package in R (R Core Team 2017). Given the fact that the response variable was overdispersed (the variance greater than the mean), a further model was fitted using the quasipoisson family as an underlying distribution for the response variable. In the quasipoisson regression model, the dispersion parameter is estimated from the data rather than the value defined by a Poisson distribution (Quinn and Keough 2002). A second solution to the observed overdispersion problem was to fit a negative binomial GLM (with the glm.nb function of the MASS package - Venables and Ripley 2002) (Crawley 2013, Quinn and Keough 2002). The negative binomial density in constrast to the Poisson density, includes a second parameter (theta, the clumping parameter) which is estimated from the data and uses the same link function as the Poisson regression (the log link function) as a function of the mean.

The quasipoisson model and the negative binomial model behave both more or less the same in the diagnosis plots. The plots showing the predicted values and the observed values along the slope range for both models (quasipoisson and negative binomial) indicate a better fit for the negative binomial model with the prior information from our technical stuff (Pasca C) that the beaver dams occur mainly in the 0 to 1 terrain slope and cease to apper from a 5 degree terrain slope upwards. This information will be further used in the Bayesian analysis presented at the end of the report.

Table 1: coefficients of the models, standard errors of coeff., P-values

Plots: plots showing the predicted values and the observed values along the slope range for both models (quasipoisson and negative binomial)

Interpretation of the regression coefficients for the negative binomial model: for 0 terrain slope we have about 114 beaver dams (the intercept in this case brings an important amount of information). The GLM indicates that for each unit increase in terrain slope we have a decrease in the log number of dams of about 0.262.

Interpretation of the regression coefficients for the quasipoisson model: for 0 terrain slope we have about 105.5 beaver dams (again the intercept brings a high amout of information). For each unit increase in terrain slope we have a decrease in the log number of dams of about 0.231.

2) Fitted and observed probability distribution of the beaver dam counts along the terrain slope range

Similar to modelling the counts of beaver dams along a terrain slope range, another negative binomial model was fitted, this time using as response variable the observed probabilities of occurrence of beaver dams along a terrain slope range. The observed probability distribution of the occurrence of beaver dams along the terrain slope and the fitted probability distribution are presented in the following plot:

Plot: Fitted and observed probability distribution of the beaver dam counts along the terrain slope range

Table: coefficients, se, P-values of the fitted model

3) Using the Poisson-Gamma model to predict future counts

a) The Poisson-Gamma model

In order to obtain a posterior predictive mass function for the prediction of future occurrences of beaver dams along the terrain slope, a Poisson-Gamma model was fitted. The Poisson-Gamma model, also known as the conjugate analysis of Poisson data, consists of: a conjugate Gamma prior density (representing the prior beliefs over the intensity parameter lambda), a Poisson likelihood (Poisson distributed observed variable) and a Gamma posterior density (updating our beliefs over lambda via the Bayes Rule) (Jackman 2009). According to the observed data the lambdahatMLE (Maximum Likelihood Estimation of lambda, the intensity parameter of the Poisson distribution) is 26.105, thus the likelihood is: beaver\_dams ~ Poisson(26.105). From the expertise of our technical stuff (Pasca C), it is assumed that the prior average count a/b=0.5, returnining a prior Gamma(a,b) density: p(lambda)~Gamma(9.5, 19). This defines the following posterior density: p (lambda|beaver\_dams)~Gamma(505.5, 38). The exact calculations of the parameters of the probability densities can be found in the R script of the present analysis.

The previously obtained posterior density was further used to find out the posterior predictive mass function, which is a negative binomial density, in order to make predictions on future dam counts along the terrain slope range.

Plots: the likelihood, prior and posterior densities

b) Posterior predictive mass function for the Poisson-Gamma model

The posterior predictive mass function for a Poisson distributed count variable with prior/posterior beliefs over lambda represented by a Gamma density is a negative binomial distribution characterized by theta (the success probability) and a parameter a (the success count) (Jackman 2009).

Using the Bayes Rule, the obtained predictive mass function is:

p(future\_beaver\_dams, beaver\_dams)~NB(505.5, 0.974359).

Because the overdispersion in the observed data, there is a strong difference between the posterior predictive mass function and the prior density. In order to make applicable the results, the sources of overdispersion need to be addressed. Overdispersion is a phenomenon with many sources, but in this case, the errors might lie in the way the observed variables were measured (low resolution of the DEM) or in the unexplained variation due to additional variables not included in the model (i.e. vegetation cover). Future research should aim at generating a viable predictive mass function for the counts along the terrain slope range. So, it is important to reduce these two data-distorting sources in the future to be able to exploit the Poisson-Gamma model for predictive purposes.

4) References

1. Kabacoff R I, 2015: R in action - Data analysis and graphics with R, Second edition, Manning

2. Crawley M J, 2013: The R book, Second edition, Wiley

3. Jackman S, 2009: Bayesian analysis for the social sciences, Wiley

4. Quinn G P, Keough M J, 2002: Experimental design and data analysis for biologists, CUP

5. R Core Team, 2017, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https ://www.R-project.org/

6. https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/glm.nb.html (08.06.2017)

7. Tesileanu R, 2016: The occurrence of beaver dams in relation to features of the water stream and vegetation characterizing the area,

Technical Report INCDS

8. Venables, W N, Ripley B D, 2002: Modern Applied Statistics with S, Fourth edition, Springer

9. Wang J, Zamar R, Marazzi A, Yohai V, Salibian-Barrera M, Maronna R, Zivot E, Rocke D, Martin D, Maechler M, Konis K, (2017):

robust: Port of the S+ "Robust Library". R package version 0.4-18, https://CRAN.R-project.org/package=robust

10. http://www.statmethods.net/advgraphs/probability.html (10.06.2017)

Author: Roxana Tesileanu (roxana.te@web.de) INCDS Romania