Species abundance patterns in an ecosystem simulation studied through Fisher's logseries

Didier Devaurs [a,*], Robin Gras [b]

[a] Know-Center GmbH and Knowledge Management Institute, Graz University of Technology, Inffeldgasse 21a, 8010 Graz, Austria
[b] School of Computer Science and Biological Department, University of Windsor, 401 Sunset Avenue, Windsor, ON, N9B 3P4, Canada

A B S T R A C T

We have developed an individual-based evolving predator–prey ecosystem simulation that integrates, for the first time, a complex individual behaviour model, an evolutionary mechanism and a speciation process, at an acceptable computational cost. In this article, we analyse the species abundance patterns observed in the communities generated by our simulation, based on Fisher's logseries. We propose a rigorous methodology for testing abundance data against the logseries. We show that our simulation produces coherent results, in terms of relative species abundance, when compared to classical ecological patterns. Some preliminary results are also provided about how our simulation is supporting ecological field results.

Keywords:
Species abundance distribution
Fisher's logseries
Fisher's α
Individual-based model
Fuzzy cognitive map

1. Introduction

Ecological modelling is a still growing field, at the crossroad between theoretical ecology, mathematics and computer science. Ecosystem models aim to characterise the major dynamics of ecosystems, in order to synthesise the understanding of such systems, and to allow predictions of their behaviour. Because natural ecosystems are very complex (in terms of number of species and of ecological interactions) ecosystem models typically simplify the systems they are representing to a limited number of components. This simplification allows for the development of computer-aided ecosystem simulations that are tractable. One of the main interests of such ecosystem simulations is that they offer a global view of the evolution of the system, which is difficult to observe in nature. However, the scope of ecosystem simulations has always been limited by the computational possibilities of their time. Today, it is possible to run simulations that are more complex than what has ever been done before.

The global objective of our work is to develop such a powerful ecosystem simulation, and to try to gain some knowledge about natural ecosystems thanks to it. We have created a generic platform able to simulate complex ecosystems with intelligent agents interacting and evolving in a large and dynamic environment [24]. We have chosen to implement an individual-based model, built upon the predator–prey paradigm. The novelty of our model stems from the fact that each agent behaviour is modeled by a fuzzy cognitive map (FCM), and evolves during the simulation. The FCM of each agent is unique, and is the outcome of the evolution process going on throughout the simulation. To our knowledge, FCMs had never been used before neither in a large scale individual-based model of an ecosystem, nor in an evolutionary context. The notion of species is also implemented, in a way that species emerge from the evolution of agents. To our knowledge, our system has been the first one allowing to model the links between speciation and individual behaviour.

* Corresponding author. Tel.: +43 699 136 156 00.
E-mail addresses: ddevaurs@know-center.at, ddevaurs@student.tugraz.at (D. Devaurs), rgras@uwindsor.ca (R. Gras).

doi:10.1016/j.simpat.2009.09.012

Our simulation produces a large amount of data that can be grouped in three categories: (i) some global statistics, such as the numbers of individuals and of species, or the quantity of resources available; (ii) some information relative to each species, such as its average FCM, or the number of individuals it contains; (iii) some data relative to each agent, such as its level of energy, or its age. Thanks to this data, we plan to investigate macroevolutionary processes, such as the emergence of species, and specific ecological issues, such as species abundance patterns or invasive species. Our preliminary results have shown coherent behaviours of the whole simulation, with the emergence of strong correlation patterns observed also in field ecosystems [24]. This is what we need to analyse even further, in order to show that our simulation can be used to understand natural ecosystems, as well as to make interesting and valid predictions.

In this article, we present a thorough study of the relative species abundance patterns that can be observed in the communities generated by our simulation. Typically, the relative abundance of species is described by the way N individuals are partitioned into S species, i.e. by the distribution of individuals into species. Our main objective is to show that the distributions found in our communities are similar to those that can be observed in field communities. This is very important, in the sense that our simulation is really designed to study speciation issues, and that we first expect from it to be coherent with the already established ecological results, before allowing us to answer new questions.

We have chosen to focus on relative species abundance because this is an issue that ecologists have been studying for several decades [23,40,46,70], and that still remains a central theme of ecology [3,12,73]. We can thus benefit from an extensive expertise, and from numerous well established results. Understanding the species abundance structure of a community is crucial, because it represents a synthesis of the interactions (such as competition, or co-operation) that take place between species [70]. Species abundance patterns are also directly correlated with other characteristics of communities, such as species body size structures, or species range size patterns [23]. Ultimately, through that kind of analyses, one of the main objectives of theoretical ecology is to find means for predicting species diversity, by identifying the mechanisms that regulate it [40].

Among the typical species abundance patterns observed in nature, the most intuitive is the following one: in a sample of a community, a majority of species represented are rare and a few are common. To study that kind of patterns, ecologists have developed mathematical models expressing the species abundance distribution of a sample of individuals. If such a model fits an observed species abundance distribution, it can help to identify the factors that contribute to this distribution, and more generally, to understand the biological processes that regulate species diversity. We will focus here on the model that pioneered the field, and that is still considered as a classical one: Fisher's logseries [21]. The logseries has been widely applied to approximate species abundance distributions [40,70], and one parameter of the model, the so-called Fisher's α, is often used to compare communities in terms of species richness [40,46,51].

The logseries allows for the computation of the whole species abundance distribution of a sample, from the numbers of individuals and species it contains. In order to determine whether the calculated distribution is a good approximation of the one observed in the sample, one has to estimate the goodness-of-fit between both distributions. This was traditionally done by mean of a by-eye analysis, or of a basic χ^2 test. This appears to be not satisfactory enough, because the goodness-of-fit cannot be always estimated in the same way. To solve this problem, we propose several formal definitions for the notion of good fit, each one being associated with a specific situation. From that, we develop a rigorous procedure for evaluating the goodness-of-fit of the logseries. This procedure can be performed in an automatic way, which has enabled us to apply it on many samples taken automatically from our communities.

To give an overview of the results yielded by our study, we can say that: first, the logseries shows a good fit to the distributions observed in relatively small samples, which has been also classically noticed by ecologists in field studies. Second, the logseries provides better results in species-rich than in species-poor communities, both in our simulation and in natural ecosystems. Third, when large samples are taken from our communities, the logseries does not present a good fit, which is also a well established result for field communities. On the other hand, the observed abundance patterns are similar in our communities and in natural ones. Fourth, and more generally, the evolution of the goodness-of-fit of the logseries according to sample size is the same in our communities and in natural communities. We observe a negative correlation whose slope depends on the species richness of the community, which had never been documented before. Finally, the evolution of Fisher's α according to sample size is the same in our communities as in natural ones.

As a by-product, our study has also extended some ecological field results. First, we have validated the only model proposed in the literature for the evolution of Fisher's α according to sample size. Second, we have discovered a dichotomy between species-rich and species-poor communities (with a threshold of 240 species) in terms of the evolution of the goodness-of-fit of the logseries according to sample size.

The rest of the article is organized as follows: Section 2 is dedicated to the presentation of our ecosystem simulation. However, due to the complexity of the whole simulation, we give here only a summary of its principal properties. For further details, the interested reader is referred to [24]. Section 3 consists of an extensive presentation of Fisher's logseries (with its derivation and its ecological interpretations) and of Fisher's α, as well as some general ecological notions (community, species abundance distribution). In Section 4, we propose our formal definitions of good fit, as well as our procedure for evaluating the goodness-of-fit between the logseries and an observed species abundance distribution. In Section 5, we describe all the experimental results we have obtained while examining the species abundance patterns found in the communities of our simulation. Finally, in Section 6, we draw the conclusions of this study, and present some prospects.

2. Our individual-based evolving predator–prey ecosystem simulation using fuzzy cognitive maps as a behaviour model

Our global aim was to develop a generic platform able to simulate complex ecosystems with "intelligent" agents interacting and evolving in a large and dynamic environment [24]. For that, we have chosen to use an individual-based model, and to implement the predator–prey paradigm. In a run of the simulation, time is divided into discrete steps (as in a digital clock) called time steps, composed of a set of common phases (cf. Section 2.2.3 and Table 1). Each agent's behaviour is modeled by a fuzzy cognitive map [36]. This enables an agent to evaluate (with memory) its environment (for example its distance to food or to a possible mate, or its energy level) and its internal state (for example fear, hunger, or curiosity), and to choose among several possible actions (for example escaping, eating, or breeding). Furthermore, agents' behaviours are allowed to evolve (at the phylogenetic scale) over the time steps of the simulation.

2.1. General concepts founding our simulation

2.1.1. Individual-based modelling

In the area of ecosystem simulation, individual-based modelling is a bottom-up approach allowing for the consideration of the traits and behaviour of individual organisms. Instead of modelling an entire ecosystem as a whole, individual-based models aim to "treat individuals as unique and discrete entities" [25]. By modelling organisms with varying characteristics (such as age, mating preferences, role in the ecosystem) the properties of the system that the individuals represent can begin to emerge from their interactions. Another important aspect of individual-based models is their generality. Their advocates are driven by paradigmatic motivations where such models may be used as a way of formulating general theories of ecology, rather than as specific and pragmatic tools [25].

While the use of individual behaviour has been included in many models during recent decades, the individual-based modelling approach is exponentially increasing as the cost to purchase and operate a machine capable of running time-consuming simulations reduces. This has brought interesting benefits, for example, in forest ecology, fish-recruitment modelling and spatial heterogeneity depiction [15]. Yet, few attempts have been made to simulate a complete and complex ecosystem. An example of such a system is the platform Echo [28], which also includes an evolutionary mechanism. However, the organisms in Echo are very simple, and are not provided with any behaviour model.

2.1.2. Predator–prey models

An important property we wanted to integrate into our simulation was the fact that agents would have to develop efficient behaviours to be able to survive in their environment. We have therefore chosen a predator–prey model in which behaviours of prey and predators have to evolve simultaneously to give them abilities to survive. A direct consequence of this choice is that our ecosystem encompasses individuals belonging to two trophic levels. Some other predator–prey models have already been proposed, such as the one in [65] for example. Yet, this particular agent model is dedicated to represent schooling behaviours, and the evolution is an offline mechanism using a genetic algorithm.

2.1.3. Fuzzy cognitive maps

As an agent's behaviour model is crucial to creating complex interacting agents, we have chosen a sophisticated but efficient model called fuzzy cognitive maps (FCMs) [36]. FCMs enable the representation of complex internal concepts such as emotions and desires. They can handle temporal information and fuzzy activation levels for the concepts. Above all, FCMs aim to represent causal relationships between concepts. As such, they allow for the building of agents that are able to perceive, to make a decision, and to act. In our simulation, FCMs are not only the base for describing and computing an agent's behaviour, but also the platform for modelling evolutionary mechanisms and speciation events (cf. Sections 2.2.2 and 2.2.4). For that, we have defined a measure of similarity between FCMs, which gives us a notion of distance between two FCMs.

Formally, a FCM is a graph which contains a set of nodes C, each node C_i being a concept, and a set of edges I, each edge I_{ij} representing the influence of the concept C_i on the concept C_j. A positive weight associated with the edge I_{ij} corresponds to an excitation of the concept C_j from the concept C_i, whereas a negative weight is related to an inhibition (a null value meaning that there is no influence of C_i on C_j). An activation level a_i is associated to each concept. A FCM allows to compute the new activation levels of the concepts of an agent, based on its perception and on the current activation levels of its concepts. This computation is called the dynamic of the map and is a normalized matrix product.

We use a FCM to model an agent's behaviour (structure of the graph) and to compute the next action of the agent (dynamic of the map). In each FCM, we define three kinds of concepts: sensitive, internal and motor. The activation level of a sensitive concept is computed by performing a *fuzzification* of the information the agent perceives in the environment. For an internal concept, the activation level corresponds to the intensity of an internal state of the agent, and affects the dynamic of the map. Note that it enables to distinguish between perception and sensation: the sensation is the real value coming from the environment, and the perception is the sensation modified by the internal states. Activation levels of the motor concepts are used to determine what the next action of the agent will be. The amplitude of the chosen action is then calculated by performing a *defuzzification* of the value of the corresponding motor concept.

2.2. Description of our evolving ecosystem simulation

In our simulation, each agent possesses its own genome. Agents can mate with genetically similar individuals, and produce offspring that inherit a modified combination of the genomes of their parents. We have also implemented a speciation mechanism based on the idea of genotypic pool. Species can emerge from the evolution of individuals, or get extinct if all their members die.

2.2.1. Agents

Each agent possesses its own FCM that models its behaviour. This FCM contains sensitive concepts: *predator_close* (prey only), *predator_far* (prey only), *prey_close* (predator only), *prey_far* (predator only), *food_close*, *food_far*, *mate_close*, *mate_far*, *energy_low*, *energy_high*, *local_food_low*, *local_food_high*, *local_mate_low*, *local_mate_high*; internal concepts: *hunting* (predator only), *fear*, *hunger*, *sexual_need*, *curiosity*, *sedentarity*, *satisfaction*, *annoyance*; and motor concepts: *escape* (prey only), *search_for_prey* (predator only), *search_for_food*, *socialize*, *explore*, *rest*, *eat*, *breed*. It includes also links and weights representing the mutual influences of these concepts.

It is important to notice that the activation level of each motor concept depends on a complex and non-linear combination of excitatory and inhibitory influences from both sensitive and internal concepts. Therefore, this behaviour model enables the representation of very complex phenomena. It is also worth noting that the activation levels of an agent's concepts are never reset during its life. As the computation of the activation level of a concept involves its previous value, the evaluation of the current state of an agent incorporates all its past states. It means that an agent has a memory of its own past that will influence its future states. Eventually, it appears that an agent's behaviour dynamically depends on a complex combination of the information it currently receives from its environment, its current internal state, and all the past states it went through during its life.

The FCM of an agent (or more precisely the set of edges of its FCM, as well as their weights) represents also its genome. It is transmitted to the agent's offspring after combination with the one of the other parent, and after possible mutations. Links between concepts can appear or disappear during this process, so that the structure and complexity of the maps can change during the evolutionary process. Each agent also possesses several physical characteristics which are: its maximum and current ages, its minimum age for mating, its maximum and current speeds, its vision distance, its maximum and current levels of energy, and the amount of energy transmitted to the offspring. Energy is provided to individuals by the resources (grass or meat) they find in their environment. An agent consumes some energy each time it performs an action, proportionally to the complexity (number of edges) of its FCM. If it uses all its energy, an individual dies.

2.2.2. Species

Our simulation implements a speciation mechanism directly related to the genotypic cluster definition proposed in [41]. Our mechanism accounts for the gradualism and the fuzziness of the speciation process. Indeed, speciation is not always a sharp and clear-cut process. Traditionally, a "good" species was defined as a population showing no gene exchange with any other population, so that there was no blurring of the species borders. However, there are numerous examples, such as the one presented in [11], of populations that show a substantial reproductive isolation, but that also exchange genes to some degree with sympatric relatives.

We define a species as a set of individuals, associated with the average genetic characteristics of its members. The average FCM of a species is evaluated on the basis of the FCMs of all the agents belonging to this species. We consider that an individual can be a member of a given species if the distance between its FCM and the average FCM of this species, calculated by mean of a metric we have defined, is below a given speciation threshold. In our current implementation, this threshold is the same for all species; but it will be specific to each species in further versions. To avoid a circular definition between the average FCM of a species and the set of its members, at the beginning of the initial time step all the agents of a given category (prey or predator) have the same FCM and belong to the same species, which has thus a trivial average FCM. Each time step begins with a known set of species associated with their respective average FCMs. At the end of the time step, the species membership of all individuals is re-evaluated (or evaluated, for the newborns). For each agent, the distance between its FCM and each one of these average FCMs is calculated. If the distance to the closest average FCM (i.e. to the most similar species) is lower than the speciation threshold, then this individual belongs to the corresponding species; otherwise it forms a new species. After the new composition of the species is determined, their average FCMs are evaluated.

It is important to note that species are not defined in terms of interbreeding populations (i.e. by applying the traditional biological species concept). Conversely, the question as to whether two agents can breed does not involve their species membership; this allows some possible inter-species gene flow. More precisely, we consider that two individuals can mate if they belong to the same category (prey or predator), and if the distance between their respective FCMs is below a given reproduction threshold. The probability of success of the mating is then defined as a decreasing function of the genetic distance between these agents. Moreover, when an individual will decide to breed, it will consider as possible mates other agents that are similar enough (in terms of distance between FCMs), based on a similitude threshold which is higher than the reproduction threshold. Thus, individuals can try to breed even if the mating will fail.

2.2.3. World

Our simulation takes place in a virtual toric world composed of 1000 cells in both dimensions. Each cell can contain an unlimited number of agents. The initial numbers of prey and predators in the world are parameters of the simulation. Their

Table 1
Successive phases of a time step.

For every prey, perception of the environment
For every prey, computation of all concepts
For every prey, Execution of an action, and update of the energy level
Update of the list of prey
For every predator, perception of the environment
For every predator, computation of all concepts
For every predator, execution of an action, and update of the energy level
Update of the list of predators
Update of the list of prey
Re-evaluation of the prey species
Re-evaluation of the predator species
For every cell {Update of the grass level Update of the meat level}
For every agent, update of the age

original positions are chosen non-uniformly by forming clusters of individuals, in order to model a realistic initial world state. Each cell can also contain resources, providing agents with energy: some grass, eatable only by prey, and some meat (dead prey actually) edible only by predators. In order to create a competition for resources, the amount of resources available in a cell has been limited (by a parameter of the simulation). At the initial time step, there is no meat in the world, and the number of grass units is randomly determined for each cell. We have modeled some growth and diffusion mechanisms, in order to create a realistic evolution of the dispersion and of the amount of grass. Predators have two modes of nutrition: hunting and scavenging. When a predator kills a prey, new meat units are added in the corresponding cell, that can be eaten by the same or other predators.

The different phases of a typical time step of the simulation are summarized in Table 1. It is worth noting that, before an agent executes an action, several dynamics of its FCM are performed. Thus, a time step represents a relatively long period of time, during which an agent performs several small "tasks", while its global behaviour is directed toward the execution of a unique action. Therefore, the action chosen by the agent can be seen as a tendency. This allows us to study ontogenetic and phylogenetic histories of individuals at the same time scale. As a consequence, an agent has a quite short lifespan (in terms of number of time steps), and performs only a few dozens of actions during its life. This enables us to obtain a high level of population renewal, which is an important criterion for studying an evolutionary process.

2.2.4. Evolution

In our simulation, evolution stems from several mechanisms: mating, mutation and speciation. Since species membership of the agents is evaluated at each time step, births and deaths of individuals influence the general species composition. Thus, a species can emerge or disappear at any time step. This enables us to model the evolution of populations of individuals sharing important genetic properties.

Due to our species model, species evolution is derived directly from individual evolution. This one can occur when a breeding event happens. If the mating is successful, the two parents give birth to a unique offspring. This one inherits a combination of the genomic information of its parents, with possible mutations. The genome of an agent is defined as the set of edges, associated with their weights, of its FCM. To model the crossover mechanism, edges are transmitted by block from one parent to its offspring. More precisely, for each concept, the child inherits all the incident edges of this concept from one parent. During that process, the weights of the edges can be modified, based on a probability of mutation (which is a parameter of the simulation). Moreover, some new edges can be created (based on a probability of apparition which is a parameter of the simulation); and some old edges can be removed (if their weight becomes smaller than a given threshold).

The apparition of new edges is a very important mechanism, in the sense that new influences between concepts can emerge during the evolutionary process. This allows the apparition of more complex and potentially more adaptive behaviours. If they show a selective advantage, such behaviours will be preserved (and thus transmitted through generations) by the process of natural selection, inherent to the interaction of the individuals with their environments. As a counterpart, the possibility for edges to disappear is also fundamental. When the complexity (i.e. the number of edges) of the FCM grows, this increases the energy needs of the agent which then needs a more efficient behavioural model for being able to obtain this energy. Thus, the influence links between concepts are somehow "tested" by the evolutionary process, and removed if they appear to be not beneficial enough. This allows agents (at the phylogenetic scale) to react to changes in the environment, and to balance the interest of a complex behavioural model with its energy cost.

During the mutation process, the majority of the modifications that take place consists in small changes in the weights of a few edges. By this mechanism, the concept of neutral theory of evolution [33] is integrated in our evolutionary process. In general, a single mutation is neutral (i.e. has no effect) with respect to the behavioural model. Thus, a unique breeding event generating a mutated offspring cannot produce a very different behaviour model. It is the accumulation of neutral mutations during several generations that allows the apparition of new individual behaviours, and then of new species.

2.3. Overall analysis of our simulation

2.3.1. Correlation patterns

The first important thing to note is that, even if this simulation is a very complex and large adaptive system, the whole behaviour of the ecosystem is relatively stable (across different runs and different sets of initial parameters) and present interesting correlation patterns. We have noticed that several variables of the simulation follow oscillatory patterns, and that several of these patterns show some correlation [24]. By computing the cross correlations between the variables of the simulation, we have observed several interesting behaviours, at the population level.

First, our data show a strong correlation between the number of individuals and the number of species, for each category (prey or predator), which is coherent with already reported ecological results [46]. This relation is stronger for predators than for prey, but for both we observe that an increase in the number of individuals is followed by an increase in the number of species few time steps later.

Second, we have observed a correlated evolution of the numbers of prey, of predators and of grass units. Having implemented a predator–prey system, the correlation between the numbers of prey and predators was of course expected. When the number of prey increases, the number of predators also raises few time steps later. On the other hand, when the number of predators grows too much, the number of prey decreases few time steps later, leading to a later decrease in the number of predators. A similar resource-based relation is observed between the number of prey and the quantity of grass.

Third, there are strong correlations (i) between the number of prey and the number of *eat* actions in the whole population of prey, (ii) between the number of prey and the number of *breed* actions in the population of prey, (iii) between the number of predators and the number of *hunt* actions in the population of predators, etc. More generally, it seems that a direct correlation exists between the total number of individuals and the number of individuals choosing a specific action, which means that a quite constant proportion of individuals choose this action during the whole run of the simulation.

Finally, we have also analysed the evolution over time of the average activation levels of the different concepts of the FCM, within the whole populations of prey and predators. For the prey population, for example, we have observed correlations (i) between the average activation level of the *predator_close* sensitive concept and the numbers of prey and predators, (ii) between the number of prey and the average activation level of the motor concepts *explore* and *rest*, (iii) between the average activation level of the *satisfaction* internal concept and the numbers of prey and predators.

2.3.2. Evolution of the behavioural model

We have observed interesting behaviours also at the individual level. Both preys and predators have seen over time an increase in the complexity of their behavioural model (i.e. the number of edges of their FCM), with a slightly higher slope for predators. Even though a higher number of edges in its FCM causes an increase in the energy used by an agent at each time step, it appears that there is enough benefit from adding new influence edges between concepts and having a more complex behavioural model to compensate for the increase in energy need. This shows that an FCM is a very powerful behavioural model: it is sophisticated enough for its increase in complexity to provide a selective advantage over a reduction in energy consumption.

The appearance of a new edge in the FCM (i.e. of a new influence link between two concepts) during a run of the simulation corresponds to the appearance of a new behaviour that was not initially "programmed". This new behaviour can then be "tested" by the process of natural selection and maintained if it shows some selective advantage. For example, we have observed the appearance of a positive influence of the *predator_far* sensitive concept on the *eat* motor concept of the prey. This means that, when there is no nearby predator, a prey will rather not move and eat. Another example is the appearance in the prey FCM of a negative influence of the *local_food_high* sensitive concept on the *fear* internal concept. This means that, when the level of food is high locally, the level of fear of a prey will have tendency to decrease. Other examples of new behaviours observed among preys are: (i) when no possible mate is close, the level of fear increases; (ii) when the level of sexual need increases, the level of sedentarity decreases; (iii) the action of breeding decreases the level of annoyance...

2.3.3. Studying the speciation process

It is interesting to note that, among the new behaviours that emerge during a run of the simulation, such as those presented in Section 2.3.2, some are constantly maintained by the evolutionary process and spread among a large number of species. More generally, this illustrates that our simulated evolutionary ecosystem allows the discovery, the conservation and the propagation of new important behaviours, and that the species is an efficient vector of this propagation. It would be particularly interesting to study this phenomenon with respect to the underlying phylogeny deriving from the successive speciation events, in order to analyse how important evolutionary "discoveries" influence the speciation process.

Before making any prediction from the data generated by our simulation, we would like to validate it in a more rigorous way than by only analysing intuitive ecological patterns, such as those presented in Sections 2.3.1 and 2.3.2. As our work will be specifically centred on studying the speciation process, we have chosen to focus this validation on the appearance of new species and on the distribution of individuals into different species. Among the ecological patterns observed in nature and mentioned in the literature, the species abundance distribution patterns are undoubtedly among the most studied and well-document. They have been studied by ecologists in several ways, but a classical one is to use Fisher's logseries. This will provide some guidelines for our study.

3. Fisher's logseries

3.1. Species abundance distribution in a sample of a community

First we need to precise what is meant by "community". In nature, it is obviously not possible to examine all the individuals of all the species found over a large area. While studying for example the abundance of several species, it is even meaningless to compare species from different groups, such as insects and birds for example, because of the huge difference in orders of magnitude. Classically, ecological studies are confined to a limited area and to a certain definite group within a trophic level, which constitutes the examined community; it can be for example, the birds of a country, the insects of a small region, or the trees of a forest [70].

Even with this restricted definition, a community will generally not be studied as a whole, but through randomly taken samples (as opposed to biased samples, such as museum collections). Some other problems then arise, such as human sampling errors, or the question as to which properties of the sample are also properties of the community and which are not; an ecological pattern found in a sample is not necessarily a miniature reproduction of a pattern of the community [70].

Among the various ecological patterns that can be observed in a sample, we focus here on the relative abundance of the species represented in the sample. Consider a sample of individuals taken from a mixed community containing several species. The species abundance distribution graph of this sample is a chart where the number of species is plotted against the number of individuals per species (cf. Fig. 1). In other words, the species abundance distribution of this sample is the series giving the number of species that contain one individual, the number of species that contain two individuals, three individuals, and so on.

It has been widely observed by ecologists that in a sample of a community, species are far from being equally abundant, even under uniform conditions. Actually, a majority of species represented in the sample are rare while a few are common [21,23,40]. More precisely, in many cases, more species are represented by one individual than by two, more by two than by three, and so on [68,70]. After plotting the species abundance distribution graph of the sample, what is observed is a "hollow curve" [68,70], also described as a J-shaped curve [30] (cf. Fig. 1). In order to further the analysis of this pattern, ecologists have tried to put mathematical relations behind this apparent order in the relative abundance of the different species represented in a sample. The first mathematical model of species abundance distribution was proposed in 1943 by Fisher [21].

3.2. Fisher's logseries as a species abundance distribution model

Given a sample of a community, Fisher has proposed a series expressing the species abundance distribution of this sample [21]. Let N and S be, respectively, the numbers of individuals and of species in the sample. If n_i is the number of species that are represented by i individuals ($i \in \mathbb{N}^*$) in the sample, then $\forall i \in \mathbb{N}^*, n_i = \alpha x^i/i$. The series is thus represented by

$$\alpha x,\ \alpha\frac{x^2}{2},\ \alpha\frac{x^3}{3},\ldots \quad (\text{or sometimes, and equivalently, by } n_1,\ n_1\ \frac{x}{2}\ , n_1\ \frac{x^2}{3},\ldots, \quad \text{with } n_1 = \alpha x)$$

where α and x, the two parameters of the model, satisfy the equations

$$S = \alpha ln\left(1+\frac{N}{\alpha}\right) \quad \text{and} \quad x = \frac{N}{N+\alpha}.$$

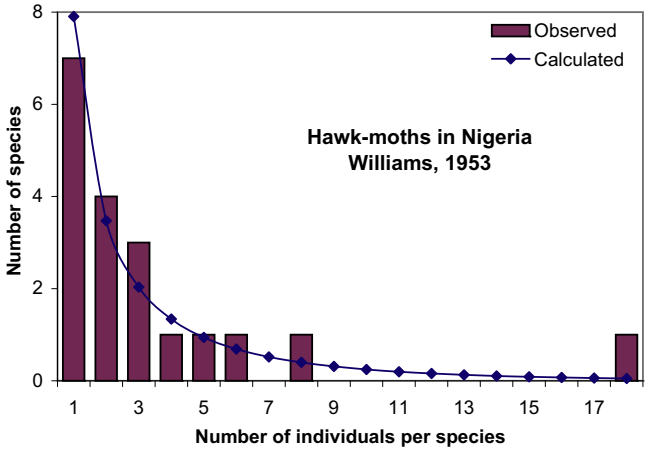


Fig. 1. Species abundance distribution of a sample of an hawk-moths community, as observed by Williams, and as approximated by Fisher's logseries; $N = 65$ (number of individuals) and $S = 19$ (number of species) [69,70].

Thus, it is sufficient to know N and S, for α and x to be calculated, and for the whole species abundance distribution of the sample to be known. The first parameter, α, is constant for all samples from a given community (it is a characteristic of the community and not of the sample). As it is also positively correlated with the total number of species in the considered community, it has been named the "index of diversity" of the community. The second parameter, x, is such that $0 < x < 1$, and varies from one sample to another. In fact as the size of the sample increases, x approaches 1 (and thus α is the upper limit of the number of singleton species n_1). A key property of the logseries distribution is that it is its own sampling distribution. In other words, if a community's species abundances are arranged according to a logseries, then any sample from it also shows a logseries distribution of species abundance (with the same value of α but a lower value of x).

One of the direct consequences of Fisher's model is known as "William's law". In [21] Williams notices that doubling the number of insects captured (in that case, by increasing the time of trapping), except for very small samples, always add about 30 species. Indeed, in the relation $S_N = \alpha \ln(1 + N/\alpha)$, if N is large compared with α, 1 can be neglected in comparison with N/α. Thus, for large samples, $S_N \approx \alpha \ln(N/\alpha)$, and $S_{2N} - S_N \approx \alpha \ln(2)$. So the number of species added to a large sample by doubling it is constant. This results graphically in a straight-line relation between the number of species and the logarithm of the number of individuals.

Fisher's logseries has been widely used to approximate the species abundance distribution of a sample of a community [8,16,17,21,31,32,40,56,59,68,70]. For example in [69,70] Williams apply it to a sample of hawk-moths captured with a light-trap in Nigeria (cf. Fig. 1). Here, $N = 65$ and $S = 19$, which gives as approximations of α and x, 9 and 0.88, respectively. From that, the whole series can be calculated: 7.9, 3.5, 2, 1.3, 0.9, 0.7, 0.5, 0.4, 0.3, 0.25, 0.2, 0.16, 0.13, 0.1, 0.09, 0.07, 0.06, 0.05. Note that it is a discrete series, although it has been traditionally plotted as a continuous curve, in order to better visualize its shape. As a compromise, we draw a continuous line on which the discrete points of the series are highlighted (cf. Fig. 1). The question then arises as to how to ascertain whether the calculated series is a good approximation of the observed distribution or not. In the literature, observations have often been considered to be fitted or not by the logseries simply on the basis of a by-eye fit [8,21,68,70]. Statistically, the goodness-of-fit has been also widely evaluated using a basic χ^2 test [16,17,21,31,32,40,56]. In the sequel, we will propose more precise definitions for the evaluation of the goodness-of-fit (cf. Section 4).

3.3. Derivation of the logseries

Fisher's logseries is based on the postulate that the relative abundance of species in a sample can be described by compounding a Gamma distribution of the actual species abundances in the community with a Poisson sampling distribution of individuals from each species (assuming that individuals are randomly sampled from the community). This produces an expected distribution of observed species abundances which is a negative binomial distribution. The logseries is then the limiting form of the negative binomial distribution derived for communities containing a large number of species.

More precisely, Fisher evolved the logseries in the following way [21]. The number of species represented by i individuals in the sample, as given by the negative binomial distribution is

$$n_i = \frac{(k+i-1)!}{(k-1)!\ i!} \times \frac{p^i}{(1+p)^{k+i}}, \quad i \geqslant 1. \tag{1}$$

The parameter p is proportional to the size of the sample. The parameter k (which is an intrinsic property of the community) measures inversely the variability of the species densities. When heterogeneity increases, k becomes small and approaches its limiting value, zero. Thus, in highly diverse communities, k is often statistically indistinguishable from zero. Actually, it is not zero, since the total number of species is finite. If, however, we put $k = 0$ in the previous expression, write x for $p/(1 + p)$, so that x stands for a positive number less than unity, varying with the size of the sample, and replace the constant factor $(k - 1)!$ in the denominator, by a new constant factor, α, in the numerator, we have

$$n_i = \alpha \frac{x^i}{i}, \quad i \geqslant 1. \tag{2}$$

The total numbers of species and of individuals expected are then

$$S = \sum_{i \geqslant 1} \alpha \frac{x^i}{i} = -\alpha \ln(1-x) \quad \text{and} \quad N = \sum_{i \geqslant 1} \alpha x^i = \frac{\alpha x}{1-x}. \tag{3}$$

This reasoning led to the frequent criticism from biologists that fitting the logseries presumes an infinite pool of species available for capture. Kempton and Taylor [32] have solved this ambiguity by proposing a more formal derivation, making it clear that Eq. (2) is simply the limit of Eq. (1), when $k \to 0$.

3.4. Ecological interpretations of Fisher's logseries

The logseries is a pure statistical model for species abundance distributions; it is expressed entirely in the language of mathematical statistics. As such it originally says nothing about the ecology of the species of the community, and what limit their abundances. However, some biologists have tried to relate this model to ecological properties of the studied

communities. For example, Kempton notices that species abundance distributions of samples from stable communities are generally fitted by the logseries, contrary to those from unstable communities [31,32]. Williams proposes to express biologically the logseries as "nothing succeeds like success" [70]; in other words, the higher level of abundance a species has reached, the easier it will be for it to take one step higher. Both interpretations are not supported enough to be given more consideration.

A more interesting idea has been proposed by May [43], and promoted by other ecologists [40,51]. According to May [43], the logseries is a distribution characteristic of a relatively simple community whose dynamics is dominated by some single factor, and where division of the available niche volume proceeds in a strongly hierarchical fashion, known as the niche-preemption process: the most successful species preempts a fraction f of the niche, the next one a fraction f of the remainder, and so on. The statistical expression of this model can be derived, for example, by assuming that the niche-preemption mechanism stems from the fact that species arrive at successive random time intervals and proceed to preempt a fraction f of the remaining niche before the arrival of the next one [7].

Despite being seducing and coherent with the logseries distribution, such an ecological interpretation cannot be really validated. A serious obstacle is that numerous other stochastic mechanisms that might account for a sample's species abundance distribution, besides the one described above, lead to the logseries distribution [43]. These models, which have no apparent relationship, are reviewed in [7,66]. Unfortunately, an empirical species abundance distribution cannot by itself give evidence on how to choose among them. In the light of the foregoing it seems more reasonable not to draw any ecological conclusion from the fact that the logseries fits or not some ecological data, and to restrict the use of this model to purposes of diversity comparison, empirical prediction, and the like.

3.5. Fisher's α as an index of diversity

A very common question in ecology is how to compare communities in terms of species richness, i.e. of diversity. The simplest way would be to use the number of species as an index of diversity: if the total number of species of two communities was known, it would be easy to recognize the richest one. However, this is rarely true in practice, and this question is generally answered through the comparison of two samples containing only a few species from their respective community pools. In that case, it is not relevant to measure the diversity of a community by evaluating the number of species represented in the available sample, since this number is strongly related to the size of the sample and to the proportion of the community this one represents. It seems obvious that a good index of diversity should be independent of sample size. Another important feature of a good diversity statistic should be its ability to discriminate between communities that are not very different [31,32,40,56]; in other words, a good index of diversity should behave consistently across samples of a given community, and respond to differences between communities.

Since, according to Fisher's theoretical model, it is independent of sample size and positively correlated with the total number of species in the community, α should be a good candidate for measuring diversity. However, despite some studies showing a constant value of α for a few samples of different sizes taken from a given community [32,46,70], this constancy of α seems to be only theoretical. When a more thorough study of its behaviour is performed, it is observed that α is correlated with sample size, whether the analysis is theoretical [46,56] or carried out on natural communities [13]. On the other hand, several studies have shown that α is less sensitive to changes in sample size than any other index of diversity proposed in the literature [13,40,45,51,56]. Moreover, the dependence of α on sample size is similar across communities: the between-communities ratio of α is less influenced by changes in sample size than α itself [13]. Finally, when α varies, it does so in a consistent way, so that a correction factor can be estimated by modelling the relationship between α and N (the size of the sample) with a function of the form: $\alpha = pN^q$ [13].

What makes α a good index of diversity is also its good discriminant ability: α is the best community discriminant among all diversity statistics proposed in the literature [32,40,56] (the discriminatory ability of each index being measured by calculating its between- to within-communities mean square ratio [32,56]). The robustness of α is explained by the fact that it is less sensitive than other indices to changes in the abundance of very common or very rare species; α is influenced mainly by the frequencies of the species of medium abundance, less subject to fluctuations than the frequencies of the species of low or high abundance [32,56]. α can thus be interpreted as the slope of the cumulative species abundance curve at the mid-range of abundance, where the moderately common species reveal the genuine richness of the community [32], for example by not considering vagrant species but only resident ones [40].

Another interesting property of α is that, if the logseries distribution provides a good fit to all samples of the considered communities, α gives a complete ordering on intrinsic diversity among communities [31]. It is important to note that, even when the logseries does not fit the observations, α remains a robust and consistent index of diversity, for ordering the communities [56]. The only disadvantage of α is that it is based purely on S and N. Thus it cannot discriminate between samples where S and N remain constant but where there is a change in the species abundance distribution. But this is mainly an academic issue, as it is very unlikely that any real data will behave in this way [40]. As a precaution, the value of α should not be considered reliable for too small samples (i.e. $N < 100$) because of its very high variation in such samples [13]. Once α has been calculated, its confidence interval can be determined, by evaluating its variance [2]:

$$Var(\alpha) \approx \frac{\alpha}{-\ln(1-x) - x}. \tag{4}$$

α has been widely recommended as a universal index of diversity [13,32,40,46,51,56]. Originally it was not much employed, perhaps because its calculation was quite tedious without a computer [46] (even with the help of Williams's nomograms [21,70]). Nowadays, α is often used to compare communities in terms of species richness, by eliminating the effect of sample size [8,12,13,45,46,50]. In that context, it has shown a strong positive correlation with habitat heterogeneity (expressed in terms of its effect on the fauna) [8,45]. α is also used for extrapolation, i.e. to predict the number of species in samples larger than a given available sample, as long as this does not go beyond regions of uniform environment [13].

3.6. Discussion of Fisher's model

A limit of Fisher's model is that it might happen that compounding several logseries distributions does not produce a logseries distribution [2]. Consider a given community as an association whose components can be seen as sub-communities. Assume that the species abundance distribution of each sub-community follows a logseries, with its own parameters. The species abundance distribution of the global community consists of a mixture of several logseries, but cannot be fitted by a single logseries, for too many very rare and very common species, and too less species of medium abundance are found.

Recall that the logseries is a limiting case of the negative binomial distribution, when $k \to 0$. However consistent with the observations this approximated model might be, Fisher himself admitted that there was no theoretical basis for supposing that k was actually zero [21]. The value of k should rather be considered finite, though perhaps very small, even up to the point of being negligible; but that k can be ignored should be derived from the data themselves [21]. Thus, one could think that the negative binomial distribution should be a good generalized form of the logseries (with three parameters instead of two), useful especially when the latter does not fit the observations. However, the negative binomial distribution has been rarely successfully fitted to species abundance data.

Another generalized form of Fisher's logseries has been proposed in [30]. It is based on the Beta distribution of the second kind, instead of the Gamma distribution, and it is defined by three parameters (instead of two for the logseries). Like the logseries' one, its graphical representation is always J-shaped. When the logseries provides a good fit to the data, this generalized model shows little improvement; but when the observed distribution is very skewed, it gives a better fit [30]. Despite its benefits, the generalized logseries has never been used by ecologists. Thus we will not give it anymore consideration.

4. Fitting species abundance distributions with the logseries

As already illustrated by Fig. 1, once a sample of a community is available, we can determine its species abundance distribution, and compare it with the logseries distribution. As mentioned in Section 3.2, the first methods proposed for estimating the goodness-of-fit between an observed abundance distribution and the calculated logseries were a simple by-eye test and a χ^2 test. Let us compare the examples presented in Figs. 1 and 2a, with respect to these techniques. Visually, the fit provided by the logseries is good for both samples. For the sample of moths (cf. Fig. 1) $\chi^2 = 22.35$, which gives $P = 0.17$ (since the number of degrees of freedom, d, is equal to 17); thus the χ^2 is not significant (i.e. $P \geqslant 0.05$) and the fit is good. For the sample of ants (cf. Fig. 2a) $\chi^2 = 100.35$, which gives $P = 2 \times 10^{-11}$ (since $d = 24$); thus the χ^2 is significant (i.e. $P < 0.05$) and the fit is not good. This result is quite counter-intuitive if it follows the by-eye analysis: the difference of the fit between both examples is not graphically manifest. Intuitively, the fit seems slightly worse for the sample of ants, but not up to disqualify it from providing a good fit.

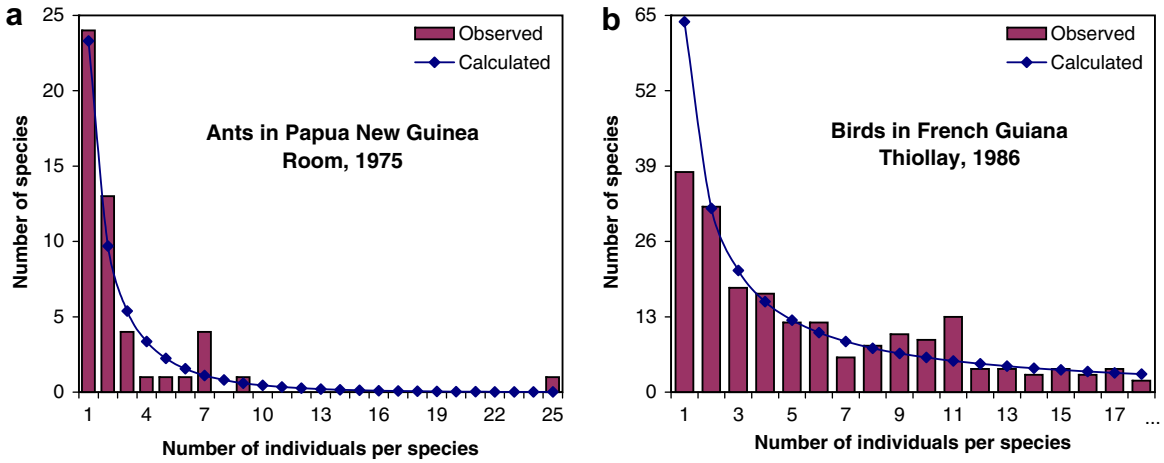


Fig. 2. (a) Species abundance distribution of a sample of ants in Papua New Guinea ($N = 139$ and $S = 50$) [45]. (b) First part of the species abundance distribution of a sample from a bird community in a rain forest area of southern French Guiana ($N = 8507$ and $S = 315$) [58].

Table 2
Characteristics of the field samples we have studied (N and S are the numbers of individuals and of species; when an example encompasses several samples, we give the approximate mean values of N and S).

Number	Description	N	S	Reference(s)
1	Sample of butterflies, Malaya	3306	501	[21]
2	Samples of soil Collembola, Alps	≈280	≈16	[1,70]
3	Samples of breeding birds, Białowieża National Park, Poland	≈100	≈20	[61]
4	Samples of hawk-moths, Ibadan, Nigeria	≈80	≈20	[69,70]
5	Sample of spiders, Tetney, England	43	20	[22,70]
6	Sample of breeding birds, West Virginia, USA	99	24	[52]
7	Sample of moths, Egham, Surrey, England	840	144	[14]
8	Sample of Lepidoptera, Hockeridge Wood, England	1486	112	[27]
9	Samples of ants, Popondetta, Papua New Guinea	≈130	≈40	[45]
10	Sample of Tipulidae, Haileybury College, England	192	28	[70]
11	Sample of birds, Monk's House observatory, England	23,938	140	[20,70]
12	Trees of the Cocoli 4 ha forest plot, Panama	8179	169	[12]
13	Sample of filamentous fungi, Beaumaris, UK	7861	33	[59]
14	Sample of butterflies, La Selva Lodge, Ecuador	883	91	[17]
15	Sample of Lepidoptera, Rothamsted, England	6815	197	[68,70]
16	Sample of butterflies, Jatun Sacha, Ecuador	6690	130	[16]
17	Trees of the Luquillo 16 ha forest plot, Puerto Rico	85,368	145	[60]
18	Trees of the Bukit Timah 4 ha forest plot, Singapore	13,360	321	[39]
19	Sample of Lepidoptera, Maine, USA	55,539	349	[19,70]
20	Sample of birds, French Guiana	8507	315	[58]
21	Breeding and wintering birds of the Czech Republic	≈40,000,000	≈180	[29]
22	Trees of the Sherman 6 ha forest plot, Panama	≈22,000	≈230	[12]
23	Trees of the Yasuní 50 ha forest plot, Ecuador	≈143,000	≈1120	[62]
24	Trees of the Fushan 25 ha forest plot, Taiwan	114,511	110	[55]
25	Trees of the Huai Kha Khaeng 50 ha forest plot, Thailand	78,444	287	[9]
26	Sample from a benthic community, Hong-Kong	10,142	139	[49]

The problem comes from the well-known property of the χ^2 test of not being statistically valid when applied on too small values. An observed abundance of 1 (as n_{25}^{obs} in Fig. 2a), when compared for example with a calculated value of 0.011 (as n_{25}^{calc} in Fig. 2a) produces a large term in the χ^2 and makes it significant. However, as we want to evaluate the fit between a distribution having integer values and another having real values, 0.011 could be considered as a good approximation of 1 in terms of species abundance (expressed only by integer values in nature). An observed distribution should not be considered not fitted by the logseries based only on that kind of differences.

It is clear that a simple χ^2 test is not robust enough to distinguish between fitted and not fitted samples. To cope with this issue, we propose several formal definitions associated with variants of the concept of *good fit*, each one enabling us to take into account a particular case, such as the one illustrated by Fig. 2a. Based on these specific definitions we define a more global notion of good fit, and we develop a procedure for evaluating the goodness-of-fit of the logseries. It is important to note that the first definition (cf. Section 4.1) is a prerequisite (i.e. a necessary condition) that will be used in all the following ones (cf. Sections 4.2–4.5). We have developed these definitions by studying a wide range of field examples found in the literature (and listed in Table 2). For sake of simplicity, in the sequel we will refer to these samples by giving their numbers only. Our objective is to develop a goodness-of-fit test that would be positive for Samples 4–18 (because they are classically considered as being *fitted* by the logseries), and negative for the other samples (since they are considered as being *not fitted* by the logseries).[1]

4.1. Acceptable fit

In order to provide a sound test for the goodness-of-fit, the first step is to take into account the importance of n_1, the number of singleton species, in a species abundance distribution. First, n_1 is historically meaningful: it has always been an important statistic of a sample, and the logseries had to give good approximations of n_1 to be validated [21]. Second, n_1 has a particular status among n_i's in Fisher's model: it can be considered as a measure of diversity [70], since it approaches α in large samples. Finally, from a practical standpoint, ignoring the role of n_1 can lead to unacceptable results of a goodness-of-fit test. This is illustrated by the sample presented in Fig. 2b. If we consider a basic χ^2 test, we find that the fit is good, because the χ^2 is not significant. However, it is graphically evident that this sample does not present a good fit to the logseries, since the estimation of n_1 is not satisfactory. The problem comes from the fact that the χ^2 is not significant even if a few values of the abundance distribution are not well approximated. That is not an issue as long as n_1 is not concerned, but this has serious consequences otherwise.

Based on these considerations, before defining what is a good fit, we define what we call an "acceptable fit". We say that the logseries distribution provides an *acceptable fit* to an observed species abundance distribution if the absolute difference

[1] In other words, our test should show a good fit for Samples 4–18 and a bad fit for the other samples.

between the observed and the calculated values of n_1 is less than 15% of the observed value, i.e. if $|n_1^{obs} - n_1^{calc}| \leqslant 0.15\, n_1^{obs}$. This value of 15% stems from the examples found in the literature, and allows to discriminate between samples that are classically considered as being fitted or not by the logseries. For example, the distributions depicted in Fig. 2b do not present an acceptable fit, since the error on n_1 is 68%. On the other hand, Fig. 2a shows an acceptable fit, with an error on n_1 of 2.9%. More generally, Samples 1–18 (cf. Table 2) show an acceptable fit to the logseries, and Samples 19–26 do not (which means that they cannot be fitted by the logseries). We now need to provide definitions more specific than the one of the acceptable fit, in order to separate Samples 1–3 from Samples 4 to 18.

4.2. Basic good fit

Based on the notion of acceptable fit, we can define what we call a "basic good fit". We say that the logseries distribution provides a *basic good fit* to an observed species abundance distribution, if it presents an acceptable fit, and if a basic χ^2 test applied on both distributions gives a χ^2 that is not significant (i.e. $P \geqslant 0.05$). For instance, the distributions depicted in Fig. 1 present a basic good fit, since the error on n_1 is 12.9% and $P = 0.17$ (for $\chi^2 = 22.35$ and $d = 17$). Actually, only Samples 4–6 (cf. Table 2) show a basic good fit to the logseries. This is due to the already mentioned property of the χ^2 test of not being statistically valid when applied on too small values. We propose several ways to cope with this issue.

4.3. Pseudo basic good fit

When examining the samples that do not provide a basic good fit to the logseries, it is very often noticed that the observed species abundance distribution is long-tailed, and that its last term is preceded by a series of null values (cf. Fig. 2a). This observation has given us the idea to simply reduce the abundance distributions, by ignoring their last parts (as $(n_i)_{10 \leqslant i \leqslant 25}$ in Fig. 2a) which visually already present a good fit, and to perform the χ^2 test on the first terms of the series.

This technique of series reduction has led us to define the notion of "pseudo basic good fit". If we consider an observed and a calculated species abundance distributions, we define their associated reduced distributions as the ones constructed by removing their respective last parts, defined according to the observed distribution, as the sub-series containing the last term and the preceding null values. We say that the logseries distribution provides a *pseudo basic good fit* to an observed species abundance distribution, if it presents an acceptable fit, and if a basic χ^2 test applied on their associated reduced distributions gives a χ^2 that is not significant (i.e. $P^{red} \geqslant 0.05$). For example, the distributions depicted in Fig. 2a present a pseudo basic good fit, because the error on n_1 is 2.9% and $P^{red} = 0.12$ (since, after reduction, $\chi^2 = 12.68$ and $d = 8$). More generally, among the samples presenting an acceptable fit that did not show a basic good fit (i.e. Samples 1–3 and 7–18), Samples 7–10 provide a pseudo basic good fit to the logseries. Even though this technique allows to fit some samples that had been rejected previously, it is not general enough to deal with all particular cases.

4.4. By-class good fit

Regarding the problem of statistical invalidity of the χ^2 test when applied on too small values, another solution consists in grouping into classes the terms of a usual species abundance distribution, in order to produce a grouped species abundance distribution. The use of this technique comes from the observation that, in ecological frequency distributions, the variation of frequencies is geometric and not arithmetic (which can be inferred from a long-tailed distribution on an arithmetic scale) [70]. Analyzing some geometrically varying data is more convenient and meaningful if they are transferred on a logarithmic scale [32,56]. The naive approach would be to use the base 2 for the logarithm, but this presents the disadvantage to violate the independence of data points [70]. Traditionally, a base 3 logarithm is used to transform the abundance data. This is done by grouping data into "$\times 3$ classes" $(C_k)_{k \geqslant 0}$: class C_k has its center at 3^k, and its edges at $3^k/2$ and $3^{k+1}/2$ [70]. When used with integer values, it gives the following classes: class C_0 contains only 1; class C_1 contains 2, 3 and 4; class C_2 contains integers from 5 to 13; class C_3 contains integers from 14 to 40; class C_4 contains integers from 41 to 121; etc.

We have associated with this grouping process the concept of "by-class good fit". We say that the logseries distribution provides a *by-class good fit* to an observed species abundance distribution, if it presents an acceptable fit, and if a basic χ^2 test applied on both grouped distributions gives a probability P greater than 0.25. This value of 0.25 derives from the field examples found in the literature, and allows to distinguish between samples that are classically considered as being fitted or not by the logseries. For example, the species abundance distribution of Sample 14 (cf. Fig. 3a) does not provide a basic good fit to the logseries, since $P = 6 \times 10^{-6}$ (for $\chi^2 = 178.1$ and $d = 103$). But its grouped species abundance distribution (cf. Fig. 3b) presents a by-class good fit to the logseries, because the error on n_1 is 12.7% and $P = 0.86$ (since $\chi^2 = 1.29$ and $d = 4$). More generally, among the samples presenting an acceptable fit that did not show a basic good fit, Samples 8–14 provide a by-class good fit to the logseries.

4.5. Pseudo by-class good fit

Sometimes, even the use of this grouping technique is not sufficient to avoid the presence of very small abundance values and the following problem of applying a χ^2 test on them. This can be circumvented by defining larger classes, such as $\times 5$ or $\times 7$ classes [70]. However, this is not recommended, since too much information might be lost in the logarithmic compres-

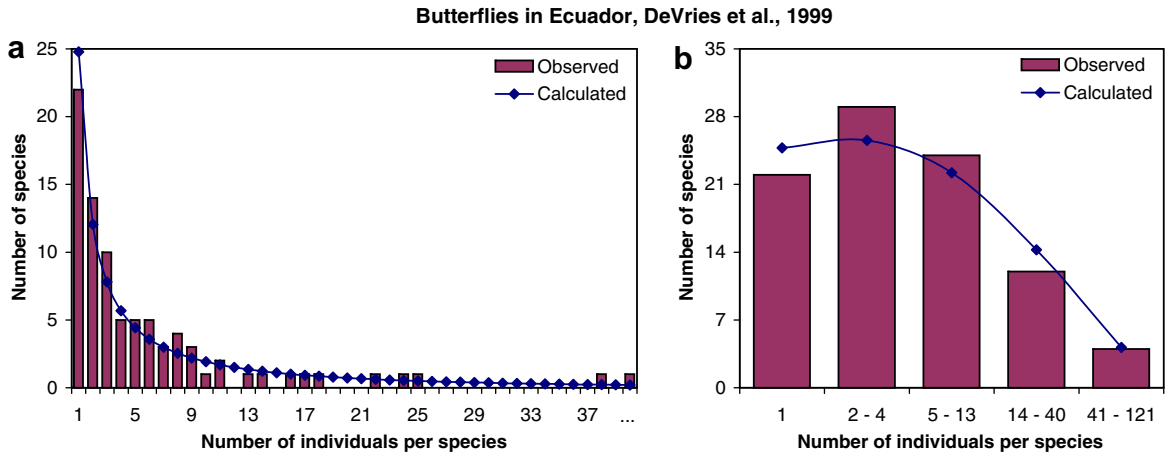


Fig. 3. Species abundance of Sample 14. (a) First part of the distribution ($(n_i)_{1\leqslant i\leqslant 40}$) on the arithmetic scale; the following terms different from zero are $n_{67} = n_{71} = n_{81} = n_{104} = 1$. (b) Distribution on the logarithmic scale.

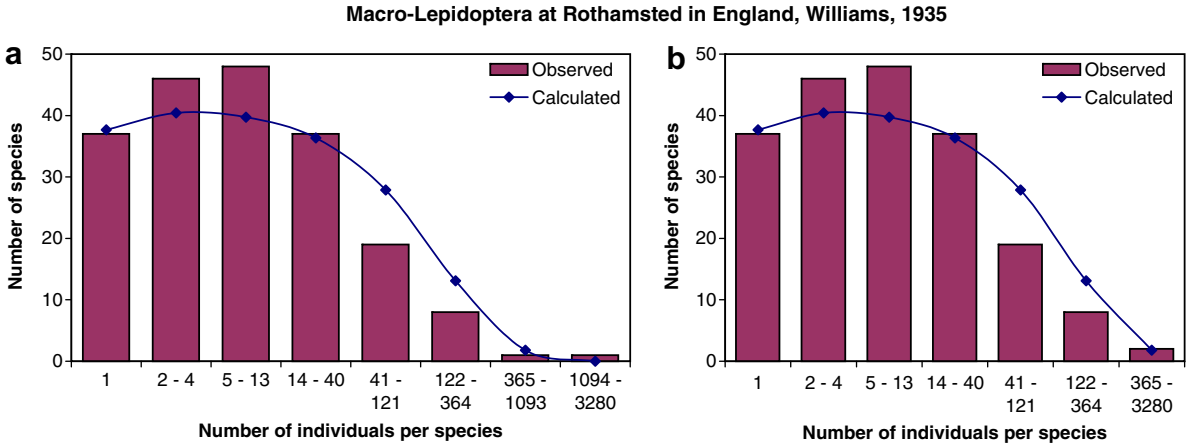


Fig. 4. (a) Grouped species abundance distribution of Sample 15. (b) The same distribution after having combined the last two classes.

sion process, and too few abundance classes might remain to ensure the validity of any statistical test. Knowing that this issue concerns in general only the ending of the grouped species abundance distributions (cf. Fig. 4a), a better solution is to combine the values of the last two classes, in order to form a wider class with a larger value [40].

This idea has given birth to the notion of "pseudo by-class good fit". We say that the logseries distribution provides a *pseudo by-class good fit* to an observed species abundance distribution, if it presents an acceptable fit, and if a basic χ^2 test applied on both grouped distributions, where the last two classes have been combined, gives a probability P^{comb} greater than 0.25. For instance, Fig. 4a illustrates the grouped species abundance distribution of Sample 15, which does not provide a by-class good fit to the logseries, since $P = 10^{-15}$ (for $\chi^2 = 85.6$ and $d = 7$). Once the last two classes are combined (cf. Fig. 4b) the same distribution shows a pseudo by-class good fit, since $P^{comb} = 0.29$ (for after combination $\chi^2 = 7.4$ and $d = 6$). More generally, among the samples presenting an acceptable fit but showing neither a basic good fit nor a by-class good fit, Sample 7 as well as Samples 15–18 provide a pseudo by-class good fit to the logseries.

4.6. Evaluating the goodness-of-fit to the logseries

In order to organize these definitions of particular fits to the logseries, we can say that: (i) the acceptable fit is a prerequisite of the other notions of fit, (ii) the pseudo basic good fit and the by-class good fit are alternatives to the basic good fit when it does not work, and (iii) the pseudo by-class good fit is an alternative to the by-class good fit. This hierarchy is illustrated by Fig. 5. It is important to note that these definitions are not mutually exclusive. For instance, Sample 9 shows both a pseudo basic good fit and a by-class good fit; Sample 7 presents both a pseudo basic good fit and a pseudo by-class good fit.

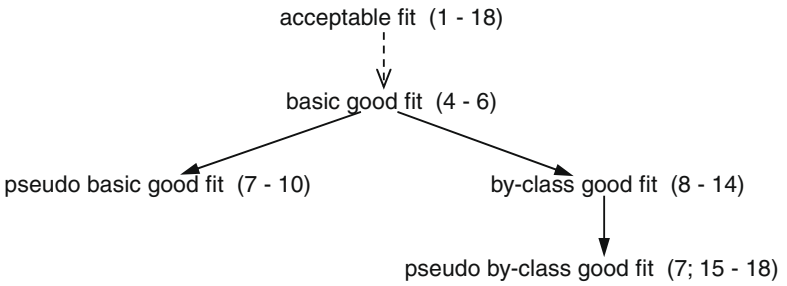


Fig. 5. Organization of the different definitions of fit, with their associated lists of samples.

From all these particular definitions we can specify a more general concept of "good fit". We say that the logseries distribution provides a *good fit* to an observed species abundance distribution, if it presents a basic good fit, or a pseudo basic good fit, or a by-class good fit, or a pseudo by-class good fit. From the list of samples associated to each particular notion of fit (cf. Fig. 5), it appears that Samples 4–18 show a good fit to the logseries, and that Samples 1–4 and Samples 19–26 do not present a good fit to the logseries. Therefore, we have met our objective: our goodness-of-fit test is general enough to recognize as positive all the field samples we have studied that are classically considered as being fitted by the logseries, and to point the others as negative ones.

Note that our definition of good fit does not explain how to perform the associated goodness-of-fit test. As a complement, we propose a procedure to estimate whether the species abundance distribution of a given sample is fitted by the logseries or not. It is useless to test all particular notions of fit on the sample. Rather, a sequence emerges from the hierarchy illustrated by Fig. 5. First, the sample has to show an acceptable fit to the logseries, otherwise it is rejected. Then, if the fit is acceptable, we test each particular fit one after the other, until the result is positive, in the following order: basic good fit, pseudo basic good fit, by-class good fit, pseudo by-class good fit. If none of them is positive, the sample is rejected.

5. Species abundance patterns in the communities of our simulation

5.1. Studied communities

Most of ecological data in the literature stem from small samples of large communities. Only a few complete censuses of communities, where all species are enumerated with their respective number of individuals, are available [73]. We list in Table 3 all complete enumerations of communities, found in the literature, that we have analyzed. They will be useful, by enabling us to compare the species abundance patterns they reveal with those found in the communities of our simulation. However, this comparison should be made with caution since, even though these censuses are classically considered as being complete by ecologists, they can also be seen as samples of larger communities populating wider geographical areas: even if we consider a whole forest, it is only a sample of a region, of a continent and of Earth. In contrast, the censuses of the communities generated by our simulation are really complete, since all individuals of the world are considered.

Remind that, according to the definition given in Section 3.1, a community is generally considered as a set of species from the same group and from the same trophic level. Thus, in the context of our simulation, it seems natural to define a community as the set of all predator species or all prey species of a given run, at a given time step. We have considered twelve com-

Table 3
Characteristics of the complete censuses of communities we have studied (N^* and S^* are the total numbers of individuals and of species; when an example covers several censuses, we give the approximate mean values of N^* and S^*).

Name(s)	Description	N^*	S^*	Ref.(s)
Quaker Run Valley	Nesting birds of Quaker Run Valley, New York, USA	14,171	80	[48,70]
Manu	Bird community of a tropical forest plot in Manu National Park, Peru	1864	245	[57]
Brown River	Bird community of a rainforest plot at Brown River, Papua New Guinea	27,112	165	[4]
Breed GB, Wint GB, Breed UK, Wint UK	Breeding and wintering birds of Great Britain and the United Kingdom	≈25,000,000	≈150	[3]
BCI 1982, BCI 1985, BCI 1990, BCI 1995, BCI 2000, BCI 2005	Trees of the Barro Colorado Island 50 ha Forest plot, Panama	≈230,000	≈305	[38]
Korup	Trees of the Korup 50 ha forest plot, Cameroon	322,833	444	[10]
La Planada 1997, La Planada 2003	Trees of the La Planada 24 ha forest plot, Colombia	≈110,000	≈230	[63]
Lambir 1992, Lambir 1997	trees of the Lambir 52 ha forest plot, Sarawak, Malaysia	≈340,000	≈1180	[37]
Mud 1988, Mud 1992, Mud 2000	Trees of the Mudumalai 50 ha forest plot, India	≈20,000	≈70	[53]
Sinharaja 1995, Sinharaja 2001	trees of the Sinharaja 25 ha forest plot, Sri Lanka	≈200,000	≈200	[26]
Paso 1987, Paso 1990, Paso 1995, Paso 2000	Trees of the Pasoh 50 ha forest plot, Peninsular Malaysia	≈300,000	≈800	[42]

Table 4
Total numbers of individuals (N^*) and of species (S^*) in our communities.

Name	Predator 1	Predator 2	Predator 3	Predator 4	Predator 5	Predator 6
N^*	15,291	2357	15,870	25,119	14,253	96,419
S^*	256	90	300	300	300	45
Name	Prey 1	Prey 2	Prey 3	Prey 4	Prey 5	Prey 6
N^*	134,538	123,574	138,283	200,117	124,005	268,136
S^*	500	237	479	197	371	18

munities, resulting from six different runs, and named, respectively, Predator i or Prey i, with $1 \leqslant i \leqslant 6$. Table 4 lists the total numbers of individuals and of species in these communities.

We have computed the species abundance distributions of a few samples taken from the communities of our simulation, as well as the associated distributions given by the logseries. We have assessed the goodness-of-fit of the logseries for each sample. In order to highlight the fact that all patterns of good fit presented in Section 4 are represented in our communities, we have chosen four samples illustrating the four patterns (cf. Fig. 6).

5.2. Replicate samples

After having estimated the goodness-of-fit of all samples, we could consider the results as properties of the communities. However, whether the fit is good or not, this might be a property of these particular samples only, and not necessarily a global property of all samples of the same size. To avoid this problem, biologists recommend using replicate samples whenever possible [74]. The idea is to take several samples (of the same size if possible) from a community, and to determine the per-

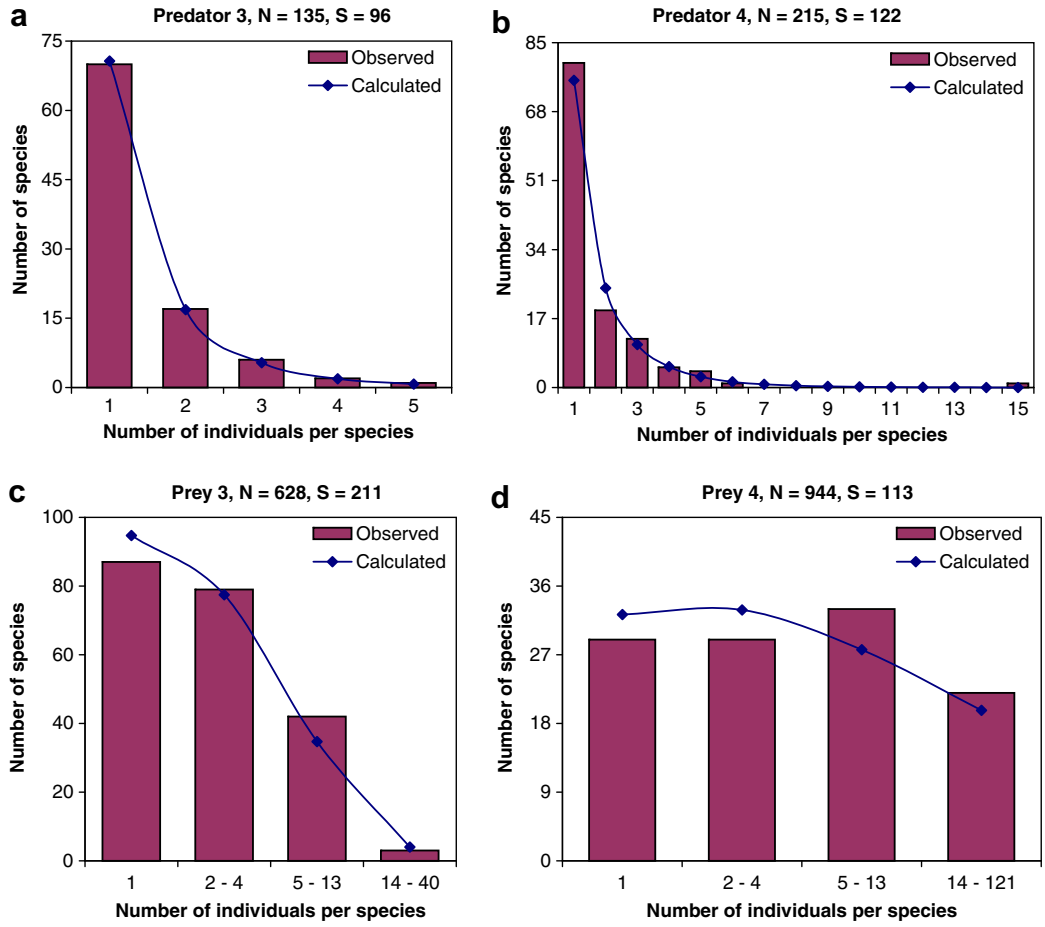


Fig. 6. Species abundance distributions of samples taken from our communities, showing: (a) a basic good fit, (b) a pseudo basic good fit, (c) a by-class good fit, (d) a pseudo by-class good fit.

Table 5
Percentage of acceptable fits (for a given sample size) in four communities of our simulation. For each community, the best sample (with the smallest error on n_1) and the worst sample (with the largest error on n_1) are presented.

	Predator 1		Prey 1		Predator 5		Prey 5	
	Best	Worst	Best	Worst	Best	Worst	Best	Worst
N (sample size)	78		399		90		770	
S	67	49	201	190	71	68	215	203
α	226	56.4	162	142	155	127	98.9	89.8
x	0.26	0.58	0.71	0.74	0.37	0.42	0.89	0.9
Calculated n_1	57.97	32.7	114.97	104.8	56.98	52.6	87.6	80.4
Observed n_1	58	28	115	76	57	46	83	44
Error on n_1 (%)	0.05	16.9	0.03	37.9	0.04	14.4	5.6	82.8
Acceptable fits (%)	99.8		71.1		100		5.8	

centage of samples that satisfy a given pattern, such as a good fit of the logseries for species abundances. The consistency of the pattern, expressed by this percentage, is then considered as a property of the community. The samples can be replicated in space [74] or in time [23,32,56,71] (depending in general on the kingdom considered: plants or animals).

To illustrate this idea, we have applied this technique to two sets of replicate samples presented in the literature. The tested pattern is: "the species abundance distribution of the sample provides a basic good fit to the logseries". The first set contains samples taken from the breeding bird community of Eastern Wood, England [23] (data gathered from [5] and [72]). The second one is constituted of samples taken from the breeding bird community of Skokholm, Wales [71]. In both cases, the samples are replicated in time, and do not have the same size. In Eastern Wood, 40% of the samples present a basic good fit, and in Skokholm the proportion is of 32%. Therefore, the pattern is weakly consistent in both sets.

One of the advantages of studying the communities generated by our simulation is that it would be trivial to replicate samples in time (by sampling at different time steps) and relatively easy to replicate samples in space (by recording the position of each individual in the world). However, we are limited by the fact that the field communities we are using for comparison do not provide any data for replicating samples neither in space nor in time. Thus, we have chosen to take from a community 1000 samples of a given size, by means of a computer routine. For each sample, individuals are drawn one by one from the community without replacement, and put back altogether before the next sample is generated. This procedure allows to test the validity of a pattern in the whole community (contrary to space sampling, which works on sub-communities) at a given time (contrary to time sampling). It has also been used by some ecologists for constructing replicate subsamples of a given size from a larger sample [17].

5.3. Fitting with the logseries the species abundance distributions observed in our communities

We start by studying the consistency of the acceptable fit to the logseries in our communities. The procedure we have defined is as follows: for each community, the size of the sample is chosen randomly, 1000 replicate samples of this size are generated, and the percentage of samples showing an acceptable fit to the logseries is calculated. Among those replicate samples, we also isolate the best one (with the smallest error on n_1) and the worst one (with the largest error on n_1). Table 5 lists the results for the communities Predator 1, Prey 1, Predator 5 and Prey 5. It appears from the comparison of the samples that the percentage of acceptable fits seems to be negatively correlated with sample size. We study this relation more thoroughly in Section 5.4.

We have also tested the consistency of the basic good fit to the logseries on our communities, by taking 1000 replicate samples of a randomly chosen size, and calculating the percentage of samples showing a basic good fit. For each community, among all samples providing an acceptable fit to the logseries, we have also isolated the best one (with the lowest χ^2) and the worst one (with the highest χ^2). Fig. 7 presents the results found for the community Prey 2, for which sample size was 363. Among all replicate samples, 29% showed an acceptable fit, and 26% a basic good fit. Among samples presenting an acceptable fit, the best one showed a basic good fit (for $\chi^2 = 6, d = 14$, and $P = 0.97$, cf. Fig. 7a) contrary to the worst one (for $\chi^2 = 46.7, d = 27$, and $P = 0.01$, cf. Fig. 7b). However, the latter presents a pseudo basic good fit to the logseries, because $P^{red} = 0.64$ (since, after reduction, $\chi^2 = 12.5$ and $d = 15$). It shows also a by-class good fit to the logseries, since $P = 0.52$ (for $\chi^2 = 2.27$ and $d = 3$). Actually, for this particular sample size, all samples that present an acceptable fit also show a good fit to the logseries.

More generally, Table 6 lists the results of the tests of consistency for each kind of fit in all our communities. We distinguish between six levels of fit. They are expressed as the percentages of samples showing, respectively: (1) an acceptable fit to the logseries, (2) a basic good fit to the logseries, (3) a basic good fit or a pseudo basic good fit to the logseries, (4) a basic good fit or a by-class good fit to the logseries, (5) a basic good fit or a by-class good fit or a pseudo by-class good fit to the logseries, (6) a good fit to the logseries. Our results show that, in general, the consistency of the good fit to the logseries is strong when the size of the replicate samples is small. On the other hand, it tends to get weaker when sample size increases. This is what is classically observed in field studies, when the logseries is used to approximate species abundance distribu-

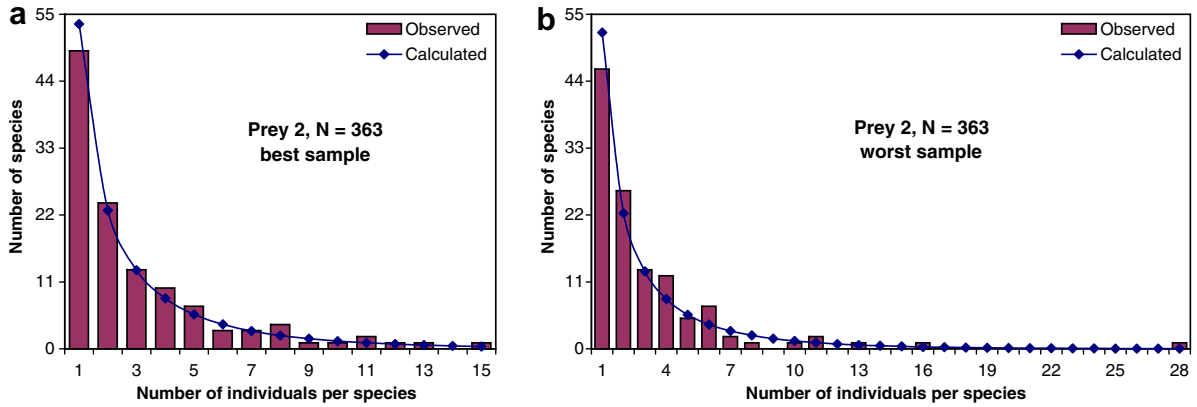


Fig. 7. Species abundance distributions of the best sample (with the lowest χ^2) and of the worst sample (with the highest χ^2) among the replicated samples of size 363 showing an acceptable fit, taken from the community Prey 2.

tions [23,68,70,73]. Therefore, with respect to that point, our communities give the same kind of results that field communities do.

Note that the percentage of acceptable fits is globally a good approximation of the percentage of good fits. This is interesting in the sense that testing automatically all the definitions of particular fits on replicate samples (in order to determine the exact percentage of good fits) is very time-consuming. If exact results are not required, the percentage of acceptable fits can provide interesting results very fast (since there is no need to compute the whole approximated distribution, or to perform a χ^2 test). Note also that the number of acceptable fits is an upper bound of the number of good fits, since the acceptable fit is a prerequisite of all other notions of fit. On the other hand, the number of basic good fits provides a lower bound of the number of good fits, since the definition of the basic good fit is the most restrictive among all levels of good fit.

5.4. Evolution of the goodness-of-fit according to sample size

In order to study more thoroughly the relation between the consistency of the good fit to the logseries and the size of the replicate samples, we have performed once again the experiment presented in Section 5.3, except that, instead of choosing randomly a particular sample size, we have allowed this one to vary. For a sample size N that is increased by 1 at each iteration (and starting with $N = 1$), we generate 1000 replicate samples of size N, and we compute the percentage of samples that show a good fit to the logseries. Fig. 8 presents the results found for the communities Prey 1 and Predator 6. The relation between both parameters is confirmed: globally, the bigger the sample, the less replicate samples show a good fit to the logseries. This negative correlation is observed in all our communities, although some fluctuations can occur and produce sometimes a slightly different curve shape.

Despite these small variations, the two charts presented in Fig. 8 are representative of two families of curves. (a) The relation found for Prey 1 shows a slow decrease of the percentage of good fits; this appears to be characteristic of the communities with a high species richness, such as Prey 1, 3, 5, and Predator 1, 3, 4, 5 (the less diverse being Predator 1, with 256 species). (b) The curve constructed for Predator 6 presents a much steeper slope; this has been found to be a property of the communities with a low diversity, such as Prey 2, 4, 6, and Predator 2, 6 (the most diverse being Prey 2, with 237 species).

Table 6
Results of the tests of consistency (expressed as percentages) for each level of fit, performed on 1000 replicate samples of size N, in all our communities.

	N	(1) Acceptable fit	(2) Basic good fit	(3) (2) + pseudo basic good fit	(4) (2) + by-class good fit	(5) (4) + pseudo by-class good fit	(6) Good fit
Predator 1	80	99.1	95.3	97.8	96.3	97.2	97.8
Prey 1	400	70.5	64.7	66.9	64.9	65.1	67
Predator 2	60	74.9	72	74.4	73.6	74.2	74.7
Prey 2	300	41.9	38.4	40.3	40.3	41	41.5
Predator 3	150	98.2	93.1	95.6	95	95.6	96
Prey 3	600	72.2	66.6	68.5	68.4	68.9	70
Predator 4	200	92.9	87.5	91.1	89.4	90.2	92
Prey 4	900	44.6	36.8	37.4	39.2	39.2	39.6
Predator 5	170	92.2	86.8	89.2	87.9	89	89.6
Prey 5	700	7.4	6.1	6.4	6.3	6.3	6.5
Predator 6	120	28.2	18.3	21.7	19.5	23.8	25.5
Prey 6	100	6.5	0.4	4.7	4.6	5.1	5.7

If the same experiment is performed on field communities, all these results are corroborated. Fig. 9 shows that the negative correlation between the percentage of good fits and sample size is satisfied for the communities Manu and Quaker Run Valley, which is also the case of all other communities listed in Table 3. The only difference between the charts presented in Figs. 8 and 9 is the scale of the sample size axis. A direct consequence of this negative correlation is that a good fit to the logseries is generally provided only by small samples of a community, which has been noticed by ecologists [23,68,70,73].

Moreover, we observe also, in the field communities, the distinction between two families of curves, each one being associated with a certain level of species richness. (a) A highly diversified community, as Manu, produces a curve with a gentle slope; this is the same for Korup, BCI 1982, 1985, 1990, 1995, 2000, 2005, Lambir 1992, 1997, and Paso 1987, 1990, 1995, 2000 (the less diverse community being Manu, with 245 species). (b) The chart constructed for Quaker Run Valley shows a steep sloped curve, which is the case of all other communities having a low species richness, such as Brown River, Breed GB, Wint GB, Breed UK, Wint UK, Mud 1988, 1992, 2000, La Planada 1997, 2003, and Sinharaja 1995, 2001 (the most diverse being La Planada 2003, with 240 species). An important repercussion of this dichotomy based on species richness is that a more diverse community provides more samples fitting the logseries, which has also been observed by ecologists [21,68,70].

On the one hand, the appearance of this dichotomy between species-rich and species-poor communities both in nature and in our simulation provides a strong validation for our simulation. On the other hand, the fact that the associated threshold of 240 species is satisfied both in nature and in our simulation tends to validate its use in field studies. Indeed, strictly speaking, only the communities of our simulation are "complete", and can allow to draw conclusions about a result related to the total number of species of a community.

5.5. Species abundance distribution patterns in large samples

It appears from our study of the evolution of the goodness-of-fit of the logseries that Fisher's model gives a good approximation of a species abundance distribution only for small samples. This seems to contradict Williams when he remarked that "samples as large as 15,609 individuals and as small as 440 individuals both agree very closely with the results calculated from Fisher's series" [21]. However, it should be noticed that he was studying an insect community comprising probably several billions of individuals, compared to what both 15,609 and 440 can be considered as small sample sizes. Indeed, Williams recognized later that the logseries would not give a good fit to abundance data if sample size was increased too much [68,70], which has been confirmed by other ecologists since then [23,73].

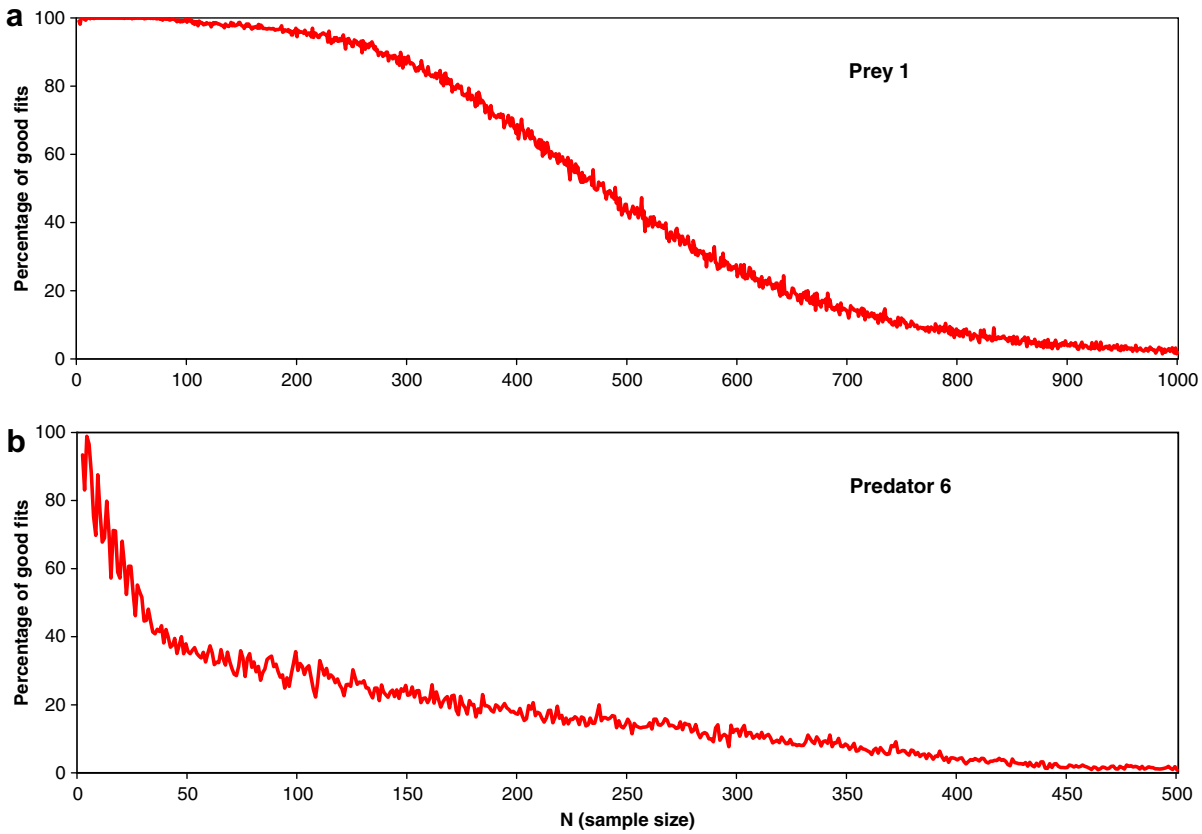


Fig. 8. Percentage of good fits vs. sample size, in two of our communities.

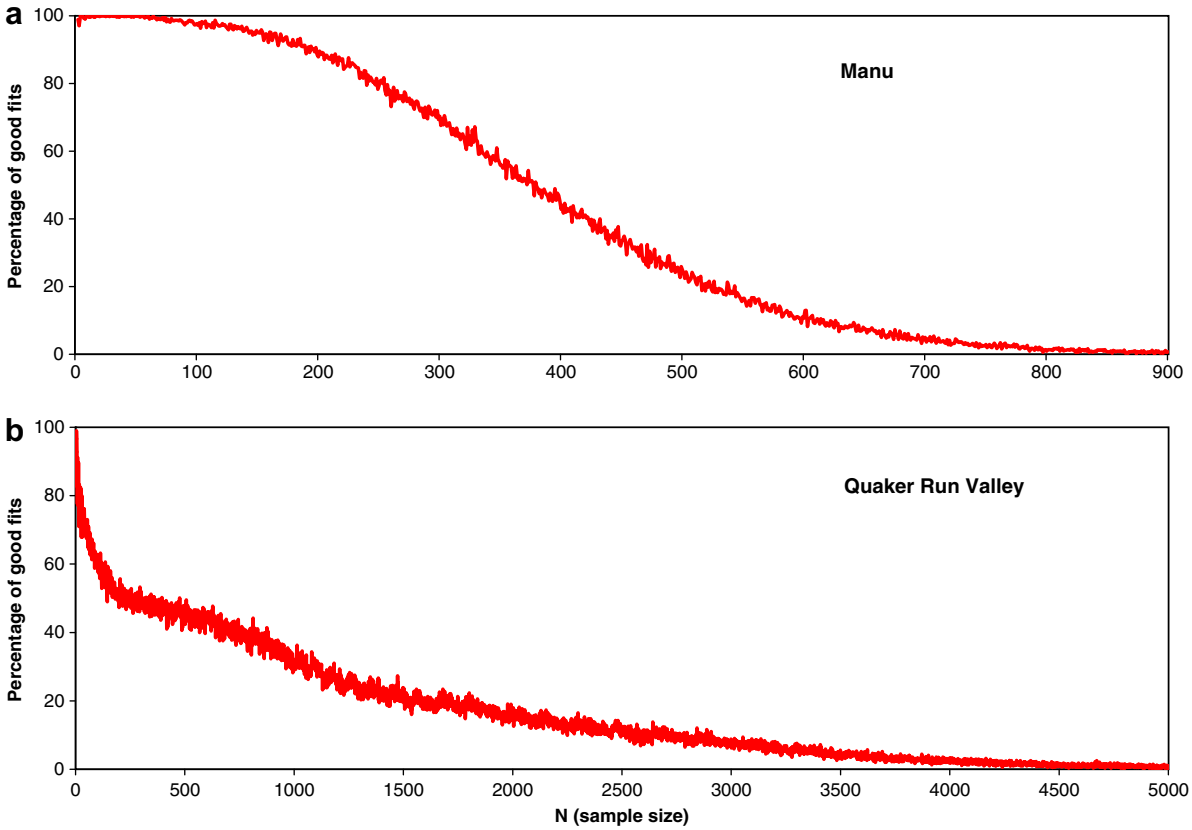
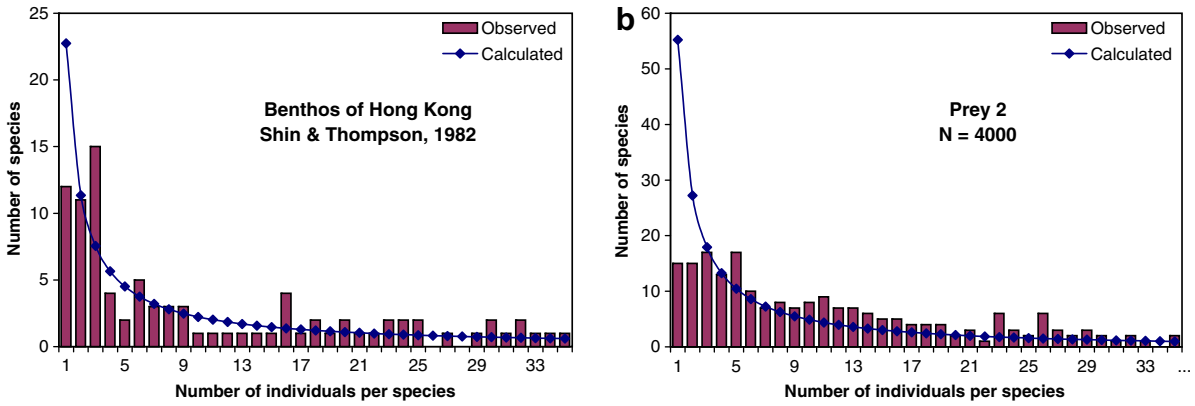


Fig. 9. Percentage of good fits vs. sample size, in two field communities.



a) Sample 26, and (b) a sample from our community Prey 2.

According to Fisher's model, on an arithmetic scale the largest number of species in a sample should always be observed for singleton species (since $n_1 > n_2 > n_3 > \ldots$), and n_1 should be about twice n_2 (since $n_2 = \alpha x^2/2 \approx \alpha x/2$, if $x \approx 1$). This was true for the small samples presented in previous sections (cf. Figs. 1, 2a, 3a, 6a and b, and 7). When larger samples are examined, whether they are taken from our communities or they result from field studies, these properties are not necessarily satisfied (cf. Fig. 2b and 10). The species abundance distributions observed are no longer J-shaped. Among the field samples listed in Table 2, Samples 19–26 cannot be fitted by the logseries for this reason.

In Fisher's model, the number of singleton species is small in very small samples; it grows when sample size increases, rapidly at first, and then more and more slowly until it approaches the value of α, which is its upper limit. This seemed not

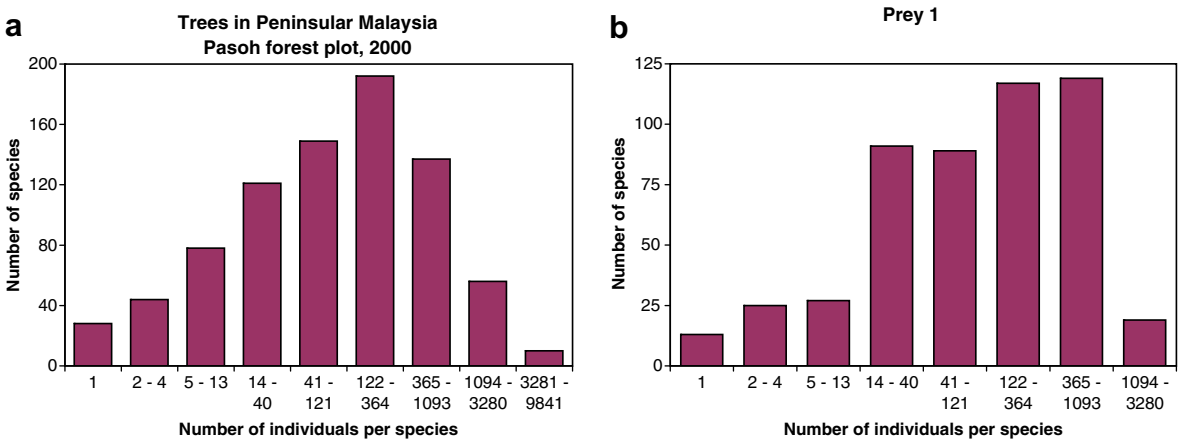


Fig. 11. Grouped species abundance distributions of (a) the field community Paso 2000 ($N = 296,134$ and $S = 815$) [42], and (b) our community Prey 1 ($N = 134,538$ and $S = 500$).

convincing for Corbet, who remarked that "it is a curious fact that the number of [singleton species] should approach a constant value with increasing [sample size]" [21]. Indeed this is not what can be observed, neither in the communities generated by our simulation nor in field communities. The comparison of Figs. 7 and 10b, which present two samples of different sizes taken from the community Prey 2, is sufficient to contradict the fact that n_1 should increase with sample size. We will not perform any further analysis of the behaviour of n_1, since it would be meaningful only if α was independent of sample size, which is not the case (cf. Sections 3.5 and 5.6).

Grouped species abundance distributions of small samples fitted by the logseries are right-skewed (cf. Fig. 3b, 4 and 6c and d), which is an intrinsic property of Fisher's model. On the contrary, abundance data of large samples and complete enumerations are generally left-skewed on a logarithmic scale (cf. Fig. 11). This is true of all our communities, and this is supported by several ecological studies (presenting various field examples, with their references) [23,73].

In the light of the foregoing observations, what is important to notice is that, even if the logseries do not fit the species abundance distributions of large samples or of complete enumerations, it fails for the same reasons in our communities and in field communities. In other words, even if they are not fitted by Fisher's logseries, the species abundance distribution patterns observed in large samples taken from the communities generated by our simulation are coherent with the patterns observed in nature. On the other hand, the patterns observed in our communities support the growing opinion that classical species abundance models cannot explain all observed field data [73].

5.6. Evolution of Fisher's α according to sample size

As already mentioned in Section 3.5, contrary to one of the main assumptions of Fisher's model, α is not independent of sample size. Charts representing the evolution of α according to the number of individuals show that it varies rapidly when sample size is small, then more and more slowly as N increases, until almost stabilising in very large samples (cf. Fig. 12). This is common to all our communities and to all field communities listed in Table 3. More precisely, they all satisfy the relation $\alpha = pN^q$, proposed in [13] for modelling the relationship between α and N. The only difference between two communities is in the values of the parameters p and q. It is important to note that the satisfaction of this relation in our communities tends to validate the work presented in [13], which was very promising but unfortunately not confirmed by any other study.

The question that arises now is: does the values of p and q from the relation $\alpha = pN^q$ say something about the community for which they are estimated? We observe that p is always positive, and that it seems related to the total number of species: globally, the richer the community, the higher the value of p. On the other hand q can have positive or negative values. This creates a dichotomy (which is different from the one mentioned in Section 5.4) between two families of relations, and thus two groups of communities. A positive value of q gives a convex curve (cf. Fig. 12). This is what is observed in all our communities, and all communities listed in Table 3, except for Mud 1988, 1992, 2000. On the other hand, a negative value of q gives a concave curve. This is what is observed in the examples presented in [13] and in the Mudumalai community. It is not clear yet what creates this distinction. All we can say is that the examples in [13] are all small samples containing only a few hundreds of individuals, and that Mudumalai forest plot is the poorest community of our study. We plan to investigate this question in our future work. In the context of such a study, the benefit of using our simulation is clear: with thousands of different communities available, whose complete composition would be known, a sound statistical analysis could be performed.

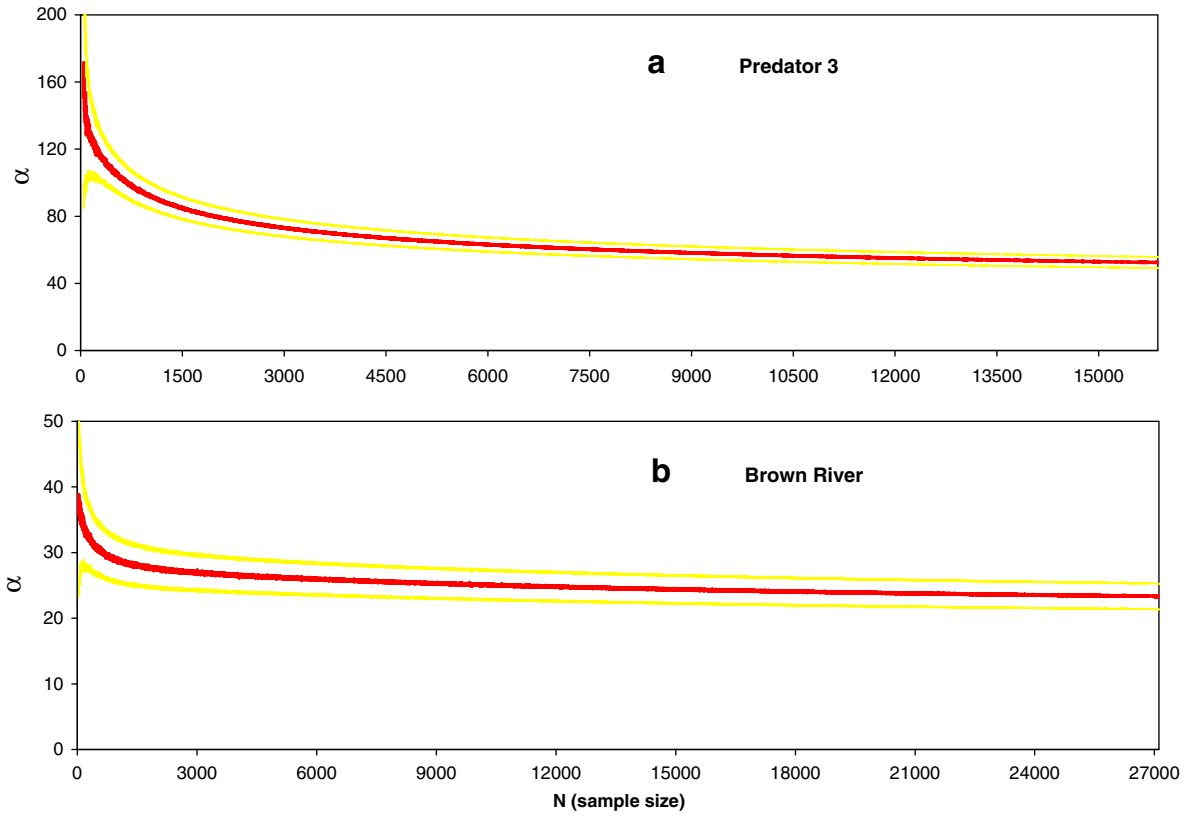


Fig. 12. α vs. sample size, (a) in our community Predator 3, and (b) in the field community Brown River. Values of α are given by the medium curve. The upper and lower curves are corrections on α, calculated with its variance (cf. Eq. (4)).

6. Conclusion and future work

In our previous work [24], we have developed an individual-based evolving predator–prey ecosystem simulation. The main characteristics of our approach are: (i) the use of a complex model, namely the fuzzy cognitive maps, to model individual behaviour, (ii) the definition of a species concept and of a speciation model based on genotypic distance, (iii) the implementation of an evolutionary mechanism allowing the genomic information of parents to be combined and mutated, (iv) the creation of a direct link between the behaviour model and the evolution mechanism. Our simulation constitutes a very complex multi-level dynamic adaptive system, involving a behavioural model allowing feedback effects and short term memory, a large number of interacting individuals, a multi-level resource system, the emergence and evolution of species, etc. Our first results have shown that the overall dynamic of the simulation at any level is coherent, and that its components present strong correlation patterns [24].

To further the validation of our simulation, we have compared the ecological patterns emerging from our ecosystem with those observed in natural ecosystems. We have focused the present study on species abundance patterns, because they are a key component of macroecological theories. To analyse these patterns, we have chosen to use Fisher's logseries, since it is one of the most classical models of species abundance distribution. These comparisons have been performed by testing the goodness-of-fit between an observed distribution and the one calculated by using the logseries. We have proposed several formal definitions for specific notions of good fit, in order to cope with some technical problems of the χ^2 test. From that, we have defined a very general concept of good fit, that can be statistically evaluated on any kind of distribution. To do so, we have developed a simple procedure that can be performed automatically by a computer routine.

The following results, that are well established in the ecological literature, are also observed in the communities generated by our simulation: (i) the logseries presents a good fit to the species abundance distributions of relatively small samples; (ii) it fails to do so for large samples and complete community enumerations; (iii) the logseries performs better on species-rich than on species-poor communities. Even though the logseries does not provide a good fit for large samples, the species abundance distribution patterns observed in our communities are similar to those observed in large samples of natural communities. Thus, at any level in sample size, our simulation gives coherent results in terms of relative species abundance, when compared with classical ecological results.

We have analysed also other species abundance patterns, that are less studied in the ecological literature. These patterns also are similar in our communities and in natural ones. The first one is given by the curve representing the evolution of

Fisher's α according to sample size, which satisfy the relation $\alpha = pN^q$. The second one, whose intuition is expressed by points (i) and (ii) in the previous paragraph, is given by the curve representing the evolution of the percentage of replicate samples showing a good fit to the logseries according to sample size, which shows a negative correlation. We have observed also that, among these negative correlations, a dichotomy appears between two kinds of patterns: (i) a highly diversified community (i.e. with more than 240 species) gives a gentle-sloped curve; (ii) a poorly diversified community (i.e. with less than 240 species) produces a steep-sloped curve.

This last result is interesting, for it allows to make predictions. If the total number of species in the community is known, one can identify what type of curve will be produced. Conversely, and more interestingly, if one considers a community for which only a few samples are available, and if these samples allow to identify what kind of curve could be constructed (by using interpolation techniques), one can obtain a lower bound or an upper bound of the total number of species in the community. Note also that, even if only one sample of the community is available, the beginning of the curve can be drawn by taking sub-samples of this sample, which can be enough to identify the curve family.

To sum up, we can say that the confrontation of the data generated by our simulation with field data has strongly validated our simulation in terms of relative species abundance. On the other hand, the study of our communities has extended ecological field results. First, we have provided a validation for the relation $\alpha = pN^q$ proposed in [13]. Second, we have discovered a dichotomy between species-rich and species-poor communities (with a threshold of 240 species) in terms of the evolution of the goodness-of-fit to the logseries according to sample size.

As a result of this study, several prospects naturally appear for our simulation. It will allow to make predictions and to analyse various important problems in theoretical biology, with a better understanding of the underlying phenomena. For example, we can test the two main biological models for conditions leading to sympatric speciation: the direct model and the indirect model for assortative mating [6,18,64], i.e. for the preference to mate with like individuals. (i) In the indirect model, a change in the ecology of species (e.g. food specialisation) affects the mating time or location, and assortative mating may evolve, as mating takes place largely among individuals sharing ecological traits. In our simulation, by adding another food resource for the prey, and creating species having different initial food preferences, we can determine whether mating preferences appear, by looking for relations between the probability of mating and individual distances. (ii) In the direct model for assortative mating, selection is supposed to operate directly on mating preferences, when the genes responsible for a trait influencing mating preferences also affect survival or fecundity [34]. We can test that as well, given that our simulation allows to compute individual fitness.

By analysing whether the heterogeneity of a population increases the number of speciation events, we can contribute to the currently widely discussed question of the relation between phenotypic plasticity and speciation [67]. Thanks to the fact that species membership of individuals is re-evaluated at each time step, we can study speciation through hybridization (when two closely related species start to interbreed and fuse back into a single species). As we have access to the exact and complete tree of life of our ecosystem we can use the genome of the individuals at a given time step to construct a phylogeny using different existing methods [47,54,75] and compare their ability to infer the correct tree of life.

We can also study what is the impact, on the average fitness, of the frequency of mutations, the number of existing species, or the physiological capacities of our agents (maximum distance of vision, maximum death age, maximum speed, etc.). We will also create a simple evolutionary mechanism applying the same mutation and crossover operations as in the original simulation but without any selection pressure. Then we could analyse, by comparing based on their fitness the agents evolved with or without selection, the importance of the selection mechanism in a world with a unique species, with several competing species, with a predator–prey hierarchy, etc. We could also compare the evolution of sexual and asexual populations [44], analyse the interactions and the diffusion of an invasive species into an existing ecosystem, study the apparition and effects of epistasis phenomena in the genome of our agents [35], etc.

References

[1] I. Agrell, Zur ökologie der Collembolen (On the ecology of Collembola), Opuscula Entomologica Supplement 3 (1941).
[2] F.J. Anscombe, Sampling theory of the negative binomial and logarithmic series distributions, Biometrika 37 (3/4) (1950) 358–382.
[3] H. Baker, D.A. Stroud, N.J. Aebischer, P.A. Cranswick, R.D. Gregory, C.A. McSorley, D.G. Noble, M.M. Rehfisch, Population estimates of birds in Great Britain and the United Kingdom, British Birds 99 (1) (2006) 25–44.
[4] H.L. Bell, A bird community of lowland rainforest in New Guinea I: composition and density of the avifauna, Emu 82 (1) (1982) 24–41.
[5] G. Beven, Changes in breeding bird populations of an oak-wood on Bookham Common, Surrey, over twenty-seven years, The London Naturalist 55 (1976) 23–42.
[6] D. Bolnik, B. Fiztpatrick, Sympatric speciation: models and empirical evidence, Annual Review of Ecology, Evolution, and Systematics 38 (2007) 459–487.
[7] M.T. Boswell, G.P. Patil, Chance mechanisms generating the logarithmic series distribution used in the analysis of number of species and individuals, in: G.P. Patil, E.C. Pielou, W.E. Waters (Eds.), Statistical Ecology, Spatial Patterns and Statistical Distributions, vol. 1, The Pennsylvania State University Press, 1971, pp. 99–130.
[8] J.A. Bullock, The investigation of samples containing many species I: sample description, Biological Journal of the Linnean Society 3 (1) (1971) 1–21.
[9] S. Bunyavejchewin, P.J. Baker, J.V. LaFrankie, P.S. Ashton, Huai Kha Khaeng forest dynamics plot, Thailand, URL <http://www.ctfs.si.edu/doc/plots/hkk/>.
[10] G.B. Chuyong, R.S. Condit, D. Kenfack, E.C. Losos, M.N. Sainge, N.C. Songwe, D.W. Thomas, Korup forest dynamics plot, Cameroon, URL <http://www.ctfs.si.edu/doc/plots/korup/>.
[11] C.A. Clarke, P.M. Sheppard, Interactions between major genes and polygenes in the determination of the mimetic patterns of *Papilio dardanus*, Evolution 17 (4) (1963) 404–413.
[12] R.S. Condit, S. Aguilar, A. Hernández, R. Pérez, S. Loo de Lao, G. Angehr, S.P. Hubbell, R.B. Foster, Tropical forest dynamics across a rainfall gradient and the impact of an El Niño dry season, Journal of Tropical Ecology 20 (1) (2004) 51–72.

[13] R.S. Condit, R.B. Foster, S.P. Hubbell, R. Sukumar, E.G. Leigh Jr., N. Manokaran, S. Loo de Lao, J.V. LaFrankie, P.S. Ashton, Assessing forest diversity on small plots: calibration using species-individual curves from 50-ha plots, in: F. Dallmeier, J.A. Comiskey (Eds.), Forest Biodiversity Research, Monitoring and Modeling, Parthenon Publishing Group, 1998, pp. 247–268.
[14] C.G.M. de Worms, A season's collecting with an electric light trap during 1929 near Egham Surrey, The Entomologist 63 (1930) 226–234.
[15] D.L. DeAngelis, W.M. Mooij, Individual-based modeling of ecological and evolutionary processes, Annual Review of Ecology, Evolution, and Systematics 36 (2005) 147–168.
[16] P.J. DeVries, D. Murray, R. Lande, Species diversity in vertical, horizontal, and temporal dimensions of a fruit-feeding butterfly community in an Ecuadorian rainforest, Biological Journal of the Linnean Society 62 (3) (1997) 343–364.
[17] P.J. DeVries, T.R. Walla, H.F. Greeney, Species diversity in spatial and temporal dimensions of fruit-feeding butterflies from two Ecuadorian rainforests, Biological Journal of the Linnean Society 68 (3) (1999) 333–353.
[18] U. Dieckmann, M. Doebeli, On the origin of species by sympatric speciation, Nature 400 (6742) (1999) 354–357.
[19] C.O. Dirks, Biological studies of Maine moths by light trap methods, The Maine Agricultural Experiment Station Bulletin 389 (1937).
[20] E.A.R. Ennion, The House on the Shore, Routledge and Kegan Paul, 1959.
[21] R.A. Fisher, A.S. Corbet, C.B. Williams, The relation between the number of species and the number of individuals in a random sample of an animal population, The Journal of Animal Ecology 12 (1) (1943) 42–58.
[22] J.A. Freeman, The distribution of spiders and mites up to 300 ft. in the air, The Journal of Animal Ecology 15 (1) (1946) 69–74.
[23] K.J. Gaston, T.M. Blackburn, Pattern and Process in Macroecology, Blackwell Science, 2000.
[24] R. Gras, D. Devaurs, A. Wozniak, A. Aspinall, An individual-based evolving predator–prey ecosystem simulation using a Fuzzy Cognitive Map as the behavior model, Artificial Life 15 (4) (2009) 423–463.
[25] V. Grimm, Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future?, Ecological Modelling 115 (2/3) (1999) 129–148
[26] C.V.S. Gunatilleke, I.A.U.N. Gunatilleke, P.S. Ashton, A.U.K. Ethugala, N.S. Weerasekera, S. Esufali, Sinharaja forest dynamics plot, Sri Lanka, URL <http://www.ctfs.si.edu/doc/plots/sinharaja/>.
[27] S.B. Hodgson, Lepidoptera at light in a chiltern beechwood, The Entomologist 70 (1937) 57–60.
[28] J.H. Holland, Hidden Order: How Adaptation Builds Complexity, Perseus Books, 1995.
[29] K. Hudec, J. Chytil, K. Šťastný, V. Bejček, Ptáci České republiky (The birds of the Czech Republic), Sylvia 31 (2) (1995) 97–148.
[30] R.A. Kempton, A generalized form of Fisher's logarithmic series, Biometrika 62 (1) (1975) 29–38.
[31] R.A. Kempton, The structure of species abundance and measurement of diversity, Biometrics 35 (1) (1979) 307–321.
[32] R.A. Kempton, L.R. Taylor, Log-series and log-normal parameters as diversity discriminants for the Lepidoptera, The Journal of Animal Ecology 43 (2) (1974) 381–399.
[33] M. Kimura, The Neutral Theory of Molecular Evolution, Cambridge University Press, 1983.
[34] M. Kirkpatrick, V. Ravigné, Speciation by natural and sexual selection: models and experiments, The American Naturalist 159 (Suppl. 3) (2002) 22–35.
[35] M. Kirschner, J. Gerhart, Evolvability, Proceedings of the National Academy of Sciences of the USA 95 (15) (1998) 8420–8427.
[36] B. Kosko, Fuzzy cognitive maps, International Journal of Man–Machine Studies 24 (1) (1986) 65–75.
[37] H.-S. Lee, S. Tan, S.J. Davies, J.V. LaFrankie, P.S. Ashton, T. Yamakura, A. Itoh, T. Ohkubo, R. Harrison, Lambir forest dynamics plot, Sarawak, Malaysia, URL <http://www.ctfs.si.edu/doc/plots/lambir/>.
[38] E.G. Leigh Jr., S. Loo de Lao, R.S. Condit, S.P. Hubbell, R.B. Foster, R. Pérez, Barro Colorado Island forest dynamics plot, Panama, URL <http://www.ctfs.si.edu/doc/plots/bci/>.
[39] S.K.Y. Lum, S.-K. Lee, J.V. LaFrankie, Bukit Timah forest dynamics plot, Singapore, URL <http://www.ctfs.si.edu/doc/plots/bukit/>.
[40] A.E. Magurran, Ecological Diversity and Its Measurement, Croom Helm, 1988.
[41] J. Mallet, A species definition for the modern synthesis, Trends in Ecology and Evolution 10 (7) (1995) 294–299.
[42] N. Manokaran, E.-S. Quah, P.S. Ashton, J.V. LaFrankie, M.N. Nur Supardi, W.A. Wan Mohd Shukri, T. Okuda, Pasoh forest dynamics plot, Peninsular Malaysia, URL <http://www.ctfs.si.edu/doc/plots/pasoh/>.
[43] R.M. May, Patterns of species abundance and diversity, in: M.L. Cody, J.M. Diamond (Eds.), Ecology and Evolution of Communities, The Belknap Press of Harvard University Press, 1975, pp. 81–120.
[44] D. Misevic, C. Ofria, R.E. Lenski, Sexual reproduction reshapes the genetic architecture of digital organisms, Proceedings of the Royal Society B: Biological Sciences 273 (1585) (2006) 457–464.
[45] P.M. Room, Diversity and organization of the ground foraging ant faunas of forest, grassland and tree crops in Papua New Guinea, Australian Journal of Zoology 23 (1) (1975) 71–89.
[46] M.L. Rosenzweig, Species Diversity in Space and Time, Cambridge University Press, 1995.
[47] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Molecular Biology and Evolution 4 (4) (1987) 406–425.
[48] A.A. Saunders, Ecology of the Birds of Quaker Run Valley, Allegany State Park, New York, New York State Museum, 1936.
[49] P.K.S. Shin, G.B. Thompson, Spatial distribution of the infaunal benthos of Hong Kong, Marine Ecology – Progress Series 10 (1982) 37–47.
[50] T.R.E. Southwood, V.K. Brown, P.M. Reader, The relationships of plant and insect diversities in succession, Biological Journal of the Linnean Society 12 (4) (1979) 327–348.
[51] T.R.E. Southwood, P.A. Henderson, Ecological Methods, Blackwell Science, 2000.
[52] R.E. Stewart, J.W. Aldrich, Breeding bird populations in the spruce region of the central appalachians, Ecology 30 (1) (1949) 75–82.
[53] R. Sukumar, H.S. Suresh, H.S. Dattaraja, R. John, N.V. Joshi, Mudumalai forest dynamics plot, India, URL <http://www.ctfs.si.edu/doc/plots/mudumalai/>.
[54] D.L. Swofford, PAUP*: phylogenetic analysis using parsimony (and other methods) 4.0 beta, Sinauer Associates (2002).
[55] Taiwan Forest Research Institute, Fushan forest dynamics plot, Taiwan, URL <http://www.ctfs.si.edu/doc/plots/fushan/>.
[56] L.R. Taylor, R.A. Kempton, I.P. Woiwod, Diversity statistics and the log-series model, The Journal of Animal Ecology 45 (1) (1976) 255–272.
[57] J.W. Terborgh, S.K. Robinson, T.A. Parker III, C.A. Munn, N. Pierpont, Structure and organization of an Amazonian forest bird community, Ecological Monographs 60 (2) (1990) 213–238.
[58] J.-M. Thiollay, Structure comparée du peuplement avien dans trois sites de forêt primaire en Guyane, La Terre et la Vie – Revue d'Écologie 41 (1) (1986) 59–105.
[59] M.R. Thomas, R.C. Shattock, Filamentous fungal associations in the phylloplane of Lolium perenne, Transactions of the British Mycological Society 87 (2) (1986) 255–268.
[60] J. Thompson, N. Brokaw, J.K. Zimmerman, R.B. Waide, E.M. Everham III, D.A. Schaefer, Luquillo forest dynamics plot, Puerto Rico, United States, URL <http://www.ctfs.si.edu/doc/plots/luquillo/>.
[61] L. Tomiałojć, T. Wesołowski, W. Walankiewicz, Breeding bird community of a primaeval temperate forest (Białowieża National Park, Poland), Acta Ornithologica 20 (3) (1984) 241–310.
[62] R. Valencia, R.S. Condit, R.B. Foster, K. Romoleroux, G. Villa Muñoz, J.-C. Svenning, E. Magård, M. Bass, E.C. Losos, H. Balslev, Yasuní forest dynamics plot, Ecuador, URL <http://www.ctfs.si.edu/doc/plots/yasuni/>.
[63] M.I. Vallejo, C. Samper, H. Mendoza, J. Tupac Otero, La Planada forest dynamics plot, Colombia, URL <http://www.ctfs.si.edu/doc/plots/laPlanada/>.
[64] S. Via, Sympatric speciation in animals: the ugly duckling grows up, Trends in Ecology & Evolution 16 (7) (2001) 381–390.
[65] C.R. Ward, F. Gobet, G. Kendall, Evolving collective behavior in an artificial ecology, Artificial Life 7 (2) (2001) 191–209.
[66] G.A. Watterson, Models for the logarithmic species abundance distributions, Theoretical Population Biology 6 (2) (1974) 217–250.

[67] M.J. West-Eberhard, Developmental Plasticity and Evolution, Oxford University Press, 2003.
[68] C.B. Williams, The relative abundance of different species in a wild animal population, The Journal of Animal Ecology 22 (1) (1953) 14–31.
[69] C.B. Williams, Notes on a small collection of Sphingidae from Nigeria, The Nigerian Field 19 (1954) 176–179.
[70] C.B. Williams, Patterns in the Balance of Nature, Academic Press, 1964.
[71] M.H. Williamson, The land-bird community of Skokholm: ordination and turnover, Oikos 41 (3) (1983) 378–384.
[72] M.H. Williamson, Are communities ever stable?, in: A.J. Gray, M.J. Crawley, P.J. Edwards (Eds.), Colonization Succession and Stability, Blackwell Scientific Publications, 1987, pp. 353–371.
[73] M.H. Williamson, K.J. Gaston, The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution, Journal of Animal Ecology 74 (3) (2005) 409–422.
[74] J.B. Wilson, Methods for fitting dominance/diversity curves, Journal of Vegetation Science 2 (1) (1991) 35–46.
[75] Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood, Molecular Biology and Evolution 24 (8) (2007) 1586–1591.