Model selection and issues

Current models

Initial models

Continuing from the GAMs being run previously for juvenile proportion, models for family size also used GAMs and were run on the SOVON flock level data, which were untabled to get family level data. An effect of this was the duplication of flock attributes (flock size, position, etc.) as many times as there were families in the flock. The initial models were run using `mgcv::gam`, with the random effects specified as random intercepts using the smoothing basis "re". The results were then checked by running the same model using the function `mgcv::gamm`, which assumes nestedness or hierarchical structure in the random effects. In this case, the random effects were specified so that observer was "nested" within habitat type which was nested within breeding year.

The models differed in their results, with flock size and longitude found to be significant effects by the simpler `gam` which assumed i.i.d. gamm did not find these significant effects.

Running the `gamm` on the geese.org data produced results opposed to the SOVON data, with both flock size and longitude significant effects. This might be explained by the difference in predictors used.

Checking for spatial autocorrelation

The semivariogram of the residuals was plotted using functions from two different packages, `nlme::Variogram` and `gstat::variogram`. Both failed to show any spatial autocorrelation. Were this to be found, F. Dormann et al. (2007) suggest that an implementation of the `gls` method in R - which allows for the specification of a correlation structure - should be able to account for it. This method is called internally by `gamm` as part of its `lme` function, making `gamm` an effective way of dealing with it as well.

Model diagnostic plots

Without exception, the model diagnostic plots are far from ideal. This prompted a look at two other ways of modelling family sizes. It might be that the assumption of a Poisson error distribution made when fitted these generalised models is not a good one. Transforming the data (advised against by O'Hara and Kotze (2010)) did not improve the diagnostic plots. The main issue might be that the family size is very strictly bounded between 1 and 10 and forced to take integer values. The following approaches were tried to help with this.

Using family size counts

The SOVON dataset presents family sizes as counts for each flock of the number of families of size x present. These counts are also integers bounded by 0, but otherwise have no upper limit. A model using these counts as a response, with a smooth function of family size, the number of families, the flock size, and a smooth function of longitude as fixed effects was run. The model found all predictors to be significant. The diagnostic plots were not improved. This could be attributed to the large number of zeroes present when families did not take high values. This model can't be used with the geese.org data.

Using a multinomial logistic regression

Since the family sizes can only take 10 values, I tried modelling them as polychotomous categorical responses using the `MCMCglmm` function. This is time-consuming, and not to be pursued unless there's a very good reason.

Issues

1. A Poisson distribution may not be the right one. A quasipoisson distribution was tried and did not help the model. A negative binomial error structure could be tried.
2. Three possible response variables are possible: the number of juveniles (absolute family size) in each family, the mean number of juveniles in a family per flock, and the counts of families of size x in a flock. A choice other than family size prevents the comparison of the SOVON and geese.org datasets.
3. The predation index is not used in any of these models and may need to be accounted for, since families in high predation years may be smaller than the mean even prior to migration.
4. The predation index is a mean value calculated as the mean of ordinal values from a differing number of locations in each year. When coerced to a factor, it either takes as many unique values as there are years, or has to be classified into intervals which lose the differences between successive years, since only peaks and troughs are really different from each other.
5. Duration of winter, calculated as the number of days between the autumn and spring migration peaks, may be an important predictor of family size. Family splitting, or juvenile dispersal, may be a function of biological processes (increasing body mass of the juveniles leading to increased within-family competition for food, for example), which may be proxied by time. All models agree that the number of days since the start of winter is a significant predictor of family size. In longer winters, the probability of detecting smaller families is thus increased (unless juvenile dispersal is also reduced by conditions correlated with long winters).

An arbitrary equation to test MathJax. $y = s(x) + s(z^{a/b})$

F. Dormann, C. et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. - Ecography 30: 609–628.

O'Hara, R. B. and Kotze, D. J. 2010. Do not log-transform count data. - Methods in Ecology and Evolution 1: 118–122.