Rapport de stage

17 Février 2020 — 15 Août 2020

Théorie des valeurs extrêmes dans le cadre des mélanges de Poisson

Étudiant :
Samuel Valiquette

Superviseurs :
Frédéric Mortier
Jean Peyhardi
Gwladys Toulemonde

Date de soutenance :
9 Juillet 2020

Jury :
Élodie Brunel-Piccinini
Gilles Ducharme
Éric Marchand
Jean-Michel Marin

RÉSUMÉ :
En écologie, les modèles de distribution d'espèces sont communément utilisés pour analyser l'abondance et la distribution des espèces. Une approche classique, consiste à supposer que ces données sont distribuées selon une loi de Poisson. Or plusieurs facteurs, biotiques ou abiotiques, conduisent à de la surdispersion. Celle-ci se traduit soit par des excès de zéros ou/et des valeurs extrêmes. Cette situation viole l'hypothèse d'égalité entre l'espérance et la variance. Pour surmonter cette difficulté, les modèles de Poisson en mélange forment une solution élégante. Il existe cependant une infinité de lois de mélange. Ce travail présente une stratégie qui permet de mieux cibler les distributions potentielles en analysant le comportement en queue des observations. Plus précisémment, on propose d'appliquer la théorie des valeurs extrêmes, traditionnellement employée dans le cas de variables aléatoires continues, au cas des lois de Poisson en mélange. On énonce les conditions qui assurent ou non que le domaine d'attaction de la loi de mélange se transmet au mélange qui en résulte. On démontre que si la densité de mélange a un comportement « dit gamma », alors le mélange résultant ne possèdera aucun domaine d'attraction, mais sera « proche » de celui de Gumbel. Ces densités incluent par exemple les lois gamma ou inverse gaussienne. Ces résultats sont mis à profit pour établir un arbre de décision basé sur les excès des données de comptage. Ce travail est illustré en utilisant des données d'abondances de deux espèces communes des forêts tropicales d'Afrique centrale.

MOTS-CLÉS :
Données de comptage, écologie, mélange de Poisson, modèles de distribution d'espèces, surdispersion, théorie des valeurs extrêmes.

ABSTRACT :
In ecology, species distribution models are classically used to analyse the abundance and distribution of species. A solution consists in supposing that these observations are generated from a Poisson distribution. However several factors, biotics or abiotics, may induce overdispersion. This produces excess zeros or/and extremes values. Such situation violates the assumption that the expected value is equal to the variance. To overcome such a problem, mixed Poisson distributions provide a elegant solution. A huge set of mixture distribution is available. This work presents a strategy to shrink the choice of the possible distributions. It is based on an analysis of the tail behavior of the observations. Precisely, we propose to apply extreme value theory, usually used for continuous random variables, to Poisson mixtures. We introduce conditions that indicate when the domain of attraction of the mixed distribution is preserved or not by the final mixture. We also show that if a density with a 'gamma behavior' is used for the mixed distribution, then the mixture won't have any domain of attraction, but will be 'close' to the Gumbel domain. Such densities include the gamma and the inverse Gaussian. These results are used to established a decision tree based on the excess of the count data. We illustrate this strategy by applying it to an abundance data set of two common tree species from Central African rainforests.

KEY WORDS :
Count data, ecology, extreme value theory, overdispersion, poisson mixture, species distribution models.

Remerciements

J'aimerais tout d'abord exprimer ma reconnaissance à mes superviseurs Frédéric Mortier, Jean Peyhardi et Gwladys Toulemonde. Ce projet étant ma première expérience de recherche, j'ai été privilégié d'avoir reçu leurs conseils expérimentés. Leurs différents domaines de recherche m'ont également permis d'élargir ma vision de la problématique et de diversifier les approches prises pour arriver à une solution. Malgré les distances causées par la pandémie du COVID-19, je n'ai jamais eu l'impression d'être seul dans ce travail grâce à eux.

Je tiens également à remercier Éric Marchand et Élodie Brunelle-Piccinni pour avoir concrétiser l'échange entre l'Université de Sherbrooke et celle de Montpellier. Cette opportunité n'aurait jamais eu lieu sans leur soutien. De plus, j'aimerais les remercier pour avoir été membre du jury d'évaluation de ce projet, ainsi que Gilles Ducharme et Jean-Michel Marin.

Enfin, j'exprime ma gratitude au CIRAD pour le financement de ce stage ainsi que la bourse ISM et la bourse Frontenac pour avoir soutenu mon échange à Montpellier.

1 Introduction

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. Comme première approche, 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 ou autres), 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.

Une stratégie pour prendre en compte la surdispersion repose sur l'utilisation de modèles de mélanges finis ou non (Karlis et Xekalaki, 2005). Ces derniers supposent que l'intensité de la loi de Poisson, λ, n'est plus une valeur fixe mais est, elle-même, aléatoire. Cette approche se justifie, d'un point de vue théorique, par au moins deux raisons : (i) la variance issue du mélange est toujours supérieure à celle d'une loi de Poisson ; (ii) si la loi de Poisson possède la même moyenne que son mélange, alors ce dernier aura une plus grande probabilité d'observer des zéros et des grandes valeurs. Karlis et Xekalaki listent jusqu'à 30 exemples de modèles de Poisson en mélange selon le choix de la distribution du paramètre λ. Ces modèles ont été développés pour répondre à différentes questions pratiques telles que l'analyse lexicale, l'étude des accidents, etc. Parmi les choix classiquement usités en écologie, on peut citer entre autre, la loi gamma (Greendwood et Yule, 1920), la loi lognormale (Bulmer, 1974) ou encore la loi de Bernoulli pour modéliser plus spécifiquement l'excès de zéros (Lambert, 1992). D'un point de vue général, toutes les lois dont le support est positif sont de potentielles candidates. À l'heure actuelle, il ne semble pas exister d'étude et de travaux permettant, au regard des données, de choisir la ou les distributions les mieux adaptées. Cela soulève donc la question : comment choisir cette loi de mélange ?

Classsiquement, la stratégie repose sur le choix de quelques distributions de référence et l'utilisation d'un critère de sélection (AIC ou BIC). On peut noter aussi que le choix de ces lois candidates est souvent rattaché au domaine d'application. Quelque soit ce domaine, le choix reste souvent arbitraire. Des propositions ont été faites pour mieux les choisir. Notamment, celle proposée par Lynch (1988) qui démontre que les lois de Poisson en mélange héritent de la 'forme' de la loi de mélange (cf Figure 1). Cependant, distinguer le comportement de lois n'est pas toujours simple. Comment peut-on, à l'oeil, dissocier la loi log-gaussienne centrée et de variance 1 avec la loi gamma(2,2) en Figure 1 ? Ce constat est le même quand l'on compare la lognormale avec les lois demi-Cauchy ou inverse bêta. Est-il même possible de les distinguer ? La théorie des valeurs extrêmes offre un cadre méthodologique rigoureux pour étudier, comparer et différencier ces distributions.

Mon travail a consisté à élaborer une stratégie pour choisir les lois potentiellement les plus intéressantes parmi l'ensemble des possibles. Ce travail se base sur l'extension au cadre discret de la théorie des valeurs extrêmes, théorie plus classiquement appliquée et connue dans le cas continu. Le chapitre 2 présente les bases de cette théorie. On montrera notamment pourquoi le cas discret peut être problématique. Dans le chapitre 3, ces outils théoriques seront étendus au cas des mélanges de Poisson. Finalement on propose une stratégie pour sélectionner des lois possibles au chapitre 4. Ce travail est illustré sur des données d'abondances de deux espèces de forêts tropicales.

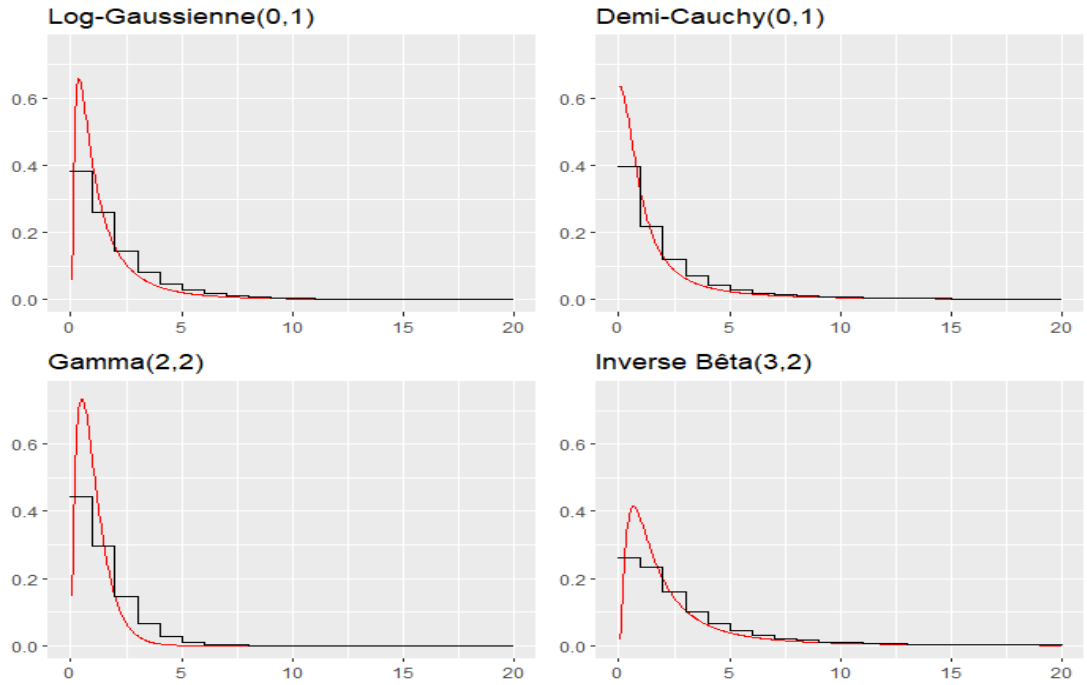

FIGURE 1 – Fonction de masse du mélange de Poisson (noir) et la densité sur λ utilisée (rouge)

2 Théorie des valeurs extrêmes

2.1 Principe

La théorie des valeurs extrêmes s'intéresse aux valeurs exceptionnelles d'un événement et tente d'extraire une tendance ou d'estimer des quantiles extrêmes. Par définition, ces événements sont rares et éloignés du centre de la distribution. Souvent ces observations sont considérées comme aberrantes et sont ignorées. Or il peut s'avèrer utile de comprendre la distribution de celles-ci. En science de l'environnement, on peut penser entre autres aux régimes des pluies, aux inondations ou encore aux événements caniculaires. Pour modéliser ces phénomènes, il est pertinent de mesurer la quantité de pluies, les débits des cours d'eau ou de la température sur une certaine période et d'inférer la distribution des valeurs extrêmes pour mieux comprendre leur fréquence d'apparition ou encore leur intensité. On peut noter, qu'en pratique, les observations sont souvent non indépendantes. Cependant la question de la modélisation de cette dépendance ne sera pas abordée dans ce rapport.

Soit $X_1, \ldots, X_n$ un échantillon de variables aléatoires indépendantes et identiquement distribuées (i.i.d) de loi F (continue ou discrète). Posons

$$M_n := \max(X_1, \ldots, X_n).$$

Notons que nous nous intéressons ici aux valeurs maximales mais il est possible de transposer ces résultats aux minimas. En effet, $\min(X_1, \ldots, X_n)$ est égal à $-\max(-X_1, \ldots, -X_n)$, donc les démarches pour M_n seront valides aussi pour les minimums. Afin de modéliser les extrêmes, il est nécessaire de connaître la fonction de répartition de M_n. Grâce aux hypothèses précédentes (i.i.d.), on a $\mathbb{P}(M_n \leq x) = F^n(x)$. Même si F était supposée connue, la fonction de répartition de M_n peut être difficile à calculer numériquement. Évidemment, dans un contexte réel F est inconnue. Pour y remédier, la théorie propose un analogue au théorème central limite pour approcher la distribution de M_n.

Soit $x_F = \sup\{x : F(x) < 1\}$, le point terminal à droite de F tel que $F(x) = 1$ pour tout $x \geq x_F$. On notera que x_F peut être fini ou non. Il est alors possible de montrer que $\lim_{n\to\infty} M_n = x_F$ presque sûrement. Ainsi, pour déduire le comportement asymptotique de M_n, il faut trouver des suites normalisantes $a_n > 0$ et $b_n \in \mathbb{R}$ telles que pour toutes valeurs de x dans le support de F on a

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = G(x)$$

où G est une fonction de répartition non dégénérée. Pour décrire cette convergence, on donne la définition suivante.

Définition 1. Soit la variable aléatoire X avec comme fonction de répartition F, on dit que F est dans un domaine d'attraction des maximums si on peut trouver des suites normalisantes $a_n > 0$ et $b_n \in \mathbb{R}$ telles que pour toutes valeurs de x dans le support de F on a une fonction de répartition non-dégénérée G telle que

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x).$$

Fisher et Tippett (1928) trouvent trois lois G possibles et Gnedenko (1943) prouve que si de telles suites existent, alors il n'y a pas d'autres lois possibles. Ces trois lois peuvent être regroupées sous une seule et même distribution nommée loi des valeurs extrêmes généralisée, ou la loi GEV pour *Generalized Extreme Value*. Il est alors possible maintenant d'énoncer le théorème fondamental des valeurs extrêmes.

Théorème 1 (Fisher et Tippett, 1928, Gnedenko, 1943)**.** *Soit $X_1, \ldots, X_n$ i.i.d. de loi F. Si F est dans un domaine d'attraction des maximums, alors il existe des suites $a_n > 0$ et $b_n \in \mathbb{R}$ tel que G est la loi GEV, notée G_γ, définie par*

$$G_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-1/\gamma}\right) & \text{pour tout } x \text{ tel que } 1+\gamma x > 0 \text{ si } \gamma \neq 0 \\ \exp\left(-e^{-x}\right) & \text{pour tout } x \in \mathbb{R} \text{ si } \gamma = 0. \end{cases}$$

On note alors l'appartenance de F au domaine d'attraction des maximums par $F \in \mathcal{D}_\gamma$.

Les trois lois données par Fisher et Tippett sont caractérisées par le signe de γ. Il y a donc trois domaines d'attraction possibles. On présente ceux-ci en table 1 et on s'y référera pour le reste de ce travail.

Valeur de γ	Domaine d'attraction
$\gamma < 0$	Weibull
$\gamma = 0$	Gumbel
$\gamma > 0$	Fréchet

TABLE 1 – Domaine d'attraction selon le signe de γ

Exemple 1. Supposons que X suit une exponentielle de paramètre $\lambda > 0$, alors $F(x) = 1 - e^{-\lambda x}$ pour $x \geq 0$. En posant $a_n = \lambda^{-1}$ et $b_n = \lambda^{-1}\log(n)$, on a que

$$\begin{aligned}\lim_{n\to\infty} F^n(a_n x + b_n) &= \lim_{n\to\infty}\left(1 - \frac{e^{-\lambda a_n x}}{e^{\lambda b_n}}\right)^n \\ &= \lim_{n\to\infty}\left(1 - \frac{e^{-x}}{n}\right)^n \\ &= \exp\left(-e^{-x}\right).\end{aligned}$$

La loi exponentielle appartient donc au domaine de Gumbel, $F \in \mathcal{D}_0$.

2.2 Méthode des excès (POT)

On introduit une deuxième approche nommée la méthode des excès, ou *Peaks over Treshold* (POT) en anglais. Celle-ci offre un point de vue différent aux extrêmes et sera utile notamment dans la section 4 qui présente la stratégie de sélection. Soit $X_1, \ldots, X_n$, des variables aléatoires i.i.d. Plutôt que de regarder le maximum sur cet ensemble, on retient seulement les valeurs qui dépassent un certain seuil u. Les éléments du sous-ensemble correspondent aux excès de l'échantillon. Précisément, un excès est défini par la variable

$$Y = \begin{cases} X - u & \text{si } X > u \\ \emptyset & \text{sinon.}\end{cases}$$

La fonction de répartition des excès conditionnel au fait que $X > u$ est égale à

$$F_u(y) = \mathbb{P}\left(Y \leq y | X > u\right) = \frac{F(u+y) - F(u)}{1 - F(u)}.$$

On note la fonction de survie par $\overline{F}(x) = 1 - F(x)$. La fonction de survie associée à F_u est donc donnée par

$$\overline{F_u}(y) = \frac{\overline{F}(u+y)}{\overline{F}(u)}.$$

Lorsque le seuil u est suffisamment grand, il est possible d'approcher la fonction de survie $\overline{F_u}$ par celle d'une loi de Pareto généralisée (GPD) :

$$\overline{H}_{\gamma,\sigma}(y) = \begin{cases} \left(1 + \gamma\frac{y}{\sigma}\right)^{-1/\gamma} & \text{si } \gamma \neq 0 \\ \exp\left(-\frac{y}{\sigma}\right) & \text{sinon}\end{cases}$$

dont le support est $\mathbb{R}^+$ si $\gamma \geq 0$ ou $\left[0; -\frac{\sigma}{\gamma}\right]$ si $\gamma < 0$, où σ et γ sont les paramètres d'échelle et de forme respectivement. Plus formellement, il est possible de faire le lien avec le théorème 1 en utilisant la distribution GEV.

Théorème 2 (Pickands, 1975)**.** *Soit $X_1, \ldots, X_n$ i.i.d. de loi F avec comme point terminal x_F, alors pour des suites $a_n > 0$ et $b_n \in \mathbb{R}$ on a*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = G_\gamma(x)$$

si et seulement si

$$\lim_{u\to x_F} \sup_{y\in[0;x_F-u]} \left|\overline{F_u}(y) - \overline{H}_{\gamma,\sigma(u)}(y)\right| = 0.$$

On remarque que le paramètre de forme de la GPD γ est le même que celui de la GEV.

2.3 Caractérisation aux domaines d'attraction

Les domaines d'attraction sont au centre de ce travail. Il est donc important d'introduire quelques résultats qui les caractérisent. Comme vu en définition 1, l'appartenance à un domaine d'attraction dépend de l'existence des suites normalisantes a_n et b_n. Or cette existence est fortement liée au comportement de la fonction de survie. Les résultats suivants permettent de comprendre pourquoi.

Théorème 3 (Leadbetter et al., 1983)**.** *Soit une loi avec comme fonction de répartition F, $\tau \in [0, \infty]$ et une suite réelle u_n. Alors*

$$n\overline{F}(u_n) \to \tau \Leftrightarrow F^n(u_n) \to e^{-\tau}.$$

Théorème 4 (Leadbetter et al., 1983)**.** *Soit une fonction de répartition F et $\tau \in (0, \infty)$. Il existe une suite u_n satisfaisant la relation du théorème 3 si et seulement si*

$$\lim_{x \to x_F} \frac{\overline{F}(x)}{\overline{F}(x-)} = 1$$

avec x_F le point terminal de F et $F(x-)$ la limite à gauche de F.

Il est important de noter que u_n n'est pas nécessairement de la forme $a_n x + b_n$. Cependant si on peut montrer que la limite au théorème 4 n'existe pas pour n'importe quelles suites u_n, nécessairement il n'existe pas de suites normalisantes. Dans cette situation, la loi F ne pourra pas être dans un domaine d'attraction. On présentera plusieurs exemples de cette situation en Section 2.4.

Comme mentionné en introduction, l'hypothèse de ce travail repose sur l'idée que la théorie des valeurs extrêmes peut être utilisée pour mieux cibler les lois de mélanges dans le cadre de l'analyse de données de comptage. Dans un cadre général, la loi peut être continue ou discrète et avoir un support fini ou infini. Dans mon étude, je me restreindrai à des lois de mélange continues dont le support est $\mathbb{R}^+$. Ce choix implique que $x_F = \infty$ et que seuls les domaines d'attraction de Fréchet et Gumbel seront considérés. En effet, une condition nécessaire pour qu'une loi soit dans le domaine de Weibull est que celle-ci soit à queue finie. En se restreignant à $x_F = \infty$, on rejette donc ce domaine pour le reste de ce travail.

2.4 Valeurs extrêmes discrètes

Jusqu'à présent, aucune hypothèse n'a été faite sur la nature continue ou discrète de F. Alors que dans le cas continu, il est généralement possible d'identifier le domaine d'attraction, cela peut s'avérer compliqué, voire impossible, dans le cas discret.

Exemple 2. Prenons la version discrétisée de la loi exponentielle : la distribution géométrique. Cette dernière admet pour fonction de répartition $F(n) = 1 - (1-p)^n$ avec $n \in \mathbb{N}$, $x_F = \infty$ et $p \in (0, 1)$. Puisque c'est une loi discrète, $F(n-) = F(n-1)$ et par le théorème 4 :

$$\begin{aligned} \lim_{n \to \infty} \frac{\overline{F}(n)}{\overline{F}(n-1)} &= \lim_{n \to \infty} \frac{(1-p)^n}{(1-p)^{n-1}} \\ &= (1-p) < 1. \end{aligned}$$

Contrairement à l'exemple 1 où on a démontré que la loi exponentielle est dans $\mathcal{D}_0$, on a démontré que la loi géométrique n'a aucun domaine d'attraction. Ainsi, discrétiser une loi continue lorsque celle-ci est dans un domaine d'attraction ne préserve pas nécessairement cette propriété. Pour plus de résultats sur ce sujet, voir Shimura (2012).

Dans le contexte de ce travail, on regarde seulement les lois discrètes avec comme support $\mathbb{N}$. Une condition nécessaire pour une variable aléatoire discrète d'être dans un domaine d'attraction requiert la définition de queue longue.

Définition 2. La fonction de répartition F d'une variable aléatoire X est à queue longue, qu'on note $F \in \mathcal{L}$, si

$$\lim_{x\to\infty} \frac{\overline{F}(x+1)}{\overline{F}(x)} = 1.$$

Anderson (1970) remarque que si une variable aléatoire discrète est dans un domaine d'attraction, nécessairement la loi est à queue longue. Cette propriété concorde avec la démonstration de la loi géométrique vue précédemment. Précisément, on a le résultat suivant :

Théorème 5 (Anderson, 1970)**.** *Soit X une variable aléatoire discrète avec comme fonction de répartition F, une condition nécessaire pour que $F \in \mathcal{D}_\gamma$ avec un certain $\gamma \geq 0$ est $F \in \mathcal{L}$.*

Plusieurs mélanges de Poisson ne sont pas à queue longue. En effet, on aura souvent que

$$\lim_{n\to\infty} \frac{\overline{F}(n+1)}{\overline{F}(n)} = L \in (0,1).$$

Anderson (1970) et Shimura (2012) démontrent que des fonctions de survie discrètes satisfaisant cette limite proviennent d'une distribution continue dans $\mathcal{D}_0$ qui a été discrétisée. En revenant à l'exemple de la loi géométrique, on constate que c'est bien le cas. Cette propriété implique que si on ajuste une GEV (ou GPD) avec paramètre $\gamma = 0$ aux maximums (ou excès), ce sera une approximation raisonnable. Ce constat sera utilisé pour choisir une loi de mélange.

Shimura (2012) démontre également que toute loi continue à queue longue qui est dans le domaine de Fréchet ou de Gumbel restera dans son domaine une fois discrétisée. De plus, toutes les lois dans le domaine Fréchet sont à queue longue. Alors la propriété du domaine d'attraction est toujours préservée lorsqu'on discrétise des lois dans Fréchet. La section 3 est dédiée à l'étude de cette propriété d'héritage dans le contexte spécifique des lois de Poisson en mélange.

3 Mélanges de Poisson

On applique maintenant la théorie des valeurs extrêmes aux mélanges de Poisson. Pour construire un critère de sélection basé sur cette théorie, on souhaite avoir des liens entre le domaine d'attraction du paramètre λ de la loi Poisson et le mélange final. Pour ce faire, on introduit quelques notations et définitions utilisées pour le reste de ce travail. Ensuite on présente des conditions pour que le mélange de Poisson reste dans un domaine d'attraction. On termine ensuite en s'intéressant à des résultats et des exemples pour les cas Fréchet et Gumbel. Comme mentionné en section 2.3, on s'intéresse seulement à ces deux domaines car λ sera restreint au support $\mathbb{R}^+$.

3.1 Préliminaire

On notera par X la variable aléatoire qui a pour distribution un mélange de Poisson. Précisément on dit que X est un mélange de Poisson si $X|\lambda \sim \mathcal{P}(\lambda)$ et $\lambda \sim F$ où $\mathcal{P}(\lambda)$ est la loi Poisson de paramètre λ et F est une fonction de répartition. On notera F_X et p_X les fonctions de répartition et de masse de la variable X.

Il faut également préciser ce que "$\sim$" signifie selon les situations. Dans un contexte de variables aléatoires, cette notation signifie que la variable suit une loi donnée. Dans le cas où on a des fonctions, "$\sim$" indique une équivalence asymptotique. C'est-à-dire deux fonctions f et g sont asymptotiquement équivalentes si $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$ et on note cette relation par $f \sim g$.

Une propriété importante pour les distributions dans le domaine de Fréchet est celle de variation régulière. Cette définition sera utilisée dans la section 3.3.

Définition 3. Une fonction $f : \mathbb{R}^+ \to \mathbb{R}^+$ est à variation régulière dans un voisinage de ∞ avec indice $\alpha \in \mathbb{R}$, qu'on note $f \in \mathcal{RV}_\alpha$, si pour $x > 0$

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^\alpha.$$

Si $\alpha = 0$, on dit que f est à variation lente ($f \in \mathcal{RV}_0$). On remarque que si $f \in \mathcal{RV}_\alpha$ alors $f(x) = x^\alpha L(x)$ avec $L(x) \in \mathcal{RV}_0$.

Finalement, on définit ce qu'est une loi à comportement Gamma/Pareto. Ce type de distribution sera utilisée pour développer des résultats concernant les domaines d'attraction dans les mélanges.

Définition 4. Soit une variable aléatoire X ayant comme densité f, on dit que f a un comportement gamma si

$$f(x) \sim C(x)x^\alpha e^{-\beta x}$$

où $C(x)$ est localement bornée sur $(0, \infty)$ et à variation lente, $\beta > 0$ et $\alpha \in \mathbb{R}$. Pour $\beta = 0$, la densité f a un comportement Pareto avec $\alpha < -1$ si

$$f(x) \sim C(x)x^\alpha.$$

3.2 Préservation du domaine d'attraction

On étudie maintenant les situations où le domaine d'attraction de λ est conservé pour le mélange. Pour cela, on rappelle deux conditions suffisantes de Von Mises pour qu'une distribution soit dans le domaine de Fréchet ou de Gumbel.

Théorème 6. *1re Condition de Von Mises*
Supposons qu'une variable aléatoire X a comme fonction de répartition F absolument continue avec densité positive f avec $x_F = \infty$. Si pour $\gamma > 0$ on a

$$\lim_{x\to\infty} \frac{xf(x)}{1-F(x)} = \frac{1}{\gamma},$$

alors $F \in \mathcal{D}_\gamma$.

Théorème 7. *3e Condition de Von Mises*
Supposons qu'une variable aléatoire X a comme fonction de répartition F avec une 2e dérivée négative f' pour tout x avec $x_F = \infty$. Si on a

$$\lim_{x\to\infty} \frac{f'(x)(1-F(x))}{f^2(x)} = -1$$

ou de façon équivalente

$$\lim_{x\to\infty} \frac{d}{dx}\left(\frac{1-F(x)}{f(x)}\right) = 0,$$

alors $F \in \mathcal{D}_0$.

On peut enfin présenter le résultat clé pour les mélanges de Poisson qui a été démontré par Perline (1998).

Théorème 8 (Perline, 1998)**.** *Soit F la fonction de répartition du paramètre λ avec comme support $\mathbb{R}^+$. Supposons la densité associée f deux fois continue dérivable pour x assez grand. Si on a que :*

1. *F satisfait la 1re condition de Von Mises, alors $F_X \in \mathcal{D}_\gamma$ pour $\gamma > 0$.*
2. *F satisfait la 3e condition de Von Mises et le taux de défaillance est tel que pour $x \to \infty$ et un $\delta \geq \frac{1}{2}$,*

$$\frac{f(x)}{1-F(x)} = o(x^{-\delta}),$$

alors $F_X \in \mathcal{D}_0$.

Plusieurs lois continues dans le domaine de Fréchet satisfont la 1re condition de Von Mises. Donc, en général, utiliser une loi dans ce domaine pour λ permettra au mélange de rester dans celui-ci. Dans l'autre cas, il existe plusieurs distributions qui respectent la 3e condition de Von Mises, mais pas celle sur le taux de défaillance. Comme dans le travail sur les lois discrétisées de Shimura (2012), l'utilisation de lois dans $\mathcal{D}_0$ est plus problématique. Il est important de noter cependant que la condition sur le taux de défaillance est seulement une condition suffisante. L'absence de domaine d'attraction pour F_X n'est pas garantie si les conditions sont violées.

Pour s'assurer de l'absence de domaine, on va montrer que si on utilise une densité f avec un comportement gamma, alors le mélange de Poisson ne possèdera aucun domaine d'attraction. Pour ce faire, on utilisera une équivalence sur la fonction de masse p_X démontrée par Willmot (1990).

Lemme 1 (Willmot, 1990)**.** Supposons X un mélange de Poisson tel que λ a une densité f possédant un comportement gamma, alors on a pour $n \to \infty$

$$p_X(n) \sim \frac{C(n)}{(1+\beta)^{\alpha+n+1}} n^{\alpha}.$$

De plus, on utilisera le résultat de Stolz-Cesàro pour une limite d'un ratio de suites.

Lemme 2 (Stolz et Cesàro)**.** Soit deux suites A_n et B_n pour tout $n \in \mathbb{N}$. Supposons $\lim_{n\to\infty} A_n = \lim_{n\to\infty} B_n = 0$ et que B_n est strictement monotone. Si on a

$$\lim_{n\to\infty} \frac{A_{n+1} - A_n}{B_{n+1} - B_n} = L,$$

alors

$$\lim_{n\to\infty} \frac{A_n}{B_n} = L.$$

On peut enfin énoncer et démontrer l'absence de domaine d'attraction concernant les lois avec un comportement gamma.

Théorème 9. *Supposons X un mélange de Poisson tel que λ a une densité f possédant un comportement gamma, alors la fonction de répartition F_X n'appartiendra pas à $\mathcal{D}_\gamma$, et ce, pour tout $\gamma \in \mathbb{R}$.*

Démonstration. On rappelle que pour qu'une distribution discrète soit dans un domaine d'attraction, il faut que celle-ci soit à queue longue (Théorème 5). Il suffit donc de montrer que

$$\lim_{n\to\infty} \frac{\overline{F}_X(n+1)}{\overline{F}_X(n)} = L < 1.$$

Pour ce faire, on étudie la limite définie par Stolz-Cesaro :

$$\begin{aligned}
\lim_{n\to\infty} \frac{\overline{F}_X(n+2) - \overline{F}_X(n+1)}{\overline{F}_X(n+1) - \overline{F}_X(n)} &= \lim_{n\to\infty} \frac{p_X(n+2)}{p_X(n+1)} \\
&= \lim_{n\to\infty} \frac{C(n+2)}{C(n+1)} \frac{1}{1+\beta} \left(\frac{n+2}{n+1}\right)^{\alpha} \\
&= \frac{1}{1+\beta} \lim_{n\to\infty} \frac{C(n+2)}{C(n+1)}
\end{aligned}$$

où on a utilisé le lemme 1 à la deuxième égalité. On étudie maintenant la limite du ratio restant. Parce que C est une fonction à variation lente, on utilise la représentation de Karamata pour démontrer la convergence. On rappelle que toute fonction à variation lente peut être représentée par :

$$C(x) = c(x) \exp\left[\int_1^x t^{-1}\eta(t)dt\right]$$

avec $x > 0$, $c(x)$ et $\eta(x)$ des fonctions de $\mathbb{R}^+$ dans $\mathbb{R}^+$ telles que $\lim_{x\to\infty} c(x) = c > 0$ et $\lim_{x\to\infty} \eta(x) = 0$ (Bingham et al., 1987). La limite peut donc s'écrire comme

$$\lim_{n\to\infty} \frac{C(n+2)}{C(n+1)} = \lim_{n\to\infty} \exp\left[\int_{n+1}^{n+2} t^{-1}\eta(t)dt\right].$$

Puisque la fonction $\eta(x)$ tend vers 0 et est dans $\mathbb{R}^+$, alors pour tout $\varepsilon > 0$, il existe un $N \in \mathbb{N}$ tel que pour tout $n \geq N$ on a

$$0 < \eta(n) < \varepsilon.$$

Donc pour $n \geq N$,

$$0 \leq \int_{n+1}^{n+2} t^{-1}\eta(t)dt \leq \varepsilon \int_{n+1}^{n+2} t^{-1}dt = \varepsilon \log\left(\frac{n+2}{n+1}\right) \leq \varepsilon \log(2).$$

Ainsi l'intégrale tend vers 0 et donc

$$\lim_{n\to\infty} \frac{C(n+2)}{C(n+1)} = 1.$$

Finalement, puisque $\overline{F}_X$ est strictement décroissante et tend vers 0, le théorème de Stolz-Cesàro (lemme 2) permet de conclure que

$$\lim_{n\to\infty} \frac{\overline{F}_X(n+1)}{\overline{F}_X(n)} = \frac{1}{1+\beta} \leq 1$$

et on a l'égalité si et seulement si $\beta = 0$. Autrement dit, on a la propriété de queue longue seulement si la densité est à comportement Pareto. □

3.3 Domaine d'attraction Fréchet ($\gamma > 0$)

Comme mentionné, le résultat de Perline permet d'utiliser plusieurs lois dans le domaine de Fréchet qui satisfont la 1re condition de Von Mises et de rester dans celui-ci. Il en est de même pour les mélanges utilisant une densité avec un comportement Pareto. Premièrement, on observe que

$$\lim_{t\to\infty} \frac{f(xt)}{f(t)} = \lim_{t\to\infty} \frac{C(xt)}{C(t)} \left(\frac{xt}{t}\right)^{\alpha} = x^{\alpha}.$$

Donc $f \in \mathcal{RV}_\alpha$ avec $\alpha < -1$ et on peut montrer que la fonction de survie $\overline{F} \in \mathcal{RV}_{\alpha+1}$. Une condition nécessaire et suffisante pour que $F \in \mathcal{D}_\gamma$ avec $\gamma > 0$ est que $\overline{F} \in \mathcal{RV}_{-\frac{1}{\gamma}}$ (Resnick, 1987). Dans ce cas, en posant $\gamma = -(1+\alpha)^{-1}$, on est dans le domaine de Fréchet et il est possible que le mélange Poisson reste dans un domaine d'attraction quelconque grâce au théorème 9 puisque $\beta = 0$. Il ne reste qu'à montrer que c'est le cas et ce domaine est celui de Fréchet.

Théorème 10. *Soit X un mélange de Poisson tel que λ a une densité f avec un comportement Pareto (donc $\alpha < -1$), alors pour $\gamma = -(1+\alpha)^{-1}$, F et F_X sont dans $\mathcal{D}_\gamma$.*

Démonstration. La preuve utilise plusieurs propriétés des fonctions à variation régulière. Pour plus de détails, voir Resnick (1987). Si $\gamma = -(1+\alpha)^{-1}$, alors $\alpha = -(\gamma^{-1}+1)$. Comme $f \in \mathcal{RV}_\alpha$ où $\alpha < -1$, on a que $\overline{F} \in \mathcal{RV}_{\alpha+1}$ et donc $F \in \mathcal{D}_\gamma$. On a aussi que

$$\lim_{x\to\infty} \frac{xf(x)}{1-F(x)} = -\alpha - 1 = \frac{1}{\gamma}.$$

Alors la 1re condition de Von Mises est satisfaite et on peut conclure que $F_X \in \mathcal{D}_\gamma$ par le théorème 8 de Perline. □

On peut démontrer le théorème 10 d'une autre façon en utilisant des techniques employées par Anderson (1970) et Hitz et al. (2017). Cette démonstration est présentée en annexe.

3.4 Domaine de Gumbel ($\gamma = 0$)

Les lois dans le domaine Gumbel sont plus problématiques que celles dans Fréchet. Comme Perline (1998) l'a démontré pour le cas Gumbel, il faut une condition suffisante sur le taux de défaillance qui n'est pas satisfaite par plusieurs lois usuelles. On peut utiliser une technique analytique pour rapidement évaluer si une loi satisfait la 3e condition de Von Mises, mais ne respecte pas la condition sur le taux de défaillance.

Théorème 11. *Soit X une variable aléatoire avec une densité f dérivable. Si pour $c > 0$ on a $f'(x) \sim -cf(x)$, alors la 3e condition de Von Mises est satisfaite (donc $F \in \mathcal{D}_0$) et pour un $\delta \geq \frac{1}{2}$ quelconque*

$$\lim_{x\to\infty} \frac{x^\delta f(x)}{1 - F(x)} = \infty.$$

Démonstration. En regardant la 3e condition de Von Mises, on a avec l'hypothèse sur la densité que

$$\begin{aligned}\lim_{x\to\infty} \frac{f'(x)(1 - F(x))}{f^2(x)} &= \lim_{x\to\infty} \frac{-c(1 - F(x))}{f(x)} \\ &= c \lim_{x\to\infty} \frac{f(x)}{f'(x)} \\ &= c \lim_{x\to\infty} \frac{f(x)}{-cf(x)} = -1\end{aligned}$$

où on a utilisé l'Hôpital en deuxième égalité. On a montré que la loi est dans $\mathcal{D}_0$ et que $\frac{f(x)}{1-F(x)} \to c > 0$ lorsque $x \to \infty$. On a donc que

$$\lim_{x\to\infty} \frac{x^\delta f(x)}{1 - F(x)} = c \lim_{x\to\infty} x^\delta = \infty$$

ce qui démontre que la condition sur le taux de défaillance du théorème 8 n'est pas satisfaite. □

En section 3.2 on a démontré au théorème 9 que si on utilise une densité $f(x) \sim C(x)x^\alpha e^{-\beta x}$ et que $\beta > 0$ (donc $\alpha \in \mathbb{R}$), alors le mélange de Poisson n'aura pas de domaine d'attraction. Précisément, on a démontré que le mélange de Poisson résultant est tel que $F_X \notin \mathcal{L}$. Cependant, on a montré que

$$\lim_{x\to\infty} \frac{\overline{F_X}(x + 1)}{\overline{F_X}(x)} = \frac{1}{1 + \beta} \in (0, 1).$$

Anderson (1970) et Shimura (2012) ont démontré que des distributions discrètes satisfaisant cette limite proviennent de lois continues dans le domaine de Gumbel qu'on a discrétisées. De plus, même si de telles lois ne sont pas dans le domaine de Gumbel, Anderson démontre qu'il est raisonnable d'ajuster les maximas par une Gumbel. Rigoureusement, il démontre le résultat suivant.

Théorème 12 (Anderson, 1970)**.** *Soit X une variable aléatoire discrète avec comme support $\mathbb{N}$ et une fonction de répartition F. Alors pour $\xi > 0$, pour tout x et une certaine suite de constantes b_n,*

$$\lim_{n\to\infty} \frac{\overline{F}(n + 1)}{\overline{F}(n)} = e^{-\xi}$$

si et seulement si

$$\begin{aligned}\liminf F^n(x + b_n) &\geq \exp\left[-e^{-\xi(x-1)}\right] \\ \limsup F^n(x + b_n) &\leq \exp\left[-e^{-\xi x}\right]\end{aligned}$$

En posant $\xi = \log(1+\beta)$, le théorème 12 est applicable aux mélanges de Poisson qui utilisent une densité à comportement gamma. Cette collection de résultats permettront d'établir une stratégie de sélection en section 4.

3.5 Exemples de mélanges

On s'intéresse maintenant à plusieurs exemples de lois possibles pour le mélange Poisson et on utilisera les théorèmes développés pour conclure l'existence ou non d'un domaine d'attraction pour le F_X résultant. Tous les exemples qui suivent auront $\mathbb{R}^+$ comme support.

3.5.1 Loi Fréchet

L'intensité λ suit une Fréchet de paramètre $\alpha > 0$ si elle a pour densité

$$f(x) = \alpha x^{-\alpha-1} e^{-x^{-\alpha}}.$$

Parce que $C(x) = \alpha e^{-x^{-\alpha}}$ est dans $\mathcal{RV}_0$ et est bornée sur $(0,\infty)$, on a les conditions nécessaires du théorème 10 et on conclut que F et F_X sont dans $\mathcal{D}_{\frac{1}{\alpha}}$.

3.5.2 Loi Demi-Cauchy

Soit C une variable aléatoire suivant une Cauchy de paramètres $\mu \in \mathbb{R}$, $\sigma > 0$ et avec comme fonction de répartition G et de densité

$$g(x) = \frac{1}{\pi\sigma\left(1 + (\frac{x-\mu}{\sigma})^2\right)}.$$

Le paramètre λ suit une Demi-Cauchy si $\lambda = |C|$. Les fonctions de répartition et de densité de λ sont les suivantes :

$$\begin{aligned} F(x) &= \mathbb{P}\left(|C| \leq x\right) \\ &= \mathbb{P}\left(-x \leq C \leq x\right) \\ &= G(x) - G(-x), \end{aligned}$$

$$\text{et } f(x) = g(x) + g(-x)$$

Parce que $(1 + \frac{(x+\mu)^2}{\sigma^2})^{-1} \sim \left(\frac{x}{\sigma}\right)^{-2}$, on a que $f(x) \sim \frac{2\sigma}{\pi} x^{-2}$, un comportement Pareto avec $\alpha = -2$. Par le théorème 10, on peut conclure F et F_X seront dans $\mathcal{D}_1$.

3.5.3 Loi Weibull

À ne pas confondre avec le domaine d'attraction, la loi Weibull avec paramètres de forme $\alpha > 0$ et d'échelle $\beta > 0$ a comme fonctions de survie et de densité :

$$F(x) = 1 - \exp\left(-\frac{x^\alpha}{\beta^\alpha}\right),$$
$$\text{et } f(x) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x^\alpha}{\beta^\alpha}\right).$$

À l'exception du cas où $\alpha = 1$, le densité n'est pas à comportement Gamma. On peut donc seulement vérifier si les conditions de Perline sont respectées. La dérivée de la densité est

$$f'(x) = f(x)\left(\frac{\alpha-1}{x} - \frac{\alpha}{\beta^\alpha}x^{\alpha-1}\right) \sim -\frac{\alpha}{\beta^\alpha}x^{\alpha-1}f(x)$$

et ainsi la 3e condition de Von Mises est respectée car

$$\lim_{x\to\infty}\frac{f'(x)(1-F(x))}{f^2(x)} = \lim_{x\to\infty}\frac{-\frac{\alpha}{\beta^\alpha}x^{\alpha-1}\exp\left(-\frac{x^\alpha}{\beta^\alpha}\right)}{f(x)} = -1.$$

Pour la condition sur le taux de défaillance, on fixe $\delta = \frac{1}{2}$ et on a que

$$\lim_{x\to\infty}\frac{x^\delta f(x)}{1-F(x)} = \lim_{x\to\infty}\frac{\alpha}{\beta^\alpha}x^{\alpha-\frac{1}{2}}.$$

La limite tend vers 0 si $\alpha \in (0, \frac{1}{2})$. Pour un tel paramètre de forme, F et F_X sont dans $\mathcal{D}_0$. Pour $\alpha \geq \frac{1}{2}$, il n'est pas garanti d'être encore dans le domaine de Gumbel.

3.5.4 Loi lognormale

Perline (1998) présente la lognormale comme un exemple dans Gumbel qui reste dans ce domaine après le mélange. Par définition, une loi lognormale de paramètres $\mu \in \mathbb{R}$ et $\sigma > 0$ a comme fonction de répartition

$$F(x) = \Phi\left(\frac{\log x - \mu}{\sigma}\right)$$

où Φ est la fonction de répartition de la normale standard. En posant ϕ la densité de la normale standard, on a que la densité et sa dérivée sont

$$f(x) = \frac{1}{\sigma x}\phi\left(\frac{\log x - \mu}{\sigma}\right),$$
$$f'(x) = -\frac{f(x)}{x}\left(1 + \frac{\log x - \mu}{\sigma^2}\right).$$

Comme pour la loi de Weibull, la lognormale n'a pas une densité à comportement gamma. On regarde donc si les conditions de Perline sont satisfaites. Une propriété utile pour les limites à venir est l'équivalence asymptotique suivante :

$$1 - \Phi(x) \sim \frac{\phi(x)}{x}.$$

La 3e condition de Von Mises est satisfaite car

$$\lim_{x\to\infty}\frac{f'(x)\,(1-F(x))}{f^2(x)} = -\lim_{x\to\infty}\frac{\sigma^2 + \log x - \mu}{\log x - \mu} = -1$$

et pour $\delta = \frac{1}{2}$, la condition sur le taux de défaillance est satisfaite aussi puisque

$$\lim_{x\to\infty}\frac{x^\delta f(x)}{1-F(x)} = \lim_{x\to\infty}\frac{\log x - \mu}{\sigma^2\sqrt{x}} = 0.$$

Ainsi la lognormale est bien une loi qui permet à F et F_X d'être dans $\mathcal{D}_0$.

3.5.5 Lois Kummer Type II, Gamma et Bêta Type II

L'utilisation de la loi Kummer Type II peut être intéressante car elle donne une fonction de masse fermée. De plus, cette loi a comme cas particuliers la Gamma et la Bêta Type II.

Définition 5. Une variable aléatoire Y suit une Kummer Type II , notée $Y \sim \mathcal{K}_2(a, b, c, \sigma)$, si elle a pour densité

$$f(y) = \frac{\sigma^b y^{a-1} \exp\left(-\frac{cy}{\sigma}\right)}{\Gamma(a)\psi(a, 1-b, c)(y+\sigma)^{a+b}}$$

avec comme support $\mathbb{R}^+$ et $a, c, \sigma > 0$, $b \in \mathbb{R}$ et ψ la fonction hypergéométrique de type II définie par

$$\psi(\alpha_1, \alpha_2, \alpha_3) = \frac{1}{\Gamma(\alpha_1)} \int_0^\infty t^{\alpha_1 - 1}(1+t)^{\alpha_2 - \alpha_1 - 1} e^{-\alpha_3 t} dt$$

pour $\alpha_1, \alpha_3 > 0$ et $\alpha_2 \in \mathbb{R}$.

On remarque les propriétés suivantes de cette distribution :

1. Si $b = -a$, alors $Y \sim Ga\left(a, \frac{c}{\sigma}\right)$, une loi Gamma qui a pour densité

   $$f(x) = \frac{\beta^a}{\Gamma(a)} x^{a-1} e^{-\beta x}$$

   où $\beta = \frac{c}{\sigma}$.

2. Pour $b > 0$ et lorsque $c \to 0$, on peut montrer que

   $$\psi(a, 1-b, 0) = \frac{\Gamma(b)}{\Gamma(a+b)}$$

   ce qui démontre que $Y \sim \mathcal{B}_2(a, b, \sigma)$, une loi Bêta Type II qui a pour densité

   $$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\sigma^b x^{a-1}}{(x+\sigma)^{a+b}}.$$

Maintenant on suppose que $\lambda \sim \mathcal{K}_2(a, b, c, \sigma)$ et on calcule la fonction de masse du mélange de Poisson X.

$$p_X(n) = \int_0^\infty \frac{\lambda^n e^{-\lambda}}{n!} f(\lambda) d\lambda = \frac{\sigma^b}{n!\Gamma(a)\psi(a, 1-b, c)} \int_0^\infty \frac{\lambda^{a+n-1}}{(\lambda+\sigma)^{a+b}} e^{-\lambda\left(\frac{c}{\sigma}+1\right)} d\lambda.$$

En faisant un changement de variable, on peut montrer que

$$\int_0^\infty \frac{\lambda^{a+n-1}}{(\lambda+\sigma)^{a+b}} e^{-\lambda\left(\frac{c}{\sigma}+1\right)} d\lambda = \Gamma(a+n)\sigma^{n-b}\psi\left(a+n, 1-b+n, \sigma+c\right).$$

On conclut que la fonction de masse est

$$p_X(n) = \frac{\Gamma(a+n)}{\Gamma(a)n!} \sigma^n \frac{\psi(a+n, 1-b+n, \sigma+c)}{\psi(a, 1-b, c)}.$$

En revenant sur nos cas particuliers, on a :

1. Si $b = -a$, $\lambda \sim Ga\left(a, \frac{c}{\sigma}\right)$ et on peut montrer que $X \sim NB\left(a, \frac{c}{c+\sigma}\right)$, une loi binomiale négative.
2. Si $b > 0$ et $c \to 0$, alors $\lambda \sim \mathcal{B}_2(a, b, \sigma)$ et on a que X a comme fonction de masse

$$p_X(n) = \frac{\Gamma(a+n)\Gamma(a+b)}{\Gamma(a)\Gamma(b)n!}\sigma^n\psi(a+n, 1-b+n, \sigma).$$

On peut montrer que la dérivée de la densité est égale à

$$f'(x) = f(x)\left(\frac{a-1}{x} - \frac{a+b}{x+\sigma} - \frac{c}{\sigma}\right) \sim -\frac{c}{\sigma}f(x)$$

et par le théorème 11 on a que la Kummer Type II est dans $\mathcal{D}_0$ et la condition de Perline sur le taux de défaillance n'est pas satisfaite.

Pour conclure l'analyse de cette loi, on utilise le théorème 10 pour vérifier s'il existe un domaine d'attraction au mélange de Poisson. Parce que $\lambda \sim \mathcal{K}_2(a, b, c, \sigma)$, on a

$$\begin{aligned} f(x) &= \frac{\sigma^b}{\Gamma(a)\psi(a, 1-b, c)} \frac{x^{a-1}}{(x+\sigma)^{a+b}} \exp\left(-\frac{cx}{\sigma}\right) \\ &\sim \frac{\sigma^b}{\Gamma(a)\psi(a, 1-b, c)} x^{-b-1} \exp\left(-\frac{cx}{\sigma}\right) \end{aligned}$$

où on a utilisé le fait que $(x+\sigma)^{a+b} \sim x^{a+b}$. La densité f a un comportement gamma et on peut conclure que F_X ne sera pas dans un domaine d'attraction. Par contre, lorsque $b > 0$ et $c \to 0$, on aura que $\lambda \sim \mathcal{B}_2(a, b, \sigma)$ et cette fois F_X pourrait être dans un domaine d'attraction. En effet, avec ces paramètres la densité f aura un comportement Pareto. Plus précisément, on aura que F et F_X seront dans $\mathcal{D}_{\frac{1}{b}}$.

3.5.6 Lois Inverse Gaussienne Généralisée, Inverse Gaussienne et Gamma Inverse

Le paramètre λ suit une Inverse Gaussienne généralisée si la densité est

$$f(x) = \frac{\left(\frac{a}{b}\right)^{p/2}}{2K_p(\sqrt{ab})} x^{p-1} \exp\left[-\frac{ax}{2} - \frac{b}{2x}\right]$$

avec $x > 0$, $a, b > 0$, $p \in \mathbb{R}$ et K_p la fonction modifiée de Bessel de type II.

On peut montrer qu'on a les cas particuliers suivants :

1. Pour $p = -0.5$, on a la loi Inverse Gaussienne.
2. Pour $b \to 0$, on a la loi Gamma.
3. Pour $a \to 0$, on a la loi Gamma Inverse.

En dérivant la densité, on obtient que

$$f'(x) = f(x)\left(\frac{p-1}{x} + \frac{b}{2x^2} - \frac{a}{2}\right) \sim -\frac{a}{2}f(x).$$

Par le théorème 11, la troisième condition de Von Mises est satisfaite, mais la condition sur le taux de défaillance ne l'est pas. Finalement, on remarque que la fonction $\exp\left[-\frac{b}{2x}\right]$ est dans $\mathcal{RV}_0$ et est

strictement croissante et tend vers 1 pour $x \to \infty$. Alors cette fonction est bornée par 1 et donc localement bornée sur $(0, \infty)$. En posant

$$C(x) = \frac{\left(\frac{a}{b}\right)^{p/2}}{2K_p(\sqrt{ab})} \exp\left[-\frac{b}{2x}\right],$$

on a que la densité de λ a la forme

$$f(x) = C(x)x^{p-1} \exp\left[-\frac{ax}{2}\right].$$

Alors un mélange de Poisson utilisant une loi Inverse Gaussienne généralisée n'est pas dans un domaine d'attraction par le théorème 9. Cependant, lorsque $a \to 0$, notre mélange utilise une loi Gamma-Inverse et la densité aura un comportement Pareto. Dans ce cas F et F_X seront dans le domaine de Fréchet.

3.5.7 Résumé des exemples

On conclut cette section en présentant une table des exemples précédents qui résume les domaines d'attraction pour la loi sur λ et pour le mélange Poisson X. On a présenté plusieurs exemples où on n'a pas de domaine d'attraction pour X lorsqu'on utilise des lois dans Gumbel. Par contre, comme mentionné en section 3.4, ces lois ont des maximums « proches » du domaine d'attracion de Gumbel. On notera cette distinction par ≈Gumbel dans la table.

Loi de mélange	Domaine (Loi)	Mélange de Poisson	Domaine (Mélange)
Fréchet	Fréchet	Poisson-Fréchet	Fréchet
Demi-Cauchy	Fréchet	Poisson Demi-Cauchy	Fréchet
Gamma Inverse	Fréchet	Poisson Gamma Inverse	Fréchet
Bêta Type II	Fréchet	Poisson Bêta II	Fréchet
Log-Gaussienne	Gumbel	PLG	Gumbel
Weibull(α, β)	Gumbel	Poisson-Weibull	Gumbel (si $\alpha < \frac{1}{2}$)
Gamma(α,β)	Gumbel	Binomiale Négative	≈Gumbel
Inverse-Gaussienne	Gumbel	Sichel	≈Gumbel
Inverse-Gaussienne Générale	Gumbel	PGIG	≈Gumbel
Kummer Type II	Gumbel	PK II	≈Gumbel

TABLE 2 – Domaines d'attraction pour les exemples

4 Stratégie de sélection

4.1 Estimation du maximum de vraisemblance

Maintenant qu'on a plusieurs outils théoriques concernant les valeurs extrêmes et les mélanges de Poisson, on présente une stratégie pour sélectionner quelques lois de mélange adéquates. Pour ce faire, on suppose qu'on possède un ensemble de données de comptage et on procède à une analyse des extrêmes. Comme présenté en section 2.2, une approche possible est de considérer les excès pour un seuil assez élevé. Plusieurs stratégies existent pour choisir un tel seuil, dans ce travail on le fixe tel qu'une proportion donnée des observations soit au dessus de celui-ci.

On rappelle que la distribution des excès conditionnelle au fait qu'ils soient positifs peut être approchée par la Pareto Généralisée (GPD) qui a pour fonction de survie

$$\overline{H}_{\gamma,\sigma}(y) = \begin{cases} \left(1+\gamma\frac{y}{\sigma}\right)^{-1/\gamma} & \text{si } \gamma \neq 0 \\ \exp\left(-\frac{y}{\sigma}\right) & \text{sinon.} \end{cases}$$

Lorsque $\gamma \neq 0$, la densité de la GPD est

$$h_{\gamma,\sigma}(y) = \frac{1}{\sigma}\left(1+\frac{\gamma}{\sigma}y\right)^{-(1+\frac{1}{\gamma})}.$$

Supposons maintenant qu'on a fixé notre seuil et qu'on a n excès $Y_1, \ldots, Y_n$ de nos données de comptage. On souhaite identifier le domaine d'attraction, donc il faut estimer le paramètre de forme γ. Pour ce faire, il faut maximiser la log vraisemblance définie par

$$l(\sigma,\gamma) := -n\log\sigma - \left(1+\frac{1}{\gamma}\right)\sum_{i=1}^{n}\log\left(1+\gamma\frac{y_i}{\sigma}\right).$$

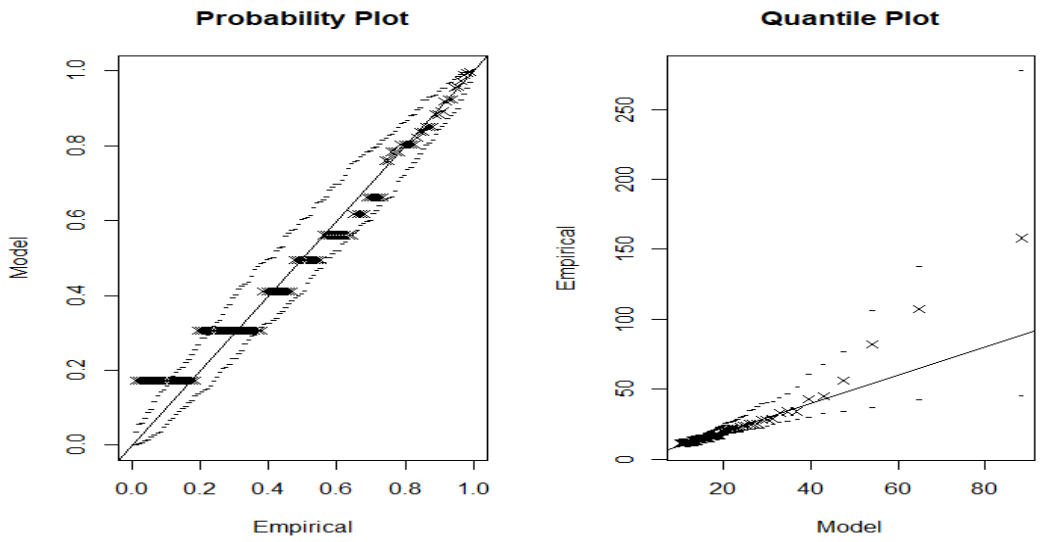

FIGURE 2 – Diagnostic graphique de l'ajustement GPD des excès Poisson-Fréchet

Le package *evd* sur R permet de produire l'estimation désirée. Une fois que c'est fait, on pourra analyser si l'ajustement est adéquat grâce aux graphiques qui comparent la fonction de répartition

(ou quantile) empirique et théorique. Comme exemple, on simule 10000 observations d'une Poisson-Fréchet avec un paramètre de forme $\alpha = 2$. En théorie on devrait avoir une estimation proche de $\gamma = \frac{1}{2}$. En fixant un seuil pour avoir au moins 100 excès, donc le 99^{e} percentile empirique, on obtient une estimation $\hat{\gamma} = 0.4383$ avec comme écart-type 0.1216. Graphiquement, on peut conclure que l'ajustement est adéquat (cf Figure 2).

4.2 Traitement du cas Gumbel

Plusieurs mélanges de Poisson dans les exemples en section 3.5 ne sont pas dans Gumbel, mais sont proches de ce domaine grâce aux résultats d'Anderson (1970) et Shimura (2012). Il est donc raisonnable de penser que si on possède des réalisations de ces lois et qu'on les transforme en variables aléatoires continues, alors l'ajustement des excès sera dans Gumbel aussi.

Une stratégie possible serait d'ajouter un bruit aléatoire aux données discrètes et de tester si l'ajustement d'une GPD avec $\gamma = 0$ est adéquat. En fixant le paramètre $\gamma = 0$, la densité change pour

$$h_{0,\sigma}(y) = \frac{1}{\sigma} \exp\left(-\frac{y}{\sigma}\right)$$

et donc la log vraisemblance devient cette fois

$$l(\sigma) := -n \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{n} y_i.$$

On test cette approche en simulant 10000 observations d'une Poisson-Gamma(1,1), c'est-à-dire une binomiale négative. Avec un seuil similaire à la simulation précédente et en posant $\gamma = 0$, on observe en figure 3 que la GPD n'est pas du tout adéquate pour les données. Bien sûr, on n'est pas surpris puisqu'en théorie la négative binomiale n'a pas de domaine d'attraction.

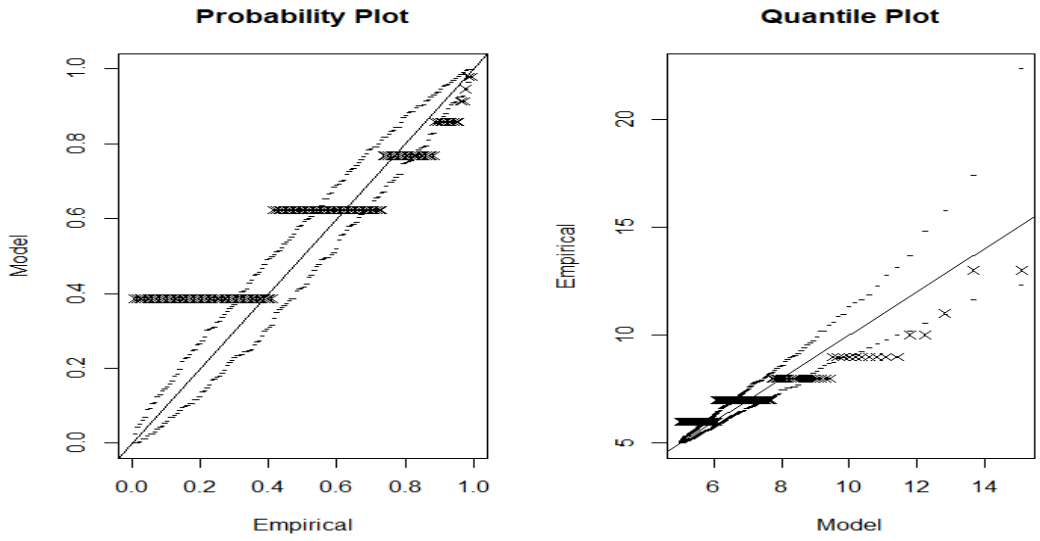

FIGURE 3 – Diagnostic graphique de l'ajustement GPD des excès binomiale négative

Maintenant, pour chaque observation discrète on ajoute un bruit aléatoire uniforme sur $(-0.5, 0.5)$ pour se ramener à une variable continue. En appliquant l'estimation de vraisemblance à ces données, on a une meilleure approximation des excès en Figure 4. Il est donc possible de traiter les cas Gumbel malgré l'absence de domaine d'attraction.

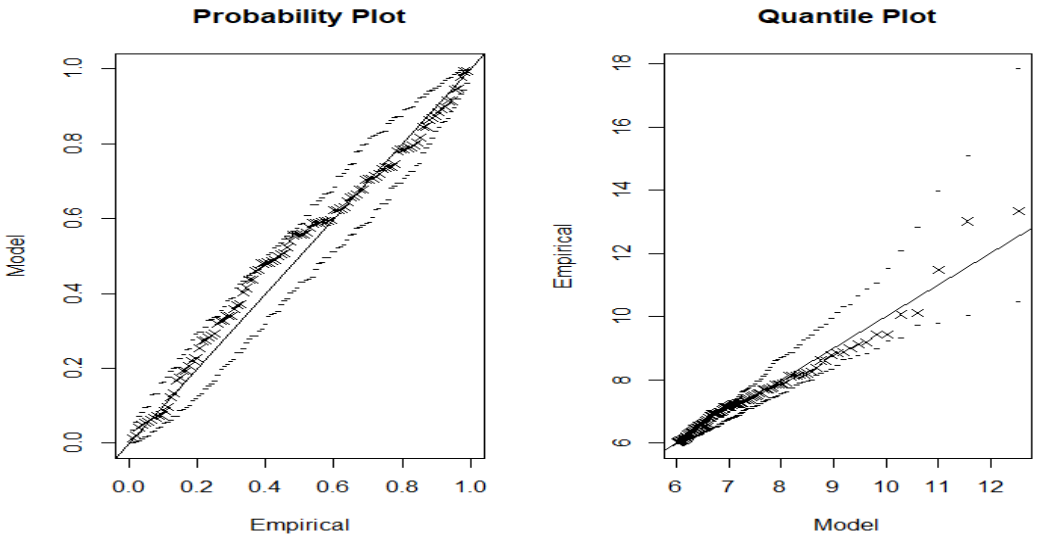

FIGURE 4 – Diagnostic graphique de l'ajustement GPD des excès binomiale négative avec bruits aléatoires

4.3 Arbre de décision pour la loi de mélange

On peut enfin avoir un méthode de sélection de la loi de mélange basée sur la théorie des extrêmes. On suppose encore une fois qu'on a des données de comptage indépendantes et identiquement distribuées. En premier lieu, on teste si la loi Poisson est appropriée pour les données. Si c'est le cas, le mélange est inutile. Broek (1995) et Yang et al. (2010) présentent des méthodes pour faire ce type de test.

On suppose maintenant qu'on a bien une surdispersion dans les données de comptage. Il suffit de fixer un seuil assez élevé pour les observations et appliquer la méthode des excès pour identifier un domaine d'attraction. On a deux situations possibles : l'ajustement des excès est graphiquement adéquat ou non. Dans le premier cas, on pourra regarder le paramètre de forme γ et tester si celui-ci est proche de 0 ou non. Par exemple, si on a une estimation de γ relativement proche de 0 avec un grand écart-type, on peut fixer $\gamma = 0$ et voir si le modèle reste adéquat. Une fois qu'on a identifié un domaine d'attraction, on peut sélectionner une loi de mélange qui respecte les conditions de Perline appropriées. Dans le cas où $\gamma > 0$, il suffit de prendre une loi dans Fréchet, par exemple la demi-Cauchy. Pour $\gamma = 0$, seulement les lois avec la 3^e condition de Von Mises et la condition sur le taux de défaillance peuvent être sélectionnées. La lognormale serait un bon choix par exemple.

Pour le deuxième cas, on peut appliquer la technique présentée à la section 4.2 pour voir si on a des données de comptage proches du domaine de Gumbel. En ajoutant un bruit aléatoire uniforme sur $(-0.5, 0.5)$ aux observations et en réajustant les excès à une GPD de paramètre $\gamma = 0$, on teste si le modèle est cette fois adéquat. Si c'est le cas, on pourra utiliser des lois avec une densité à comportement gamma. En effet, on sait que de telles lois donnent des mélanges de Poisson proches

de Gumbel grâce au théorème 12 démontré par Anderson. Si ce n'est pas le cas, on ne pourra pas utiliser la théorie des valeurs extrêmes pour faire un choix éclairé. Il faudra donc prendre une des approches classiques présentées en introduction. La stratégie établie jusqu'à présent peut être visualisée avec l'arbre de décision en Figure 5 et sera utilisée sur des données en section 4.4.

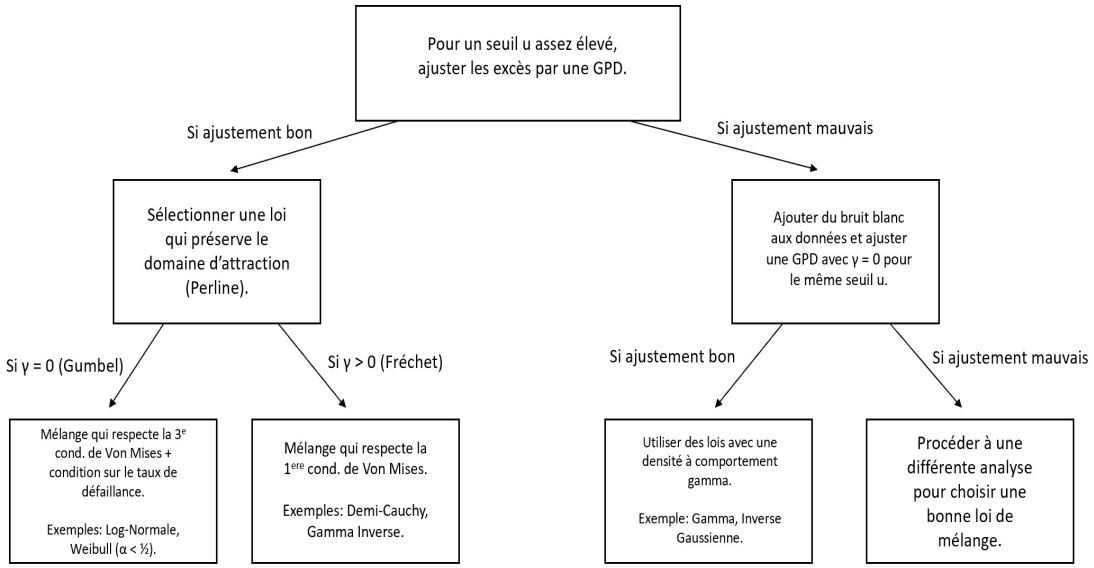

FIGURE 5 – Arbre de décision pour sélectionner des lois de mélange potentielles

4.4 Application

Pour évaluer l'approche proposée, on s'est intéressé à l'abondance de deux espèces forestières communes présentes en Afrique Centrale : le *Musanga Cecropioides* (Parasolier) et le *Macaranga spp.* Ces deux espèces pionnières sont communes aux forêts tropicales humides et leurs présences sont le marqueur de perturbations anthropiques récentes. Les abondances de chacune de ces deux espèces ont été obtenues par l'échantillonage de 1571 quadrats de 10 km $\times$ 10 km sur l'ensemble des forêts d'Afrique Centrale.

La stratégie établie (cf section 4.3), suggère que le Musanga serait potentiellement mieux modélisé par une loi de mélange dont le domaine d'attraction appartient à celui de Gumbel, alors qu'une loi appartenant au domaine de Frechet serait prévilégiée, selon notre approche, pour l'analyse de l'abondance du Macaranga. Pour étudier la loi de distribution des valeurs extrêmes de ces deux espèces, l'approche des excès a été utilisée (cf section 2.2).

Cas du Musanga : Pour le Parasolier, le seuil est estimé à $u = 60$. Cela conduit à l'analyse de 98 excès (cf Figure 6). L'ajustement d'une loi GPD par maximum de vraisemblance permet d'estimer le paramètre de forme $\hat{\gamma} = -0.022$, avec un écart-type de 0.077. L'intervalle de confiance contient donc 0. On ajuste donc les excès avec une GPD de paramètre $\gamma = 0$. Graphiquement, l'ajustement de ces excès semble bon (cf Figure 7). On peut conclure que la distribution des abondances du Parasolier appartient au domaine d'attraction de Gumbel. Donc, choisir une ou des lois de mélange dont le domaine d'attraction appartient à celui de Gumbel (par exemple la loi lognormale, Weibull avec un paramètre de forme plus petit que $\frac{1}{2}$), semble pertinent.

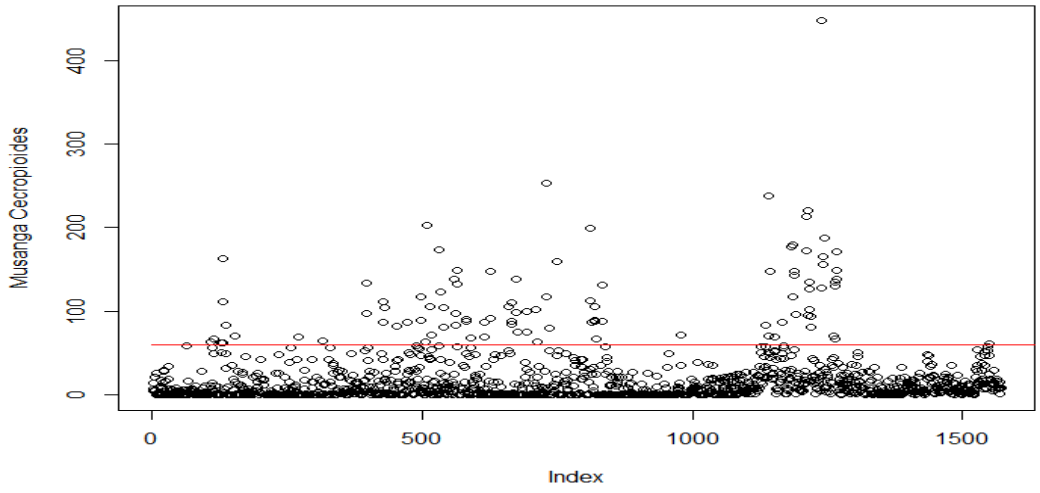

FIGURE 6 – Comptage des Musanga Cecropioides avec un seuil de 60

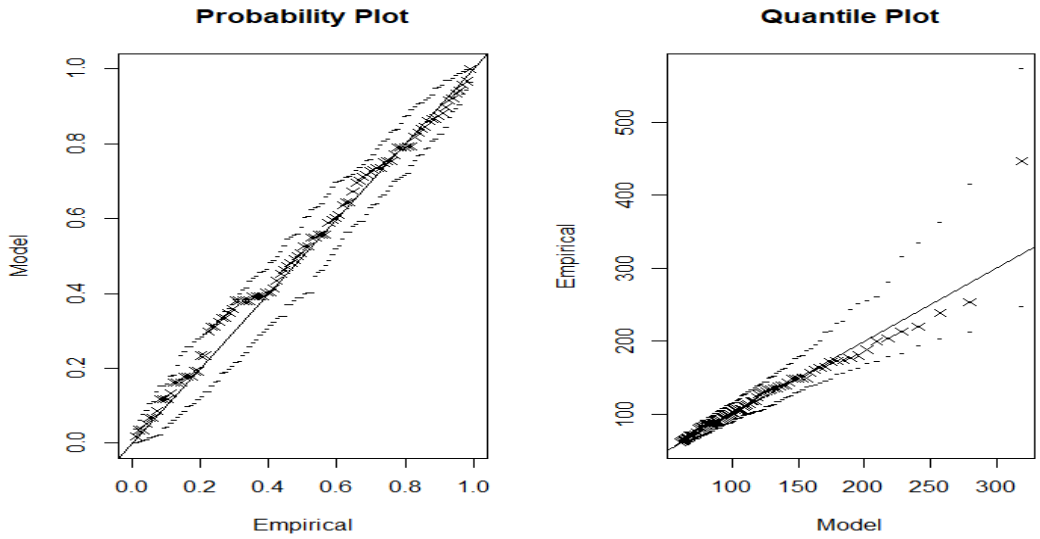

FIGURE 7 – Ajustement graphique des excès des Musanga Cecropioides

Cas du Macaranga : En utilisant la même approche, on estime $u = 24$ (cf Figure 8), et le paramètre de forme estimé $\hat{\gamma} = 0.102$ avec un écart-type de 0.085. Contrairement au cas du Musanga, l'intervalle de confiance ne contient pas 0. Par les mêmes arguments, il est raisonnable de penser que $\gamma > 0$ pour la GPD. La Figure 9 montre que la répartition et les quantiles empiriques s'ajustent bien à une loi dont le domaine d'attraction est celui de Fréchet. On peut donc en conclure que choisir une loi de mélange dont le domaine d'attaction appartient au domaine de Fréchet serait plus adapté. On peut suggérer comme loi de mélange une demi-Cauchy ou une inverse-gamma.

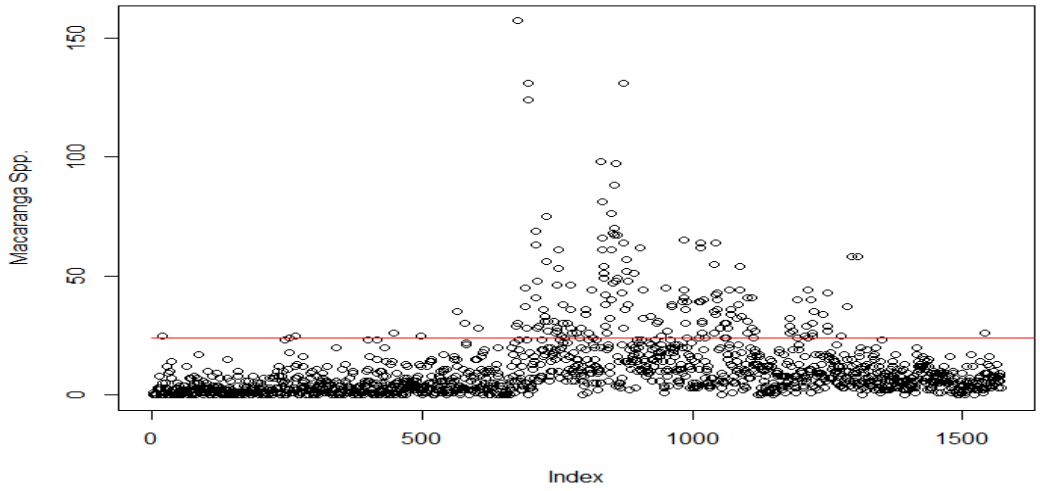

FIGURE 8 – Comptage des Macaranga Spp. avec un seuil de 24

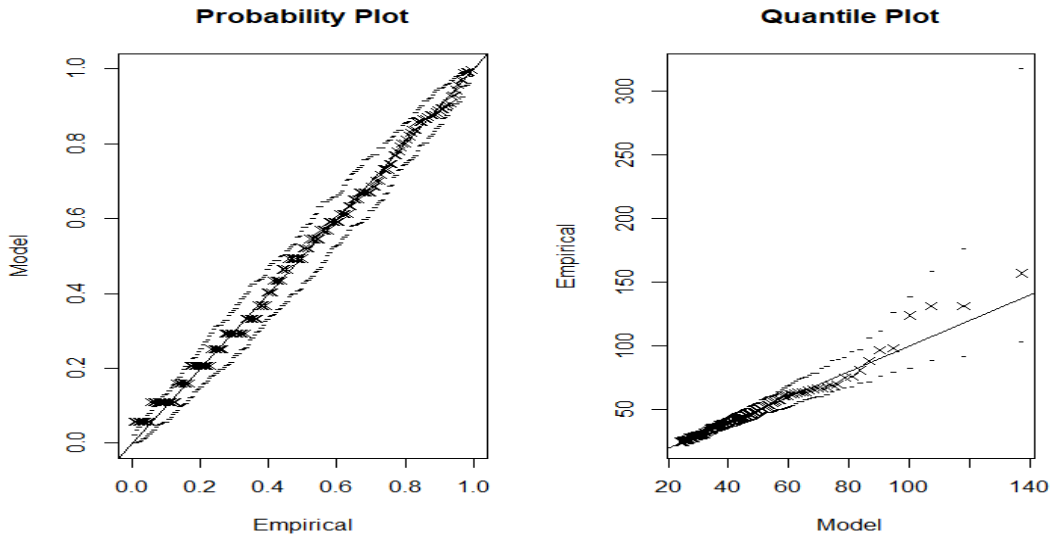

FIGURE 9 – Ajustement graphique des excès des Macarangas Spp.

5 Conclusions et perspectives

L'analyse des données de comptage sous la loupe de la théorie des valeurs extrêmes a permis d'identifier trois classes de Poisson en mélange : Fréchet, Gumbel et approximativement Gumbel. En se basant sur cette classification, on a établi une stratégie sur le comportement des excès pour sélectionner des lois *a priori* sur λ. Comme mentionné en introduction, le choix des lois se base classiquement sur des densités possédant une forme similaire à celle de la distribution empirique. Or les distributions sélectionnées peuvent être similaires. Par exemple on avait remarqué que la gamma et la lognormale pouvaient avoir une forme indiscernable. Or grâce à l'étude des excès et des domaines d'attraction, il est possible de les distinguer.

Il reste cependant beaucoup de travail à faire. En effet, il faudra valider la méthode grâce à plusieurs simulations de données de comptage et de s'assurer qu'au final la méthode ajuste bien les observations. De plus, il serait intéressant de prouver que toute utilisation de lois dans le domaine de Fréchet sur λ permet au mélange d'être dans ce domaine. En effet, Perline (1998) a démontré une condition suffisante, mais celle-ci n'englobe pas toutes les distributions dans ce domaine. Il y a donc un intérêt théorique à démontrer si c'est toujours le cas.

À long terme, il est nécessaire d'introduire des covariables dans cette démarche. En effet, on a utilisé la méthode seulement sur les données de comptage dans ce travail. Or, il serait pertinent d'exploiter toutes les informations dont on dispose pour sélectionner le mélange. Finalement, l'approche présentée concerne seulement des observations univariées. La question qu'on se pose alors est la suivante : Est-il possible de généraliser la stratégie pour le cas multivarié ? Il faudra donc tenter de trouver des liens similaires entre la théorie des extrêmes et les mélanges de Poisson, mais cette fois pour des données multivariées.

Ce travail de stage a été financé par le projet GAMBAS (ANR-18-CE02-0025).

Références

C.W. Anderson : Extreme value theory for a class of discrete distributions with applications to some stochastic processes. Journal of Applied Probability, 7:99–113, 1970.

N.H. Bingham, Goldie C.M. et J.L. Teugles : Regular Variation. Cambridge University Press, 1987.

J. Broek : A score test for zero inflation in a poisson distribution. Biometrics, 51:738–743, 1995.

M.G. Bulmer : On fitting the Poisson lognormal distribution to species abundance data. Biometrics, 30:101–110, 1974.

R.A. Fisher et L.H.C. Tippett : Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proceedings Cambridge Philos. Soc., 24(2):180–190, 1928.

B.V. Gnedenko : Sur la distribution limite du terme maximum d'une série aleatoire. Annals of Mathematics, 44:423–453, 1943.

M. Greendwood et G.U. Yule : 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, 83:255–279, 1920.

Adrien Hitz, Richard Davis et Gennady Samorodnitsky : Discrete Extremes. arXiv e-prints, page arXiv :1707.05033, juillet 2017.

D. Karlis et E. Xekalaki : Mixed poisson distributions. International Statistical Review, 73:35–58, 2005.

D. Lambert : Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics, 34:1–14, 1992.

M.R. Leadbetter, G. Lindgren et H. Rootzén : Extremes and Related Properties of Random Sequences and Processes. Springer, 1983.

J. Lynch : Mixtures, generalized convexity and balayages. Scandinavian Journal of Statistics, 15:203–210, 1988.

R. Perline : Mixed poisson distributions tail equivalent to their mixing distributions. Statistics and Probability Letters, 38:229–233, 1998.

J. Pickands : Statistical inference using extreme order statistics. The Annals of Statistics, 3:119–131, 1975.

S.I. Resnick : Extreme Values, Regular Variation and Point Processes. Springer, 1987.

T. Shimura : Discretization of distributions in the maximum domain of attraction. Extremes, 15 (3):299–317, 2012.

G.E. Willmot : Asymptotic tail behaviour of poisson mixtures with applications. Advances in Applied Probability, 22:147–159, 1990.

Z. Yang, J.W. Hardin et C.L. Addy : Score tests for zero-inflation in overdispersed count data. Communications in Statistics - Theory and Methods, 39:2008–2030, 2010.

6 Annexe

On donne une preuve complémentaire au théorème 10 en utilisant des techniques utilisées dans Anderson (1970) et Hitz et al. (2017). Premièrement, on présente le prolongement continu de la fonction de masse p_X introduite par Anderson. Voici comment il procède : tout d'abord on pose pour $n \in \mathbb{N}$

$$h(n) = -\log(1 - F_X(n))$$

et on utilise une interpolation linéaire pour avoir une version continue de h

$$h_c(x) := h(\lfloor x \rfloor) + (x - \lfloor x \rfloor)(h(\lfloor x + 1 \rfloor) - h(\lfloor x \rfloor))$$

avec $x \geq 1$. Ainsi, on a une version continue de F_X définie par

$$F_c(x) = 1 - e^{-h_c(x)}$$

et on peut rendre la fonction de masse p_X continue en posant

$$p_c(x) = F_c(x) - F_c(x - 1).$$

Deuxièmement, on utilise un résultat démontré par Hitz et al. (2017) concernant le domaine de Fréchet pour des lois discrètes. Supposons une variable aléatoire discrète X avec comme fonction de masse p_X et Y une variable continue possédant une fonction de survie $\overline{F_Y}$. De plus, posons que $Y \in \mathcal{D}_{\frac{\gamma}{1+\gamma}}$, le domaine d'attraction d'indice $\frac{\gamma}{1+\gamma}$ avec $\gamma > 0$. Alors si pour $k \in \{d, d+1, \ldots\}$ avec $d \in \mathbb{N}$ et $c > 0$ on a $p_X(k) = c\overline{F_Y}(k)$, Hitz et al. démontrent que $X \in \mathcal{D}_\gamma$. Pour prouver le théorème 10, on tente de construire une variable aléatoire continue Y similaire pour la fonction de masse du mélange de Poisson utilisant une densité à comportement Pareto sur λ.

Une telle construction utilisera la version continue de $p_X(n)$, notée $p_c(x)$, qu'on a introduit. Tout d'abord on étudie le comportement asymptotique de $p_c(x)$. On a par le lemme 1 que $p_X(n) \sim C(n)n^\alpha$ avec $\alpha < -1$ et C une fonction à variation lente localement bornée sur $(0, \infty)$. Aussi, on a démontré que $p_X(n+1) \sim p_X(n)$ dans la démonstration du théorème 9. Alors on peut dire que

$$p_X(n+1) \sim C(n)n^\alpha.$$

Pour la suite, on remplace les entiers n par $\lfloor x \rfloor$, la partie entière de $x \in \mathbb{R}^+$. Parce que notre mélange de Poisson X est éventuellement décroissant, il existe un x_0 tel que $p_X(\lfloor x \rfloor)$ devient strictement décroissante. Pour tout $x \geq x_0$, on remarque que $p_c(x)$ possède l'inégalité suivante :

$$p_X(\lfloor x + 1 \rfloor) \leq p_c(x) \leq p_X(\lfloor x \rfloor).$$

En divisant les trois valeurs par $C(\lfloor x \rfloor)\lfloor x \rfloor^\alpha$ et en utilisant l'équivalence asymptotique, on obtient par le théorème du sandwich que

$$\lim_{x \to \infty} \frac{p_c(x)}{C(\lfloor x \rfloor)\lfloor x \rfloor^\alpha} = 1.$$

On montre maintenant que $C(\lfloor x \rfloor)\lfloor x \rfloor^\alpha \sim C(x)x^\alpha$. On utilise encore la représentation de Karamata de $C(x)$ pour démontrer cette équivalence et on utilise le fait que $x \sim \lfloor x \rfloor$.

$$\lim_{x \to \infty} \frac{C(x)}{C(\lfloor x \rfloor)} \left(\frac{x}{\lfloor x \rfloor} \right)^\alpha = \lim_{x \to \infty} \exp \left[\int_{\lfloor x \rfloor}^{x} t^{-1} \eta(t) dt \right].$$

Pour tout $\varepsilon > 0$, il existe $x_\varepsilon \geq 1$ tel que pour tout $x \geq x_\varepsilon$ on a

$$0 < \eta(x) < \varepsilon.$$

Alors pour tout x tel que $\lfloor x \rfloor \geq x_\varepsilon$, on a

$$0 \leq \int_{\lfloor x \rfloor}^{x} t^{-1}\eta(t)dt \leq \varepsilon \int_{\lfloor x \rfloor}^{x} t^{-1}dt = \varepsilon \log\left(\frac{x}{\lfloor x \rfloor}\right) \leq \varepsilon \log(2).$$

Ceci implique que l'intégrale tend vers 0, donc la limite de l'exponentielle tend vers 1. On a alors l'équivalence

$$p_c(x) \sim C(x)x^\alpha.$$

On a tout ce qui faut pour construire la variable aléatoire Y. On peut définir pour tout $x \geq x_0$ la fonction de survie

$$\overline{F}_Y(x) = \frac{p_c(x)}{p_c(x_0)}$$

car $p_c(x)$ devient strictement décroissante pour un tel support. Parce que $p_c(x) \sim C(x)x^\alpha$, on a que $\overline{F}_Y(x) \in \mathcal{RV}_\alpha$ ce qui implique $Y \in \mathcal{D}_{-\frac{1}{\alpha}}$ (Resnick, 1987). En posant $\gamma = -\frac{1}{1+\alpha} > 0$, on a que $Y \in \mathcal{D}_{\frac{\gamma}{1+\gamma}}$. Puisque $p_X(n) = p_c(n)$, on a $p_X(n) = p_c(x_0)\overline{F}_Y(n)$ pour tout $n \geq x_0$. Par Hitz et al. (2017), on peut conclure que le mélange de Poisson X est dans Fréchet.