Multivariate regression models for zero-inflated and over-dispersed count data.

Context

Abrupt climate change or land-use changes are triggering species extinctions, range shifts and biotic homogenization. The development of models able to predict the future of biodiversity as well as ecosystem services is now a critical endeavor, as stated by the recently published report of IPBES. Recently, tremendous progress has been made in extending species distribution models (SDMs) to Joint Species Distribution Models (JSDMs). These advances allow the integration of ecological mechanisms more efficiently by considering how species dependencies shape community composition. On the other hand, modeling multivariate count data is not an easy task especially taking into account zero-inflation and over-dispersion as well as spatial or time dependencies.

Methods and objectives

Different approaches have been proposed to model multivariate count data (Inouye et al 2018). One is based on convolution of probability measures. While this approach allows the direct control of dependencies between outcomes, it only permits considering positive correlations (Karlis et al 2007}. Another and more recent one is based on copula theory (Joe, 2014). While this approach is promising, its use remains subject to caution, because many of the convenient properties of copula do not carry over from continuous to discrete cases (Genest et al 2007). The other, very common approach in a Bayesian context, uses the hierarchical Bayesian framework; for example modeling Poisson intensity according to different priors. In particular, a popular approach is the multivariate Poisson-log normal distribution, which combines independent Poisson distributions with a multivariate log normal distribution for the Poisson intensities (Banerjee et al. 2004, Chib and Winkelmann, 2001). Moreover, count data present two well-known difficulties: zero-inflation and over-dispersion. In the univariate case, different models have been developed for handling either zero-inflation, such as zero-inflated or zero-truncated models (ZIM, ZTM), and over-dispersion using negative-binomial models (NBM) or generalized linear mixed models (GLMM). Methods combining ZIM, NBM and GLMM have also been proposed to address both problems (Flores et al 2009). However, applying such models without considering the potential source of zero and over dispersion can lead to erroneous conclusions and results. In particular, spatial correlations, time dependencies or both can often be at the origin of the zero-inflation and over-dispersion (Flores et al 2009, Arab 2015). In the multivariate context, few solutions have been proposed to address these issues while taking into account spatial or time dependencies.

The main objective of this PhD thesis is to provide a new class of distribution functions for count data with enough flexibility to fit zero-inflated and over-dispersed data in the uni- and multi-variate context, and encompassing classical ones such as Poisson, Negative Binomial, Poisson-log-Normal, Negative Binomial-log-Normal (Zhou et al. 2012) or Poisson-inverse Gaussian distributions (Dean et al. 1989).

The strategy is to follow the infinite mixture structure and select the distributional components of the mixture carefully in order to address both zero-inflation and over-dispersion. Combining Poisson and heavy tail distributions, which have the ability to produce values far from their means, should improve global fit with a wide spread that can capture both the zeros and the large values.

Data and case studies

The PhD student will develop methods to address these issues and will apply them to species, traits and ecosystem services data from a comprehensive collection of datasets shared with GAMBAS' partners from different ecosystems (terrestrial and aquatic, temperate and tropical). This will allow the student to collaborate with all GAMBAS' partners.

Practical Information

The PhD fellowship will start on October 1st 2019. It will be funded by the GAMBAS project involving six partners (CIRAD, IRSTEA, CNRS, MNHN, Univ de Montpellier and Paris-Sud Orsay). The PhD student will be based at Université de Montpellier.

PhD Director:
Frédéric Mortier, UPR Forêts et Sociétés, CIRAD, Campus International de Baillarguet, TA C-105/D
34398 MONTPELLIER CEDEX 5 - FRANCE
Tél. +33 (0)4 67 59 37 66
Co-supervision:
Marie Denis (CIRAD, UMR AGAP), Gwladys Toulemonde (Université de Montpellier, UMR IMAG) and Camille Coron (Université Paris-Sud, UMR 8628).

Candidate profile

The candidate should have a training in applied mathematics with a strong background in statistics and probability. He/she should also be interested in applications of mathematical tools to the applied field of ecology. The candidate should have good skills in scientific English (reading, speaking and writing) and work organization (autonomous and good reporting skills). In addition, the candidate must have good computational skills and will be expected to develop R packages.

Candidates selection

A first phase of selection will be organized in mid 2019. Candidates have to send:

- a CV,
- a motivation letter,
- two letters of recommendation,
- a Master's thesis report

by e-mail Frédéric Mortier (fmortier@cirad.fr) before June 15th 2019. Selected candidates will be interviewed between late June and early July by the researchers involved in the project.

References

1. Arab. Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. International Journal of Environmental Research and Public Health, 12:10536{10548, 2015.
2. S. Banerjee, B. Carlin, and A. Gelfand. Hierarchical modeling and analysis for spatial data, volume 101 of Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, Boca Raton, FL, 2004.
3. P. Chagneau, F. Mortier, N. Picard, and J. Bacro. Prediction of a multivariate spatial random field with continuous, count and ordinal outcomes. Biometrics, 58:345-367, 2011.
4. S. Chib and R. Winkelmann. Markov chain monte carlo analysis of correlated count data. Journal of Business & Economic Statistics, 19(4):428{435, 2001.
5. O. Flores, V. Rossi, and F. Mortier. Autocorrelation of sets zero-inflation in models of tropical saplings density. Ecological Modeling, 220:1797-1809, 2009.
6. Genest C, Nešlehová J. A primer on copulas for count data. Astin Bulletin. 2007; 37(02):475–515.
7. H. Joe. Dependence Modeling with Copulas. Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, 2014.

8. D. Karlis and L. Meligkotsidou. Finite mixtures of multivariate poisson distributions with application. Statistical Journal and Inference, 137:1942-1960, 2007.
9. M. Zhou, L. Li, D. Dunson, and L. Carin. Lognormal and gamma mixed negative binomial regression. In ICML, 2012.