# Computing the Rao's distance between negative binomial distributions. Application to Exploratory Data Analysis

Claude Manté

# Computing the Rao's distance between negative binomial distributions. Application to Exploratory Data Analysis

Claude Manté

*Aix-Marseille Université, Université du Sud Toulon-Var, CNRS/INSU, IRD, MIO,UM 110, Campus de Luminy, Case 901, F13288 Marseille Cedex 09, France*

*email: claude.mante@mio.osupytheas.fr, claude.mante@gmail.com*

## Abstract

The statistical analysis of counts of living organisms brings information about the collective behavior of species (schooling, habitat preference, etc), possibly depending on their biological characteristics (growth rate, reproductive power, survival rate, etc). This task can be implemented in a non-parametric setting, but parametric distributions, such as the negative binomial (NB) distributions studied here, are also very useful for modeling populations abundance. Nevertheless, the parametric approach is ill-suited from an exploratory point of view, because the visual distance between parameters is irrelevant. On the contrary, considering the Riemannian manifold $NB(D_{\mathcal{R}})$ of NB distributions equipped with the Rao metrics $D_{\mathcal{R}}$, one can compute intrinsic distances between species which can be considered as absolute. Unfortunately, computing this distance requires solving a second-order nonlinear differential equation, whose solution cannot be always found in an acceptable length of time with enough precision. While Manté and Kidé [1] proposed numerical remedies to these problem, we propose a geometrical one, based on Poisson approximation. It consists in superseding $A$ and/or $B$ by "equivalent" better-suited distribution(s) before computing the distance, insofar as possible. The proposed method is illustrated by displaying distributions of counts of marine species: these counts having been fitted by NB distributions, we compute the distance table $\Delta$ between species and represent $\Delta$ through multidimensional scaling (MDS).

*Keywords:* Riemannian manifold, Negative Binomial, geodesics, cut locus,

## Notations

Consider a Riemannian manifold $\mathfrak{M}$, and a parametric curve $\alpha : [a,b] \to \mathfrak{M}$. Its first derivative will be denoted $\dot{\alpha}$. A geodesic curve $\gamma$ connecting two points $p$ and $q$ of $\mathfrak{M}$ will be denoted $p \curvearrowright q$, and $p \curvearrowright s \oplus s \curvearrowright q$ will denote the broken geodesic [2] connecting $p$ to $q$ with a "stopover" at $s$. We will also consider for any $\theta \in \mathfrak{M}$ the local norm $\|V\|_g(\theta)$ associated with the metrics $\mathfrak{g}$ on the tangent space $T_\theta\mathfrak{M}$ :

$$\forall\, V \in T_\theta\mathfrak{M},\ \|V\|_{\mathfrak{g}}(\theta) := \sqrt{V^t.\mathfrak{g}(\theta).V}. \tag{1}$$

The length of a curve $\alpha$ traced on $\mathfrak{M}$ will be denoted $\mathit{\Lambda}(\alpha)$. A parametric probability distribution $\mathfrak{L}^i$ will be identified with its coordinates with respect to some chosen parametrization; for instance, we will write $\mathfrak{L}^i \equiv \left(\phi^i, \mu^i\right)$ for some negative binomial distribution. In addition, $\mathbb{R}^{+*} := ]0, +\infty[$, and $\|M\|_F$ will denote the Frobenius norm of the matrix $M$; logical propositions will be combined by using the classical connectors $\vee$ (or) and $\wedge$ (and).

## 1. Introduction

The statistical analysis of counts of living organisms brings information about the collective behavior of species (schooling, habitat preference, *etc*), possibly associated with their socio-biological characteristics (aggregation, growth rate, reproductive power, survival rate, *etc*). Such data consist in general of two-way $r \times c$ tables of counts, whose rows are associated with surveys (spatial-temporal positions, generally) and columns are associated with species. Roughly speaking, these tables can be analyzed through two different approaches. On the one hand, multivariate methods are widely used to investigate relationships between the community structure and the spatio-temporal variations of the surveys, frequently in connection with explanatory environmental variables (see for instance [3, 4] and the references therein). On the other hand, an alternative way, much earlier used in Ecological Statistics, consists in modeling the rows or the columns count distributions. In the spirit of [3], we propose an intermediary method, consisting in measuring the dissimilarity between species through the probability distribution of some characteristic, and analyzing the obtained dissimilarity table through MDS. In [3], this characteristic was the dispersion of each species while here it will be its abundance. There is a wide range of functional methods to deal with distributional data, fundamentally depending on the chosen metrics on the probabilities set. Recently, multivariate methods with a **geometric** dominance appeared in the literature, based on Riemannian structures equipping spaces of probability densities: non-parametric Fisher-Rao metrics [5], Wasserstein metrics [6]. But all these methods were designed in a non-parametric setting, for absolutely continuous distributions, while our data are discrete. In addition, ecological field data are typically characterized by a large number of zeros (problematic for most of the methods above); that is why the negative binomial (NB) distribution is widely used to model catches of animals [7, 8, 9, 10]. It is especially relevant for ecologists, because

1. it arises as a Gamma-Poisson mixture, whose parameters depend on the more or less aggregative behavior of the species, and on the efficiency [11, 12] of the trap for catching it
2. it arises as the limit distribution of the Kendall [13] birth-and-death model; in this setting, its parameters depend on the demography of the species (reproductive power, mortality, immigration rate)
3. one of its limit cases, the log-series distribution, is a natural model for collections (of animals, for instance) [14].

But while the parametric approach is quite sound from the ecological point of view (see [10] and the references therein), it is ill-suited for Exploratory Data Analysis (EDA): the visual distance between parameters of several distributions is misleading, because on the one hand it depends on the chosen parametrization and, on the other hand, because these parameters are not commensurable in general (different ecological meaning, different ranges, ...).

In a seminal paper, Rao [15] noticed that, equipped with the Fisher information metrics denoted $\mathfrak{g}(\bullet)$, a family of probabilities depending on $p$ parameters can be considered as a $p$-dimensional Riemannian manifold. The associated Riemannian (Rao's) distance between the distributions of parameters $\theta^1$ and $\theta^2$ is

$$D_{\mathcal{R}}\left(\theta^1,\theta^2\right) := \int_0^1 \sqrt{\dot{\gamma}^t\left(t\right).\mathfrak{g}\left(\gamma\left(t\right)\right).\dot{\gamma}\left(t\right)}dt \tag{2}$$

where $\gamma$ is a **segment** (minimal length curve) connecting $\theta^1 = \gamma(0)$ to $\theta^2 = \gamma(1)$ and $\dot{\gamma}(t) := \frac{d\gamma}{dt}(t)$; as any Riemannian distance, $D_{\mathcal{R}}$ is intrinsic. Naturally, Rao [15, 16] proposed to use (2) as a distance between populations or for Goodness-Of-Fit (GOF) testing, followed by a number of authors [17, 18, 19, 20, 21, 1, 22, 23, 24]. The Rao's distance between members of a common family of distributions has been calculated in a number of classical cases [25] but it cannot be obtained in a closed form, generally. In such cases, like the NB distributions (when both parameters are unknown), $D_{\mathcal{R}}$ must be obtained by numerically solving a second-order nonlinear differential equation which is frequently hard to integrate.

The outline of this study is as follows. After reminding in Section 2 essential notions of Riemannian geometry, we resume in Section 3.1 a method proposed by Manté and Kidé [1] for approximating $D_{\mathcal{R}}\left(\theta^1,\theta^2\right)$. Next, in Section 3.2 we show how Poisson approximation can be used to speed up its computation, and an application to EDA of ecological data is shown in Section 3.3. Finally, Section 4 is dedicated to conclusion and discussion.

## 2. Essential elements of Riemannian geometry

According to the fundamental theorem of Riemannian geometry [2], there is a unique symmetric connection $\nabla$ compatible with a given metrics $\mathfrak{g}$ (the Levi-Civita or Riemann connection), giving in our case the Rao's distance. It is noteworthy that other statistically sound (but not Riemannian) connections can be fruitfully considered (see Amari et al. [26]).

**Definition 1.** [2, 27] Let $\gamma : I \to \mathfrak{M}$ be a curve traced on $\mathfrak{M}$, and $\boldsymbol{D}$ be a connection on $\mathfrak{M}$. $\gamma$ is a geodesic with respect to $\boldsymbol{D}$ if its acceleration $\boldsymbol{D}_{\dot{\gamma}(t)}\dot{\gamma}(t)$ is null $\forall t \in I$.

**Theorem 2.** *Let $\gamma : I \rightarrow \mathfrak{M}$ be a geodesic* ***with respect to the metric connection*** *$\nabla$. Then $\gamma$ has constant speed in the local norm (1)*

$$\|\dot{\gamma}\|_{\mathfrak{g}} := \|\dot{\gamma}(\bullet)\|_{\mathfrak{g}} (\gamma(\bullet)) = \sqrt{\dot{\gamma}^t(\bullet) . \mathfrak{g}(\gamma(\bullet)) . \dot{\gamma}(\bullet)}$$

*and, for any $[a, b] \subseteq I$, we have:*

$$\int_a^b \sqrt{\dot{\gamma}^t(t) . \mathfrak{g}(\gamma(t)) . \dot{\gamma}(t)} dt = (b - a) \|\dot{\gamma}\|_{\mathfrak{g}} .$$

Geodesics on a $p$-dimensional Riemannian manifold with respect to $\nabla$ are solutions of the Euler-Lagrange equation [27, 2, 25]:

$$\forall\, 1 \leq k \leq p,\ \ddot{\gamma}_k(t) + \sum_{i,j=1}^{p} \mathit{\Gamma}_{i,j}^{k}\, \dot{\gamma}_i(t)\, \dot{\gamma}_j(t) = 0 \tag{3}$$

where each coefficient (some Christoffel symbol $\mathit{\Gamma}_{i,j}^{k}$) only depends on $\mathfrak{g}$ and is defined in coordinates by:

$$\mathit{\Gamma}_{i,j}^{k} := \sum_{m=1}^{p} \frac{\mathfrak{g}^{\mathrm{km}}}{2} \left( \frac{\partial \mathfrak{g}_{jm}}{\partial \theta_i} + \frac{\partial \mathfrak{g}_{im}}{\partial \theta_j} - \frac{\partial \mathfrak{g}_{ij}}{\partial \theta_m} \right) \tag{4}$$

where $\mathfrak{g}^{\mathrm{im}}$ (resp. $\mathfrak{g}_{mk}$) is some entry of $\mathfrak{g}^{-1}$ (resp. $\mathfrak{g}$). To determine the shortest path between two points of $\mathfrak{M}$, one applies the following result.

**Theorem 3.** *[27, 2] Let $p$, $q \in (\mathfrak{M}, \boldsymbol{\nabla}, \mathfrak{g})$ and suppose $\alpha : [a, b] \rightarrow \mathfrak{M}$ is a curve of minimal length connecting $p$ to $q$. Then, $\alpha$ is a geodesic.*

Nevertheless, building the segment connecting $\mathfrak{L}^1$ to $\mathfrak{L}^2$ is not straightforward, since this theorem only says that a segment is a geodesic. But a geodesic is not necessarily a segment...

**Theorem 4.** *[2, 28] Let $p = \alpha(0)$ be the initial point of a geodesic. Then there is some $0 < t_0 \leq +\infty$ such that $\alpha$ is a segment from $p$ to $\alpha(t)$ for every $t \leq t_0$ and for $t > t_0$ thereafter never again a segment from $p$ to any $\alpha(t)$ for $t > t_0$. This number $t_0$ is called the cut value of $\alpha$ and $\alpha(t_0)$ is called the cut point of $\alpha$. There are only two possible reasons (which can occur simultaneously) for $\alpha(t_0)$ to be to be the cut point of $\alpha$:*

- *there is a segment from $p$ to $\alpha(t_0)$ different from $\alpha$*
- *$\alpha(t_0)$ is the first conjugate point on $\alpha$ to $p$ (i.e. $t_0 \dot{\alpha}(0)$ is a critical point of* the exponential map *(see Definition 5 and Figures S1, S3 and S4).*

*In addition, the distance function $D_{\mathcal{R}}(p, \bullet)$ is not differentiable at $\alpha(t_0)$ [29, 2].*

*Remark* 1. No matter the cause of the phenomenon, the main point for us is that if $t_0$ is a cut value of the unit-speed geodesic $\alpha$, $\forall\, t \leq t_0$, $D_{\mathcal{R}}\left(p, \alpha\left(t\right)\right) = t$ while $\forall\, t > t_0$, $D_{\mathcal{R}}\left(\alpha\left(0\right), \alpha\left(t\right)\right) < t$. Nevertheless, $\forall \tau > 0$, $D_{\mathcal{R}}\left(\alpha\left(t_0\right), \alpha\left(t_0 + \tau\right)\right) = \tau$. This remark is the basis of the method proposed by Manté and Kidé [1] for detecting cut points.

**Definition 5.** [2] Let $\mathfrak{M}$ be a Riemann manifold and $x \in \mathfrak{M}$. The exponential map of $\mathfrak{M}$ at $x$ is $\exp_x : W_x \to \mathfrak{M}$, defined on some neighborhood $W_x$ of the origin of $T_x\mathfrak{M}$ by:

$$\exp_x\left(V\right) := \alpha_{\mathcal{B}(V)}\left(\|V\|\right) \tag{5}$$

where $\mathcal{B}\left(V\right)$ is the projection of $V$ onto the unit ball and $\alpha_{\mathcal{B}(V)}$ is the unique unit-speed geodesic in $\mathfrak{M}$ such that $\alpha_{\mathcal{B}(V)}\left(0\right) = x$ and $\dot{\alpha}_{\mathcal{B}(V)}\left(0\right) = \mathcal{B}\left(V\right)$.

*Remark* 2. If $\alpha := p \curvearrowright q$ is a segment and $V_0 := \dot{\alpha}\left(0\right)$, because of uniqueness of geodesics, $\exp_p\left(V_0\right) := \alpha_{\mathcal{B}(V_0)}\left(1\right) = q$; reciprocally, if $V_1 := -\dot{\alpha}\left(1\right)$, we have also that $\exp_q\left(V_1\right) := \alpha_{\mathcal{B}(V_1)}\left(1\right) = p$.

## 3. The special case of $NB(D_{\mathcal{R}})$

There is a large number of parametrizations for the NB distribution, and the most classical one is probably

$$P\left(X = j; \left(\phi, p\right)\right) = \begin{pmatrix} \phi + j - 1 \\ \phi - 1 \end{pmatrix} p^{j} \left(1 - p\right)^{\phi}\ j \geq 0 \tag{6}$$

with $(\phi, p) \in \mathbb{R}^{+} \times ]0, 1[$. Nevertheless, because of its orthogonality, we chose instead the parametrization used by Chua and Ong [30]:

$$P\left(X = j; \left(\phi, \mu\right)\right) = \begin{pmatrix} \phi + j - 1 \\ j \end{pmatrix} \left(\frac{\mu}{\mu + \phi}\right)^{j} \left(1 - \frac{\mu}{\mu + \phi}\right)^{\phi}, \ j \geq 0 \tag{7}$$

$(\phi, \mu) \in \mathbb{R}^{+} \times \mathbb{R}^{+}$; here, $\mu$ is the mean of the distribution and $\phi$ is the so-called “index parameter". In these coordinates, the information matrix is:

$$\mathfrak{g}(\phi, \mu) = \begin{pmatrix} G_{\phi\phi} & 0 \\ 0 & G_{\mu\mu} \end{pmatrix}$$

where $G_{\mu\mu} = \frac{\phi}{\mu(\mu+\phi)}$, while the expression of $G_{\phi\phi}$ is more complicated:

$$G_{\phi\phi} = -\frac{\mu + \phi\left(\mu + \phi\right)\left(\left({}^{\phi}\!/_{\mu+\phi}\right)^{\phi} - 1\right)\psi^{1}(\phi)}{\phi\left(\mu + \phi\right)} \tag{8}$$

where $\psi^1$ is the Trigamma function [31].

One will find in Burbea and Rao [25] the closed-form expression of the Rao's distance for a number of probability families. These authors reported that, when the index parameter of two NB distributions **is the same**, the Rao's distance is given by:

$$D_{NB(p)}\left(\left(\phi, p^1\right), \left(\phi, p^2\right)\right) := 2\sqrt{\phi}\ \cosh^{-1}\left(\frac{1-\sqrt{p^1 p^2}}{\sqrt{\left(1-p^1\right)\left(1-p^2\right)}}\right) \tag{9}$$

in the parametrization (6). Of course, if $\mathfrak{L}^1 = NB\left(\phi, p^1\right)$ (resp. $\mathfrak{L}^2 = NB\left(\phi, p^2\right)$), we have necessarily:

$$D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right) \leq D_{NB(p)}\left(\mathfrak{L}^1, \mathfrak{L}^2\right). \tag{10}$$

Due to the complexity of (8), $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ cannot be obtained in a closed-form. It must be computed by finding the numerical solution of a the Euler-Lagrange equation (3), completed in the parametrization (7) by the conditions (boundary value problem)

$$\left\{\gamma(0) = \left(\phi^1, \mu^1\right),\ \gamma(1) = \left(\phi^2, \mu^2\right)\right\}. \tag{11}$$

Geodesics can be as well be computed by solving (3) under the alternative constraints (initial value problem)

$$\left\{\gamma(0) = \left(\phi^1, \mu^1\right),\ \dot{\gamma}(0) = V \in \mathbb{R}^2\right\} \tag{12}$$

where $V$ is the initial velocity of the geodesic; this solution is associated with the exponential map at $\left(\phi^1, \mu^1\right)$.

### *3.1. Numerical approximation of $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ [1]*

From now, $\mathfrak{L}^i \equiv \left(\phi^i, \mu^i\right)$ will denote some NB distribution parametrized in the (7) system, but our purpose could be extended to any parametric family of probabilities. Firstly, all the Christoffel symbols (4) were calculated from the expression (8) of $G_{\phi\phi}$, with the help of *Mathematica* [32] . Then, the differential equation (3) was numerically solved under the the boundary conditions (11), for a number of distributions of counts of marine species whose parameters had been estimated in [10]. In most cases a solution could be found in an acceptable time (four CPU minutes), with a good numerical precision (20 digits), but was each one of the geodesics found a segment? And what about failures met in computation? We indeed had to face various problems detailed in [1], where some numerical remedies were proposed. The first one consisted in inserting a well-placed "stopover" $S$ between each pair of problematic distributions $A$ and $B$, in such a way that $D_{\mathcal{R}}(A, S)$ and $D_{\mathcal{R}}(S, B)$ could be computed in a reasonable time, while $D_{\mathcal{R}}(A, B)$ could not. Furthermore, $S$ was placed in order that $D_{\mathcal{R}}(A, S) + D_{\mathcal{R}}(S, B)$ should by a good approximation of $D_{\mathcal{R}}(A, B)$. For sake of brevity, we moved to the supplementary material useful information and illustrations about this previous work. Notice that **all references to this supplement will be preceded by an S**.

### *3.2. Contribution of geometry: making computations easier thanks to Poisson approximation*

From the numerical side, it is noteworthy that the index parameter $\phi$ often takes large values, causing difficulties in the evaluation of quantities associated with $\mathit{\Gamma}(\phi)$, like formulas (7) and (8) or the Christoffel's symbols (4).

From the statistical side, the convergence of some NB distribution $\mathfrak{L} \equiv (\phi, \mu)$ towards a Poisson distribution $\mathcal{P}$ when $\phi \to \infty$ is well-known. More precisely, Majsnerowska [33] proved that

$$d_{TV}(\mathfrak{L}, \mathcal{P}(\lambda)) \leq \Delta(\phi, \mu) := \left(1 - e^{-\mu}\right) \frac{\mu}{\phi} \tag{13}$$

where $\lambda := \frac{\phi\,\mu}{\phi+\mu}$ and $d_{TV}$ denotes the total variation distance. Thus, we can claim that $(\phi \gg \mu) \vee (\mu\ \mathit{small}) \Rightarrow \Delta(\phi, \mu)\ \mathit{small}$ and conclude that in such cases it may be quite impossible to find a difference between $\mathfrak{L}$ and $\mathcal{P}(\lambda)$, even when the index parameter is small or moderate! This fact suggests to replace the NB model by the Poisson one when both distributions are very close to each other. This is also biologically sound, since the former is well-suited for aggregative species, while the latter is associated to species with a random behavior (see [10, 3] and the references therein).

Let's focus now on the application

$$\begin{cases} \omega : \mathbb{R}^{+*} \times \mathbb{R}^{+*} \rightarrow \mathbb{R}^{+*} \\ (\phi, \mu) \mapsto \lambda \end{cases}$$

associating to each NB distribution $\mathfrak{L} \equiv (\phi, \mu)$ the corresponding limit Poisson distribution $\mathcal{P}(\lambda)$.

**Lemma 6.** *The tangent application* $T_{(\phi,\mu)}\omega = \frac{1}{(\phi+\mu)^2} \left( \begin{array}{c} \mu^2 \\ \phi^2 \end{array} \right)$ *is surjective.*

*Proof.* Let us fix some $\rho \in \mathbb{R}^{+*}$; one can easily show that the set of solutions of the equation $T_{(\phi,\mu)}\omega\,(x, y) = \rho$ is the line of equation $y = -\left(\frac{\mu}{\phi}\right)^2 x + \rho\left(1 + \frac{\mu}{\phi}\right)^2$ □

As a consequence, $\omega$ is a surjective submersion and the fiber $F_\lambda := \omega^{-1}(\lambda)$ associated with any $\lambda \in \mathbb{R}^{+*}$ is a sub-manifold of $NB(D_{\mathcal{R}})$.

**Proposition 7.** *$F_\lambda$ is defined by either equation:*

$$\begin{cases} \mu\,(\lambda; \phi) = \frac{\lambda}{1-\lambda/\phi} & : \phi > \lambda \\ \phi\,(\lambda; \mu) = \frac{\lambda}{1-\lambda/\mu} & : \mu > \lambda \end{cases}. \tag{14}$$

*Proof.* $F_\lambda := \left\{(\phi, \mu) : \frac{\phi\,\mu}{\phi+\mu} = \lambda\right\}$; thus, the strictly positive parameters $\lambda$, $\phi$ and $\mu$ are linked by the relationship $\phi\mu = \phi\lambda + \lambda\mu$, which proves that $\phi = \lambda + \lambda\frac{\phi}{\mu}$ and $\mu = \lambda + \lambda\frac{\mu}{\phi}$. Consequently, $\lambda < \min(\phi, \mu)$ and $\lim\limits_{\phi \to +\infty} \mu\,(\lambda; \phi) = \lim\limits_{\mu \to +\infty} \phi\,(\lambda; \mu) = \lambda$ □

**Lemma 8.** *One can easily verify that:*
$\forall\, \lambda \in \mathbb{R}^{+*}$, $F_\lambda := \{(\phi, \mu) \in \mathbb{R}^{+*} \times \mathbb{R}^{+*} : \omega\,(\phi, \mu) = \lambda\} \neq \emptyset$
$\forall\ (\phi, \mu) \in \mathbb{R}^{+*} \times \mathbb{R}^{+*}$, $(\phi, \mu) \in F_{\omega(\phi,\mu)}$
$\forall\ (\lambda_1, \lambda_2) \in \mathbb{R}^{+*} \times \mathbb{R}^{+*}$, $F_{\lambda_1} \cap F_{\lambda_2} = \emptyset$.

**Proposition 9.** *Suppose $\mathfrak{L} \equiv (\phi, \mu) \in F_\lambda$ and $\Delta\,(\phi, \mu) \leq \delta$, where $\delta$ is some threshold chosen for deciding whether $\mathfrak{L}$ can be identified with $\mathcal{P}\,(\lambda)$. Then, if $\mathfrak{L}' \equiv (\phi', \mu') \in F_\lambda$ is another distribution, such that $\phi' > \phi$, $\Delta\,(\phi', \mu') < \delta$ and $\mathfrak{L}'$ cannot be practically distinguished from $\mathcal{P}\,(\lambda)$ too.*

*Proof.* See Appendix 5.1 □

Let us now fix $\mathfrak{L}^0 \equiv \left(\phi^0, \mu^0\right)$; by definition $\mathfrak{L}^0 \in F_{\lambda^0}$, with $\lambda^0 = \omega\left(\phi^0, \mu^0\right)$. Suppose $\Delta\left(\phi^0, \mu^0\right) \leq \delta$; then $\mathfrak{L}^0$ could be identified with $\mathcal{P}\left(\lambda^0\right)$, as well as any distribution of the fiber whose index parameter is greater than $\phi^0$, due to the proposition above. We can now, thanks to (14), find the distribution $(\phi_*, \mu_*)\left(\lambda^0, \delta\right)$ such that

$$\phi_*\left(\lambda^0, \delta\right) := \underset{\phi:(\phi,\mu)\in F_{\lambda^0}}{\arg}\left(\Delta\left(\phi, \mu\right) = \delta\right) = \underset{\phi}{\arg}\left(\Delta\left(\phi, \mu\left(\lambda; \phi\right)\right) = \delta\right) \qquad (15)$$

and define :

$$\mathring{\mathcal{P}}\left(\lambda^0, \delta\right) := \left\{(\phi, \mu) \in F_{\lambda^0} : \phi \geq \phi_*\left(\lambda^0, \delta\right)\right\}.$$

Obviously, $\mathfrak{L}^0 \in \mathring{\mathcal{P}}\left(\lambda^0, \delta\right)$ because $\phi_* \leq \phi^0$, but we have that, more generally, when $(\phi, \mu) \in \mathring{\mathcal{P}}\left(\lambda^0, \delta\right)$, $d_{TV}\left(\mathfrak{L}, \mathcal{P}\left(\lambda^0\right)\right) \leq \delta$ and $\left|\omega\left(\phi, \mu\right) - \lambda^0\right| \approx 0$, simultaneously. Thus, in such cases, $NB\left(\phi, \mu\right)$ and $\mathcal{P}\left(\lambda^0\right)$ are **practically indiscernible** from the statistical point of view!

**Definition 10.** We will say that $\mathfrak{L} \equiv (\phi, \mu)$ is **Poisson-like** if $(\phi, \mu) \in \mathring{\mathcal{P}}\left(\omega\left(\phi, \mu\right), \delta\right)$.

We displayed on Figure 1 four examples of such NB distributions (setting $\delta = 0.01$, say). Let us now denote $\stackrel{\delta}{\equiv}$ the following relation (for $\delta$ fixed) between Poisson-like distributions:

$$\mathfrak{L}^1 \stackrel{\delta}{\equiv} \mathfrak{L}^2 \Leftrightarrow \exists\lambda : \mathfrak{L}^i \in \mathring{\mathcal{P}}\left(\lambda, \delta\right), i = 1, 2.$$

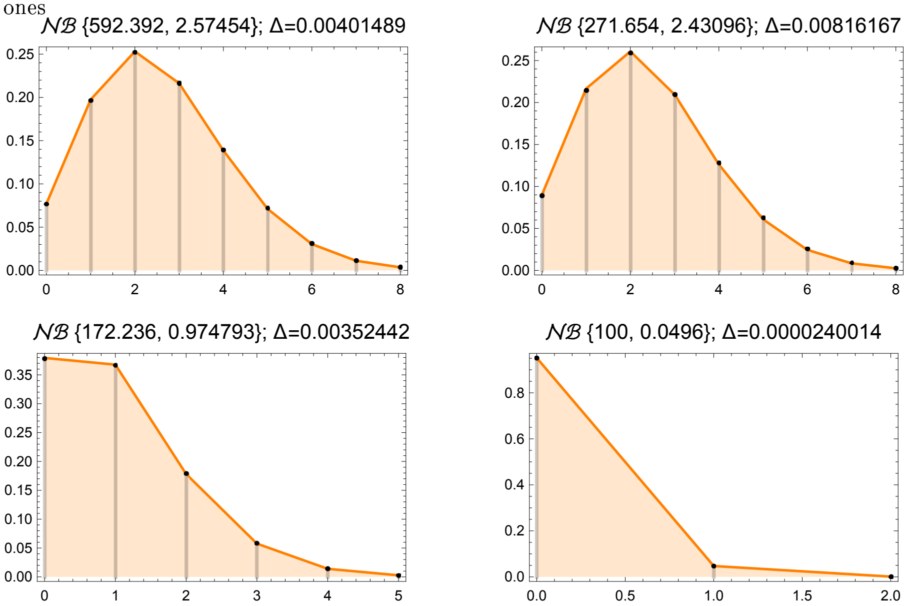
**Corollary 11.** *The relation* $\stackrel{\delta}{\equiv}$ *is an equivalence relation between Poisson-like distributions.*

*Proof.* Reflexive and symmetric properties are straightforward. Suppose now $\mathfrak{L}^1 \stackrel{\delta}{\equiv} \mathfrak{L}^2$ and $\mathfrak{L}^3 \stackrel{\delta}{\equiv} \mathfrak{L}^2$; there exists $\lambda_{1,2} : \mathfrak{L}^i \in \mathring{\mathcal{P}}\left(\lambda_{1,2}, \delta\right), i = 1, 2$ and $\lambda_{2,3} : \mathfrak{L}^i \in \mathring{\mathcal{P}}\left(\lambda_{2,3}, \delta\right), i = 2, 3$. Consequently, $\mathfrak{L}^2 \in F_{\lambda_{1,2}} \cap F_{\lambda_{2,3}}$ which is empty if $\lambda_{1,2} \neq \lambda_{2,3}$ (see Lemma 8), and these three Poisson-like distributions belong to the same fiber. Thus, $\mathfrak{L}^1 \stackrel{\delta}{\equiv} \mathfrak{L}^3$ and $\stackrel{\delta}{\equiv}$ is transitive □

Suppose $\mathfrak{L}^1 \stackrel{\delta}{\equiv} \mathfrak{L}^2$ belong to a common fiber, $F_{\lambda_{1,2}}$. Being indiscernible, these distributions should be necessarily close to each other, and it would be statistically sound to supersede $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ by $\delta$.

**Corollary 12.** *Consider now two Poisson-like distributions* $\mathfrak{L}^1$ *and* $\mathfrak{L}^2$ *belonging to different fibers,* $F_{\lambda_1}$ *and* $F_{\lambda_2}$. *After computing (thanks to formula 15)* $\tilde{\phi} := \max\left(\phi_*\left(\lambda^1, \delta\right), \phi_*\left(\lambda^2, \delta\right)\right)$, $\tilde{\mu}_1 := \mu\left(\lambda_1; \tilde{\phi}\right)$ *and* $\tilde{\mu}_2 := \mu\left(\lambda_2; \tilde{\phi}\right)$, *we can determine the Poisson-like distributions* $\tilde{\mathcal{L}}^1 := \left(\tilde{\phi}, \tilde{\mu}_1\right) \in F_{\lambda_1}$ *and* $\tilde{\mathcal{L}}^2 := \left(\tilde{\phi}, \tilde{\mu}_2\right) \in F_{\lambda_2}$. *Then, thanks to formula 9, we can easily compute* $D_{NB(p)}\left(\tilde{\mathcal{L}}^1, \tilde{\mathcal{L}}^2\right)$, *which is an upper bound for* $D_{\mathcal{R}}\left(\tilde{\mathcal{L}}^1, \tilde{\mathcal{L}}^2\right)$.

Figure 1: Four instances of Poisson-like distributions; $\Delta$ is given by Formula (13) and vertical bars are associated with NB probabilities while continuous curves are associated with Poisson ones



$\mathcal{NB}$ {592.392, 2.57454}; Δ=0.00401489

$\mathcal{NB}$ {271.654, 2.43096}; Δ=0.00816167

$\mathcal{NB}$ {172.236, 0.974793}; Δ=0.00352442

$\mathcal{NB}$ {100, 0.0496}; Δ=0.0000240014

**Corollary 13.** *Under the same conditions, an alternative strategy is possible. Suppose $\phi_1 \leq \phi_2$; thanks to formula 14 we can determine $\check{\mu}_1 := \mu(\lambda_1; \phi_2)$, such that $\check{\mathfrak{L}}^1 := (\phi_2, \check{\mu}_1) \in F_{\lambda_1}$ is Poisson-like too (because of Proposition 9) and belongs to the same class as $\mathfrak{L}^1$. Then, we can compute $D_{NB(p)}\left(\check{\mathfrak{L}}^1, \mathfrak{L}^2\right)$, which is another upper bound for $D_{\mathcal{R}}\left(\check{\mathfrak{L}}^1, \mathfrak{L}^2\right)$.*

Since $\tilde{\mathcal{L}}^1 \overset{\delta}{\equiv} \mathfrak{L}^1 \overset{\delta}{\equiv} \check{\mathfrak{L}}^1$ and $\tilde{\mathcal{L}}^2 \overset{\delta}{\equiv} \mathfrak{L}^2$ , it is quite sound to supersede $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ by $D_{NB(p)}\left(\check{\mathfrak{L}}^1, \mathfrak{L}^2\right)$ or $D_{NB(p)}\left(\tilde{\mathcal{L}}^1, \tilde{\mathcal{L}}^2\right)$. One could also measure the difference between $\mathfrak{L}^1$ and $\mathfrak{L}^2$ by $D_{\mathcal{P}}\left(\mathcal{P}(\lambda^1), \mathcal{P}(\lambda^2)\right)$ but, since $\mathcal{P}(D_{\mathcal{P}})$ is not a sub-manifold of $NB(D_{\mathcal{R}})$, there is no clear relationship between the associated Rao's distances (for instance, if $\mathfrak{L}^1 \neq \mathfrak{L}^2$ belong to the same fiber $F_\lambda$, $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right) > D_{\mathcal{P}}\left(\omega\left(\mathfrak{L}^1\right), \omega\left(\mathfrak{L}^2\right)\right) = 0$). Nevertheless, one can easily prove that

$$\begin{cases} D_{NB(p)}\left(\tilde{\mathcal{L}}^1, \tilde{\mathcal{L}}^2\right) = \mathcal{C}\left(\tilde{\mu}_1, \tilde{\mu}_2\right)\ D_{\mathcal{P}}\left(\mathcal{P}(\lambda^1), \mathcal{P}(\lambda^2)\right) \\ D_{NB(p)}\left(\check{\mathfrak{L}}^1, \mathfrak{L}^2\right) = \mathcal{C}\left(\check{\mu}_1, \mu_2\right)\ D_{\mathcal{P}}\left(\mathcal{P}(\lambda^1), \mathcal{P}(\lambda^2)\right) \end{cases} \tag{16}$$

where $\mathcal{C}\left(\check{\mu}_1, \mu_2\right)$ and $\mathcal{C}\left(\tilde{\mu}_1, \tilde{\mu}_2\right) \geq 1$ are given by the function defined hereunder (up to a simple change of parametrization).

**Lemma 14.** *$\mathcal{C}$ is defined in the parametrization (6) by*

$$\mathcal{C}\left(p^1,p^2\right)=\frac{\cosh^{-1}\left(\frac{1-\sqrt{p^1p^2}}{\sqrt{(p^1-1)(p^2-1)}}\right)}{\sqrt{-2\sqrt{p^1p^2}+p^1+p^2}}\geq 1$$

if $p^1\neq p^2$, and $\mathcal{C}\left(p,p\right):=1$.

*Proof.* In the parametrization (6), the mean $\mu=\frac{K\,p}{1-p}$ and thus $p=\frac{\mu}{\phi+\mu}$, while $\phi=K$. Consequently, $\lambda^i=K\,p^i$ and $D_{\mathcal{P}}\left(\mathcal{P}(\lambda^1),\mathcal{P}(\lambda^2)\right)=2\sqrt{K}\,\left|\sqrt{p^1}-\sqrt{p^2}\right|$. Then, because of formula (9), we have:

$$\frac{D_{NB(p)}\left(\check{\mathfrak{L}}^1,\check{\mathfrak{L}}^2\right)}{D_{\mathcal{P}}\left(\mathcal{P}(\lambda^1),\mathcal{P}(\lambda^2)\right)}=\mathcal{C}\left(p^1,p^2\right)$$

□

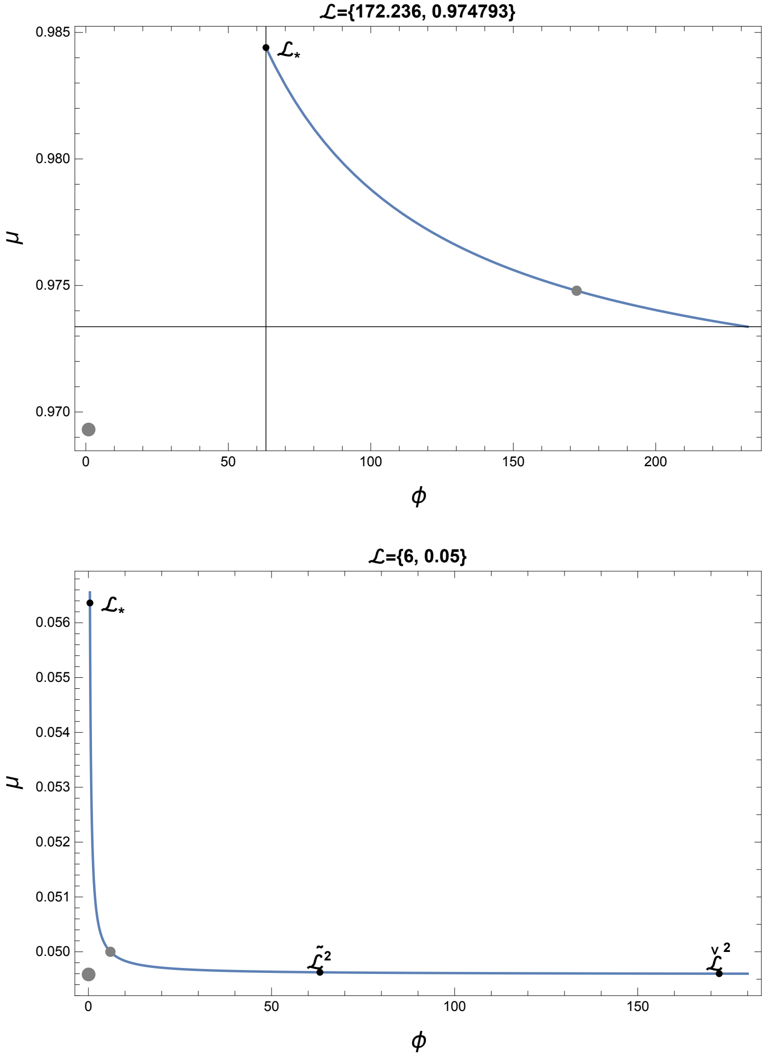
**An exemple**

Consider $\mathfrak{L}^1=(172.236,0.974793)$ and $\mathfrak{L}^2=(6,0.05)$, and lets us fix $\delta:=0.01$. Both these distributions are Poisson-like, with $\mathfrak{L}^1\in F_{0.0495868}$ while $\mathfrak{L}^2\in F_{0.969307}$. We plotted on Figure 2 interesting portions of these fibers. On each one of the panels, the big gray point (of coordinates $(\lambda,\lambda)$) corresponds to the lower bound of $\phi$ and $\mu$, while $\mathfrak{L}_*:=\left(\phi_*,\mu_*\right)(\lambda,\delta)$ is the distribution given by Equation 15. All the distributions situated on the right of $\mathfrak{L}_*$ are Poisson-like. It is the case of $\mathfrak{L}^1$ and $\mathfrak{L}^2$, represented on Figure 2 by small gray points.

We found that $\phi_*^1=63.2096<\phi^1$ and $\phi_*^2=0.412537$; thus, $\tilde{\phi}=\phi_*^1$ and $\tilde{\mathcal{L}}^1=\mathfrak{L}_*^1$, while $\tilde{\mathcal{L}}^2\neq\mathfrak{L}_*^2$. Next, in accordance with Corollary 12, we compute $D_{NB(p)}\left(\tilde{\mathcal{L}}^1,\tilde{\mathcal{L}}^2\right)\approx 1.53375$, which is rather close to $D_{\mathcal{P}}\left(\mathcal{P}(\lambda_1),\mathcal{P}(\lambda_2)\right)\approx 1.52371$. Using Formula 16, one obtains 1.53375 again.

Since $\phi_*^1=63.2096<\phi^1$ , it is also possible to determine the distribution $\check{\mathfrak{L}}^2:=\left(\phi_1,\check{\mu}_2\right)\overset{\delta}{\equiv}\mathfrak{L}^2$ (black point on the lower panel) and, in accordance with Corollary 13, to compute $D_{NB(p)}\left(\check{\mathfrak{L}}^2,\mathfrak{L}^1\right)\approx 1.52737$ (very close to $D_{\mathcal{P}}\left(\mathcal{P}(\lambda_1),\mathcal{P}(\lambda_2)\right)$ too).

Figure 2: Portions of fibers associated with two Poisson-like distributions ($\delta = 0.01$): $F_{0.0495868}$ (lower panel) and $F_{0.969307}$ (upper panel).



Out[ ]=

### *3.2.1. Application to EDA*

Let $\mathfrak{L}^1$ and $\mathfrak{L}^2 \in NB(D_{\mathcal{R}})$; three cases may be met: both of them are Poisson-like, only one of them is Poisson-like, or none of them is so.

Suppose first $\mathfrak{L}^1$ and $\mathfrak{L}^2$ belong to distinct fibers $F_{\lambda^1}$ and $F_{\lambda^2}$ and each $\mathfrak{L}^i \in \mathring{\mathcal{P}}\left(\lambda^i, \delta\right)$. Then $\mu^i \approx \lambda^i$ and we can use Corollaries 12 or 13 to build a pair of equivalent distributions, whose interdistance is easier to compute.

Suppose now $\mathfrak{L}^1$ is Poisson-like while $\mathfrak{L}^2$ is not; if $\phi^1_* \leq \phi^2$, we can again build $\check{\mathfrak{L}}^1 := \left(\phi^2, \check{\mu}^1\right)$, and four distances can be computed: $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$, $D_{\mathcal{R}}\left(\mathfrak{L}^1_*, \mathfrak{L}^2\right)$, $D_{\mathcal{R}}\left(\check{\mathfrak{L}}^1, \mathfrak{L}^2\right)$ and $D_{NB(p)}\left(\check{\mathfrak{L}}^1, \mathfrak{L}^2\right)$. On the contrary, if $\phi^1_* > \phi^2$, we cannot consider the two last ones.

**Three instances.**

To illustrate our purpose, look first at Figures S1, S3 and S4. In all these cases, only one of the pair of distributions is Poisson-like. In the first case, $\mathfrak{L}^1 = (0.00487399, 0.262591)$, and we found that $\mathfrak{L}^2 = (592.392, 2.57454) \overset{\delta}{\equiv} (3.5634, 9.13442) = \mathfrak{L}^2_*$. Since $\phi^2_* = 3.5634 \nleq 0.00487399$ we could not build $\check{\mathfrak{L}}^2 := \left(\phi^1, \check{\mu}^2\right)$ and compute$D_{NB(p)}\left(\check{\mathfrak{L}}^2, \mathfrak{L}^1\right)$, but $D_{\mathcal{R}}\left(\mathfrak{L}^2_*, \mathfrak{L}^1\right) = 3.53253$ could be computed straightforwardly (simple configuration, no cut point), while the original $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$ corresponded to an intricate configuration with a cut point, and to the upper bound 45.1321 (definitions of **simple** and **intricate configurations** are remembered in the supplementary material).

In the next case (Figure S3) $\mathfrak{L}^1 = (0.00996246, 0.121282)$, while the second distribution is the same as in the previous case. The original distance corresponded to an intricate configuration with a cut point, and to the upper bound 43.1519. We found instead $D_{\mathcal{R}}\left(\mathfrak{L}^2_*, \mathfrak{L}^1\right) = 3.48809$.

In the last case (Figure S4), $\mathfrak{L}^1 = (0.938781, 9.86571)$, and we found that $\mathfrak{L}^2 = (172.236, 0.974793) \overset{\delta}{\equiv} (63.2096, 0.984403) = \check{\mathfrak{L}}^2$. Since $\phi^2_* = 63.2096 \nleq 0.938781$, we could not compute $D_{NB(p)}\left(\check{\mathfrak{L}}^2, \mathfrak{L}^1\right)$, but we found that $D_{\mathcal{R}}\left(\mathfrak{L}^2_*, \mathfrak{L}^1\right) = 12.5294$ (an intricate configuration with an acceptable rough solution and a stopover), while the original $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$, corresponding to an intricate configuration with linear interpolation, gave rise to the upper distance 21.351.

### *3.3. EDA of field data: representation of counts distributions of marine species*

The Mauritanian coast, situated on the Atlantic side of the northwestern African continent, embeds a wide long continental shelf of about 750 $km$ and $36000 km^2$, with an Exclusive Economic Zone (the MEEZ) of $230000 km^2$. Manté et al. [10] considered the abundance of species of fish and invertebrates collected in the MEEZ during annual scientific trawl surveys since 1997 to now. Because the spatial distribution of groundfish species is strongly influenced by the physical environment, we split this set into an optimal number (four) of subsets (typical habitats) associated with homogeneous physical conditions determined by available environmental variables (bathymetry, sedimentary type of the substrate, latitude and longitude). The counts associated with each species found in each one of the four habitats were then gathered, and fitted by a truncated NB distribution; notice that only a reduced number of species could be satisfactorily fitted in each habitat (for further information, see [10]).

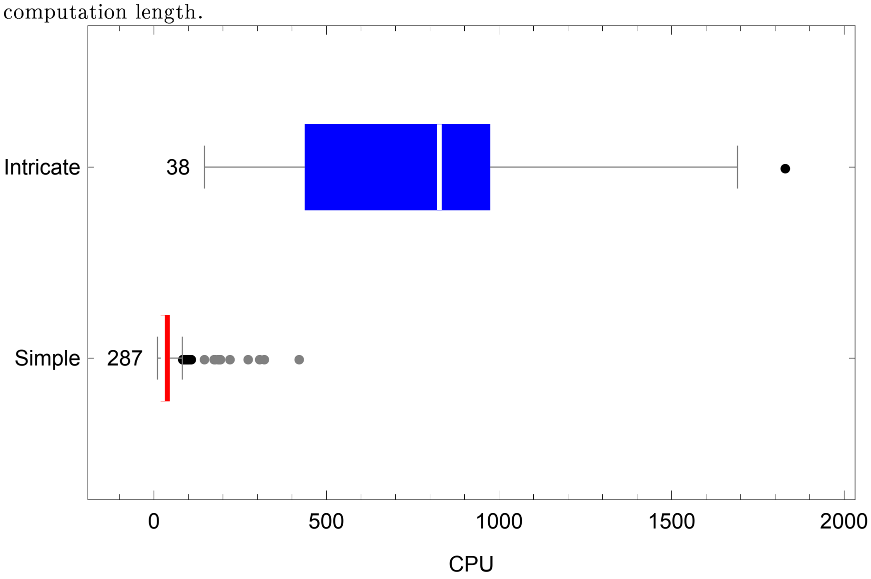
Table 1: Global results obtained in the four habitats of the MEEZ

| Habitat | Number of species | Simple | Intricate | Cut points | Poisson-like |
|---|---|---|---|---|---|
| | (well-fitted) | configurations | (Rough, Linear) | | distributions |
| $C1$ | 30 | 400 | (35,0) | 1 | 3 |
| $C2$ | 19 | 147 | (24,0) | 0 | 3 |
| $C3$ | 26 | 309 | (16,0) | 0 | 5 |
| $C4$ | 26 | 304 | (21,0) | 0 | 1 |

#### *3.3.1. Benefits of Poisson approximation*

Processing these data, we found in [1] an overwhelming proportion of simple configurations (more than 70%), while numerical cut points were quite rare. In the intricate cases, the rough solution was generally accepted (more than 90% of occurrences). We displayed on Figure 3 statistics about the computational cost of the $26 \times 26$ distance matrix corresponding to $C4$. Among the 325 distances computed, 287 were simple cases, with a median computation cost of 25"; the remaining 38 cases were intricate, with a median cost of 826". Superseding each Poisson-like distribution $\mathfrak{L}$ by the corresponding $\tilde{\mathcal{L}} \overset{\delta}{\equiv} \mathcal{L}$ (see Section 3.2),

Figure 3: Statistics of the computational burden for processing (without Poisson approximation) species collected in zone 4 of the MEEZ: number of distances of each type, box-plots of computation length.



we found that the proportion of simple configurations was greater than 90%, excepted for the second type of habitat, $C2$ (85%) (see Table 1). Furthermore, in the intricate cases, the rough solution was always accepted. In our previous study, numerical cut points were rare (less than a pair per class), but now we detect a single numerical cut point (see Table 1)! Poisson-like distributions were quite rare, but they were often so “pathological” that their replacement by equivalent NB distributions changed a lot the results. This is illustrated by the statistics displayed on Figure 4. Among the 325 distances computed, 305 were simple cases, with a median computation cost of 20”; the remaining 20 cases were intricate, with a median cost of 404”. So, the introduction of Poisson approximation simultaneously increased the number of simple configuration, approximately divided by two the number of intricate configurations, and approximately divided by two the computational cost of the corresponding distances.

Figure 4: Statistics of the computational cost for processing the same species as in Figure 3, with Poisson approximation.

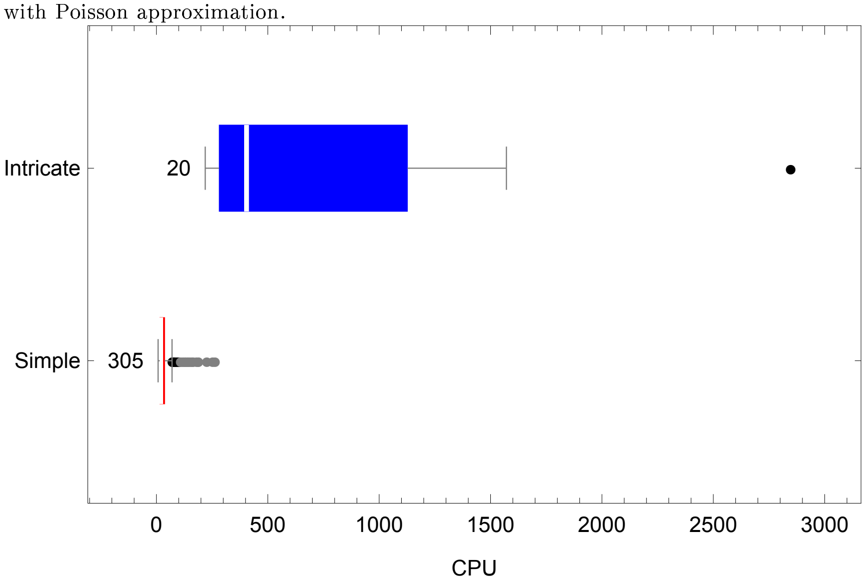
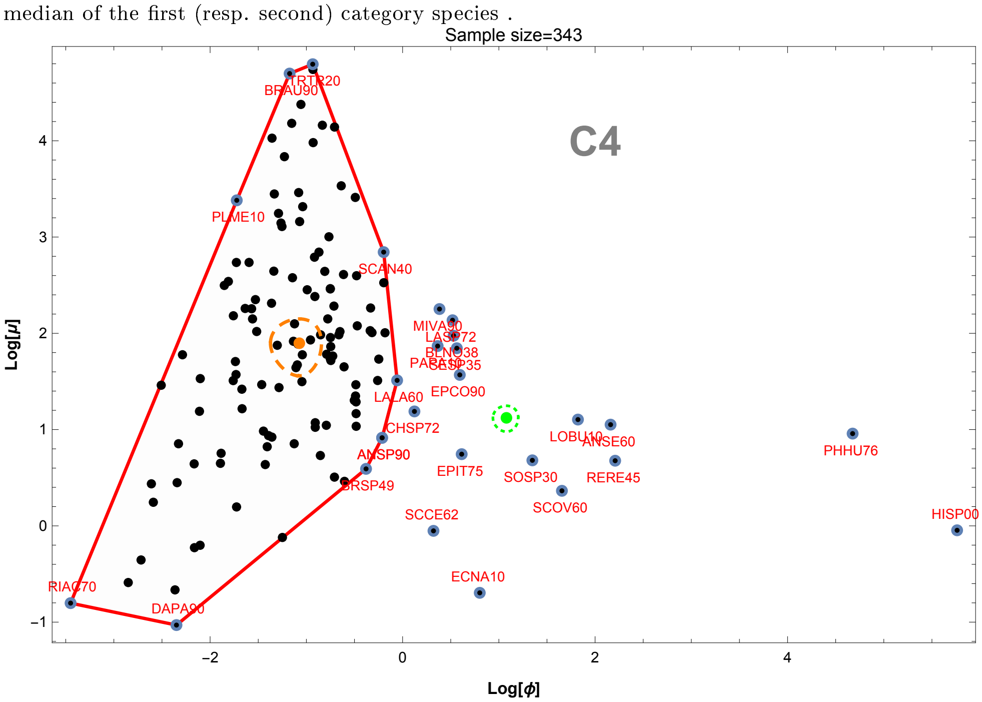


Figure 5: Species collected in zone 4 of the MEEZ; the green dotted (resp. orange dashed) closed curve corresponds to the confidence region of 0.99 level associated with the spatial median of the first (resp. second) category species .



### *3.3.2. Parametric representation of species from the habitat C4*

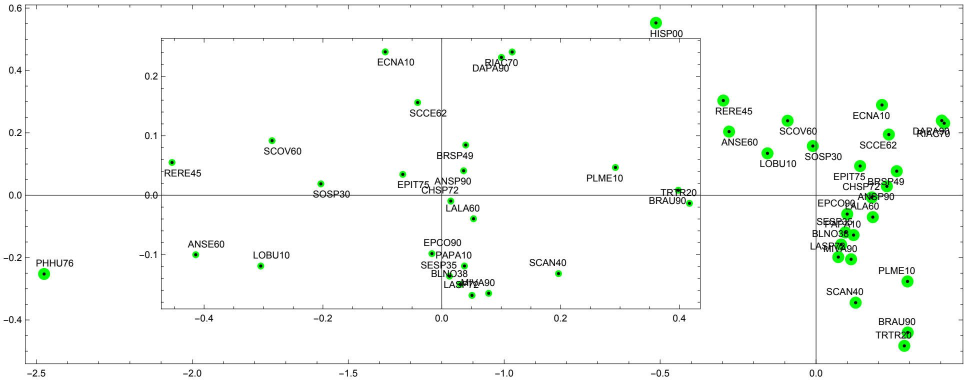
We display on Figure 5 the estimated parameters of the counts distribution of species sampled in the zone 4 of the MEEZ. This region is of paramount importance: it is a high plankton productivity area, supporting a large variety of fish communities, with many commercial species that sustain fishing activities. We distinguished two categories of species on Figure 5, according to the index parameter: $\phi > 1$ or $\phi \leq 1$. Species belonging to the second category being very numerous, we only kept for MDS those which are situated on the convex envelope on the associated cloud. Twenty-six species were selected this way (while 138 species are represented) for computation of the Rao's distances and subsequent MDS of the table of distances. We also plotted on Figure 5 theoretical confidence ellipses centered on the spatial medians [34] of both species categories.

### *3.3.3. Representation of the Rao's distance table*

Only one of the 26 species retained in C4 was Poisson-like: "HISP00" (*Hippocampus sp*). Consequently, the distance table was defined this way: $\varDelta_{i,j} = D_{\mathcal{R}}\left(\mathfrak{L}^i, \mathfrak{L}^j\right)$, excepted when one of these species was "HISP00" (index $h$, say). In these cases, we set instead:

$\varDelta_{h,j} = \min\left(D_{\mathcal{R}}\left(\mathfrak{L}^h, \mathfrak{L}^j\right), D_{\mathcal{R}}\left(\tilde{\mathcal{L}^h}, \mathfrak{L}^j\right)\right)$. The median distortion of the Schoenberg exponential transformation of $\Delta$ (see Section 5.2) denoted $\varDelta^{Exp}$, was 0.854805, which is slightly lower than the one corresponding to the traditional Additive Constant method $\varDelta^{AC}$, 0.872421. Nevertheless, since $\left\|\varDelta - \eta^{AC}\,\varDelta^{AC}\right\|_F$ was much smaller than $\left\|\varDelta - \eta^{Exp}\,\varDelta^{Exp}\right\|_F$ (where $\eta^{Exp}$ minimizes the Frobenius norm), $\varDelta^{AC}$ has been chosen. Notice that the ratio $c^*/\overline{\varDelta^2}$ was 2.302: the additive constant method greatly altered the distance matrix. MDS of the 26 species is represented on Figure 6. Clearly, the counts of the species "HISP00" and "PHHU76" (*Physiculus huloti*, a type of cod) are distributed in a very special manner, which was not so obvious on Figure 5. In the inset region we represented the other 24 species separately analyzed by MDS of the restricted table $\varDelta^{AC}$ (no point from the first MDS is hidden). In this case, the median distortion associated with the Schoenberg transformation of the distances table

Figure 6: MDS of the Rao's distance between the 26 selected species (big points); the inset graph corresponds to the same analysis, performed after removing both the species "HISP00" and "PHHU76". .



$\varDelta^{Exp}$ was 0.401832, greater than the one corresponding to $\varDelta^{AC}$ (0.298623), while the relative perturbation of distances was high $\left(c^{*}/\overline{\varDelta^{2}} = 1.2529\right)$.

## 4. Results and discussion

Following Rao [15], a number of authors used $D_{\mathcal{R}}\left(.,.\right)$ in various statistical settings: either exploratory methods [16, 17, 18, 21, 1, 20, 35] or hypothesis testing problems [23, 24]. Motivated by the analysis of a large data set of marine species counts collected in the MEEZ, we developed a parameter-free method to compare species counts distributions in the setting of the Riemannian manifold $NB(D_{\mathcal{R}})$ of negative binomial distributions, equipped with $D_{\mathcal{R}}$.

We focused first [1] on numerical problems met in computing $D_{\mathcal{R}}\left(\mathfrak{L}^{1}, \mathfrak{L}^{2}\right)$: lengthy computations could result from the presence of a cut point on the geodesic $\mathfrak{L}^{1} \curvearrowright \mathfrak{L}^{2}$, requiring to determine a stopover $S$ somewhere between these distributions. $D_{\mathcal{R}}\left(\mathfrak{L}^{1}, \mathfrak{L}^{2}\right)$ is then approximated by the upper bound $D_{\mathcal{R}}\left(\mathfrak{L}^{1}, S\right) + D_{\mathcal{R}}\left(S, \mathfrak{L}^{2}\right)$. In Section 3.2 we show how Poisson approximation can be used to evaluate more efficiently $D_{\mathcal{R}}\left(\mathfrak{L}^{1}, \mathfrak{L}^{2}\right)$ when one (at least) of the distributions involved is "Poisson-like". Superseding original NB distributions by equivalent ones (in the sense of Definition 10), we could obtain lower upper bounds of the distances than with the former strategy, with a lower computational cost. In addition, this refinement enabled us to get around most numerical issues (numerical cut points, unsuitable rough solutions).

Afterwards, an application to EDA of ecological data is shown. Unfortunately, we had to restricts ourselves to a small number of species, because of the computational cost of $D_{\mathcal{R}}$; therefore, future research should focus on speeding up more the computation of $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$.

## 5. Appendices

### *5.1. Proof of Corollary 9*

Notice first that on the fiber $F_\lambda$, because of (14), the expression of $\Delta$ defined in (13) is $\frac{\lambda\left(e^{\frac{\lambda\phi}{\lambda-\phi}}-1\right)}{\lambda-\phi}$, giving: $\frac{\partial\Delta}{\partial\phi}\left(\phi\right) = \frac{\lambda\left(e^{\frac{\lambda\phi}{\lambda-\phi}}\left(\lambda^2+\lambda-\phi\right)-\lambda+\phi\right)}{(\lambda-\phi)^3}$. Since $\phi > \lambda$, the denominator of this expression is always negative while the numerator is clearly positive, excepted potentially if $\left(\lambda^2+\lambda-\phi\right) < 0$. Substituting $\lambda^2+\lambda+\zeta$ to $\phi$ (with $\zeta > 0$) in the equation, we get a simpler expression for $\frac{\partial\Delta}{\partial\phi}$:

$$-\frac{\lambda\left(-\zeta\, e^{\frac{\zeta}{\zeta+\lambda^2}-\lambda-1}+\zeta+\lambda^2\right)}{\left(\zeta+\lambda^2\right)^3}$$

whose sign depends on the sign of $\left(1-e^{\frac{\zeta}{\zeta+\lambda^2}-\lambda-1}\right)$. Since the only solutions of $\frac{\zeta}{\zeta+\lambda^2}-\lambda-1=0$ are $\lambda=0\wedge\zeta\neq0$ and $\zeta=-\left(\lambda+\lambda^2\right)\wedge\lambda\neq0$, this expression is negative, and $\left(1-e^{\frac{\zeta}{\zeta+\lambda^2}-\lambda-1}\right)\geq0$. Consequently, $\frac{\partial\Delta}{\partial\phi}\left(\phi\right)$ is negative and $\Delta\left(\phi,\mu\right)$ is a decreasing function of $\phi$ on a fiber.

### *5.2. Pre-processing distance tables for MDS*

There are several methods for making a distance matrix like $\mathit{\Delta}$ Euclidean (*i.e.* find a close distance matrix which can be **exactly** represented in an Euclidean space) - see for instance [36]. The simpler one is the Additive Constant (AC) one [37], consisting in adding an optimal positive perturbation $c^*$ to all the extra-diagonal terms of $\mathit{\Delta}^2$. But other pre-processing methods are worth considering [36]: one can search for the smallest positive $\gamma_0$ such that the power $\mathit{\Delta}^\gamma$ is Euclidean for $\gamma\leq\gamma_0$ [38], or the smallest positive $\gamma^*$ such that $1-e^{-\gamma\,\Delta}$ is

Euclidean (Exp method). It is noteworthy that both these transformations belong to the class of Schoenberg transformations introduced in Data Analysis by Bavaud [39]. We chose the last one, since it is bounded and rectifiable, *i.e.* a finite length curve is transformed into another finite length curve - fractal curves are not rectifiable [40], for instance. Thus, two element-wise transformations of $\varDelta$ were considered:

$$\begin{cases} \varDelta^{AC}_{i,j} := \sqrt{\varDelta^2_{i,j} + c^*} & i \neq j \\ \varDelta^{Exp}_{i,j} := \frac{1-\exp(-\lambda^* \varDelta_{i,j})}{\lambda^*} & i \neq j \end{cases} \tag{17}$$

Naturally, these perturbation should be as small as possible. In the case of $\varDelta^{AC}$ the ratio $c^*/\overline{\varDelta^2}$, where $\overline{\varDelta^2}$ denotes the mean squared distance, is a straightforward and natural criterion. Remember now that since the set of Euclidean matrices of a given size is a convex cone, the solutions proposed in (17) can be easily improved by looking for an optimal $\eta$ minimizing the Frobenius norm $\|\varDelta - \eta\,\varDelta^\bullet\|_F$. **In the applications, the original table $\varDelta^\bullet$ will be systematically superseded by this optimum.** In addition to the specific index $c^*/\overline{\varDelta^2}$ for AC, it is interesting to consider statistics of local distortions. Benasseni *et al.* [36] proposed the criterion $\max_{i\neq j}\left(\frac{\varDelta^\bullet_{i,j}}{\varDelta_{i,j}}\right)/\min_{i\neq j}\left(\frac{\varDelta^\bullet_{i,j}}{\varDelta_{i,j}}\right)$ which was ill-suited for our data, since $\varDelta_{i,j}$ was frequently very small. We instead computed for each distance obtained from (17) the index

$$\omega_{i,j} := \begin{cases} 0 & if\ \varDelta^\bullet_{i,j} = \varDelta_{i,j} \\ 1 - \frac{\varDelta_{i,j}}{\varDelta^\bullet_{i,j}} & if\ \varDelta^\bullet_{i,j} \neq \varDelta_{i,j} \end{cases}.$$

Then, each list $\varOmega$ of distortions was described by its kernel density estimate, its average and its median.

## Acknowledgments

We thank the Mauritanian Institute of Oceanographic Research and Fisheries (especially S. O. Kidé) and the Department of Cooperation and Cultural Action of the Embassy of France in Mauritania for their support for this study. We also thank all scientists who contributed to field surveys and data collection.

## References

## References

[1] C. Manté, S. O. Kidé, Approximating the Rao's distance between negative binomial distributions. Application to counts of marine organisms, in: Proceedings of COMPSTAT 2016, A. Colubi, A. Blanco and C. Gatu., Oviedo (Spain), 2016, pp. 37–47. URL: https://hal.archives-ouvertes.fr/hal-01357264.

[2] M. Berger, A Panoramic View of Riemannian Geometry, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[3] C. Manté, J. P. Durbec, J. C. Dauvin, A functional data-analytic approach to the classification of species according to their spatial dispersion. Application to a marine macrobenthic community from the Bay of Morlaix (Western English Channel), Journal of Applied Statistics 32 (2005) 831–840.

[4] S. O. Kidé, C. Manté, L. Dubroca, H. Demarcq, B. Mérigot, Spatio-Temporal Dynamics of Exploited Groundfish Species Assemblages Faced to Environmental and Fishing Forcings: Insights from the Mauritanian Exclusive Economic Zone, PLOS ONE 10 (2015) e0141566.

[5] A. Srivastava, I. Jermyn, S. Joshi, Riemannian Analysis of Probability Density Functions with Applications in Vision, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[6] V. Seguy, M. Cuturi, Principal Geodesic Analysis for Probability Measures under the Optimal Transport Metric, in: C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28, Curran Associates, Inc., 2015, pp. 3312–3320.

[7] C. I. Bliss, R. A. Fisher, Fitting the Negative Binomial distribution to biological data, Biometrics 9 (1953) 176–200.

[8] M. F. O'Neill, M. J. Faddy, Use of binary and truncated negative binomial modelling in the analysis of recreational catch data, Fisheries Research 60 (2003) 471–477.

[9] L. Vaudor, N. Lamouroux, J.-M. Olivier, Comparing distribution models for small samples of overdispersed counts of freshwater fish, Acta Oecologica-International Journal of Ecology 37 (2011) 170–178.

[10] C. Manté, S. O. Kidé, A.-F. Yao-Lafourcade, B. Merigot, Fitting the truncated negative binomial distribution to count data. A comparison of estimators, with an application to groundfishes from the Mauritanian Exclusive Economic Zone, Environmental and Ecological Statistics 23 (2016) 359–385. doi:10.1007/s10651-016-0343-1, wOS:000382017300002.

[11] R. A. Fisher, A. Corbet, C. B. Williams, The relation between the number of species and the number of individuals in a random sample of an animal population, J. anim. Ecol 12 (1943) 42–58.

[12] F. J. Anscombe, Sampling theory of the negative binomial and logarithmic series distributions, Biometrika 36 (1950) 358–382.

[13] D. G. Kendall, On some modes of population growth leading to R. A. Fisher's logarithmic series distribution, Biometrika 35 (1948) 6–15.

[14] C. B. Williams, The Logarithmic Series and its application to biological problems, Journal of Ecology 34 (1947) 253–272.

[15] C. R. Rao, Information and the Accuracy Attainable in the Estimation of Statistical Parameters, Resonance-Journal of Science Education 20 (2015) 78–90.

[16] C. R. Rao, Comment to Kass' paper, Statistical Science 4 (1989) 229–231.

[17] K. M. Carter, R. Raich, W. G. Finn, A. O. Hero, FINE: Fisher Information Nonparametric Embedding, Ieee Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 2093–U195.

[18] G. Galanis, P. C. Chu, G. Kallos, Y.-H. Kuo, C. T. J. Dodson, Wave height characteristics in the north Atlantic ocean: a new approach based on statistical and geometrical techniques, Stochastic Environmental Research and Risk Assessment 26 (2012) 83–103.

[19] C. T. J. Dodson, Some illustrations of information geometry in biology and physics, in: J. Leng, W. W. Sharrock (Eds.), Handbook of research on computational science and engineering, Engineering Science Reference, 2012, pp. 287–315.

[20] M. Cubedo, A. Minarro, J. M. Oller, A dissimilarity based on relevant population features, Journal of Statistical Planning and Inference 143 (2013) 346–355.

[21] I. Ilea, L. Bombrun, C. Germain, I. Champion, R. Terebes, M. Borda, Statistical Hypothesis Test for Maritime Pine Forest Sar Images Classification Based on the Geodesic Distance, in: 2015 IEEE International Geoscience and Remote Sensing Symposium (igarss), Ieee, New York, 2015, pp. 3215–3218.

[22] R. E. Kass, The geometry of asymptotic inference, Statistical Science 4 (1989) 188–234.

[23] M. Menendez, D. Morales, L. Pardo, M. Salicru, Statistical Tests Based on Geodesic Distances, Applied Mathematics Letters 8 (1995) 65–69.

[24] M. Cubedo, J. M. Oller, Hypothesis testing: a model selection approach, Journal of Statistical Planning and Inference 108 (2002) 3–21.

[25] J. Burbea, C. R. Rao, Informative geometry of probality spaces, Expo. Math. 4 (1986) 347–378.

[26] S.-i. Amari, H. Nagaoka, D. Harada, Methods of information geometry, number 191 in Translations of mathematical monographs, nachdr. ed., American Math. Soc. [u.a.], Providence, RI, 2007.

[27] A. Gray, Modern differential geometry of curves and surfaces with Mathematica, 2nd ed ed., CRC Press, Boca Raton, 1998.

[28] M. P. d. Carmo, Riemannian geometry, Mathematics. Theory & applications, Birkhauser, Boston, 1992.

[29] J. I. Itoh, T. Sakai, Cut loci and distance functions, Math. J. Yokohama Univ. 49 (2007) 65–92.

[30] K. C. Chua, S. H. Ong, Test of misspecification with application to negative binomial distribution, Computational Statistics 28 (2013) 993–1009.

[31] M. Abramowicz, I. A. Stegun, Handbook of mathematical functions with formulas, graphs and mathematical tables, Knovel, London, 2002.

[32] W. R. Inc., Mathematica, Version 11.3, 2017. Champaign, IL, 2018.

[33] M. Majsnerowska, A note on Poisson approximation by w-functions, Applicationes mathematicae 25 (1998) 387–392.

[34] R. Serfling, Nonparametric multivariate descriptive measures based on spatial quantiles, Journal of Statistical Planning and Inference 123 (2004) 259–278.

[35] G. Lebanon, Metric learning for text documents, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (2006) 497–508.

[36] J. Benasseni, M. B. Dosse, S. Joly, On a General Transformation Making a Dissimilarity Matrix Euclidean, Journal of Classification 24 (2007) 303–304.

[37] F. Caillez, The analytic solution of the Additive Constant problem, Psychometrika 48 (1983) 305–308.

[38] S. Joly, G. Le Calvé, Etude des puissances d'une distance, Statistique et Analyse des Données 11 (1986) 30–50.

[39] F. Bavaud, On the Schoenberg transformations in data analysis: Theory and illustrations, Journal of Classification 28 (2011) 297–314.

[40] C. Tricot, Courbes et dimension fractale, 2. ed., Springer, Berlin, 1999.