Quantitative Life Science

Prof. *Samir Suweis*
University of Padova

Written by: *Francesco Manzali*

Academic Year 2020/21

Compiled on October 22, 2020

Contents

Introduction

If you find any mistake/typo/missing thing, please write at *francesco.manzali@studenti.unipd.it* (or even just for feedback).

Francesco Manzali, 30/09/2020

Aknowledgments

Introduction

(Lesson 1 of 30/09/2020)
Compiled: October 22, 2020

Several models and techniques from physics can be used to study *life*, leading to important insights. In this view, organisms are regarded as **complex systems**, i.e. lots of *simple* nodes that are interacting with each other in a *network*, so that their global behavior exhibits *new properties* that the single parts do not have. This is, in essence, the concept of **emergence**.
The advantage of the complex system's framework lies in its **generality**. It can be applied to model population dynamics, for example the size of predator and prey species, but also neurons in the brain, or genes in cells — since in all of these cases we are dealing with many units interacting with each other.
Nonetheless, all of this would not lead to much insight if the *emergent properties* were extremely complex. Fortunately, this is not always the case: sometimes, the interaction of many parts in complex ways leads to surprisingly **simple patterns**, which moreover appear in very different systems. This suggests underlying common principles, things that make all complex systems similar, which are exactly the type of general "fundamental" knowledge sought by physicists, and that can (hopefully) be captured by **simple models**.

Part I

Ecosystems

Chapter 1

Dynamics of Single Species

Population Dynamics

We will start our discussion of complex dynamics by focusing on modeling the evolution of a single number: the size of a species' population. This apparently *simple* question is the core of the field of **Population Dynamics**.
Historically, the first model of population growth was developed by Fibonacci, and involved an exponentially increasing function: the famous Fibonacci's sequence models the uncontrolled growth of a group of rabbits. Clearly this is not very realistic — and the first corrections accounting for the *constraints* to growth were suggested by Malthus several years later.

1.1 Deterministic Models

We will now follow a similar course, starting by considering a simple organism, such as a bacterium. In general, bacteria are a single celled form of life ranging in size from 10 µm to 100 µm, consisting of a membrane that separates them from the environment, which contains some molecular machinery allowing replication. A bacterium can *grow* by converting external chemical nutrients into *biomass*, i.e. "parts of itself". This process of chemical transformation needed to sustain growth and life is denoted as **metabolism**.

1.1.1 Exponential growth

Let's suppose that the environment is favorable and rich in nutrients. In this case, a bacterium will regularly grow in biomass until it reaches a **critical size**, after which it *reproduces* by splitting into two bacteria.
So, if nothing changes, the total number of bacteria N will double after a certain time T, needed for a newly born bacterium to grow up to its critical size. Mathematically, we can describe the evolution of N as a function of time t with a differential equation:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = N\mu \Rightarrow \frac{\mathrm{d}N}{N} = \mu\,\mathrm{d}t \Rightarrow N(t) = N(0)e^{\mu t}$$

a. Exponential growth of bacteria

where μ is the time needed for N_t to increase by a factor of e, and it is related to the **doubling time** $T = \ln 2/\mu$, leading to:

$$N(t) = 2^{t/T} N(0)$$

For example, in the case of Escherichia Coli bacterium, $T \approx 20\,\text{min}$.

1.1.2 Logistic growth: Consumer-Producer model

Constraints to growth

In this analysis, we are implicitly assuming that organisms are **eternal**, and that resources are **infinitely abundant**. A more realistic model needs to consider that organisms die with a certain rate d, and that the amount R of resources changes with time (usually decreasing).
Experimentally, we observe that the population grows exponentially at the start, but then stabilizes around a certain size, called the environment's **carrying capacity**.

Carrying capacity

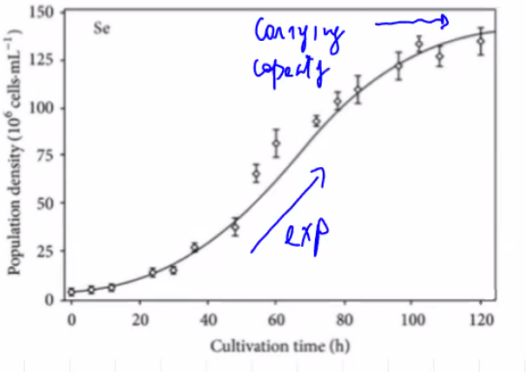


Figure (1.1) – Population size (per unit volume) of a bacterium species as function of time.

b. Consumer-Resource model

A model able to describe the emergence of the carrying capacity is MacArthur's **Consumer-Resource model**, which consists of two coupled differential equations. The first involves N_t (the *consumer* species):

$$\frac{\mathrm{d}N}{\mathrm{d}t} = N\Big(\underbrace{bcwR}_{\substack{\text{growth}\\ \text{rate}}} - \underbrace{d}_{\substack{\text{death}\\ \text{rate}}}\Big) \tag{1.1}$$

Here, the growth of N_t is driven by the *growth rate* term, which is the product of the frequency of births (b), and the rate of conversion of resources R into biomass, which is a chemical transformation with **metabolic efficiency** $w < 1$. The remaining constant c is characteristic of the species being examined. Often, we group bw in a single coefficient γ, called the **yield** coefficient:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = N(\gamma cR - d) \tag{1.2}$$

Finally, the term d, called the *death rate*, contributes to *decrease* N_t, counterbalancing the growth rate.

The second equation of MacArthur's model describes how resources R evolve over time. *Resources* are here considered to be **biotic** (organic), i.e. they are

different species (e.g. plants), which are consumed by the first (the *consumer*). Normally, R would increase up to a certain value (since they would still energy, e.g. sunlight, which is limited), according to a **logistic curve**:

$$\frac{\mathrm{d}R}{\mathrm{d}t} = R\Big[\underbrace{r^*\left(1-\frac{R}{k_c}\right)}_{\substack{\text{Logistic}\\ \text{growth}}} - \underbrace{cN_t}_{\substack{\text{Used by}\\ \text{Population}}}\Big] \tag{1.3}$$

The first term gives the logistic growth, which stabilizes at the environment's maximum resource density k_c. However, resources are used by the consumer species, and so they decrease by cN_t, with c and r^* being constants.
In summary, the Consumer-Resource models is completely described by (1.2) and (1.3):

(Lesson 2 of 01/10/2020) Compiled: October 22, 2020

$$\begin{cases} \frac{\mathrm{d}N_t}{\mathrm{d}t} = N_t(\gamma cR - d) \\ \frac{\mathrm{d}R}{\mathrm{d}t} = R\left[r^*\left(1-\frac{R}{k_c}\right) - cN_t\right] \end{cases}$$

Consumer-Resource equations, biotic resources

Finding a general solution is not easy, and so we introduce **simplifications**. For example, let's suppose that the amount of resources R *reacts quickly* to changes in population N_t, i.e. the timescale of resource dynamics is much *faster* than that of population dynamics.
This means that when N_t changes, R will quickly reach a new equilibrium *before* N_t can change again. So, from the point of view of the Consumer equation (1.2), R is fixed at the equilibrium value, which can be found by setting the time derivative in (**??**) to 0 (**quasi-stationary approximation**):

$$\frac{\mathrm{d}R}{\mathrm{d}t} = R\left[r^*\left(1-\frac{R}{k_c}\right) - cN\right] \stackrel{!}{=} 0 \tag{1.4}$$

Quasi-stationary approximation

Let's call R^* the value of R which solves (1.4). Now $R \equiv R^*$ is fixed, and so we can replace R in the first equation with R^*:

$$\begin{cases} \frac{\mathrm{d}N_t}{\mathrm{d}t} = N_t(\gamma cR^* - d) \\ R^*\left[r^*\left(1-\frac{R^*}{k_c}\right) - cN_t\right] = 0 \end{cases}$$

We solve the second one for R^* to obtain:

$$cN_t = r^*\left(1-\frac{R^*}{k_c}\right) \Rightarrow R^* = k_c - \frac{k_c cN_t}{r^*}$$

And then substitute the expression for R^* into the first one:

$$\frac{\mathrm{d}N_t}{\mathrm{d}t} = N_t\left(\gamma c\left[k_c - \frac{k_c cN_t}{r^*}\right] - d\right) = N_t(\tilde{k} - \tilde{a}N) \tag{1.5}$$

with:

$$\tilde{k} = \gamma c k_c - d \qquad \tilde{a} = \gamma c^2 \frac{k_c}{r^*} \tag{1.6}$$

Then we divide both sides by N_t, leading to:

$$\frac{1}{N_t}\frac{\mathrm{d}N_t}{\mathrm{d}t} = \underbrace{\tilde{k}}_{\mu}\left(1 - \underbrace{\frac{\tilde{a}}{\tilde{k}}}_{1/k} N_t\right) = \mu N_t\left(1 - \frac{N_t}{k}\right) \qquad \begin{matrix}\mu = \tilde{k} \\ k = \tilde{k}/\tilde{a}\end{matrix} \tag{1.7}$$

which is the equation for the logistic curve, which can be solved by separation of variables leading to:

$$N(t) = N_0 e^{\mu t}\left(1 + \frac{N_0}{k}(e^{\mu t} - 1)\right) \tag{1.8}$$

which correctly reproduces the dynamics that is experimentally observed for bacteria.

__Abiotic__ resources: the Monod equation

In the case of **abiotic** resources, i.e. when studying organisms that feed off chemicals in the environment, the Resource equation (1.3) is replaced by the **Monod equation**, which, ignoring the interaction term, reads:

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \mu_{\max}\frac{R}{k_s + R}$$

Here $\mu_{\max}$ is the maximum growth rate of the resource (i.e. the maximum value of $\frac{\mathrm{d}R}{\mathrm{d}t}$), and k_s is the *half-velocity constant*, i.e. the value of R when $R/\mu_{\max} = 0.5$.
Note that, when $R \ll 1$, growth is linear:

$$\frac{\mathrm{d}R}{\mathrm{d}t} \approx \frac{\mu_{\max}}{k_s}R$$

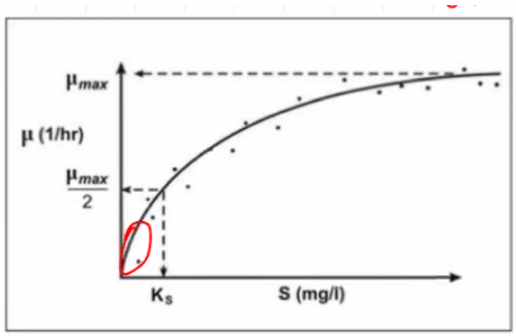


Figure (1.2) – Plot of $\frac{\mathrm{d}R}{\mathrm{d}t}$ given by the Monod equation.

Inserting back the interaction term $-cN_tR$, we get the following system:

$$\begin{cases} \frac{\mathrm{d}R}{\mathrm{d}t} = -cN_tR + \mu_{\max}\frac{R}{k_s+R} \\ \frac{\mathrm{d}N}{\mathrm{d}t} = (\gamma_c R - d)N_t \end{cases} \tag{1.9}$$

Exercise 1.1.1:

Apply the Quasi-Static Approximation to solve (1.9), and find how the solution $N(t)$ compares to (1.8).

Solution. For the Quasi-Static approximation, we suppose the resources to be constantly at equilibrium:

$$0 \stackrel{!}{=} \frac{\mathrm{d}R}{\mathrm{d}t} = \mu_{\max}\frac{R}{k_s + R} - RcN_t\Big|_{R=R^*}$$

Solving for R^* we get:

$$\mu_{\max} = cN_t(k_s + R^*) \Rightarrow R^* = \frac{\mu_{\max} - cN_t k_s}{cN_t}$$

which can then be substituted into the first equation in place of R, leading to:

$$\begin{aligned}\frac{\mathrm{d}N_t}{\mathrm{d}t} &= N_t(\gamma c R^* - d) = N_t\left(\frac{\gamma \cancel{c}}{\cancel{c}N_t}\left[\mu_{\max} - cN_t k_s\right] - d\right) = \\ &= \underbrace{\gamma\mu_{\max}}_{\tilde{a}} - \underbrace{(\gamma c k_s + d)}_{\tilde{b}} N_t\end{aligned}$$

So we are left with a linear ordinary differential equation:

$$\dot{N}_t = \tilde{a} - \tilde{b}N_t$$

The homogeneous equation (i.e. without the term $\tilde{a}$), is immediately solved by an exponential:

$$\dot{N}_t = -\tilde{b}N_t \Rightarrow N_t = Ce^{-\tilde{b}t}$$

To get the full solution we just need to add *any* particular solution, for example the stationary one, where $\dot{N}_t = 0$, which is:

$$0 \stackrel{!}{=} \dot{N}_t = \tilde{a} - \tilde{b}N_t \Rightarrow N_t = \frac{\tilde{a}}{\tilde{b}}$$

And so we arrive to:

$$N_t = \frac{\tilde{a}}{\tilde{b}} + Ce^{-\tilde{b}t} \tag{1.10}$$

C is a constant of integration, which is determined by the initial conditions. In this case, at time $t = 0$ the population is N_0, and so:

$$N_0 \stackrel{!}{=} \frac{\tilde{a}}{\tilde{b}} + C \Rightarrow C = \frac{\tilde{a}}{\tilde{b}} - N_0$$

Substituting back in (1.10):

$$N_t = \frac{\tilde{a}}{\tilde{b}} + \left(\frac{\tilde{a}}{\tilde{b}} - N_0\right)e^{-\tilde{b}t}$$

1.2 Stochastic Models

Typically, real data does not follow "clean" pattern such as the ones in the previous plots. Instead, N_t always *fluctuates* around the underlying trend. Thus, also at "stationarity", the population $N(t)$ is not fixed, but randomly rises and lowers around the carrying capacity.

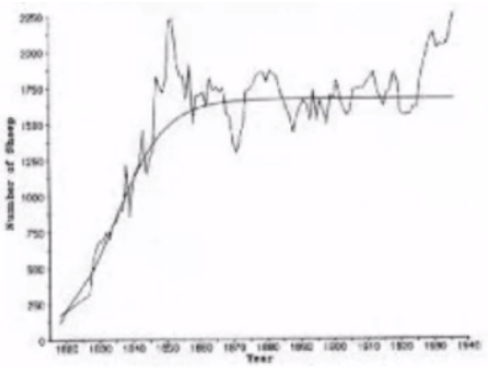


Figure (1.3) – Plot of a population's size over the years. As you can see, the trend being followed is that described by the previous models, but real data also exhibits *fluctuations*.

We can model this kind of behavior as a **stochastic process**, by treating N_t as a *collection* of random variables. The simplest non-trivial stochastic process is a **Markovian process**, where knowing the system's state (i.e. all the probabilities $\mathbb{P}[N_t = k]$) at an instant $t = t^*$ suffices for computing the distribution of each future state at $t > t^*$.

Markovian processes

A Markovian process is fully described by its **transition probabilities** $\mathbb{P}[N_t = n|N_{t^*} = n^*]$ with $t > t^*$.
If t were a discrete variable rising at steps of size Δt, then knowing all the transition probabilities for the *successive timestep* ($\mathbb{P}[N_{t+\Delta t} = k|N_t = j]$) would suffice[1], since the distribution at any future time can be reached by iterating *single step* evolutions.
However, here t is *continuous*, so we need a different approach. In general, a transition probability $\mathbb{P}[N(t + \mathrm{d}t) = k|N(t) = j]$ will be some (complex) function $f(j, k, \mathrm{d}t)$ of the end-points k and j and the size of the step $\mathrm{d}t$.
If we let $\mathrm{d}t = 0$, then there is no time for any transition to happen, and so:

$$\mathbb{P}[N(t) = k|N(t) = j] = \delta_{kj}$$

Then, we can Taylor expand f around $\mathrm{d}t = 0$:

$$\mathbb{P}[N(t + \mathrm{d}t) = k|N(t) = j] = \delta_{kj} + q_{kj}\,\mathrm{d}t + O(\mathrm{d}t^2) \qquad \mathrm{d}t \to 0$$

Transition rates

The linear coefficients q_{kj} are the so-called **transition rates** of the process, which provide the generalization of the transition probabilities over a single

[1] ∧Assuming the transition probabilities to be constant for all timesteps, which is usually the case.

timestep for the case with continuous time. In the following, we will use a more expressive notation, making clear that q_{kj} involves the transition from j to k:

$$q_{kj} \equiv T(j \to k)$$

So, to model the growth of a population as a stochastic process, we just need to specify the correct *transition rates* that can replicate the behavior we are interested in.

1.2.1 Logistic growth with fluctuations

Reactions

One way to do this effectively is to imagine a physical system and list the **reactions** that change its state. In our case, let's consider a population of n individuals, each acting independently of the others. The state n, i.e. the population's size, may change in response to the birth/death of an individual, and so we examine these two cases separately.

First, each of the living individuals X may reproduce, leading to two individuals $X + X$. This will happen with some frequency b, which is called the (per capita) **rate** of the reaction, and we assume for now to be constant and independent of n (this assumption will be relaxed later on). Precisely, b is the instantaneous *expected*[2] number of times the reaction occurs *per unit time* (rate) and *per individual* (per capita). So, on average, we will have $nb\,\mathrm{d}t$ births in a tiny (instantaneous) interval of time $[t, t + \mathrm{d}t]$.
We can denote a reaction by specifying the *reactants* (X) and the products ($X + X$), connected by an arrow which specifying the *rate*. In this case:

R1: Birth

$$X \xrightarrow{b} X + X \tag{1.11}$$

Similarly, by introducing a *second population* Φ, we can model the *loss* of an individual X as a transition from X to Φ:

R2: Death

$$X \xrightarrow{d} \Phi \tag{1.12}$$

where d is the death rate for the population of X.

Note that Φ here is just a placeholder for a "population which is not that of X and that we neglect". Thus, we can also model the reverse reaction:

R3: Immigration

$$\Phi \xrightarrow{\nu} X$$

which should not be interpreted as a "dead individual coming back to life", but instead as the arrival of a new organism of the same species as X from another region. Then, ν is the **speciation immigration** rate, i.e. the rate of arrival of new individuals.

[2] ∧It is not *actually* a probability, since it can be > 1.

Probability vs Rates. Suppose that a process happens at a rate r, i.e. we expect r events occurring per unit time, all *independent* of each other. Then, in a given time interval t, we expect $\lambda = rt$ events to happen. This λ can be measured only by running the same experiment many times and averaging the results. In any single *run*, however, we will observe k events in the time t, with a probability that is given by the Poisson distribution:

$$\mathbb{P}[k \text{ events in interval } t] = \frac{(rt)^k e^{-rt}}{k!}$$

In fact, the Poisson distribution is *defined* as the distribution followed by the random variable which *counts* the occurrences of random **independent** events, happening at a rate independent of the occurrences, and not simultaneously.

So, the probability is *specific* to a certain time window, while the *rate* is just the expected number of events per unit time.

We can now use the reactions to compute the *transition rates*, by simply measuring how they change the system's state (population).
In the case of the birth reaction $X \to X + X$, population rises by 1 unit. Since the rate *per capita* is b, the total rate for the population will be nb, and so:

$$T(n \to n+1) = nb \equiv b_n \qquad n \geq 1 \tag{1.13}$$

Rate of birth

(Since there are no other reactions that contribute to this specific transition)
The reaction $\Phi \to X$ is the special case of transition $T(n \to n+1)$ where the initial state is $n = 0$:

$$T(0 \to 1) = \nu \equiv b_0 \tag{1.14}$$

Rate of immigration

By denoting ν with b_0, we can merge (1.14) and (1.13) in a single expression:

$$T(n \to n+1) = b_n \qquad \forall n \geq 0$$

Finally, we have the reaction $X \to \Phi$, where the population lowers by 1:

$$T(n \to n-1) = nd \equiv d_n \qquad n \geq 1$$

Rate of death

If we want to capture the **logistic growth** of the population, i.e. the fact that at stationarity the population size lies around the carrying capacity k, we need to choose d as linear in n. In fact, look back to (1.15):

$$\frac{\mathrm{d}N_t}{\mathrm{d}t} = N_t(\underbrace{\tilde{k}}_{b} - \underbrace{\tilde{a}N}_{d}) \tag{1.15}$$

Here, the *change* of N_t is driven by a growth factor $\tilde{k}$, which we identify with the *per capita* birth rate b (which is indeed constant, since $\tilde{k} = \gamma c k_c - d$ from (1.6)), and the loss factor $\tilde{a}N$, which is linear in the population's size n, and which we identify with d. Then we can rewrite it as a function of the carrying

capacity k. Recall from (1.7) that $k = \tilde{k}/\tilde{a} \Rightarrow \tilde{a} = \tilde{k}/k$, and since $\tilde{k} = b$ we have:

$$d = \tilde{a}n = \frac{\tilde{k}b}{k} = \frac{bn}{k}$$

Thus we arrive at the rate:

$$T(n \to n-1) = nd = \frac{b}{k}n^2 \equiv d_n \tag{1.16}$$

All other possible transition have a *transition rate* of 0. For example:

$$T(n \to n+2) = 0$$ *Other rates*

This is because there is no *reaction* which allows this kind of transition. A rise in 2 of n can be realized by *repeating* other transition, e.g. by considering 2 concurrent births. However, assuming independent births, the probability of that happening is negligible:

$$\mathbb{P}[N_{t+\mathrm{d}t} = n+2 | N_t = n] = \mathbb{P}[\text{two births in } \mathrm{d}t] = (T(n \to n+1)\,\mathrm{d}t)^2 = O(\mathrm{d}t^2)$$

Since T are the *linear coefficients*, here T must be 0.

What about the *null* transition rate $T(n \to n)$? This is clearly non-zero, and we can derive it from the other rates. The idea is that, the sum over all the transition probabilities from a state n to any state k (including n) must be 1. In other words, the system must *always* be in some state, and so *total probability is conserved*. Then:

$$1 \overset{!}{=} \sum_{k \in \mathbb{N}} \mathbb{P}[N_{t+\mathrm{d}t} = k | N_t = n] = \sum_{k \in \mathbb{N}} T(n \to k)\,\mathrm{d}t = T(n \to n)\,\mathrm{d}t + \sum_{k \neq n} T(n \to k)\,\mathrm{d}t$$

And rearranging we get:

$$T(n \to n)\,\mathrm{d}t = 1 - \sum_{k \neq n} T(n \to k)\,\mathrm{d}t \tag{1.17}$$ *Rate of no change*

So, if probability is conserved, it is sufficient to specify all the transition rates *to other states* to completely define the stochastic process.

We are now ready to compute how the population distribution *evolves* over time. Let's denote with $P_n(t)$ the probability of the population still being n at time t:

$$P_n(t) \equiv \mathbb{P}[N_t = n]$$

After a tiny amount of time $\mathrm{d}t$, the distribution will be $P_n(t + \mathrm{d}t)$. For this we need to consider just two possibilities:

- Either the system was in some other state $k \neq n$ (with probability $P_k(t)$) and it moved to n at $t + \mathrm{d}t$
- Or the system was already in state n (with probability $P_n(t)$) and remained there at $t + \mathrm{d}t$

So:

$$P_n(t+\mathrm{d}t) = \sum_{k\neq n} \mathbb{P}[N_{t+\mathrm{d}t} = n | N_t = k] P_k(t) + \mathbb{P}[N_{t+\mathrm{d}t} = n | N_t = n]$$

Then we rewrite it in terms of known transition rates:

$$= \sum_{k\neq n} T(k \to n)\,\mathrm{d}t\, P_k(t) + T(n \to n) P_n(t) =$$

$$\underset{(1.17)}{=} \sum_{k\neq n} T(k \to n) P_k(t)\,\mathrm{d}t + \left(1 - \sum_{k\neq n} T(n \to k)\,\mathrm{d}t\right) P_n(t)$$

All that's left is to rearrange the terms and divide by $\mathrm{d}t$, forming a difference quotient:

$$\frac{P_n(t+\mathrm{d}t) - P_n(t)}{\mathrm{d}t} = \sum_{k\neq n} \Big[\underbrace{T(k \to n) P_k(t)}_{\text{Inward flow}} - \underbrace{T(n \to k) P_n(t)}_{\text{Outward flow}} \Big]$$

Here we notice that the *change* in a state's probability is given by a difference of the *probability flow* entering the state and that exiting it.
Finally, we take the limit $\mathrm{d}t \to 0$, so that the left term collapses into a first derivative:

$$\dot{P}_n(t) = \sum_{k\neq n} \Big[T(k \to n) P_k(t) - T(n \to k) P_n(t) \Big] \qquad \textit{Master Equation (general)}$$

In our case, only the $n \to n \pm 1$ transitions are possible, and so we have:

$$\begin{aligned} \dot{P}_n(t) &= T(n-1 \to n) P_{n-1}(t) + T(n+1 \to n) P_{n+1}(t) \qquad (1.18) \\ &\quad - [T(n \to n+1) + T(n \to n-1)] P_n(t) = \\ &= b_{n-1} P_{n-1}(t) + d_{n+1} P_{n+1}(t) - [b_n + d_n] P_n(t) \quad \forall n \geq 1 \qquad (1.19) \end{aligned}$$

Master Equation (Consumer-Producer)

This is the so-called **Master Equation** of the stochastic process (**birth-death** model). By integrating it, we can know the population's size distribution at any time.
Since if $n = 0$ there is death rate (population cannot be negative), the Master Equation reduces to:

$$\dot{P}_0(t) = d_1 P_1(t) - b_0 P_0(t) \qquad n = 0$$

For simplicity, let's study the distribution at **stationarity**: we expect to find it peaked *around* the carrying capacity k. So, we impose $\dot{P}_n(t) \equiv 0\ \forall n$, and denote the stationary distribution with P_n^*.
Then, for $n = 0$:

$$d_1 P_1^* - b_0 P_0^* = 0 \Rightarrow P_1^* = \frac{b_0}{d_1} P_0^* \qquad (1.20)$$

While for $n = 1$ we have:

$$b_0 P_0^* + d_2 P_2^* - (b_1 + d_1) P_1^* = 0 \Rightarrow P_2^* = (b_1 + d_1) P_1^* - b_0 P_0^* \underset{(1.20)}{=} \frac{b_0 b_1}{d_1 d_2} P_0^*$$

By reiterating this argument we can compute P_n^* $\forall n$, finding the following expression:

$$P_n^* = \prod_{j=0}^{n-1} \frac{b_j}{d_{j+1}} P_0^* = \frac{b_{n-1} b_{n-2} \cdots b_0}{d_n d_{n-1} \cdots d_1} P_0^* \qquad n > 0 \tag{1.21}$$

The only missing value is P_0^*, which can be obtained by imposing normalization:

$$\sum_{n=1}^{+\infty} P_n^* \stackrel{!}{=} 1$$

Recall that b_n (1.13) and d_n (1.16) have the following values:

$$b_n = nb \qquad d_n = \frac{b}{k} n^2 \tag{1.22}$$

Inserting them into (1.21) we get:

$$P_n^* = P_0^* \frac{b(n-1)b(n-2)\cdots \overbrace{b_0}^{\nu}}{b\frac{n^2}{k} b\frac{(n-1)^2}{k} \cdots \frac{b}{k} 1^2} = \frac{P_0^* \nu b^{n-1} (n-1)!}{\left(\frac{b}{k}\right)^n (n!)^2 n} = P_0^* \frac{\nu}{b} \frac{k^n}{n!} \frac{1}{n} \tag{1.23}$$

And to find P_0^* we impose normalization: *Stationary solution*

$$\sum_{n=1}^{\infty} P_n^* = 1 \Rightarrow P_0^* = \left(\sum_{n=1}^{\infty} P_n^* \right)^{-1} \tag{1.24}$$

Surprisingly, (1.23), when (1.24) is inserted in it, has a nice **closed analytical form**, which can be found either by hand or by using software (e.g. Mathematica):

```
Pn[n_][ν_, b_, k_] :=
 -b/(ν (EulerGamma + Gamma[0, -k] + Log[-k])) * ν/b k^n/n
   1
  ---------
  Factorial[n]
```

Figure (1.4) – Closed form expression for (1.23).

In any case, we can plot the result:

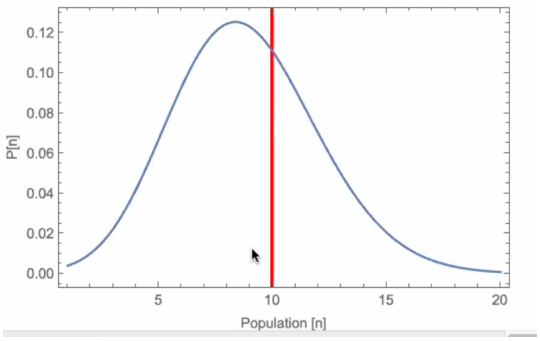


Figure (1.5) – Plot of (1.23) as function of a continuous n (for simplicity), with $\nu = 0.05$ (species immigration), $b = 5$ (birth rate) and $k = 10$ (carrying capacity). The distribution is *peaked* under the value of k (red line), but note that also values that are above k have a significant probability of being reached.

The most likely state is *under* the carrying capacity, but there is a non-zero probability for states with more individuals: fluctuations can in fact bring (temporarily) n over it!

Exercise 1.2.1:

What happens if, instead of a logistic death rate ($d_n \propto n^2$), we use a linear one?
More precisely, set:

$$\begin{aligned} b_n &= T(n \to n+1) = b\,n \qquad n > 0 \\ d_n &= T(n \to n-1) = d\,n \end{aligned} \tag{1.25}$$

with $b_0 = \nu$. How is P_n^* at stationarity for these transition rates? What happens to P_n^* if $\nu = 0$?

Solution. Substituting (1.25) into (1.21) we get the stationarity distribution:

$$\begin{aligned} P_n^* &= \frac{b_{n-1}b_{n-2}\cdots b_1 \overbrace{b_0}^{\nu}}{d_n d_{n-1}\cdots d_1} P_0^* = \frac{b(n-1)b(n-2)\cdots b(1)\nu}{d(n)d(n-1)\cdots d(1)} = \\ &= \frac{b^{n-1}\nu P_0^*}{n d^n} = \left(\frac{b}{d}\right)^n \frac{\nu}{nb} P_0^* \qquad n \geq 1 \end{aligned}$$

And for P_0^* we use the normalization condition:

$$1 \overset{!}{=} \sum_{n\geq 0} P_n^* \Rightarrow P_0^* = \left(\sum_{n=1}^{+\infty} \left(\frac{b}{d}\right)^n \frac{\nu}{nb}\right)^{-1}$$

Let's denote $b/d \equiv \xi$ for simplicity. Then recall the Taylor series for the natural logarithm:

$$\ln(1-x) = -\sum_{n=1}^{\infty} \frac{x^n}{n}$$

So:

$$P_0^* = \left(\frac{\nu}{b}\sum_{n=1}^{+\infty}\frac{\xi^n}{n}\right)^{-1} = \left(-\frac{\nu}{b}\log(1-\xi)\right)^{-1} = -\frac{b}{\nu}\frac{1}{\log(1-\xi)}$$

Putting everything together we get:

$$P_n^* = -\xi^n \frac{\cancel{\nu}}{n\cancel{b}}\frac{\cancel{b}}{\cancel{\nu}}\frac{1}{\log(1-\xi)} = -\frac{\xi^n}{n\log(1-\xi)} \qquad \xi \equiv \frac{b}{d}$$

For P_n^* to be ≥ 0, which is required for a probability, we need $\log(1-\xi) \leq 0$, meaning $0 < \xi < 1$, and so $d > b$. In this case, $\xi^n \to 0$ as $n \to +\infty$, and so does P_n^*. More importantly, $P_n^* \to +\infty$ when $n \to 0$, and so we are dealing with an *improper* distribution. Physically, this means that while some fluctuation may happen with non-zero probability, the system will be mostly **empty** ($n = 0$). This is to be expected: $d > b$ means that more individuals die than ones are born. If $b > d$, instead, P_n^* does not exist: in

fact, no stationary state is present, as the population will tend to *explode* to infinity.

Finally, if we set $\nu = 0$, then $P_n^* = 0 \ \forall n$, which is not a probability distribution, as it cannot be normalized. Physically, if there is no species immigration and $b < d$, then the population will be non-zero only for a finite time, and then $n = 0$ forever, meaning that the probability of state 0 should be 1. So, in that case, we could define a stationary distribution as a Dirac delta $\delta(n)$.

Stochastic Doubling

In the previous section, we modelled the growth of bacteria as happening at a fixed *rate*. This was motivated by observing that bacteria grow until a certain *critical size*, and then reproduce. Then, assuming that the rate of growth of a bacterium and the critical size are fixed, the birth rate will be *constant* over time.

However, in reality this is not so simple. In fact, while bacteria usually grow in size at about the same rate, the critical size at which reproduction occurs differs from generation to generation (fig. 1.6).

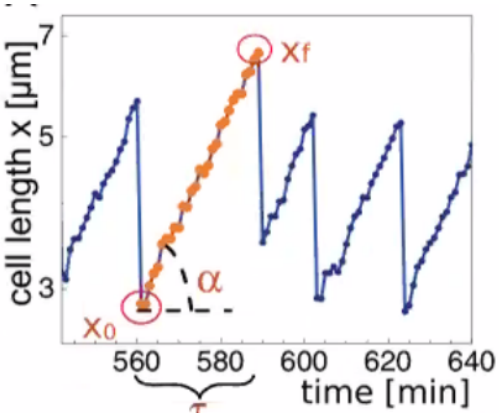


Figure (1.6) – Evolution of a bacterium length over time. Note that it grows until a threshold, and then splits, "plummeting" to a smaller size. However, the *threshold* varies at each generation, and so does the time needed for splitting. Taken from *Concerted control of Escherichia coli cell division* by Osella et al.

Let's briefly see how we can model such process. First, consider a cell born at $t = 0$. The rate of cell division is denoted as $h_d(x(t), t, x_0, \alpha)$, where x_0 is the size at birth, $x(t) = x_0 e^{\alpha t}$ is the size at time t, and α is the growth rate. Then, we are interested in computing the *cumulative probability* $P_0(t|x_0, \alpha)$ of a cell *not having divided* up to time t, given it started at size x_0 and it is growing at rate α.
To compute P_0 as a function of t, we can inspect how it evolves over a *tiny* interval of time $\mathrm{d}t$, leading to the following Master Equation:

$$P_0(t + \mathrm{d}t\,|x_0, \alpha) = P_0(t|x_0, \alpha)\,[1 - h_d(x(t), t, x_0, \alpha)\,\mathrm{d}t] \tag{1.26}$$

In fact, the only way for a cell to *not having divided* up to $t + \mathrm{d}t$ is to *first* not divide up to t, and then not divide in the interval $[t, t + \mathrm{d}t]$. Since h_d is the

division rate, $h_d(x(t), t, x_0, \alpha)\,\mathrm{d}t$ is the probability of the cell's dividing during $[t, t+\mathrm{d}t]$, and so we take $1 - h_d$ for the probability of **not** dividing.

Empirically, we observe that α is uncorrelated with both x_0 and x_f (i.e. the final cell's size, immediately before division). So, we can neglect α and write:

$$h_d(x(t), t, x_0, \alpha) \approx h_d(x, t)$$

Rearranging (1.26) and taking the limit $\mathrm{d}t \to 0$ leads to a differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} P_0(t|x_0, \alpha) = -h_d(x, t) P_0(t|x_0, \alpha)$$

which has the following *formal* solution:

$$P_0(t|x_0, \alpha) = \exp\left(-\int_0^t \mathrm{d}s\, h(x(s), s)\,\mathrm{d}s\right) \tag{1.27}$$

P_0 so computed is useful to compute the probability $P(t|x_0, \alpha)\,\mathrm{d}t$ of a cell dividing in $[t, t+\mathrm{d}t]$, which is just the probability of *not having divided* up to t, and then divide in $[t, t+\mathrm{d}t]$:

$$P(t|x_0, \alpha)\,\mathrm{d}t = P_0(t|x_0, \alpha) h_d(x(t), t)\,\mathrm{d}t \tag{1.28}$$

Substituting in the solution from (1.27) and dividing by $\mathrm{d}t$ we get:

$$P(t|x_0, \alpha) = h_d(x, t) \exp\left(-\int_0^t \mathrm{d}s\, h(x(s), s)\,\mathrm{d}s\right) = -\frac{\mathrm{d}}{\mathrm{d}t} P_0(t|x_0, \alpha)$$

Similarly, we can compute the probability of division as function of size x instead of time, since $x(t)$ is a monotonic function.
The steps are mostly the same. We denote with $h_d^*(x, t(x))\,\mathrm{d}x$ the probability of a cell dividing if it has size in $[x, x+\mathrm{d}x]$. Let t be the instant at which x is reached, and $t+\mathrm{d}t$ that at which $x+\mathrm{d}x$ is reached. Then, we can link h_d^* to the h_d previously defined:

$$h_d^*(x, t(x))\,\mathrm{d}x = h_d(x(t), t)\,\mathrm{d}t$$

Thus, if we know h_d and also the derivative $\mathrm{d}x\,/\,\mathrm{d}t$, we can rearrange the above to find h_d^*. For example, if $x(t) = x_0 e^{\alpha t}$, then:

$$h_g(x) \equiv \frac{\mathrm{d}x}{\mathrm{d}t} = \alpha x$$

After this, we can repeat the same steps to find the probability of a cell dividing at x:

$$P(x|x_0, \alpha) = h_d^*(x, t(x)) \exp\left(-\int_{x_0}^x \mathrm{d}s\, h_d^*(s, t(s))\right) = -\frac{\mathrm{d}}{\mathrm{d}x} P_0(x|x_0, \alpha)$$

1.2.2 Continuum limit and the Fokker-Planck equation

In general, working in the discrete case is difficult, i.e. it is not easy to solve expressions like (1.19) when n is discrete.
On the other hand, we can use all the methods from calculus to deal with *continuous* distributions. So, it would be useful to somehow *convert* (1.19) into a continuous equivalent.
To do so, we need to consider a sort of *thermodynamic limit*, in which the system's size N gets larger and larger, because in the continuum we have access to an *infinite* number of states. Then, we need to make so that the discrete states get "closer together" as N rises. This can be done by *redefining* the state as a ratio with the system's size, which in this case is a **population density**:

$$x \equiv \frac{n}{N}$$

In this way, if we use x as state, all values will always be in $[0, 1]$ (since $0 \leq n \leq N$), and the distance between a state and the next vanishes as $N \to \infty$. In other words, we are mapping a system getting *bigger* to a lattice of fixed size getting *finer*.

In the case of the Consumer-Producer model, we identify the system's size N with the **carrying capacity**[3] k.

Then we need to rewrite (1.19) using x instead of n, which involves a *change of random variable*. Let's fix a specific (but finite) k, meaning that x is still a discrete variable. If we increase n by one ($\Delta n = 1$), then x will increase by:

$$\Delta x = \frac{\Delta n}{N} = \frac{1}{k}$$

Let's define $P(x, t)$ to be the **probability density** of the system being in state x, meaning that the system will be in state x with *probability* $P(x, t)\Delta x$. But $x = n/k$ identifies a single state n, which is observed with probability $P_n(x)$. Thus, these two probabilities are the same:

$$P(x, t)\Delta x = P_n(t) \tag{1.29}$$

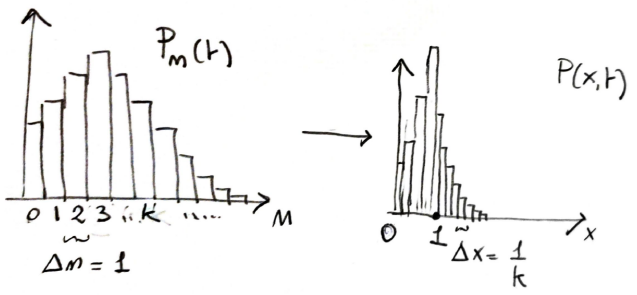


Figure (1.7) – The change of variable $x = n/k$ can be interpreted *graphically* as a "squishing" of the bin widths in the histogram. Since their area must remain the same (because probability is conserved), we have $P_n(t)\Delta n = P(x, t)\Delta x$, which is exactly (1.29).

[3] In fact, as the environment gets bigger, it can hold more individuals, meaning that also k will get bigger. Note that the system's size needs not to be the *maximum possible value* for n, just a quantity with the correct *scaling*. In fact, in this case, there is no *maximum possible state* n, but we have seen that the probability of finding $n \gg k$ decays exponentially.

For the transition rates:

$$T_n^+ \equiv T(n \to n+1) = bn = bkx = k\tilde{T}^+(x) \tag{1.30}$$
$$T_n^- \equiv T(n \to n-1) = \frac{b}{k}n^2 = \frac{b}{\cancel{k}}\cancel{k^2}x^2 = k\tilde{T}^-(x)$$

where we have gathered a factor k (it will simplify notation later on), and defined $\tilde{T}^+(x) = bx$ and $\tilde{T}^-(x) = bx^2$.

We now substitute (1.29) and (1.30) in (1.18):

$$\dot{P}_n(t) = T(n-1 \to n)P_{n-1}(t) + T(n+1 \to n)P_{n+1}(t) \tag{1.31}$$
$$- [T(n \to n+1) + T(n \to n-1)]P_n(t)$$

Note that:

$$T(n-1 \to n) = b(n-1) = bk(x - \Delta x) = k\tilde{T}^+(x - \Delta x)$$

and similarly $T(n+1 \to n) = k\tilde{T}^-(x + \Delta x)$. So, at the end we get:

$$\frac{\partial}{\partial t}P(x,t)\cancel{\Delta x} = k\Big[\tilde{T}^+(x - \Delta x)P(x - \Delta x, t)\cancel{\Delta x} + \tag{1.32}$$
$$+ \tilde{T}^-(x + \Delta x)P(x + \Delta x, t)\cancel{\Delta x} +$$
$$- [\tilde{T}^+(x) + \tilde{T}^-(x)]P(x,t)\cancel{\Delta x}\Big]$$

Now comes the tricky part. We *assume* that both $P(x,t)$ and $\tilde{T}^{\pm}(x)$ vary *smoothly* with x, meaning that we can effectively treat them as *differentiable* in x. This is clearly an *approximation* if k is finite, since normally x varies in steps of $\Delta x = 1/k$, but here we need x to be a real variable.
This allows to make Taylor expansions for the terms $\tilde{T}^{\pm}(x \pm \Delta x)P(x \pm \Delta x, t)$ as follows:

$$\tilde{T}^{\pm}(x \pm \Delta x)P(x \pm \Delta x) = \tilde{T}^{\pm}(x)P(x) \pm \Delta x\frac{\partial}{\partial x}[\tilde{T}^{\pm}(x)P(x)] +$$
$$+ \frac{1}{2}\Delta x^2\frac{\partial^2}{\partial x^2}[\tilde{T}^{\pm}(x)P(x)] + O(\Delta x^3)$$

In principle, this expansion will involve an *infinite* number of terms. However, often just the first two suffice. There are various *rationales* for this choice [1]: for example, just 2 terms are needed to model an evolution driven by a deterministic drift and some gaussian stochastic noise (Langevin equation).

So, by expanding all terms in (1.32) to second order, ignoring all higher orders, we get:

$$\frac{\partial}{\partial t}P(x,t) = k\Big[\cancel{\tilde{T}^+(x)P(x,t)} - \frac{\partial}{\partial x}\big(\tilde{T}^+(x)P(x,t)\big)\Delta x + \frac{1}{2}\frac{\partial^2}{\partial x^2}\big(\tilde{T}^+(x)P(x,t)\big)\Delta x^2 +$$
$$+ \cancel{\tilde{T}^-(x)P(x,t)} + \frac{\partial}{\partial x}\big(\tilde{T}^-(x)P(x,t)\big)\Delta x + \frac{1}{2}\frac{\partial^2}{\partial x^2}\big(\tilde{T}^-(x)P(x,t)\big)\Delta x^2 +$$
$$- \cancel{\tilde{T}^+(x)P(x,t)} - \cancel{\tilde{T}^-(x)P(x,t)}\Big] =$$

$$= k\Delta x \frac{\partial}{\partial x}\Big[P(x,t)\underbrace{(\tilde{T}^-(x)-\tilde{T}^+(x))}_{A(x)}\Big] + \frac{1}{2}k\Delta x^2\frac{\partial^2}{\partial x^2}\Big[P(x,t)\underbrace{(\tilde{T}^+(x)+\tilde{T}^-(x))}_{B(x)}\Big]$$

Recalling that $\Delta x = 1/k$ we have $k\Delta x = 1$, and so:

$$\frac{\partial}{\partial t}P(x,t) = -\frac{\partial}{\partial x}[P(x,t)A(x)] + \frac{1}{2\cancel{k}}\frac{\partial^2}{\partial x^2}[P(x,t)B(x)] \tag{1.33}$$

Fokker-Planck equation

where we absorbed the remaining $1/k$ in the definition of $B(x)$:

$$A(x) = -(\tilde{T}^-(x)-\tilde{T}^+(x)) \underset{(1.30)}{=} -(bx^2-bx) = bx(1-x) \tag{1.34}$$

$$B(x) = \frac{1}{k}\Big(\tilde{T}^-(x)+\tilde{T}^+(x)\Big) \underset{(1.30)}{=} \frac{1}{k}(bx^2+bx) = \frac{bx(x+1)}{k}$$

Equation (1.33) is called the **Fokker-Planck equation**. In summary, it is the *continuum* version of the Master Equation (1.19), involving a population density $x = n/k$ in place of the absolute population n, and obtained using a second order Taylor expansion.

1.2.3 The Langevin equation

The Fokker-Planck equation is a *deterministic differential equation* describing how the **probability distribution** of states evolves over time.
Physically, it describes the evolution of an *ensemble* of systems: if we simulate a huge number of populations, all with the same parameters, they will have different evolutions due to random fluctuations, but the *fraction* of systems that have a population density in $[x, x+\mathrm{d}x]$ at time t will be given exactly by $P(x,t)\,\mathrm{d}x$ (in the limit of an *infinite* ensemble).

An equivalent description can be derived by instead following a *single* population. In this case, a change in population density $\mathrm{d}x$ is given by a *drift* term $A(x)\,\mathrm{d}t$ and a stochastic fluctuation $\sqrt{B(x)}\,\mathrm{d}W$:

$$\mathrm{d}x_t = A(x_t)\,\mathrm{d}t + \sqrt{B(x)}\,\mathrm{d}W_t \tag{1.35}$$

where $\mathrm{d}W_t = \xi_t\,\mathrm{d}t$, with ξ_t being a random variable with 0 mean, unit variance $\langle\xi_t^2\rangle = 1$ and *no memory*:

$$\langle\xi_t\xi_{t'}\rangle = \delta(t-t')$$

The precise *meaning* of (1.35) is the following:

$$x_t = x_0 + \int_0^t \mathrm{d}t'\,A(x_t) + \int_0^t \sqrt{B(x_t)}\,\mathrm{d}W_t \tag{1.36}$$

where the latter integral is **not** the usual Riemann integral, but instead a **stochastic integral**. To explain what that is, we need to take a little detour.

The idea is to define a Wiener process (Brownian motion) W_t as a set of random variables $\{W_t\}$ with $0 \le t \le T$ and the following properties:

1. W_t is continuous in t and $W_0 = 0$

2. For each fixed t, $W_t \sim \mathcal{N}(0, t)$. Moreover, if $t \neq t'$, then W_t and $W_{t'}$ are **independent**.

3. An increment $W_{t+s} - W_s$ is **independent** of the starting point W_s (i.e. knowing the current state, there is no way to foresee how it will evolve). Moreover, $W_{t+s} - W_s$ has variance t.

Then, we use $W(t)$ as an *integrating function* to define a Riemann-Stieltjes (RS) integral as follows:

$$\int_a^b f(t)\,\mathrm{d}W_t = \lim_{n\to\infty} \sum_{i=0}^{n-1} f(c_i)[W_{t_{i+1}} - W_{t_i}] \qquad a = t_0 < t_1 < \cdots < t_n = b \tag{1.37}$$

where the $\{t_i\}_{i=0,\ldots,n}$ are a partition of the interval $[a, b]$, and $c_i \in [t_i, t_{i+1}]$. Basically, this is the same as the *usual* Riemann integral, but instead of having $(t_{i+1} - t_i)$ as the weight for $f(c_i)$, we use the increment of the Wiener process between these two instants.

This works well if the function f does not depend on stochastic components such as x_t or W_t. In this case, we say that the system is subjected to **additive noise**.
Unfortunately, this is not our case, since a dependence on x_t appears in (1.36), meaning that we are dealing with **multiplicative noise**. For this, (1.37) is problematic, since *how* we choose a specific c_i inside each sub-interval $[t_i, t_{i+1}]$ will change the result[4]!
Thus, it is needed to fix an **interpretation**, i.e. a rule to choose the c_i. In the literature, two possibilities are the most common:

- Itô's prescription, where c_i is taken to be the left most edge (t_i)

- Stratonovich's prescription, where c_i is the middle-point of the sub-interval, i.e. $(t_i + t_{i+1})/2$.

Each of them has pros and cons. Itô's prescription is easier to work with, since *by definition* of the Wiener process, W_t and $W_{t+\mathrm{d}t} - W_t$ are independent, which often can be used to simplify computations. This comes at a cost: all the commonly used rules of calculus *do not* work with Itô integrals!
On the other hand, Stratonovich prescription allows to salvage all common integration formulas from Riemann calculus. However, in this case the function f may not be independent of the increment, rendering several problems effectively intractable.

In the following, we will stick with Itô's calculus. For our purposes, the most important result is that there exists a **correspondance** between the Langevin

[4] For some explicit examples of that, see the notes of the *Models of Theoretical Physics* course.

equation (1.35) and the Fokker-Planck equation. Explicitly, given the following stochastic differential equation, in the Itô prescription:

$$\mathrm{d}x_t = A(x_t)\,\mathrm{d}t + \sqrt{B(x)}\,\mathrm{d}W_t$$

it can be shown[5] that the probability distribution $p(x,t)$ of x_t evolves according to:

$$\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x}[A(x)p(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[B(x)p(x,t)]$$

In the case of the Consumer-Producer model (1.34) we get:

$$\mathrm{d}x_t = bx_t(1-x_t)\,\mathrm{d}t + \sqrt{\frac{x_t(1+x_t)}{k}}\,\mathrm{d}W_t$$

Recalling that $\mathrm{d}W_t = \mathrm{d}t\,\xi_t$, we can divide by $\mathrm{d}t$ and take the limit $\mathrm{d}t \to 0$, resulting in the following (purely formal) expression:

$$\dot{x}_t = bx_t(1-x_t) + \sqrt{\frac{x_t(1+x_t)}{k}}\xi_t$$

In practice, this allows to numerically simulate the evolution of a given population, so that the statistics, when averaged over *many realizations*, will be the "correct ones" given by the Fokker-Planck equation.
This can be done, for example, through the Euler-Maruyama method, which is the generalization of the Euler algorithm to the stochastic case. The idea is to *discretize* time in a uniform grid $\{t_i\}_{i=0,\dots,n}$, set the initial state x_0, and iteratively *evolve* it step by step:

$$x_{i+1} = x_i + A(x_i)\Delta t + \sqrt{B(x_i)}\Delta W_i$$

where $\Delta W_i \sim \mathcal{N}(0,\Delta t)$ is randomly generated.

Exercise 1.2.2:

> Do the same for $T(0 \to 1) = \nu \neq 0$, and for both $\nu \neq 0$ and $T(n \to n-1) = dn$.

[5] ∧A proof would need some more results from Itô's calculus, for example the fact that $\mathrm{d}W^2 = \mathrm{d}t$. A discussion can be found in the notes for the *Models of Theoretical Physics* course.

Chapter 2

Dynamics of Multiple Species

(Lesson 3 of 07/10/2020)
Compiled: October 22, 2020

An **ecosystem** can contain tens of thousands of different species, all competing for survival and interacting in non-trivial ways.
In this chapter, we will continue the discussion about population dynamics, and show that the basic **birth-death stochastic process** discussed in the previous sections can be naturally generalized to the case of multiple species, and it is flexible enough to account for interesting effects of the *macro* world, i.e. it is not limited just to describing bacteria cultures.

2.1 Density Dependence

To apply the **birth-death** Master Equation (1.19) to ecosystems, we first need to generalize it to M species. So, starting from:

$$\frac{\mathrm{d}P_n}{\mathrm{d}t} = b_{n-1}P_{n-1}(t) + d_{n+1}P_{n+1}(t) - (b_n + d_n)P_n(t)$$

we label all variables with a (s) to identify each species:

Birth-Death ME for multiple species

$$\frac{\mathrm{d}P_n^{(s)}}{\mathrm{d}t} = b_{n-1}^{(s)}P_{n-1}^{(s)}(t) + d_{n+1}^{(s)}P_{n+1}^{(s)}(t) - (b_n^{(s)} + d_n^{(s)})P_n^{(s)}(t) \qquad s = 1, \dots, M \tag{2.1}$$

All the parameters $b_n^{(s)}$ and $d_n^{(s)}$ uniquely determine the distribution of each species' population at stationarity (1.21):

$$P_n^{*,(s)} = C_s \prod_{j=0}^{n-1} \frac{b_j^{(s)}}{d_{j+1}^{(s)}} \tag{2.2}$$

In general, the b_n and d_n will depend on all the species populations $n^{(s)}$ in some complex way.
For now, let's suppose they depend only on the size n of the population they are describing. We have seen that by choosing (1.22):

$$b_n = nb \qquad d_n = \frac{b}{k}n^2 \tag{2.3}$$

we get a *logistic curve* for the population's growth. More generally, we can consider the following *Taylor expansions* for the *per capita* birth/death rates, i.e. b_n and d_n normalized by the population's size n:

$$\begin{aligned} \frac{b_n^{(s)}}{n} &= b_s + \frac{r_s}{n} + O\left(\frac{1}{n^2}\right) \\ \frac{d_n^{(s)}}{n} &= d_s + \frac{u_s}{n} + O\left(\frac{1}{n^2}\right) \end{aligned} \tag{2.4}$$

If we keep only the 0-th order coefficients, we obtain a *pure* random walk, with growth rate b_s and death rate d_s. If we instead stop at first order, we get *per capita* rates that depend on the population's size. In other words, this means that the probability of an individual giving birth or dying is affected by the total number of individuals of the same species in the environment (**density dependence**).

Density dependence

Practically, fits of experimental data regarding trees in Amazon rainforests suggests a *negative value* for r_s, and that u_s may be neglected:

$$\begin{aligned} b_n^{(s)} &= r_s + b_s n \\ d_n^{(s)} &= d_s n \end{aligned} \tag{2.5}$$

A suggested explanation is the **Hanzen-Connell** hypothesis. The idea is that a tree attracts its natural enemies, such as species-specific parasites, which start inhabiting the nearby area. Thus, seedlings that fall *too close* to a tree will be subject to these predators, and may not survive. As a consequence, the areas with the highest density of trees of the same species will also be the ones where a new seed will find it harder to grow, leading to the density dependence of the *per capita* birth rate.

Hanzen-Connell hypothesis

With the choice (2.5), the solution at stationarity (2.2) is a *negative* binomial distribution (assuming r_s integer):

$$P_n^{*,(s)}(r_s, \xi_s) = \frac{1}{1-(1-\xi_s)^{r_s}} \binom{n + r_s - 1}{n} \xi_s^n (1-\xi_s)_s^r \qquad \xi_s = \frac{b_s}{d_s} \tag{2.6}$$

Note that, by including density dependence, we are *indirectly* accounting for the effect of interactions between species (here, for example, with the parasites that attack seedlings falling near a tree of the same species). Nonetheless, the equations are specific for each species: in a sense we are "integrating" the interactions in a sort of common "mean field", which then *independently* interacts with each population[1].

2.2 Neutral Theory

The generalized birth-death process (2.1) works well in principle, but as it is formulated requires a *huge* number of parameters to be estimated. In fact,

[1] ∧Here we are really *approximating* a complex multivariate probability distribution $P^*(n_1, \ldots, n_M)$ with one that is *separable*, i.e. can be expanded as $P^{*,(1)}(n_1) \cdots P^{*,(M)}(n_M)$. Clearly this won't be able to capture *all behaviors*, but practically works well enough, and makes a really complex problem mathematically tractable.

in general all the expansion's terms from (2.4) are *specific* to a given species s. Since an ecosystem involves ten of thousands of species, this becomes very impractical in practice.

Fortunately, species in nature that lie at the *same trophic level* (i.e. they are at the same "level" on the food chain), appear to evolve the same, with no difference in birth rates nor death rates. In other words, we can *neglect* the s index in the previous equations, dramatically reducing the number of free parameters, and still get a good description for an ecosystem.

Neutral Theory ***assumption***

The idea that different species appear "the same" is the thesis of **Neutral Theory**, proposed by Stephen P. Hubbel in 2001. Within this view, two species evolve as two different *realizations* of the same stochastic process. In other words, they are two random paths generated by the same algorithm with the same parameters, but just a different *random seed*. Thus, the evolution of a species is solely driven by *stochastic drift*.

Neutral Theory finds empirical validation in data from *huge*, complex ecosystems with *very high biodiversity*, such as rainforests. This is because only when considering a system with *lots* of species, all with *lots* of individuals can we neglect the *details*, "average everything" and focus on the "big picture".

So, according to neutral theory, P_n^* given by (2.6) is not just the probability density of a specific species having n individuals, but also the probability density of **any** species having n individuals, which is (approximately) the fraction of species having n individuals in an ecosystem. So, if we consider M species, we expect MP_n^* of them to have n individuals. This can be measured by making a *census* of all the species in a given area. Then we can plot the number $y(n)$ of species having n individuals, and fit the graph with (2.6) (multiplied by M) to find the missing parameters r_s and ξ_s.

In practice, *most* of the species will be *rare*, in the sense that they have *few* individuals. Conversely, large populations are rare, and so $y(n) \sim 0$ for large n, which can create problems with fitting.

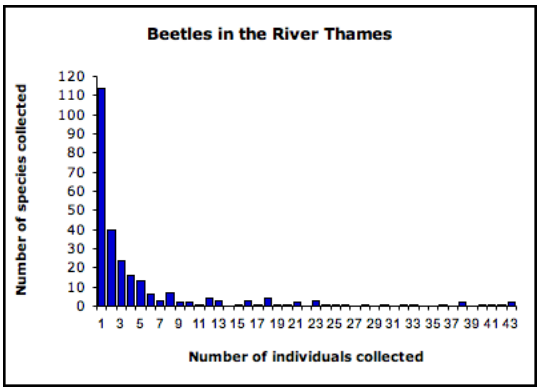


Figure (2.1) – Plot of the number of species $y(n)$ that have a given number of individuals n in the considered sample. Note that the height of bins decays *quickly*, and that for large n data is *sparse*.

However, when studying the distribution of population sizes, we are not much interested in the probability of finding a species with exactly 151 individuals,

but in general how many populations there are in the range of *hundreds* of individuals. This suggests the use of a *logarithmic binning*, i.e. "group" bins in common **abundance classes** that get exponentially larger and larger. For example, we can consider the first i-th bin to contain all the $n \in [2^i, 2^{i+1})$. Equivalently, we can plot $\log_2 n$ on the x axis rather than n (**Preston plot**).

This procedure solves the problem of bins getting "rarefied", and makes fitting easier. The number of species inside each abundance class is the so-called **Relative Species Abundance** (RSA). When normalized to unitary area (i.e. plotting *fractions* instead of *counts*), it is also called the **Species Abundance Distribution** (SAD).
By fitting it, we can validate both the stochastic model and the neutrality assumption:

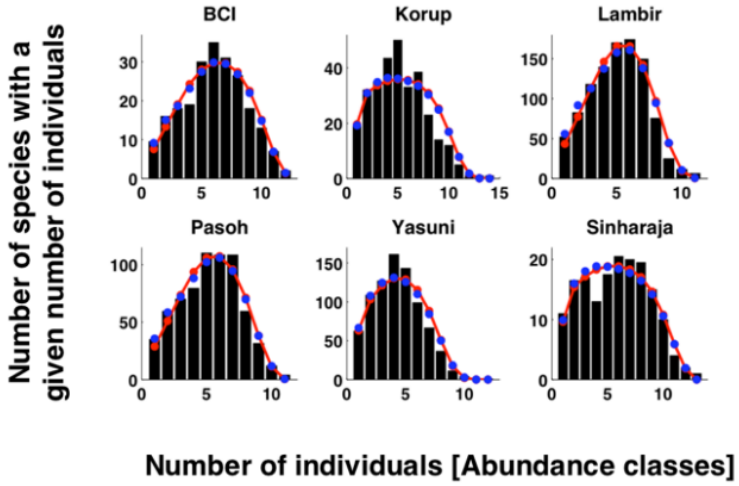


Figure (2.2) – Preston plot of the Relative Species Abundance for several forests, fitted by the negative binomial function (2.6) found by adding *density dependence* in a birth-death process. So, on the x axis we have $\log_2(n)$, and on the y axis the number $y(n)$ of species in $[2^n, 2^{n+1}]$, represented as a bin. Recall that here we are considering *separable* distributions (first approximation), a limited expansion of the birth/death rates (second approximation), and all species to be the same (neutrality assumption). Nonetheless, the predicted function is able to fit nicely the available data.

2.3 Scaling

Since most species are *rare*, to correctly measure their abundance we need to consider a very **large** sample, otherwise we are simply missing their contribution. This means that the RSA is **scale dependent**: we get a different plot depending on the size of the considered sample.

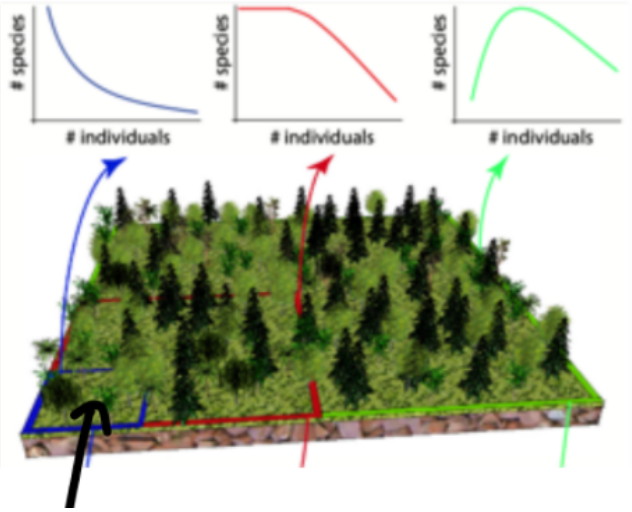


Figure (2.3) – If we look in a small patch of a forest (blue rectangle), we will inevitably miss most of the *rarer* species. Since a very high number of species are rare, we will get a high abundance of species with a very low population, leading to the "hollow curve" plot at the left. Note that this does not mean that most species have so few individuals (which is not the case if they are not in the process of extinction), but simply that we are missing most of the population size of the rare species due to the limited sample. If we study a larger area (red), we see that the abundance of low population species stabilizes, and finally lowers when we consider the whole forest (green). What it's happening is that we are observing new individuals of the same *rare* species, and so some *area* from the low abundance bins is moving towards the higher abundance, moving the left peak to the middle of the graph.

Sample vs Population

In practice, we cannot identify *all trees* in the Amazon rainforest (they are $\sim 10^{11}$), and so we need to start from a *small patch* and *infer* properties about the entire population. However, we cannot just *linearly scale* our observations: the species abundance on limited area *is not representative* of the global one, since we are missing many *rare* species.

This means also that the value of the parameters r, ξ *depends* on the sample size (**scale**) of the data being fitted. Fortunately, the negative binomial can fit the RSA at *any* scale[2]:

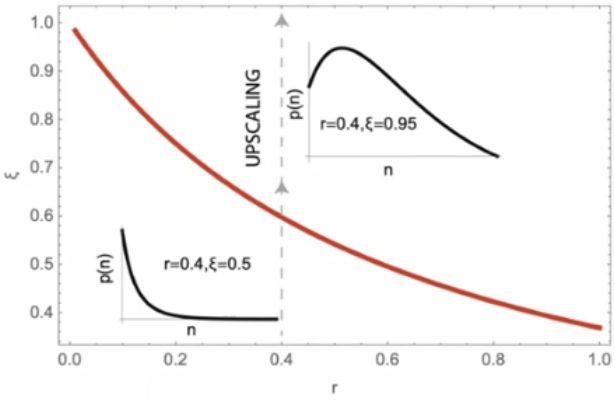


Figure (2.4) – The negative binomial can reconstruct the plots in (2.3). In particular, for any fixed value of parameter r, there is a critical value $\xi^*(r)$ (red curve) at which the plot *changes shape* from a monotonic decreasing curve ($\xi < \xi^*(r)$, left small plot) to a "log-normal" shape ($\xi > \xi^*(r)$, right small plot).

[2]∧This is thanks to its self-similarity property, as we will show later on.

Is there a way to *relate* r, ξ measured at different scales?

Let's suppose we are measuring the population sizes of tree species in a forest. Let A be the **area** of the entire forest. We can divide it in patches (**quadrants**) of size $a \ll A$, occupying a fraction $p \equiv a/A$ of the total forest.
Then we measure all the S_i species living in i-th patch, and gather their populations in a vector $\boldsymbol{n}_i^p$:

$$\boldsymbol{n}_i^p = \{n_1^{(i)}, n_2^{(i)}, \dots, n_{S_i}^{(i)}\}$$

So $n_j^{(i)}$ is the number of individuals of population j that are found in the i-th quadrant of the forest. From $\boldsymbol{n}_i^p$ we can compute the distribution $P_n^*(p)$ of species sizes at scale p, from which we can plot the RSA.

To *improve* the results, and get some information about the *variability* of species in the forest, we usually measure $\boldsymbol{n}_i^p$ for several patches $i \in \mathcal{I}$. In general $\mathcal{I}$ will contain a very small set of all the possible quadrants, meaning that we have still no direct information about the total population of the forest. However, we can now do *averages*, and for example compute the average number of species $S^{(p)}$ in quadrants at scale p:

$$S^{(p)} = \frac{1}{M}\sum_{i=1}^{M} S_i^{(p)}$$

where $M = |\mathcal{I}|$ is the number of measured quadrants.

Given this available information, we are searching for a way to construct the RSA at the *global* level, i.e. finding $P_n^*(1)$.

The scaling problem

$$\text{Measure } \{\boldsymbol{n}_i^p\}, \text{ i=1,...,M} \Rightarrow P_n^*(p) \xrightarrow{?} P_n^*(1)$$

Let's focus on a single species, and reverse the problem. Suppose we know that a species has n individuals in total (i.e. at scale 1). What is the probability $P_p(k|n,1)$ that in a random quadrant at scale p we will find only $k \leq n$ individuals?

Sampling assumption

Assuming that all species are **well mixed**, all quadrants have the same statistics, meaning that all species are equally distributed over the entire forest (they do not "cluster" in specific quadrants). In other words, we can imagine this situation as if we had n individuals and placed them at random in the forest, as if launching *darts* on a *grid* target. The probability of an individual being placed in a *specific* quadrant is p, i.e. the fraction of area occupied by that quadrant. The probability of this happening k times over n trials is given by the **Bernoulli distribution**:

$$P_p^*(k|n,1) = \binom{n}{k} p^k (1-p) \tag{2.7}$$

The probability of a species having n individuals in total is $P_n^*(r,\xi)$, which is the distribution we want to find. If we do not know n, the probability $P_p(k)$ of finding k individuals of a species in a quadrant is obtained by marginalization:

$$P_p^*(k) = \sum_{n=1}^{\infty} P_p(k|n,1) P_n^*(r,\xi) \underset{(2.7)}{=} \sum_{n=1}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} P_n^*(r,\xi)$$

Since $P_n^*(r,\xi)$ is a **negative binomial** distribution, this infinite sum with binomial weights leads back to a negative binomial!
Intuitively, this is because we are *filtering*[3] the product of the negative binomial P_n^* with an additional binomial trial.
So we have:

$$P_p^*(k) \propto \binom{k+\tilde{r}_p-1}{k} \tilde{\xi}_p^k = c\mathrm{NB}(\tilde{r}_p, \tilde{\xi}_p) \tag{2.8}$$

Negative Binomial self-similarity

where c is a normalization constant, NB denotes a negative binomial, and:

$$\tilde{r}_p = r \qquad \tilde{\xi}_p = \frac{p\,\xi}{1-\xi(1-p)} \tag{2.9}$$

***Scaling** relations*

This is the so-called **self-similarity property** of the Negative Binomial distribution.

The parameters $\tilde{r}_p$ and $\tilde{\xi}_p$ can be computed by fitting the RSA at the scale p, which we can measure. Then, applying (2.9), we can obtain r, ξ, and so an estimate of the global distribution $P_n^*(r,\xi)$ at scale 1.

Moreover, we can estimate also the *total number of species* S at scale 1. The idea is that $P_p^*(0)$ is the probability of a species *not appearing* in the sample, which is just the number of species missing in the sample normalized by the total number of species S (since P_p^* is normalized):

$$P_p^*(0) = \frac{S - S^{(p)}}{S} = 1 - \frac{S^{(p)}}{S}$$

Rearranging we get:

$$S = \frac{S^{(p)}}{1-P_p^*(0)} \underset{(2.8)}{=} S^{(p)} \frac{1-(1-\xi)^r}{1-(1-\tilde{\xi}_p)^r} \tag{2.10}$$

Since $S^{(p)}$ is known from data and ξ has been estimated from scaling, we can compute also S.

2.3.1 Results

Let's see some applications of the above techniques [2], and how they compare with other methods for *upscaling* local information to estimates for the entire population.

[3] $\wedge$Consider a trial with success probability ξ. You repeat it until r *failures* happen, and count all the successes that you get. The distribution of these counts is exactly the Negative Binomial $\mathrm{NB}(r,\xi)$. Suppose now that, after each success, you try a second trial, this time with probability p of success. You keep the original success if also this second trial succeeds, and otherwise you discard it. This is *binomial filtering*. It turns out that you can *integrate* the second trial in the first, by just using instead of ξ a different (lower) probability $\tilde{\xi}$.

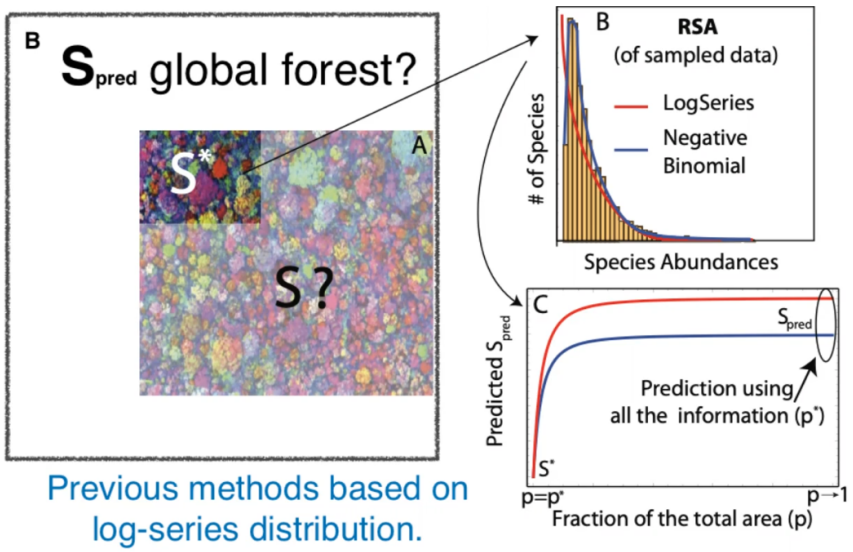


Figure (2.5) – A small quadrant of a forest is measured, finding S^* species. Its RSA is plotted, and fitted with a Negative Binomial (blue curve) to get r and $\tilde{\xi}_p$. Then $\tilde{\xi}_p$ can be used to find ξ at higher scales (and in particular at scale 1), and from that S can be computed through (2.10). A previous method based on a LogSeries fit (red line) is considered for comparison.

Clearly the new method leads to different predictions, but without measuring the entire forest we can not be sure about their quality. One way to test for that is to consider a measured sample as *the entire population*, then divide it in (sub)quadrants, select few of them, and try to reconstruct the original information only from them. In this way, we can compare the upscaling with the ground truth:

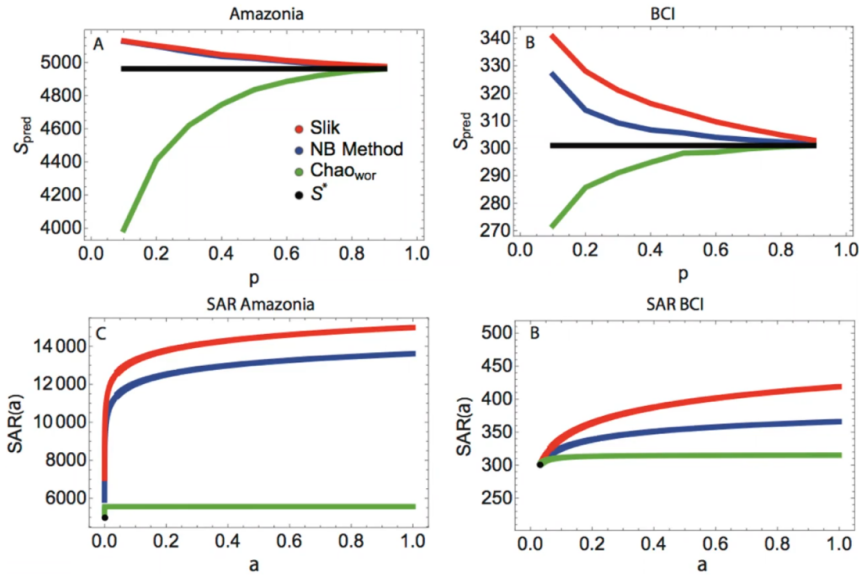


Figure (2.6) – Upscaling from a subsample to the original sample (top plots), and from a sample to the entire population (bottom plots). Three different methods are compared: the Negative Binomial presented above (blue line), Slik (red) and Chao (green). The black line represents the ground truth, when it is known.

This framework can be used also to **propose** new measurements, i.e. suggest if

the current samples are "good enough" for **reliably** upscaling to the population, or more data is needed.
The idea is to use **bootstrap**:

1. We start with a sample with S^* species at scale p^*, and *upscale* it to the entire population $p = 1$, where the number of species is S.

2. Using binomial sampling, we can *downscale* S to the p^* scale again. Since randomness is involved, we will get a new sample that is different from the starting one, with a population $\tilde{S}$.

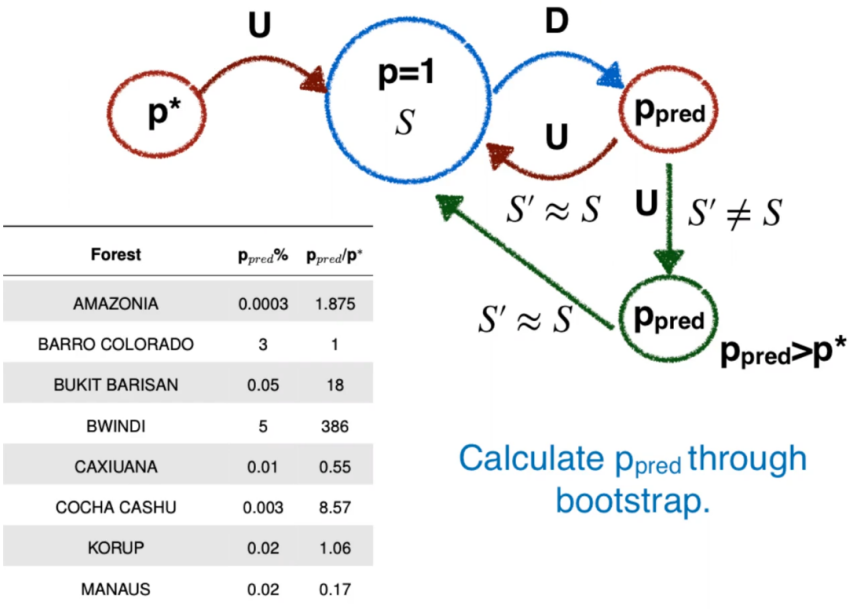
3. Now we can *upscale* again the new sample to full scale $p = 1$, getting a new estimate for the population's size S'.

4. If the sample size p^* encoded "sufficient information", we expect S' to be close to the original estimate S, i.e. $S \approx S'$ (clearly this must be measured on *average* over many trials, mapping the distribution of deviations).

   However, if S' and S are significantly different, then p^* was not a sufficient scale to reliably reconstruct the population. So, we can select a new scale p_2^*, and restart from step 2, this time *downscaling* to p_2^* instead of p^*. Then we repeat the upscaling and the comparison. If the estimates are still not close enough, we will consider an even bigger p_3^*, and so on, until we find the p_{pred} that "encodes sufficient information" for our needs.

   Similarly, if the original p^* was "good enough", we could consider a *lower* p_2^*, and repeat the above until we found the *minimum* p_{pred} needed for our purposes.

The final p_{pred} is the needed scale for a local measurement to *reliably* lead to good estimates for the entire populations (up to some arbitrary precision). This means that we need data *at least* at p_{pred} scale to make good inferences.
Finally, we compute the ratio p_{pred}/p^*, where p^* is the scale of available data. If it is > 1, this means that *we need more data* ($p_{\text{pred}} > p^*$). If it is ≈ 1, then the available data is sufficient. Otherwise, if it is < 1, it means that not only the available data is sufficient, but that it is *too much*, and we wasted resources that could have been used to map other forests instead.



Forest	p_{pred}%	p_{pred}/p^*
AMAZONIA	0.0003	1.875
BARRO COLORADO	3	1
BUKIT BARISAN	0.05	18
BWINDI	5	386
CAXIUANA	0.01	0.55
COCHA CASHU	0.003	8.57
KORUP	0.02	1.06
MANAUS	0.02	0.17

Figure (2.7) – Diagram of the bootstrap method for computing the minimum scale p_{pred} needed to reliably reconstruct global population S. A table comparing p_{pred} with the p^* of available data is reported below.

2.3.2 Binary data

Most of the time, however, we do not have abundances of an ecosystem, since they are costly to measure. It is easier to measure the presence/absence of species (binary data) inside a quadrant: *(Hint for project)*

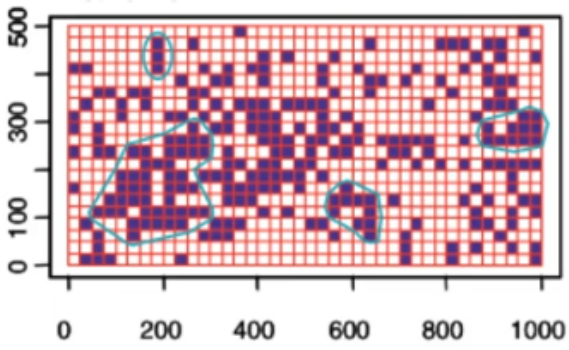


Figure (2.8) – Grid representing a forest, where a species is found only in the colored squares, but not in the white ones. The *abundance* of individuals of that species inside each square is not known.

In this case, we cannot reconstruct the RSA. However, we can plot instead of the *abundancy*, the *area occupied*, leading to the **Relative Species Occupancy** plot:

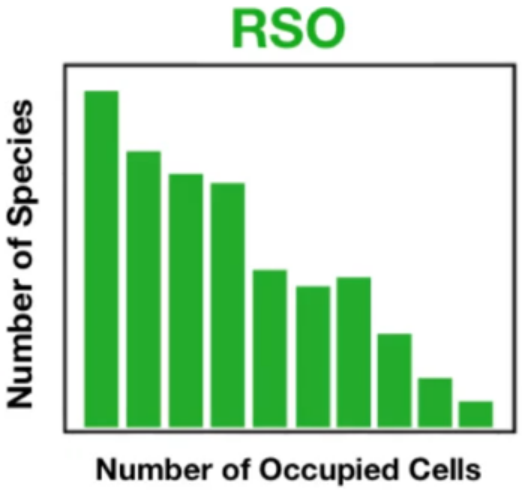


Figure (2.9) – Plot of the count $y(a)$ of species that occupy a cells in the grid. This is similar to RSA, but instead of plotting the absolute *number of individuals*, we are considering the *area inhabited* by a species (the two are correlated, but not the same).

As before, the RSO is scale dependent. Then, given a local measurement, can we infer global properties by using methods similar to the previous ones?

Let's denote with A the full area, and with a^* that of the measured sample. We can sub-divide a^* in smaller quadrants a, measure the number of species $S_a^{(i)}$ inside each of them and average:

$$\langle S_a \rangle = \frac{1}{M} \sum_{i=1}^{M} S_a^{(i)}$$

If we do this for different values of a, we can plot $\langle S_a \rangle$ for $a < a^*$, reconstructing the **Species Accumulation Curve** (SAC):

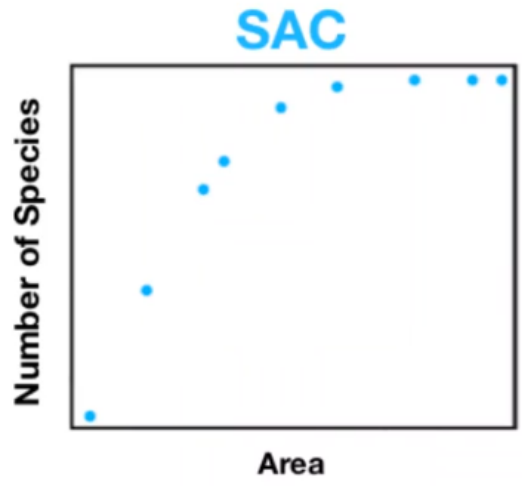


Figure (2.10) – Species Accumulation Curve

From the negative binomial model we can construct an analytical form for the SAC [3], that can be used for fitting. Then, from the results we can predict an RSO that can be compared with the one from real data.

2.3.3 Criticality

We can define the **coefficient of variation** as the relative fluctuation of population around the mean:

$$\sigma = \frac{\sqrt{\langle (n - \langle n \rangle)^2 \rangle}}{\langle n \rangle}$$

It can be shown that σ diverges (in the thermodynamic limit) as $\xi \to 1$ and $r \to 0$, and so this is a *critical point* for the system. Surprisingly, from fits of real data we find $r \approx 0$ and $\xi \approx 1$:

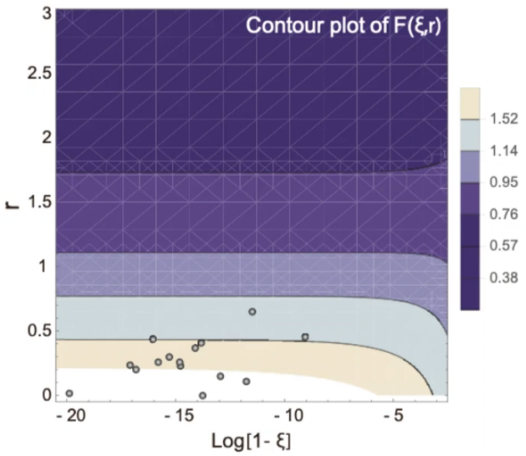


Figure (2.11) – Values of ξ and r in real data are very close to the critical point.

This means that, indeed, stochastic fluctuations *dominate* the behavior of real forests.

2.3.4 Phenomenological Renormalization Group

Bibliography

[1] Davin Lunz. "On continuum approximations of continuous-time discrete-state stochastic processes of large system size". working paper or preprint. May 2020. URL: https://hal.inria.fr/hal-02560743.

[2] Anna Tovo et al. "Upscaling species richness and abundances in tropical forests". *Science Advances* 3.10 (2017). DOI: 10.1126/sciadv.1701438. eprint: https://advances.sciencemag.org/content/3/10/e1701438.full.pdf. URL: https://advances.sciencemag.org/content/3/10/e1701438.

[3] Anna Tovo et al. "Inferring macro-ecological patterns from local species' occurrences". *bioRxiv* (2018). DOI: 10.1101/387456. eprint: https://www.biorxiv.org/content/early/2018/08/08/387456.full.pdf. URL: https://www.biorxiv.org/content/early/2018/08/08/387456.

Analytical index