HAL open science

Fitting the truncated negative binomial distribution to count data. A comparison of estimators, with an application to groundfishes from the Mauritanian Exclusive Economic Zone

Claude Manté, Oumar, Saikou Kidé, Anne-Françoise Yao, Bastien Mérigot

▶ **To cite this version:**

Claude Manté, Oumar, Saikou Kidé, Anne-Françoise Yao, Bastien Mérigot. 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, 2016, 23 (3), pp.359-385. 10.1007/s10651-016-0343-1 . hal-01292224

HAL Id: hal-01292224

https://hal.science/hal-01292224

Submitted on 22 Mar 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

FITTING THE TRUNCATED NEGATIVE BINOMIAL DISTRIBUTION TO COUNT DATA: A COMPARISON OF ESTIMATORS. APPLICATION TO GROUNDFISHES FROM THE MAURITANIAN EXCLUSIVE ECONOMIC ZONE.

CLAUDE MANTÉ, SAÏKOU OUMAR KIDÉ, A.F. YAO, BASTIEN MÉRIGOT

Abstract. A frequent issue in the study of species abundance consists in modeling empirical distributions of repeated counts by parametric probability distributions. In this setting, it is desirable that the chosen family of distributions is flexible enough to take into account very diverse patterns, and that its parameters possess clear biological/ecological meanings. This is the case of the Negative Binomial distribution, chosen in this work for modeling counts of marine fishes and invertebrates. This distribution depends on a vector $(K, \mathfrak{P})$ of parameters, and ranges from the Poisson distribution (when $K \to +\infty$) to Fisher's log-series, when $K \to 0$. Besides, these parameters have biological/ecological interpretations detailed in the literature and reminded hereafter.

We focus on the comparison of three estimators of K, $\mathfrak{P}$ and the parameter α of Fisher's log-series, revisiting a nice paper of Rao (1971) about a three-parameter unstandardized variant of the Negative Binomial distribution. We investigate the coherency of values of the parameters resulting from these different estimators, with both real count data collected in the Mauritanian Exclusive Economic Zone during the period 1987-2010 and realistics simulations of theses data.

In the first case, we first built homogeneous lists of counts (replicates), by gathering observations of each species with respect to "typical environments" obtained by clustering the sampled stations. The best estimation of $(K, \mathfrak{P})$ was generally obtained by Penalized Minimum Hellinger Distance Estimation. Interestingly, the parameters of most of the correctly sampled species seem compatible with a classical birth-and-dead model of population growth with immigration of Kendall (1948).

1. Introduction

Ecological data frequently consist 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, such tables can be analyzed through two different approaches. On the one hand, multivariate methods are widely used to investigate relationships between the community structure (columns) and the spatio-temporal variations of the surveys (rows), frequently in connection with explanatory environmental variables. On the other hand, an alternative way, much earlier in Ecological Statistics, consists in modeling the rows or the columns count distributions. Most authors focus on fitting the rows of such tables, because distributions of counts are closely associated with biodiversity (Magurran , 2005) or stochastic abundance models for communities (Watterson , 1974; Diserud , 2001). For example, the log-series (LS) introduced by Fisher, Corbet and Williams (1943) is standard for evaluating or modeling biodiversity (Taylor, Kempton and Woiwod ,

1976; Magurran , 2005; Watterson , 1974; Diserud , 2001). Besides, the LS has been also used to model the columns of such tables; for instance, Williams (1947) (see also Bliss and Fisher (1953)) reported that it fitted well the number of lice per head of prisoners, as well as the number of fleas on rats, and Quenouille (1949) reported that the number of bacteria in a colony is also well-fitted by this series. Notice finally that the LS was obtained by Fisher, Corbet and Williams (1943) as a limit case of the Negative Binomial distribution (NBD), which is also classically used for fitting count data (Bliss and Fisher , 1953; Elliot , 1979; Vaudor, Lamouroux and Olivier , 2011).

From another side, Kendall (1948) has shown that both NBD and LS can be obtained as stable solutions of a single-species population growth process. As a consequence, estimating the parameters of the distribution of counts of each species (columns) would bring information about its dynamics. Notice that this task could be also tackled from a non-parametric viewpoint, since the collective behavior of wild species can be inferred from the Multivariate Analysis of the empirical distribution function of counts of individuals (Manté, Durbec and Dauvin , 2005). Nevertheless, we think that the parametric approach of Kendall (1948), with sound parameters (reproduction rate, immigration rate, mortality, etc.) is more informative for ecologists than a purely exploratory approach.

Outline: we first remind in Section 2 different stochastic mechanisms generating the NBD or its limit case, the LS. Then, we examine in Sections 3&4 the relationships between NBD and LS, in connection with a nice paper of Rao (1971), and compare thoroughly statistical estimation methods designed for these distributions. The data are described and analyzed in Section 5: constitution of samples (replicates) and assessment of the estimators. Finally, the ecological results are commented in Section 6, and Section 7 is dedicated to coclusion and discussion.

2. Biological interpretations of the parameters of the Negative Binomial and Log Series Distributions

We remind in this Section different stochastic mechanisms generating the NBD or its limit case, the LS.

2.1. NBD and LS: two models for collections.

Boswell and Patil (1970) have given twelve different mechanisms generating the NBD, and two mechanisms generating the Truncated Negative Binomial Distribution (TNBD). Two of these mechanisms sound well-adapted for ecological tables of counts. The first one is the well-known Gamma-Poisson model, reminded in greater detail hereunder. The second one obtains the TNBD as the equilibrium group-size distribution of a system of difference equations (see also (Cohen , 1972)). The LS is afterward obtained as a zero-truncated Poisson mixture, or as a group-size distribution (Boswell and Patil , 1971). These theoretical results explain why the same distributions are well-suited for modeling rows and columns of ecological tables, and why LS is well-suited for fitting a variety of frequency biological series recorded at different taxonomic level (species, genera, family,...) (Williams , 1944).

To our knowledge, Williams (1944, 1947, 1952) was the first to notice that LS is naturally associated with the grouping of random counts. He distinguished two cases (Williams , 1947), corresponding to rows or columns of our table.

(1) "In a randomized collection of individual insects (as, for example, a number of moths caught in a light trap) which are later classified into species, the catch is randomized on the individuals, and in addition to an increase in the size of the sample will bring in new individuals to species already represented, i.e. new units in old groups."
(2) "If, on the other hand, collections of rats are made, and the number of fleas on each rat counted, then an increase in the number of rats examined will not add any fleas to the rats already counted, i.e. all the new units will be in new groups. In this case, the sample is randomized by groups."

Remark 1. We will study data of the second category, and trawls will play the part of rats in the second example above. It is worth noting that Williams (1947, p. 263) has shown that in this case the first parameter of the $LS(\alpha, x)$ distribution should increase with the total number of counts, denoted β. More precisely, if L similar lists of length β_0 of a common distribution $LS(\alpha_0, x_0)$ are merged we will have on the one hand $\alpha_0 \approx \frac{f_1}{x_0}$, where f_1 denotes the frequency of counts represented by one individual in any of the L lists, and on the other hand $\alpha \approx \frac{L\, f_1}{x_0} \approx L\, \alpha_0 \approx \beta\, \frac{\alpha_0}{\beta_0}$. Thus, roughly speaking, α should be proportional to the number of merged series or, in other words, is is inseparably a measure of sampling redundancy (associated with each species), and the propensity to clumping of this species.

2.2. **The Gamma-Poisson model.** This model is standard for counts associated with ecological surveys (rows) (Fisher, Corbet and Williams , 1943; Diserud , 2001). Each random count obeys a Poisson distribution, whose random intensity obeys $\gamma(K, \mathfrak{P})$. Suppose $\mathfrak{P}$ has been fixed; then, the more K is close to zero, the more the probability density of $\gamma(K, \mathfrak{P})$ is concentrated near zero (for instance, its median belongs to $\left]\frac{\max\left(0, K-\frac{1}{3}\right)}{\mathfrak{P}}, \frac{K}{\mathfrak{P}}\right[$ (Chen and Rubin , 1986)). Thus, the more K is close to zero, the more a sample of $NBD(K, \mathfrak{P})$ will consist of small integers, and a great number of individuals collected in a survey should be split into a large number of rare species, and fewer and fewer common species. That is why $\frac{1}{K}$ is sometimes considered as an index of diversity or of aggregation (Fisher, Corbet and Williams , 1943; Elliot , 1979; Taylor, Woiwod and Perry , 1979), depending on the context, and K is considered as an intrinsic parameter (Rao , 1971), with a biological meaning (this point of view has been contested by Taylor, Woiwod and Perry (1979)).

As for $\mathfrak{P}$, Rao (1971) considered that increasing by a multiplying factor $a > 0$ the probability of a fish being caught increases the same way $\mathfrak{P}$, since $a\, \gamma(K, \mathfrak{P}) \sim \gamma(K, a\, \mathfrak{P})$. Thus, "the parameter $\mathfrak{P}$ depends on the efficiency of the trap" (here: the trawl). This fact was already mentioned by Fisher, Corbet and Williams (1943) and Anscombe (1950), who underscored that the "efficiency of the trap" must include time of exposure (standardization of the data).

In conclusion, according to this model, K is an **intrinsic** biological characteristic of the organism of interest; since $\mathfrak{P}$ is related to the efficiency of the trap for catching the species considered, it is **intrinsic** too.

2.3. **A population growth model for a single species .** Kendall (1948) proposed a birth-and-dead model of population growth with immigration, starting with no individual at time $-T$ (large) and leading to a Negative Binomial distribution of the population size at each time $t \geq T$. The first parameter of this distribution is $K = \frac{\iota}{\rho}$, where ι is the immigration incidence and ρ is the reproductive power of

the species (by binary fission). Consequently, $K \approx 0$ when the immigration is negligible, or when the reproduction power is important; of course, the last condition brings to mind aggregation.

The second parameter is, at time t, $\mathfrak{P}_t := \frac{\rho(\exp(t(\rho-\mu))-1)}{\rho-\mu}$, where μ denotes the mortality incidence of the species. The variations of $\mathfrak{P}_t$ depends on the ratio $\frac{\rho}{\mu}$; if $\frac{\rho}{\mu} > 1$, $\mathfrak{P}_t$ grows exponentially, as well as the population size. If $\frac{\rho}{\mu} < 1$, $\mathfrak{P}_t$ is again a growing function of t, but $\lim_{t\to+\infty} \mathfrak{P}_t = \frac{\rho}{\mu-\rho} > \frac{\rho}{\mu}$. As a consequence, if $\mu \gg \rho$ (in case of overfishing, for instance), $\lim_{t\to+\infty} \mathfrak{P}_t \approx \frac{\rho}{\mu}$ should be small, and we should have for such a species e of parameters $(K_e, \mathfrak{P}_e)$:

$$Log\left(\mathfrak{P}_e\right) \approx -Log\left(K_e\right) + Log\left(\frac{\iota_e}{\mu_e}\right). \tag{2.1}$$

According to this model, K is also an **intrinsic** biological characteristic of the organisms of interest, but $\mathfrak{P}$ asymptotically results simultaneously from an **intrinsic** property of the species (its reproductive power) and from **mixed** (intrinsic/extrinsic) factors: immigration and mortality, which can simultaneously depend on the species and on the environment.

3. From NBD to ULSD: the three-parameter model of Rao

Fisher, Corbet and Williams (1943) fitted observed frequency of species by LS, which depends on the parameters α and x, which are estimated by solving the equations:

$$\begin{aligned} S &= -\alpha\, Log\left(1-x\right) \\ N &= \tfrac{\alpha\, x}{(1-x)} \end{aligned} \tag{3.1}$$

where S denotes the observed number of species and N the total number of individuals. Fisher derived these equation from the expression of the density of the NBD:

$$P\left(NBD\left(K,\mathfrak{P}\right)=n\right) = \frac{\mathfrak{P}^{\mathfrak{n}}}{\left(1+\mathfrak{P}\right)^{n+K}} \begin{pmatrix} K+n-1 \\ K-1 \end{pmatrix} \tag{3.2}$$

where $K > 0$ and $\mathfrak{P} > 0$ (a number of other parametrization are classically used (Bliss and Fisher , 1953; Boswell and Patil , 1970; Vaudor, Lamouroux and Olivier , 2011)). Defining $x := \frac{\mathfrak{P}}{1+\mathfrak{P}}$ and letting K converge towards zero, Fisher, Corbet and Williams (1943) found that the expected number of species with $n > 0$ individuals should be $\frac{\alpha}{n}x^n$. Thus, actually, the LS "is not a probability distribution, but a model for means" (Watterson , 1974): it is in fact an unstandardized distribution, denoted $ULSD$ by Rao (1971).

Remark 2. Other authors (Quenouille , 1949; Boswell and Patil , 1971; Taylor, Kempton and Woiwod , 1976) considered instead the normalized series (LSD), which depends only on x, since in this case $\alpha = -1/\ln\left(1-x\right)$ is not to be estimated.

Rao (1971) noted that Fisher's demonstration was not correct (see also Boswell and Patil (1971, p. 101)) and proved that, if a vector of counts of length β results from the Gamma-Poisson model, we would actually have:

$$E\left(f_r\right) = \alpha \frac{\mathfrak{P}^{\mathfrak{r}}}{\left(1+\mathfrak{P}\right)^{r+K}} \begin{pmatrix} K+r-1 \\ K-1 \end{pmatrix} \tag{3.3}$$

where f_r denotes the frequency of counts represented by $r \geq 1$ individuals and $\alpha = K\,\beta$. Thus, the distribution of average frequencies would be that of an **unstandardized zero-truncated** distribution, named $UNBD\left(K, \mathfrak{P}, \alpha\right)$ by Rao (1971). Afterward, under the conditions

$$\begin{gathered} (\beta, K) \to (\infty, 0) \\ \beta\, K = \alpha \end{gathered} \tag{3.4}$$

he obtained the $ULSD\left(\alpha, x\right)$.

To estimate the parameters of $UNBD\left(K, \mathfrak{P}, \alpha\right)$, Rao (1971) proposed a pseudo-maximum likelihood method, whose system of equations is:

(3.5)

$$\begin{gathered} S = \sum_{r=1}^{\mathcal{R}} f_r = \alpha \frac{1 - \left(\left(1+\mathfrak{P}\right)^{-K}\right)}{K} \\ N = \sum_{r=1}^{\mathcal{R}} r\, f_r = \alpha\, \mathfrak{P} = \alpha \tfrac{x}{1-x} \\ \sum_{r=1}^{\mathcal{R}} f_r \sum_{i=1}^{r-1} \frac{1}{K+i} = \alpha\, Log\left(1+\mathfrak{P}\right) \tfrac{1-\left(\left(1+\mathfrak{P}\right)^{-K}\right)}{K} + \tfrac{\alpha}{K^2}\left(-1 + \left(1+\mathfrak{P}\right)^{-K}\left(1 + K\, Log\left(1+\mathfrak{P}\right)\right)\right) \end{gathered}$$

where $\mathcal{R}$ **is the largest observed count**. Notice that the second equation of systems (3.5) and (3.1) are identical and that $\lim_{K\to 0} \alpha \frac{1-\left(\left(1+\mathfrak{P}\right)^{-K}\right)}{K} = \alpha\, Log\left(1+\mathfrak{P}\right) = -\alpha\, Log\left(1-x\right)$. As a consequence, when K is small enough, one shouldn't discern much differences between fitting a vector of counts by the zero-truncated distributions $TNBD\left(K, \mathfrak{P}\right)$, $UNBD\left(K, \mathfrak{P}, \alpha\right)$ or $ULSD\left(\alpha, \frac{\mathfrak{P}}{1+\mathfrak{P}}\right)$. It is interesting to compare issues from these models, since

- we can use two different methods for estimating K: the MLE for $TBND$ (Wyshak , 1974), and the system (3.5) for $UNBD$
- $\mathfrak{P}$ being a parameter common to all the distributions, there are three ways for estimating it: systems (3.5) or (3.1), and MLE for $TBND$
- we can use two methods for estimating α, by solving either (3.5) or (3.1); but notice that we cannot expect a common value of this parameter when conditions (3.4) are not fulfilled (approximately, at least); consequently, it is interesting to investigate whether or not the Williams-Rao's condition $\alpha \approx K\,\beta$ was actually fulfilled by the data (real, or simulated).

4. A robust estimator of $(K, \mathfrak{P})$

It is well-known that MLE suffers from several weaknesses: it can be biased, neither its uniqueness nor its existence is guaranteed and it is not robust in general, because its influence function at the model is sensible to outliers or aberrant data (Basu, Shioya and Park , 2011; Simpson , 1987). In addition, its computational cost can be excessive: Bliss and Fisher (1953), for instance, claimed that MLE of the

parameters of $NBD(K, \mathfrak{P})$, is "practicable and rapid when the largest observation does not exceed 20 or 30" (this weakness has mostly disappeared now; nevertheless, see Section 5.3).

While estimating the $ULSD(\alpha, x)$ parameters doesn't cause any problem, since equation (3.1) is quite easy to solve numerically, we encountered a number of convergence issues when fitting $TNBD(K, \mathfrak{P})$ or $UNBD(K, \mathfrak{P}, \alpha)$ with MLE or pseudo-MLE, like Vaudor, Lamouroux and Olivier (2011) or Anscombe (1950). Consequently we turned ourselves to another method: the Minimum Hellinger Distance Estimator (MHDE).

4.1. **The robustness of Minimum Distance Estimators.** In the seventies, Beran (1977) established the consistency and asymptotic efficiency of MHDE for absolutely continuous distributions, as well as its minimax robustness. Ten years after, Simpson (1987) proved that it has similar desirable properties (asymptotic normality and high breakdown point) for probabilities supported by $\mathbb{N}$, like NBD or LSD. More recently, Basu, Shioya and Park (2011) extended Minimum Distance Estimation to a large family of statistical disparities including the Hellinger distance. They proved (Basu, Shioya and Park , 2011, p. 43-45) that Minimum Distance Estimators have the same influence function at the model as MLE (consequently, all of them are first-order efficient). To compare first-order efficient estimators, Rao (1962) introduced second-order efficiency and proved that in the case of multinomial distributions, the MLE is second-order efficient, contrarily to several classical estimation methods (minimum chi-square, minimum discrepancy, minimum Kullback-Leibler divergence, and MHDE). Among these alternative methods, the MHDE was the best one (Rao , 1962, p. 51). Nevertheless, as pointed by Mandal and Basu (2013), the bias of Minimum Distance Estimators is generally greater than the one of MLE. That is why these authors, to keep the robustness and decrease the bias of such estimators, introduced some penalization on the "inliers" (cells with less data than expected under the model). Anyway, there is indeed no unbiased estimator for the parameters of NBD (Wang , 1996), and the unicity of the MHDE is not guaranteed...

4.2. **The estimator.** Let $\mathcal{R}$ be the largest observation in a vector of zero-truncated counts, and $p = \{p_1, \cdots, p_{\mathcal{R}}\}$ be the associated proportions. Denoting

$$\Pi_{(K,\mathfrak{P})}(r) := \frac{P(NBD(K, \mathfrak{P}) = r)}{1 - \frac{1}{(1+\mathfrak{P})^K}}$$

the probability density associated with the $TNBD$, we can define the Hellinger distance between the probabilities p and $TNBD(K, \mathfrak{P})$:

(4.1)

$$d_H(p, TNBD(K, \mathfrak{P})) := \frac{1}{\sqrt{2}} \sqrt{\sum_{r=1}^{\mathcal{R}} \left(\sqrt{p_r} - \sqrt{\Pi_{(K,\mathfrak{P})}(r)}\right)^2 + \sum_{r>\mathcal{R}} \Pi_{(K,\mathfrak{P})}(r)}$$

To neutralize the influence of empty cells on d_H (see the right part of Formula 4.1), a (twice, squared) penalized form of d_H has been introduced (Basu, Shioya and Park , 2011, Section 6.2):

$$PHD_h(p, f_\theta) := 2\sum_{r:p_r>0}\left(\sqrt{p_r} - \sqrt{f_\theta(r)}\right)^2 + h\sum_{r:p_r=0} f_\theta(r) \tag{4.2}$$

where h is positive and f_θ is a probability distribution belonging to some given family of probabilities supported by $\mathbb{N}$. We adopted the default value recommended by these authors: $h = 1$.

5. The MEEZ Data analysis

5.1. **Data description.** The Mauritanian coast, situated on the Atlantic side of the northwestern African continent, embeds a wide long continental shelf of about $750\,km$ and $36000\,km^2$ (see Figure 5.1) with an Exclusive Economic Zone (MEEZ) of $230000\,km^2$. The study area extends of 16°05 N in the South with the border of Senegal and up to 20°36 N in the North with the Western Sahara area. This study focuses on the analysis of abundance of fish and invertebrates data collected during annual scientific trawl surveys performed by oceanographic vessels, *N'Diago* until 1996 and with *Al Awam* since 1997 to now on the continental shelf (<200 m depth). These IMROP (Institut Mauritanien de Recherches Océanographiques et des Pêches) vessels have similar performances. The sampling strategy and the observation protocol remained the same during the 24 years of the study. The sampling method consists in a random stratified sampling design (Bergerard, Domain and Richer de Forges (1983); Domain (1986)).

We analyzed 48 demersal fish surveys taking account the number of those stations which trawling is more than or equal to 60 hauls, a total of 4589 stations are retained. These surveys conducted between 1987 to 2010 and covered the entire Mauritanian continental shelf (see Figure 5.1). Trawling speed varied between 2.5 and 3.95 knots, and the duration of fishing ranged from 15 to 40 minutes. All the species (fish and invertebrate) captured in a given station were identified, counted and then recorded on the database. Abundance data were standardized per half an hour of trawling in order to adjust variability in trawling duration. In addition, each station has been characterized by supplementary environmental variables: bathymetry, sedimentary type of the substrate, latitude and longitude.

Groundfish assemblages properly sampled in the MEEZ were composed on 543 species, belonging to 322 genera and 176 families.

The set of counts associated with each species sampled in the MEEZ consist of a mass of spatio-temporal observations. Because the spatial distribution of groundfish species is strongly influenced by the physical environment (Lamouroux et al. , 1999; Gaertner & al. , 1999; Johnson & al. , 2013), we split each one of these sets into an appropriate number of subsets (replicates), associated with homogeneous physical conditions (typical habitats). Then, for each species in each typical habitat, we estimated from these replicates the parameters of the distributions $TNBD(K, \mathfrak{P})$, $UNBD(K, \mathfrak{P}, \alpha)$ and $ULSD\left(\alpha, \frac{\mathfrak{P}}{1+\mathfrak{P}}\right)$. Afterward, we compared the estimators and determined the best one.

5.2. **Constituting habitats and replicates.** The sampled stations were distributed in a vast zone of various geographical, bathymetric and sedimentary specifics. We established a typology of trawl stations according to their bathymetry (denoted B) and sedimentary nature (denoted S), defining **typical habitats**.

More specifically, each station is associated with a vector (i, j, B, S), whose two first coordinates are longitude and latitude. We underscore that to any sampled position (i, j) is associated a whole set $\omega_{(i,j)}$ (of size $N_{(i,j)}$) of environmental characteristics vectors, corresponding to all the stations sampled in places confounded with (i, j), because of lack of precision in the position of boats:

$$\omega_{(i,j)} := \left\{ \left(B_{m_{(i,j)}}, S_{m_{(i,j)}} \right) : 1 \leq m_{(i,j)} \leq N_{(i,j)} \right\}.$$

Consequently, classical methods designed for spatial data (kriging, thin-plate splines, regression, *etc.*) cannot be used. We instead consider that we face a continuous random field of probability distributions, associating to each position (x, y) a probability distribution $\Omega_{(x,y)}$ such that $\omega_{(i,j)}$ is a sample of $\Omega_{(i,j)}$. Thus, to any position (i, j) is associated a probability density describing the local distribution of environmental characteristics. Our problem consists in classifying such functions under contiguity constraints. Dabo-Niang, Yao, Pischedda, Cuny and Gilbert (2010) proposed a method for clustering such data, which has been adapted for our purpose. In short, it consists in

(1) choosing a distance $\Delta(P, Q)$ between probabilities: we chose the discrepancy metric (Gibbs and Su , 2002):

$$\Delta(P, Q) = \sup_{\mathfrak{b} \subseteq \mathcal{D}} |f_P(\mathfrak{b}) - f_Q(\mathfrak{b})|,$$

where f_P and f_Q denote associated density estimates and $\mathfrak{b}$ is some closed ball of the domain of variation $\mathcal{D}$ of the environmental characteristics

(2) estimating the spatial density $\mathcal{S}(x, y)$ of environmental characteristics in the neighborhood of (x, y), through a kernel method (Dabo-Niang, Hamdad, Ternyck and Yao , 2014)

(3) detecting the set $\{\Omega_m, 1 \leq m \leq M\}$ of **functional modes** of $\mathcal{S}(x, y)$ (see Dabo-Niang, S., Ferraty, F. and Vieu, P. (2007); Ferraty and Vieu (2006, Ch. 9), or Gasser, Hall and Presnell (1998)); the number M of modes is determined by successive splittings, until the obtained group can be considered as homogeneous enough (Dabo-Niang, S., Ferraty, F. and Vieu, P. , 2007; Dabo-Niang, Hamdad, Ternyck and Yao , 2014)

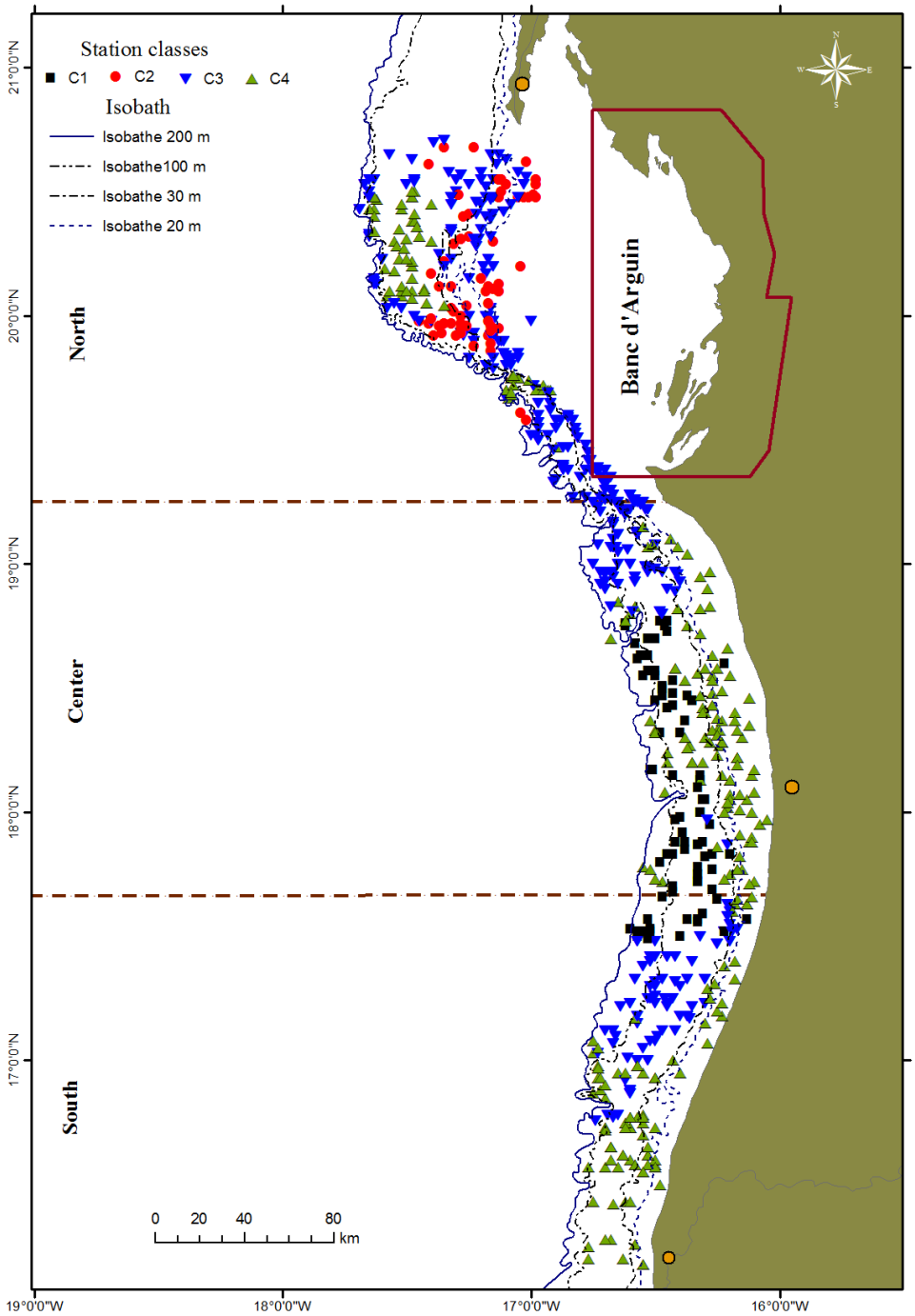
(4) building a partition of the stations by assigning the station (i, j) to the mth class if $\Delta\left(\Omega_m, \omega_{(i,j)}\right)$ is minimal.

For further details of this method, see (Dabo-Niang, Yao, Pischedda, Cuny and Gilbert , 2010; Dabo-Niang, Hamdad, Ternyck and Yao , 2014).

It was found that the optimal number of modes (or classes) for the MEEZ data was $M = 4$. The obtained typology of stations is represented on Figure 5.1. Then, for each species and each habitat, a list of counts has been constituted. We considered that such lists consisted of **replicates** sampled in similar environmental conditions.

We will focus in the next section on the "test habitat" C4: it is a sandy or sandy-muddy habitat, with depth lower than 100m. This is a zone of seasonal upwelling, while C2 is situated in a zone of permanent upwelling and neither C1 nor C3 are affected by this important phenomenon. More precisely, C4 is under the influence of two ocean currents; these currents and the profile of the continental shelf trigger an important seasonal upwelling phenomenon, from December to March. These water masses (less saline and poor in nutrients) result from the intensification of the

FIGURE 5.1. Map of trawl stations in four different station classes identified by a clustering method based on environmental variables: bathymetry, sedimentary types and geographic positions.



Guinea current in the Cap Blanc area. Consequently, C4 is a high plankton productivity area, supporting a large variety of fish communities, with many commercial species that sustain fishing activities.

5.3. **Comparing the estimators (counts from C4).** A great number of species (541) were found in this habitat, but many of them were rarely observed. More precisely, 240 species were observed less than 6 times (in other words, their number of counts was $\beta \leq 5$), and discarded from subsequent analysis. But extremely abundant species cause problems too! Suppose for instance that for some species the maximum observed count was $\mathcal{R} = 95211$ (a real case). Then, solving the system (3.5) or minimizing (4.2) is practically impossible (excessive time and memory consumption). While Bliss and Fisher (1953) recommended that $\mathcal{R}$ should be less than 30, we fixed 3000 as a maximum. Thus, in the estimation steps of TNBD we used, for extremely abundant species, truncated count vectors of length $\underline{\mathcal{R}} \leq \min(3000, \mathcal{R})$, while the genuine value $\mathcal{R}$ could be kept for evaluating goodness of fit. Among the 301 species kept, 46 species were extremely numerous, and we fitted their truncated counts.

We will now compare the results obtained on this habitat with four estimators:

(1) the classical estimator for the parameters of $ULSD(\alpha, x)$ (Fisher, Corbet and Williams , 1943)
(2) the classical MLE for the parameters of $TNBD(K, \mathfrak{P})$ (Wyshak , 1974)
(3) the pseudo-MLE for $UNBD(K, \mathfrak{P}, \alpha)$ (Rao , 1971)
(4) the MHDE for $TNBD(K, \mathfrak{P})$, obtained by globally minimizing the PHD_h (4.2), with $h = 1$ and $f_\theta := \Pi_{(K,\mathfrak{P})}$.

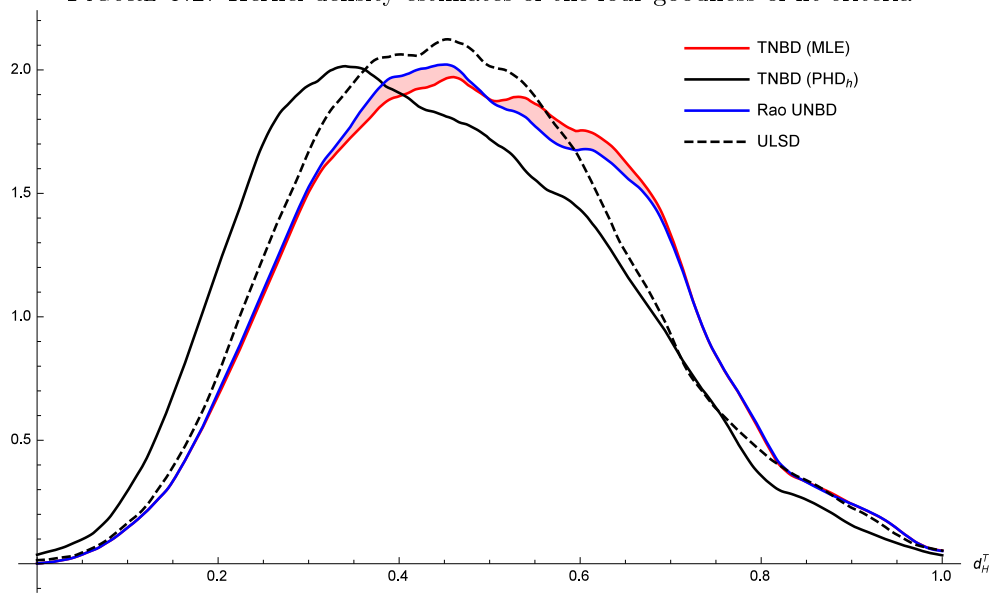
5.3.1. *A global insight: goodness of fit statistics per estimator.* The quality of fit was evaluated by the truncated Hellinger distance:

$$[0,1] \ni d_H^T(p, f_\theta) := \frac{1}{\sqrt{2}} \sqrt{\sum_{r=1}^{\mathcal{R}} \left(\sqrt{p_r} - \sqrt{f_\theta^{\mathcal{R}}(r)} \right)^2} \tag{5.1}$$

where $f_\theta^{\mathcal{R}}(r) := \frac{f_\theta(r)}{1 - f_\theta(0) - \sum_{r > \mathcal{R}} f_\theta(r)}$ (same notations as in Section 4). To each one of the 301 retained species, e (say), are associated four goodness of fit criteria: $d_H^T\left(p^e, \Pi^e_{(K,\mathfrak{P})_{MLE}}\right)$, $d_H^T\left(p^e, \Pi^e_{(K,\mathfrak{P},\alpha)_{Rao}}\right)$, $d_H^T\left(p^e, \Pi^e_{(K,\mathfrak{P})_{PHD_h}}\right)$ and $d_H^T\left(p^e, \Pi^e_{(\alpha,x)_{ULSD}}\right)$. We plotted kernel density estimates of these criteria (sample size=301) on Figure 5.2; the reader can see that, in general, the goodness of fit of the MLE for $TNBD$ and $UNBD$ are nearly indistinguishable (this point is examined in more details hereunder), while $ULSD$ is slightly better and the PHD_h for $TNBD$ is the best. More precisely, the best estimator was PHD_h for $TNBD$ (260 species), followed by $ULSD$ (32 species), MLE for $TNBD$ (4 species) and pseudo-MLE for $UNBD$ (5 species).

5.3.2. *Coherency of estimations of* $(K, \mathfrak{P})$ *obtained from MLE or pseudo-MLE.* Notice first that theses estimators are associated with very different model since, like LS, pseudo-MLE is "a model for means" (Watterson , 1974). There were only negligible differences between these two estimators, in general. On Figure 5.3, we plotted the estimations of both the parameters of the $TNBD$. We can see that,

FIGURE 5.2. Kernel density estimates of the four goodness of fit criteria



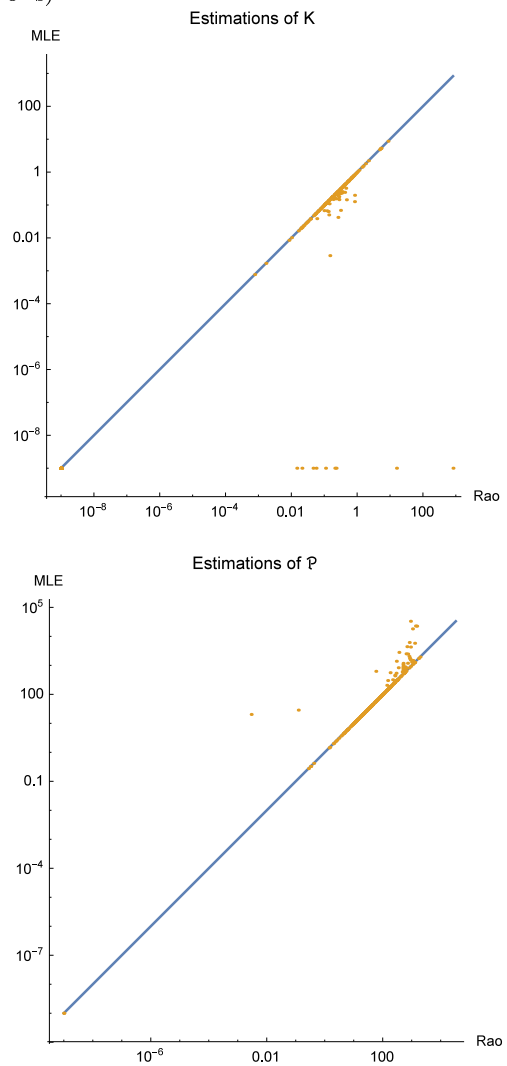
for most of the species, the estimations are coherent with each other and that, generally $K_{MLE} \leq K_{Rao}$ while $\mathfrak{P}_{MLE} \geq \mathfrak{P}_{Rao}$. More precisely, 262 species (87% of the species) were such that $|K_{MLE} - K_{Rao}| < 10^{-4}$; in theses cases $|\mathfrak{P}_{MLE} - \mathfrak{P}_{Rao}|$ was very small too. In conclusion, one may say that these estimators were quite coherent with each other. Consequently, we considered that the original MLE method was useless for our purpose and drop it, since the pseudo-MLE method give us in addition an estimation of α.

5.3.3. *Coherency of estimations of* $(K, \mathfrak{P})$ *obtained from pseudo-MLE or* PHD_h . Figure 5.4 shows that the results were very different from the previous ones: generally, $K_{PHD_h} > K_{Rao}$ and $\mathfrak{P}_{PHD_h} < \mathfrak{P}_{Rao}$. Moreover, the upper panel of this figure shows that the PHD_h seems free from convergence problems: there are a number of points on the vertical axis corresponding to "pathological" species, whose pseudo-MLE converged towards negative (or complex) values of K_{Rao}; in such cases, we arbitrarily fixed $K_{Rao} = 10^{-6}$. The reader can see on this panel that a number of these aberrant values indeed correspond to acceptable values of K_{PHD_h}.

5.3.4. *Checking the Williams-Rao's condition:* $\alpha = K\,\beta$. Notice that this equality is implicit in equation (3.3), but that it not a constraint in the root finding of the pseudo-MLE system (3.5). Consequently, the relationship $\alpha \approx K\,\beta$ must be considered as a sign of consistency of the estimation of $(K, \mathfrak{P}, \alpha)$.

Let us investigate whether this condition is at least approximately fulfilled by the C4 data, *i.e.* whether the hypothesis $\alpha - \beta K \approx 0$ (see Sections 2.1 &3) is not clearly unacceptable in general. Remember that α can be estimated from both the systems (3.1) and (3.5), while $(K, \mathfrak{P})$ can be estimated either from the system (3.5) or (see Section 5.3.3) by minimizing the PHD_h given by Formula (4.2). We have here to face an additional estimation task: because of the dubious nature

FIGURE 5.3. Simultaneous log-log plots of K and $\mathfrak{P}$ (MLE and Rao's methods).

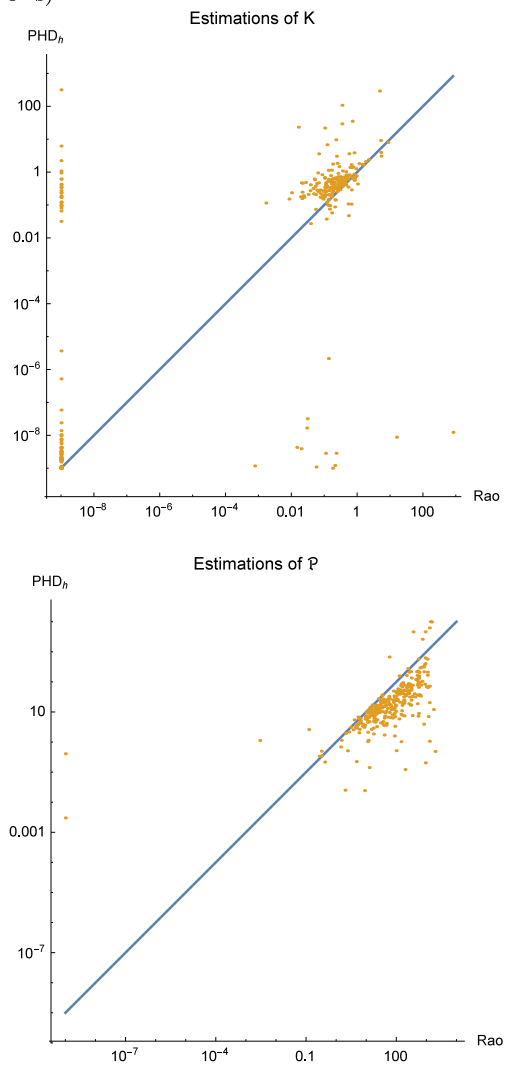


(stochastic or structural) of the collected zeros, the true value of β is unknown, while we only know the number β^+ of strictly positive counts. Then, it is classical to estimate β by:

$$\beta^{\left(\hat{K},\hat{\mathfrak{P}}\right)} := \frac{\beta^+}{\left(1 - 1/\left(1+\hat{\mathfrak{P}}\right)^{\hat{K}}\right)}$$

where $\left(\hat{K},\hat{\mathfrak{P}}\right)$ is an estimation of the TNBD parameters, obtained either from pseudo-MLE or PHD_h. The simulations carried on in Appendix 2 show that both

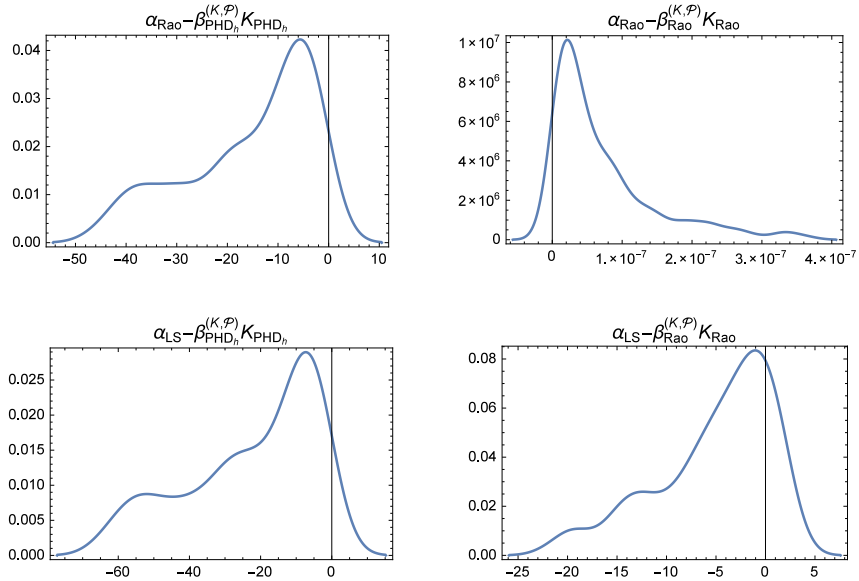
FIGURE 5.4. Simultaneous log-log plots of K and $\mathfrak{P}$ (PHD_h and Rao's methods).



these estimators give similar results, excepted in the considered "aggregative" case ($K = 10^{-4}$), where none of them estimates β well.

Notice that for about 34% of the species from the MEEZ, the estimate $\beta_{Rao}^{(K,\mathfrak{P})}$ of β was much greater than the total number of hauls; in the case of $\beta_{PHD_h}^{(K,\mathfrak{P})}$, this happened for 26% of the species. Probably, these species were aggregative ones, whose counts could not be fitted from any sample of reasonable size (see Appendix 2). We displayed on Figure 5.5 the results associated with all the estimations of α, β and $(K, \mathfrak{P})$ obtained from the MEEZ data. On this figure, we can see that $|\alpha - K\,\beta|$ was generally moderate, but was very small only when the parameters were estimated by pseudo-MLE.

FIGURE 5.5. Empirical verification of Williams-Rao's hypothesis for the MEEZ data (20% of upper and 20% of lower values discarded).



6. ECOLOGICAL RESULTS

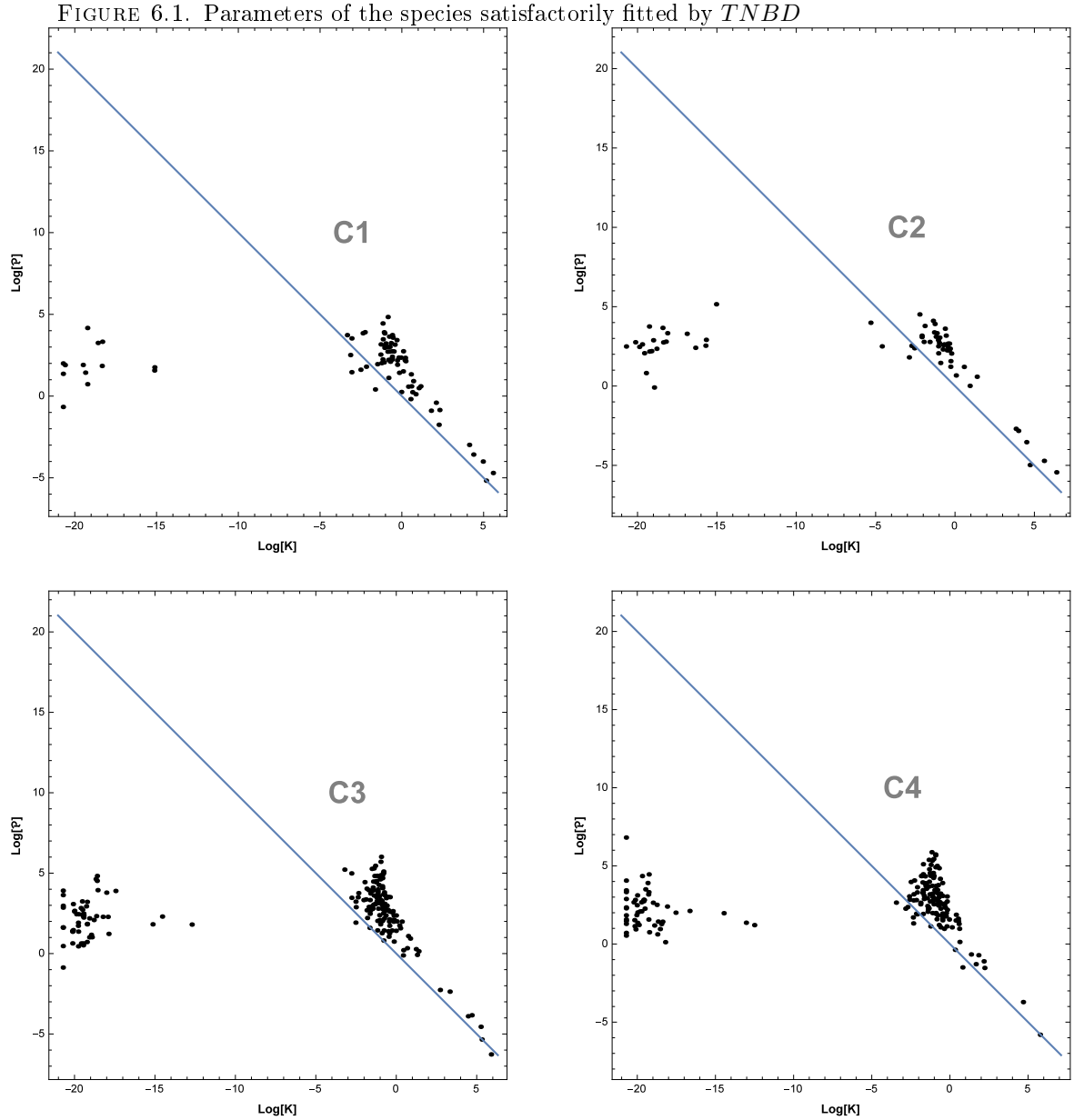
We represented on Figure 6.1 a selection of "Negative Binomial species" found in the MEEZ; these species were such that they were better fitted by some $TNBD\,(K,\mathfrak{P})_{Best}$ (the best of the three estimators, according to criterion (5.1)) than by $ULSD\,(\alpha,x)$. Since the counts of a number of species could not probably be correctly fitted by any standard distribution, we imposed in addition the constraint:

$$d_H^T\left(p^e,\Pi^e_{(K,\mathfrak{P})_{Best}}\right)\leq 0.53$$

determined from a Monte Carlo experiment detailed in Appendix 1. In other words, the species e is displayed in some panel of Figure 6.1 if its counts in the corresponding habitat were better fitted by $TNBD$ that by $ULSD$, and if the goodness-of-fit was satisfactory. We retained this way 84 species in C1, 67 in C2, 193 in C3 and 193 in C4. The habitats C3 and C4 shared about 68% of the species selected, while 42 species were common to the four habitats.

It is interesting to examine the relationships between K and $\mathfrak{P}$, in connection with the considerations of Section 2. If the right model for the data is the classical Gamma-Poisson one, the estimated parameters could be independent. On the contrary, if the right model is Kendall's one, the relationship (2.1) between the parameters could hold. Finally, in the case of the group-size model, the parameters K and $\mathfrak{P}$ would also be interrelated (Boswell and Patil , 1970, 1971), but the relationship would depend on unknown (social) groups and individual factors.

FIGURE 6.1. Parameters of the species satisfactorily fitted by $TNBD$



Such models proved their efficiency for modeling primate social dynamics (Cohen , 1972), but are they well-suited for fish populations?

On Figure 6.1, we superimposed to these estimations the line $Log\left(\mathfrak{P}_{\mathfrak{e}}\right) = -Log\left(K_e\right)$ corresponding to Kendall's model (see Section 2.3) with $\iota_e \approx \mu_e$. It is noteworthy that most of the retained species seem compatible with this model, with (in general) a positive additional term $Log\left(\frac{\iota_e}{\mu_e}\right)$ - see Formula (2.1). Notice also that in the setting of Kendall's model, $Log\left(K_e\right) \leq 0 \Longleftrightarrow K_e = \frac{\iota_e}{\rho_e} \leq 1$. Thus, for most

of the displayed species, the mortality rate should slightly exceed the immigration rate. In all the habitats, most of the selected species compatible with the Kendall's model were such that $\iota_e \approx \mu_e$, but few species were such that $\mathfrak{P} < 1$, *i.e.* $\mu \gg \rho$. In each panel there is also a minority of aggregative species discordant with Kendall's model, associated with values of K smaller than e^{-10}.

Finally, the counts of very few species were better fitted and well-represented by the $ULSD$: 13 species in C1, 6 in C2, 16 in C3 and 20 in C4.

7. Conclusion - Discussion

We investigated the performance and coherency with each other of three statistical models for overdispersed positive counts: the truncated Negative Binomial distribution $TNBD\,(K, \mathfrak{P})$, the three-parameter variant $UNBD\,(K, \mathfrak{P}, \alpha)$ of Rao (1971) and the Fisher's log-series. We focused on results obtained in the test habitat C4, but we stress that quite similar ones were obtained in C1, C2 and C3. Processing the MEEZ data, we thus found that:

(1) the Maximum Likelihood estimations of $(K, \mathfrak{P})$ for $TNBD$ and $UNBD$ were very close to each other
(2) processing either real count or simulated ones, we found similar performances of the estimators
(3) the Williams-Rao's condition: $\beta\, K = \alpha$ was roughly fulfilled by most species
(4) the penalized minimum Hellinger distance estimator of $(K, \mathfrak{P})$ for $TNBD$ performed better than the other ones, in general.

From the Ecological side, we found that, even if 543 species were sampled, it was possible to satisfactorily estimate the parameters of less than half of them, because of the rarity of most species (a general and problematic phenomenon: see Kunin, W.E. and Gaston, K.J. (1997); Manté, Durbec and Dauvin (2005); Manté, Claudet and Rebzani-Zahaf (2003)). These manageable species could be split into two categories. The first one is composed of very aggregative NB species, such that $K \approx 0$, and of species obeying a Log-series distribution. The second category consists of moderately aggregative species (the most numerous ones), obeying some distribution $TNBD\,(K, \mathfrak{P}) : K \gg 0$. It is worth noting that species of the second category seem consistent with the population growth model of Kendall (1948). Rather surprisingly, no species obeyed the Poisson distribution (*i.e.* was indifferent to the presence of fellow creatures).

In this work, we focused on truncated Negative Binomial Distributions essentially because log-series are supported by strictly positive integers, and because many important references (Rao , 1971; Kendall , 1948; Fisher, Corbet and Williams , 1943; Williams , 1944) only considered such counts. Furthermore, it is well-known that the status of observed zeros is ambiguous in ecological surveys: are they stochastic, or structural? Zero-inflated models were designed for answering this question; according to some authors (Lewin et al. , 2010) they clearly outperform classical models, while other authors do not support them (Warton , 2005); Vaudor, Lamouroux and Olivier (2011) compared a number of models for counts, and selected the zero-inflated NBD model for only 1% of the samples! Another way to deal with extra zeros is the hurdle model, consisting in modeling the zeros by a separate process. O'Neil and Faddy (2003) processed this way recreational catch data, where the number of extra zeros (no fishing) largely depends on various events (holidays, bad weather, ability of fishers, etc.) stranger to the presence of fishes.

As Vaudor, Lamouroux and Olivier (2011), we think that this model is ill-suited for scientific systematic catches.

This is also to avoid the unsolvable problem of zeros that we fitted truncated counts and gathered the data according to habitats. Indeed, if an habitat is ill-suited for a species, one should not find it frequently in this habitat. Since species which were observed less than 6 times were excluded from subsequent analysis, we should not observe mixtures of stochastic and structural zeros. In addition, notice that in generic cases (sere Appendix 2), the number of (possible) structural zeros can be satisfactorily estimated by $\beta^{(\hat{K},\hat{\mathfrak{P}})} - \beta^{+}$ while, in the case of aggregative species, the number of stochastic zeros could be really much bigger than the number of hauls (this could be named "zero-deflation")! In the latter case, how could we determine the nature of some zero? For instance, suppose a theoretical "aggregative species" was found six times (the minimum to be taken into account in the study). With $(K, \mathfrak{P}) = (0.0001, 14.43)$, we should then have theoretically $\beta \approx \dfrac{6}{K\ \ln(1+\mathfrak{P})} > 21 \times 10^3$, while the total number of hauls in this study is 4589 (1928 in C4).

Now, what about the spatio-temporal structure of the MEEZ data? The spatial side has been taken into account in a special way, through a continuous random field of environmental characteristics (see Section 5.2), to build replicates for fitting NB or LS distributions to the 543 species caught in order to investigate their collective behavior. The most important spatial feature of our data, the presence of upwellings, was then taken into account through a typology of trawls stations. Our results show that a number of the species found in the MEEZ could probably be modeled using dynamical processes. But, to our knowledge, most spatio-temporal statistical models designed for similar count data (Aidoo et al , 2015; Nielsen et al. , 2014) are too sophisticated for dealing with a large number of species. Nevertheless, our results could probably be used for parametrization of the Poisson intensity involved in spatio-temporal models focusing on species of interest (Hooten and Wikle , 2008). Spatio-temporal exploratory methods (Di Salvo, F., Ruggieri, M. and Plaia, A. , to appear) are probably better suited for dealing with a large number of species, but supplementary issues should be addressed for processing marine ecological data:

- qualitative descriptive variables (such as sedimentology) should be included in the method
- because of the major role of turbulence (more active vertically than horizontally), space is no more isotropic.

Acknowledgments

We thank the Mauritanian Institute of Oceanographic Research and Fisheries (IMROP) 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, to Jean-Pierre Durbec for helpful discussions, and to anonymous reviewers for their constructive comments.

Appendix 1: determination of a threshold for the truncated Hellinger distance

While results about the asymptotic distribution of our estimators abound, nothing is known about the distribution of the goodness-of-fit index (see Formula (5.1))

$$d_H^T\left(TNBD\left(\hat{K},\hat{\mathfrak{P}}\right), TNBD\left(K,\mathfrak{P}\right)\right) \tag{7.1}$$

where $\left(\hat{K},\hat{\mathfrak{P}}\right)$ is an estimate of $(K,\mathfrak{P})$. In order to determine from d_H^T the species which were correctly fitted, we performed a Monte Carlo study. It consisted in generating a sample of the statistics (7.1) for each one of the three estimators used, from a population of "Negative Binomial species" similar to the genuine population of the C4 habitat considered as a reference structure. This study is detailed hereunder.

The reference distribution of $(K,\mathfrak{P})$. We plotted on Figure 7.1 the minimum PHD_h estimates of the vector $(K,\mathfrak{P})$ associated with the species collected in the C4 habitat. About 35.6% of the species were associated with very small values of the first parameter ($\hat{K} \leq e^{-10}$); discarding these species, we could fit a bi-dimensional log-normal distribution of parameters (μ_B,Σ_B) to the remaining vectors of estimates, whose confidence ellipsoids are also represented on Figure 7.1. Neither $Log\left(\hat{K}\right)$ nor $Log\left(\hat{\mathfrak{P}}\right)$ strictly obeyed a normal distribution, but this model was retained for sake of simplicity, since the corresponding 95% confidence region widely covers the data (see Figure 7.1). As for the discarded species, we postulated that $Log\left(\hat{\mathfrak{P}}\right)$ could be considered as obeying some Gaussian distribution, $\mathcal{N}\left(\mu_S,\sigma_S\right)$.

Generating a "population" consistent with the reference distribution. To build a sample of $d_H^T\left(TNBD\left(\hat{K},\hat{\mathfrak{P}}\right), TNBD\left(K,\mathfrak{P}\right)\right)$ for counts having the same overall characteristics as the C4 data, we generated random counts of 300 "NB species", whose random parameters obeyed the mixture distribution

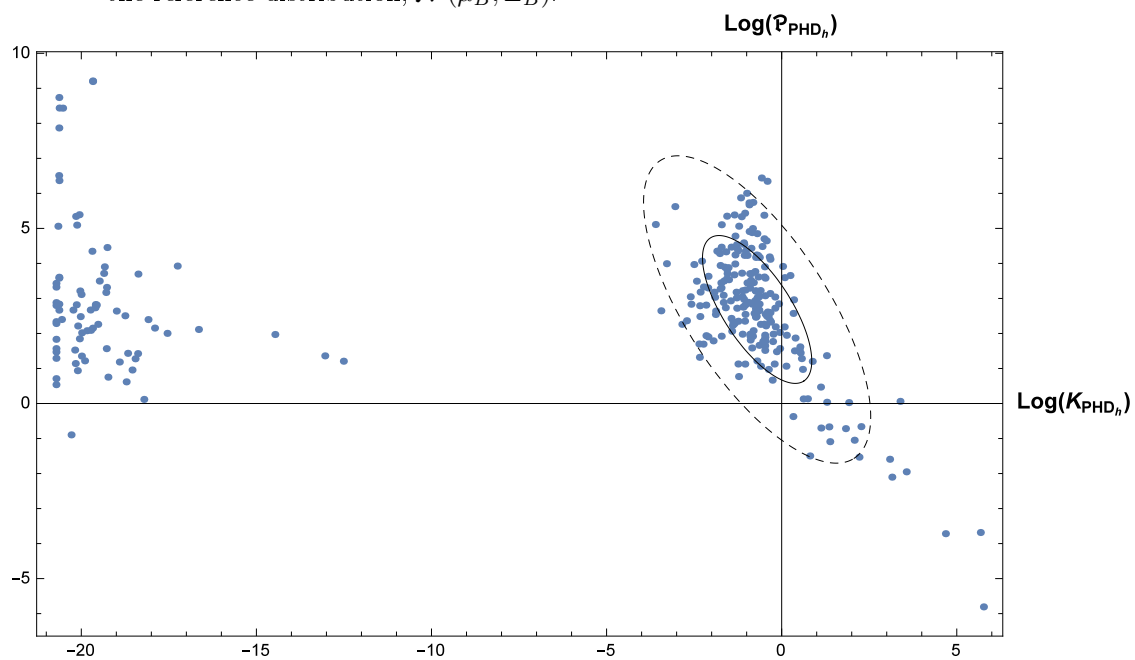
$$\mathcal{M} := 0.356\,\mathcal{U}\left(\left[e^{-12},e^{-8}\right]\right) \otimes \mathcal{LN}\left(\mu_S,\sigma_S\right) + 0.644\,\mathcal{LN}\left(\mu_B,\Sigma_B\right)$$

where $(\mu_S,\sigma_S) = (-0.701062, 2.68525)$ and (μ_B,Σ_B) were estimated from the C4 data. Practically, the parameters (k,p) of each one of the species was first drawn according to $\mathcal{M}$; then a sample of $\beta = 3000$ (or $\beta = 6000$ when $k \leq e^{-8}$) counts obeying $NBD\left(k,p\right)$ was drawn. The simulated data were then processed the same way as the MEEZ ones.

On Figure 7.2, we superimposed to the parameters of the species (estimated by PHD_h), confidence ellipsoids of the reference distribution $\mathcal{LN}\left(\mu_B,\Sigma_B\right)$ and of the distribution $\mathcal{LN}\left(\widehat{\mu_B},\widehat{\Sigma_B}\right)$, whose parameters were estimated from the independent drawns of $\mathcal{M}$. It is worth noting that in this case, there was no significant difference between the empirical distribution of $\left(\hat{K},\hat{\mathfrak{P}}\right)$ and the reference distribution $\mathcal{LN}\left(\mu_B,\Sigma_B\right)$ (P-values: Cramer-Von Mises =0.544087; Pearson $\chi^2 = 0.523489$).

Results. About 30% (93) of the "species" were observed less than 6 times, and discarded. The goodness-of-fit density estimates for the remaining ones are plotted on Figure 7.3. The reader can see that in the case of genuine TNB distributions, the

FIGURE 7.1. Fit of the estimated parameters for the C4 data. The ellipsoids correspond to 50% and 95% confidence regions for the reference distribution, $\mathcal{N}(\mu_B, \Sigma_B)$.



performances of the three estimators are very close to each other, while goodness-of-fit by $ULSD$ is very different. Among the 207 remaining species, only 6 (3%) were better fitted by ULSD, while the best estimator for $TNBD$ was PHD_h (185 species: 89% of the total), followed by MLE (10 species) and pseudo-MLE (6 species). Thus almost all "aggregative species" were discarded, due to their rarity.

The quantile of order 0.95 of the goodness-of-fit associated with PHD_h was 0.531096; consequently, we considered that 0.53 is an appropriate threshold for d_H^T, which should not be passed by genuine Negative Binomial species. This threshold has been used in Section 6.

APPENDIX 2: THE WILLAMS-RAO'S CONDITION AND THE ESTIMATION OF β

Notice that the equality $\alpha = K\,\beta$ is implicit in equation (3.3), but that it not a constraint in the root finding of the pseudo-MLE system (3.5). Consequently, this relationship must be considered as a sign of consistency of the estimation of $(K, \mathfrak{P}, \alpha)$. If in addition, the first condition of (3.4) is fulfilled, we should also have: $\dfrac{\alpha_{LS}}{K\,\beta} \approx \dfrac{\alpha_{Rao}}{K\,\beta} \approx 1$. We investigated the validity of this relationship by performing 50 Monte Carlo experiments with "Negative Binomial species" random draws, in each one of four typical cases:

- the "mean" one: $(K, \mathfrak{P}) = (1.193, 73.15)$ is the mean of the bivariate Log-normal distribution fitting the C4 habitat non-aggregative species (see Section 5.2 and Appendix 1)

FIGURE 7.2. In black: 50% and 95% confidence ellipsoids for the reference distribution $\mathcal{N}(\mu_B, \Sigma_B)$; in gray: same confidence ellipsoids for the distribution $\mathcal{N}\left(\widehat{\mu_B}, \widehat{\Sigma_B}\right)$ obtained from the mixture distribution $\mathcal{M}$. Dots correspond to the parameters of the "NB species", estimated by PHD_h.

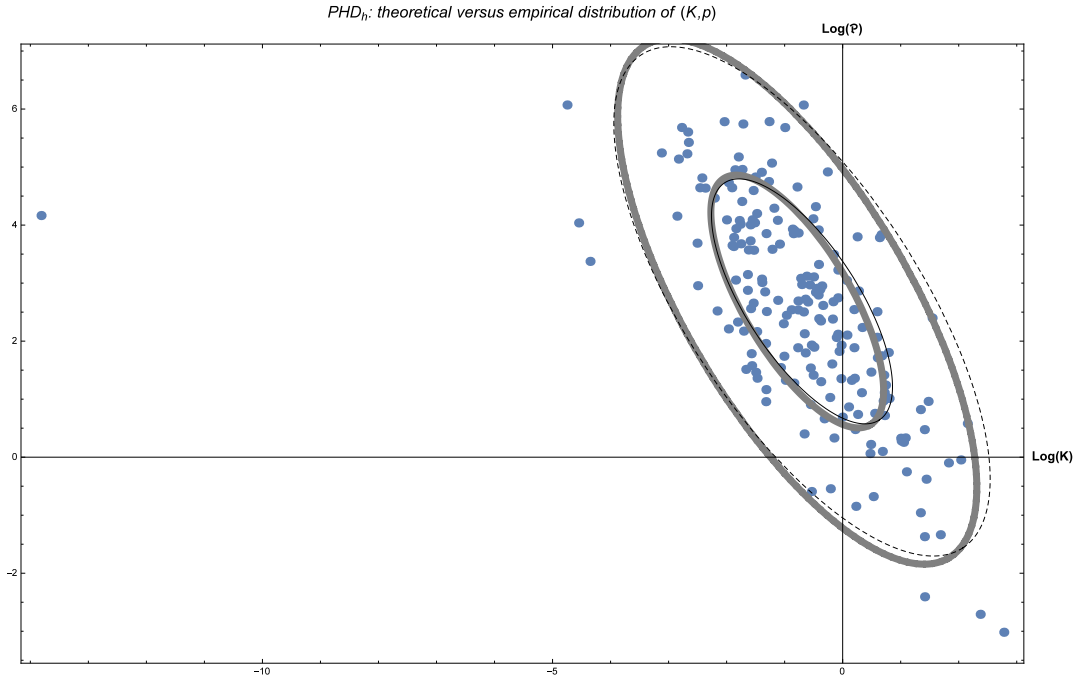


FIGURE 7.3. Simulations: density estimates of the four goodness-of-fit criteria.

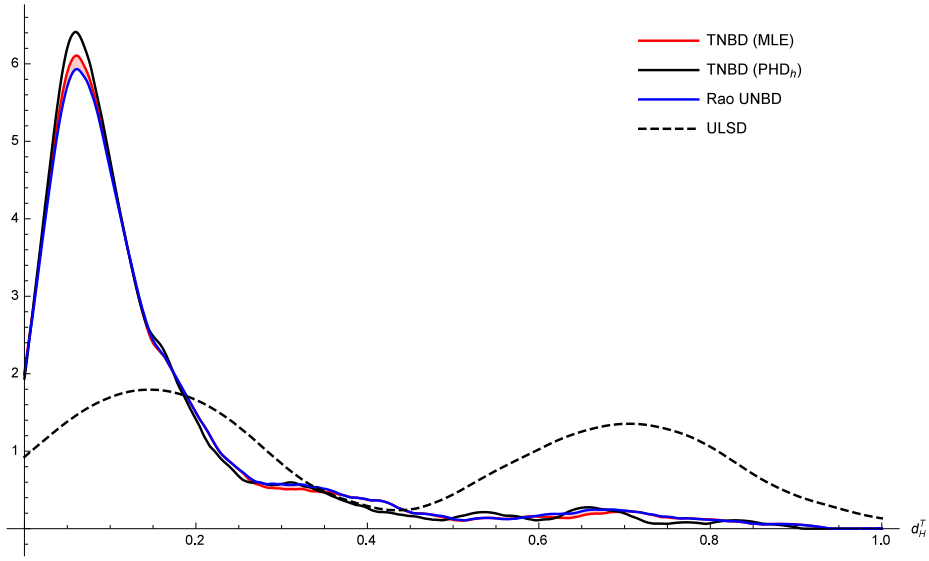


FIGURE 7.4. The Willams-Rao's condition in the bell-shaped case

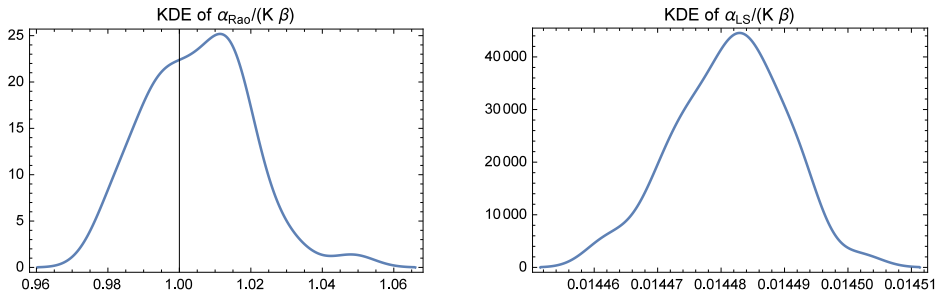
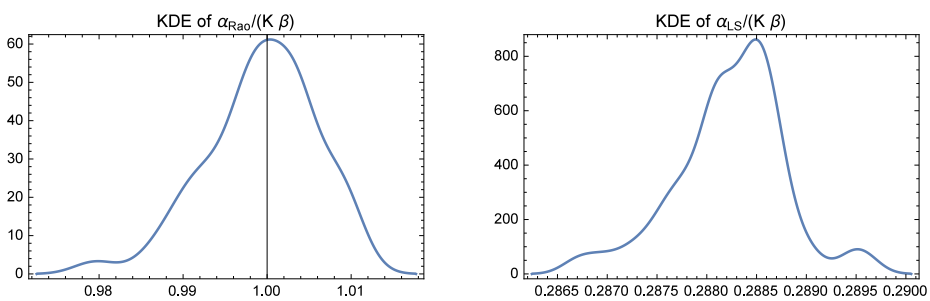


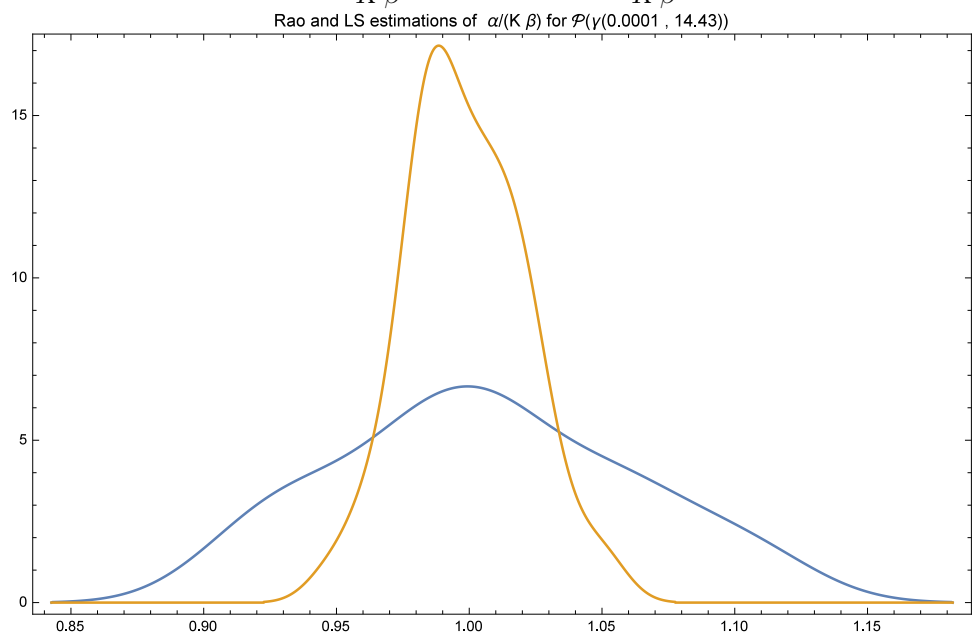
FIGURE 7.5. The Willams-Rao's condition in the "common" case



- the "common" one: $(K, \mathfrak{P}) = (0.7767, 14.43)$ is the spatial median (Serfling , 2004) of the parameters of the simulated "NB species"; in this case, we chose $\beta = 10^5$ as the sample size of each one of the 50 simulations
- an "aggregative" case: $(K, \mathfrak{P}) = (0.0001, 14.43)$, with $\beta = 10^7$
- a "bell-shaped" case: $(K, \mathfrak{P}) = (10, 14.43)$, with $\beta = 10^4$.

In these four cases, the best fit was obtained with PHD_h, and we observed that α_{LS} and α_{Rao} could be considered as normally distributed (according to the Cramer-von Mises test), and that the mean of α_{Rao} was always close to $K\,\beta$ (T test), while the relationship $\dfrac{\alpha_{LS}}{K\,\beta} \approx 1$ was verified only in the aggregative case (Figure 7.6). The equality $\alpha_{LS} = K\,\beta$ was unacceptable in the "common" case (see Figure 7.5), as well as in the bell-shaped one (Figure 7.4) and in the "mean" case (not shown).

FIGURE 7.6. The Willams-Rao's condition in the aggregative case - the blue density is $\frac{\alpha_{Rao}}{K\,\beta}$; the yellow one is $\frac{\alpha_{LS}}{K\,\beta}$



From another side, dealing with real data, we often have to face an additional estimation task: because of the dubious nature of the collected zeros (if there are), the true value of β is unknown and we only know the number β^+ of strictly positive counts. Then, it is classical to estimate β by:

$$\beta^{\left(\hat{K},\hat{\mathfrak{P}}\right)} := \frac{\beta^+}{\left(1 - 1/\left(1+\hat{\mathfrak{P}}\right)^{\hat{K}}\right)} \tag{7.2}$$

where $\left(\hat{K},\hat{\mathfrak{P}}\right)$ is an estimation of the TNBD parameters, obtained either from pseudo-MLE or PHD_h. The bell-shaped case is problem-free, because the probability of zero is extremely weak.

Let us now examine the most interesting case: the "common" one. On Figure 7.7, we plotted on the left panel KDE estimates of the 50 values of the expression (7.2) obtained with $\left(\hat{K},\hat{\mathfrak{P}},\hat{\alpha}\right) = (K_{Rao},\mathfrak{P}_{Rao},\alpha_{Rao})$ and divided by β: they were very close to 1. On the right panel, we plotted the values of (7.2) obtained with $\left(\hat{K},\hat{\mathfrak{P}}\right) = (K_{PHD_h},\mathfrak{P}_{PHD_h})$. Thus, this figure shows that all estimates of β are excellent in the "common" case; as a consequence, the results above concerning the Willams-Rao's condition (see Figure 7.5) stay valid in this case.

Quite similar results were obtained in the "mean" case, as the reader can see on Figure 7.8.

FIGURE 7.7. Estimating β: the "common" case

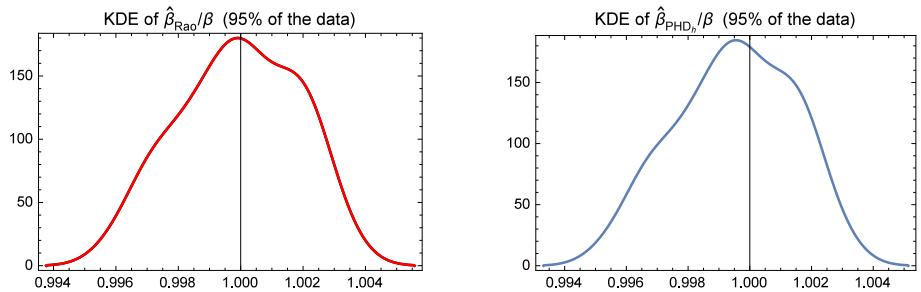
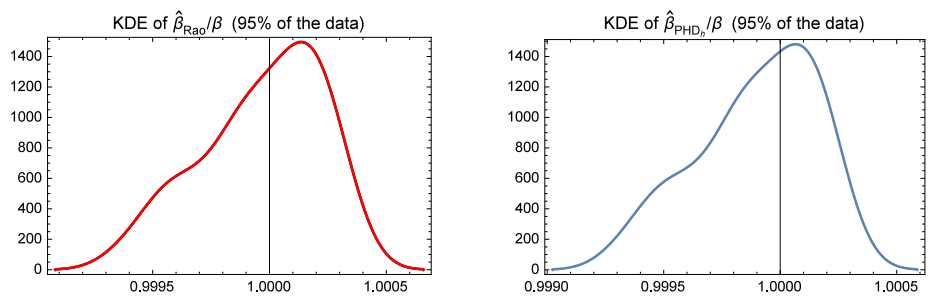
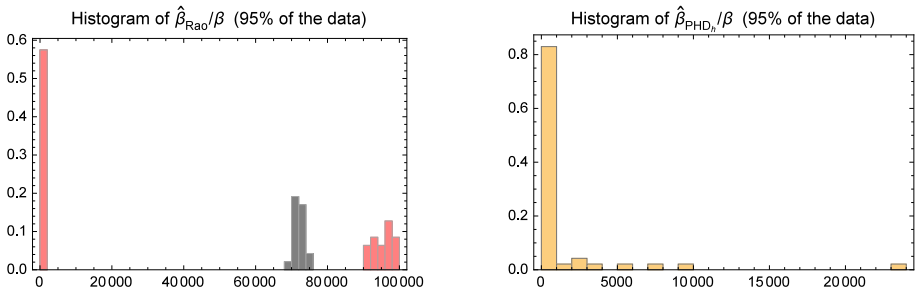


FIGURE 7.8. Estimating β: the "mean" case



Things are very different in the aggregative situation, as the reader can see on Figure 7.9: the estimator (7.2) based on $(K_{Rao}, \mathfrak{P}_{Rao}, \alpha_{Rao})$ strongly underestimates or overestimates β. More precisely, in about 40% of the samples, β was highly overestimated which is natural, since $\lim_{\hat{K}\to 0} \beta^{\left(\hat{K},\hat{\mathfrak{P}}\right)} = \dfrac{\beta^{+}}{\hat{K}\ \ln\left(1+\hat{\mathfrak{P}}\right)}$. As for the estimator (7.2) associated with $\left(\hat{K},\hat{\mathfrak{P}}\right) = (K_{PHD_h}, \mathfrak{P}_{PHD_h})$, it always underestimated β. Consequently, when K is very small, the consistency of the estimations of the parameters of $UNBD\,(K, \mathfrak{P}, \alpha)$ is questionable and the log-series model is probably more sound - even if the fit is not quite as good as with $UNBD\,(K, \mathfrak{P}, \alpha)$. This is the meaning of Figure 7.6, undoubtedly.

FIGURE 7.9. Estimating β in the aggregative case; the 5% upper values were dropped



References

Aidoo, E. N., Mueller, U., Goovarets, P. and Hyndes, G. A. (2015). Evaluation of geostatistical estimators and their applicability to characterise the spatial patterns of recreational fishing catch rates. *Fisheries Research*, 168, 20-32.

Anscombe, F.J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, 36: 358-382.

Basu, A., Shioya, H. and Park, C. (2011). *Statistical inference. The Minimum Distance approach.* Monographs on Statistics and Applied Probability 120, Chapman & Hall/CRC, Boca Raton.

Beran, R.J. (1977) Minimum Hellinger Distance Estimates for parametric models. *The Annals of Statistics*, 5: 445-463.

Bergerard P., Domain F., Richer de Forges B. (1983). Evaluation par chalutage des ressources démersales du plateau continental mauritanien. In: *Bull. Centr. Nat. Rech. Océanogr.et des Pêches, Nouadhibou, V*: 11, 217-290.

Bliss, C.I. and Fisher, R.A. (1953). Fitting the Negative Binomial distribution to biological data. *Biometrics*, 9:176-200.

Boswell, M.T. and Patil, G.P. (1970). Chance mechanisms generating the Negative Binomial distribution. In: Patil G.P. (ed), *Random counts in models and structures (Vol. 1)*. Pensylvania State University Press, pp 3-22.

Boswell, M.T. and Patil, G.P. (1971). Chance mechanisms generating logarithmic series distribution used in the analysis of number of species and individuals In: Patil G., Pielou E. and Waters W. (eds), *Statistical Ecology (Vol. 1)*. Pennsylvania State University Press, pp. 99-130.

Chen, J. and Rubin, H. (1986). Bounds for the difference between median and mean of Gamma and Poisson distributions. *Statistics & Probability Letters*, 4: 281-283.

Cohen, J.E. (1972). Markov population processes as models of primate social and population dynamics. *Theoretical Population Biology*, 3(2): 119-134.

Dabo-Niang, S., Ferraty, F. and Vieu, P. (2007). On the using of modal curves for radar waveforms classication. Computational Statistics and Data Analysis, 51(10), 4878-4890.

Dabo-Niang, S., Yao, A. F., Pischedda, L., Cuny, P. and Gilbert F. (2010). Spatial mode estimation for functional random fields with application to bioturbation problem. *Stochastic Environmental Research and Risk Assessment*, 24(4): 487-497.

Dabo-Niang, S., Hamdad, L., Ternynck, C. and Yao, A.F. (2014). A kernel spatial density estimation allowing for the analysis of spatial clustering. Application to Monsoon Asia Drought Atlas data. *Stochastic Environmental Research and Risk Assessment*, 28, 2075-2099.

Di Salvo, F., Ruggieri, M. and Plaia, A. (to appear). Functional principal components analysis for multivariate multidimensional environmental data. *Environmental and Ecological Statistics*, DOI 10.1007/s10651-015-0317-8.

Diserud, O.H. (2001). Detecting changes in diversity in a fluctuating environment based on simulation of stochastic processes. *Oceanologica Acta,* 24(5): 505-517.

Domain F. (1986). Evaluation par chalutage des ressources démersales du plateau continental mauritanien. In: Josse E. et Garcia S. (eds). *Description et évaluation des ressources halieutiques de la ZEE mauritanienne. Rapport du Groupe de travail CNROP/FAO/ORSTOM*, Nouadhibou, Mauritanie, 16–27 septembre 1985, COPACE/PACE Séries 86/37 : 248–273.

Elliot, J.M. (1979). *Some methods for the statistical analysis of samples of benthic invertebrates.* Freshwater Biological Association Scientific Publication n° 25 (2nd edn), Ambleside.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis. Theory and Practice.* Springer Series in Statistics.

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

Gaertner, J.C., Mazouni, N., Sabatier, R. and Millet, B. (1999). Spatial structure and habitat associations of demersal assemblages in the Gulf of Lions: a multi-compartimental approach. *Marine Biology*, 135: 199-208.

Gasser, T., Hall, P. and Presnell, B. (1998). Nonparametric estimation of the mode of a distribution of random curves. *J. R. Statist. Soc. B.*, 60, 4, 681-691.

Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review*, 70, 3, 419-435.

Hooten, M. B. and Wikle, C. (2008). A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics*, 15, 59-70.

Johnson, A.F., Jenkins, S.R., Hiddink, J.G. and Hinz, H. (2013). Linking temperate demersal fish species to habitat: scales, patterns and future directions. *Fish and fisheries*, 14: 256-280.

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

Kunin, W.E. and Gaston, K.J. (1997). *The biology of rarity. Causes and consequence of rare-common differences.* Chapman & Hall, London.

Lamouroux, N., Capra, H., Pouilly, M. and Souchon, Y. (1999). Fish habital preferences in large streams of southern France. *Freshwater Biology*, 42: 673-687.

Lewin, W.-C., Freyhof, J., Huckstorf, V. Mehner, T. and Wolter, C. (2010). When no catches matter: coping with zeros in environmental assessments. *Ecological Indicators*, 10, 572-583.

Magurran, A.E. (2005). Species abundance distributions: pattern or processes? *Functional Ecology*, 19: 177-181.

Mandal, A. and Basu, A. (2013). Minimum disparity estimation: improved efficiency through inliers modification. *Computational Statistics and Data Analysis*, 64, 71-86.

Manté, C., Claudet, J. and Rebzani-Zahaf, C. (2003). Fairly processing rare and common species in multivariate analysis of ecological series. Application to macrobenthic communities from Algiers harbour. *Acta Biotheoretica*, 51, 277-294.

Manté, C., Durbec, J.P. and Dauvin, J.C. (2005). 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(8), 831-840.

Nielsen, J. R., Kristensen, K., Lewy, P. and Bastardie, F. (2014). A statistical model for estimation of fish density including correlation in size, space, time and between species from research survey data. *PLOS ONE*, 9(6), 1-15.

O'Neil, M. F. and Faddy, M. J. (2003). Use of binary and truncated negative binomial modelling in the analysis of recreational catch data. *Fisheries Research*, 60, 471-477.

Quenouille, M.H. (1949). A relation between the Logarithmic, Poisson and Negative Binomial series. *Biometrics*, 5, 2: 162-164.

Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples (with discussion). *Journal of the Royal Statistical Society, Series B (Methodological)*, 24(1): 46-72.

Rao, C.R. (1971). Some comments on the logarithmic series distribution in the analysis of insect trap data In: Patil G., Pielou E. and Waters W. (eds) Statistical Ecology (Vol. 1). Pennsylvania State University Press, pp. 131-142.

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

Simpson, D.G. (1987). Minimum Hellinger Distance Estimation for the analysis of count data. *Journal of the American Statistical Association*, 82, 399: 802-807.

Taylor, L.R., Kempton, R.A. and Woiwod, I.P. (1976). Diversity statistics and the Log-Series model. *Journal of Animal Ecology*, 45(1): 255-272.

Taylor, L.R., Woiwod, I.P. and Perry, J.N. (1979). The Negative Binomial as a dynamic ecological model for aggregation, and the density dependence of K. *Journal of Animal Ecology*, 48: 289-304.

Vaudor, L., Lamouroux, N. and Olivier, J.M. (2011). Comparing distribution models for small samples of overdispersed counts of freshwater fish. *Acta Oecologica*, 37(3): 170-178.

Wang, Y. (1996). Estimation problems for the two-parameters negative binomial distribution. *Statistics & Probability Letters*,26: 113-114.

Warton, D. I. (2005). Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics*, 16, 275-289.

Watterson, G.A. (1974). Models for the logarithmic species abundance distributions. *Theoretical Population Biology*, 6: 217-250.

Williams, C.B. (1944). Some applications of the Logarithmic Series and the Index of diversity to ecological problems. *Journal of Ecology*, 32(1): 1-44.

Williams, C.B. (1947). The Logarithmic Series and its application to biological problems. *Journal of Ecology*, 34(2):253-272.

Williams, C.B. (1952). Sequences of wet and of dry days considered in relation to the Logarithmic Series. *Quaterly journal of the Royal Meteorological Society*, 73(335): 91-96.

Wyshak, G. (1974). Algorithm AS 68: A Program for Estimating the Parameters of the Truncated Negative Binomial. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 23(1): 87-91.

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

☎ (+33) 486 090 631

email: claude.mante@mio.osupytheas.fr