

# Modélisation multivariée de données de comptage en présence de surdispersion.

F. Mortier<sup>\*</sup>, J. Peyhardi<sup>†</sup> & G. Toulemonde<sup>‡</sup>

1<sup>er</sup> octobre 2019

**Mots clés :** loi de Poisson, mélange, surdispersion, valeurs extrêmes, multivarié, écologie.

## Contexte

Modéliser et prédire la distribution des espèces est un enjeu majeur pour préserver les écosystèmes naturels en particulier face au changement climatique et à l'augmentation des pressions humaines. Classiquement, l'abondance des espèces est supposée distribuée selon une loi de Poisson, dont l'intensité dépend de caractéristiques environnementales. Toutefois, en raison de différents facteurs (dispersion limitée, compétition entre espèces, etc), les données d'abondances présentent une surdispersion qui se caractérise soit par un excès de zéros, soit par des valeurs extrêmes soit les deux simultanément. Cette surdispersion viole la propriété d'égalité entre l'espérance et la variance d'une loi de Poisson. Cela se traduit par une détérioration des qualités d'ajustement rendant finalement ce simple modèle souvent mal adapté.

## Cadre mathématique et objectifs

Les modèles mélanges, finis ou non, de lois de Poisson forment une solution élégante pour gérer la question de la surdispersion (Karlis and Xekalaki, 2005). Dans le cas univarié, Greenwood and Yule suggèrent, dès 1920, de remplacer la loi de Poisson par une loi négative binomiale, celle-ci pouvant se définir comme le mélange d'une loi Poisson dont l'intensité est distribuée selon une loi gamma. Pour gérer une très forte surreprésentation de zéros, les modèles de Poisson zéro-enflés (zero-inflated Poisson, ZIP) sont largement utilisés (Lambert, 1992). Ils se définissent en termes de mélange fini, considérant que l'intensité est distribuée selon une loi de Bernoulli. Cette construction présente de nombreux intérêts. On peut citer entre autres : (i) la probabilité d'observer un zéro est plus grande que dans le cas d'une loi de poisson, (ii) la variance est structurellement supérieure dans le cas du mélange que dans le cas simple, (iii) les moments sont facilement calculables, (iv) la forme et la queue de la distribution de la loi de Poisson en mélange sont fortement reliées à celle de la loi de mélange. D'autre part, selon les lois de mélange, le cas univarié peut s'étendre au cas multivarié. Par exemple la loi de Poisson multivariée log-gaussienne (Aïtchison and Ho, 1989; Chib and Winkelmann, 2001) qui généralise la loi de Poisson log-gaussienne (Bulmer, 1974). Cela représente un atout majeur dans le contexte de la modélisation multi-spécifique. Enfin, cette construction s'interprète aussi d'un point de vue bayésien, la loi de mélange étant alors la loi *a priori* de l'intensité de la loi de Poisson. Toutefois, peu de travaux ont cherché à caractériser la loi de Poisson en mélange en fonction de la loi de mélange considérée

---

\*fmortier@cirad.fr

†jean.peyhardi@umontpellier.fr

‡gwladys.toulemonde@umontpellier.fr

et à les comparer objectivement en termes de performances pour mieux représenter et capturer la surdispersion (excès de zéros, valeurs extrêmes).

Les objectifs du stage sont (i) dans le cas univarié, d'étudier et caractériser les lois de Poisson en mélange, en se focalisant sur leur comportement en zéro et sur la structure des queues de distribution, (ii) de proposer, si cela est possible, un critère permettant d'orienter le choix de la loi de mélange selon le degré de dispersion des observations, (iii) d'amorcer une réflexion sur l'extension des lois de Poisson en mélange au cas multivarié.

## Applications

Les résultats seront testés sur simulations et appliqués à données réelles. Celles-ci concernent la distribution de 49 taxons de macro-invertébrés échantillonnés dans les rivières du sud-est de la France. Ce stage contribuera à une meilleure compréhension du fonctionnement écologique des cours d'eau, les communautés d'invertébrés étant sensibles à la qualité des eaux, ainsi qu'à l'élaboration d'indicateurs de biodiversité des cours d'eau.

## Informations pratiques

Ce stage en statistique, d'une période de quatre à six mois, se déroulera à Montpellier au sein de l'unité *Forêts et Sociétés* du CIRAD et de *l'Institut Montpelliérain Alexander Grothendieck* – IMAG - de l'université de Montpellier. L'étudiant(e) percevra une indemnité de stage de 540€/mois.

Ce stage s'insère dans le cadre du projet GAMBAS financé par l'ANR<sup>1</sup>. Ce travail pourra donner lieu à une thèse dont le financement est assuré par le projet.

## Pièces à fournir

- CV
- Lettre de motivation
- Notes de M1

## Références

- Aitchison, J. and Ho, C. (1989). The multivariate poisson-log normal distribution. *Biometrika*, 76 :643–653.
- Bulmer, M. (1974). On fitting the poisson lognormal distribution to species abundance data. *Biometrics*, 30 :101–110.
- Chib, S. and Winkelmann, R. (2001). Markov chain monte carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, 19(4) :428–435.
- Greenwood, M. and Yule, G. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society A*, 83 :255—279.
- Karlis, D. and Xekalaki, E. (2005). Mixed poisson distributions. *International Statistical Review.*, 73 :35–58.
- Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34 :1–14.

---

1. Generating Advances in Modeling Biodiversity And ecosystem Services (GAMBAS) : statistical improvements and ecological relevance of joint species distribution models, ANR-18-CE02-0025