STATISTICAL ANALYSIS OF HIGHLY CORRELATED SYSTEMS IN BIOLOGY AND PHYSICS

BY

HÉCTOR GARCíA MARTíN

Licenciado en Ciencias Fisicas,
University of the Basque Country UPV/EHU, 1999

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2004

Urbana, Illinois

STATISTICAL ANALYSIS OF HIGHLY CORRELATED SYSTEMS IN BIOLOGY AND PHYSICS

Héctor García Martín, Ph.D.
Department of Physics
University of Illinois at Urbana-Champaign, 2004
Nigel Goldenfeld, Advisor

In this dissertation, I review my work on the statistical study of highly correlated systems in three fields of science: Ecology, microbial Ecology and Physics.

Regarding Ecology, I propose an explanation for how the highly correlated distribution of species individuals, and an abundance distribution commonly observed in ecological systems, give rise to a power law dependance between a given area and the number of unique species it harbors. This is one of the oldest known ecological patterns: the power-law Species Area Rule.

As a natural extension of my studies in Ecology, I have undertaken research in a developing field: microbial Ecology. In particular, I have formed part of a multidisciplinar team with the goal of studying whether microbes can affect the formation of macroscopic structures; specifically, the calcium carbonate terraces at Yellowstone National Park Hot Springs. I have used ecological techniques to characterize the biodiversity of our study site and developed a new bootstrap method for extracting abundance information out of clone libraries. This has singled out the most abundant microorganisms and suggested a non-passive role of microorganisms in carbonate precipitation.

Simultaneously, convinced that many of the tools commonly used in Physics may have future applications outside this field, I have undertaken research in a topic of current interest in Physics, dealing with highly correlated systems: rotating Bose-Einstein condensates.

I have used finite difference techniques to solve the Gross-Pitaevskii equation to obtain the structure of a vortex in a lattice. Surprisingly, I have found that, in order to understand this structure, it is necessary to add a correction to the Gross-Pitaevskii equation which introduces a dependance on the particle scattering length.

I have also used Path Integral Monte Carlo techniques to go beyond the Gross-Pitaevskii equation to study these vortices. Interestingly, the Gross-Pitaevskii equation seems to be valid for much higher densities than expected if properly renormalized. This explains its validity for studying dense systems such as superfluid helium.

The reasonable man adapts himself to the world;

the unreasonable one persists in trying to adapt the world to himself.

Therefore, all progress depends on the unreasonable man.

George Bernard Shaw

Acknowledgments

Four years have elapsed since I joined Nigel's group and it is hard to believe how much has happened. A year ago, I was afraid my thesis title would end up being "Four unfinished projects in Biology and Physics"; but here I am, enjoying a much different outcome. This would not have happened without the help of many people in Urbana-Champaign and elsewhere:

First and foremost my advisor, Prof. Nigel Goldenfeld, who has taught me not that Science can be fun, but that it *must* be fun. A trustworthy man, gifted with the ability to separate the important details of a problem and disregard the irrelevant ones, he has transformed me into a person capable of independent and effective research. This is a skill that promises to provide me entertainment for the rest of my life. For this, I feel deeply indebted. I will probably name my first two or three kids after him[1].

The people in Nigel's group have proved to be an interesting bunch, always willing to discuss interesting problems, relevant to their research or not. I would like especially to thank John Veysey for being a remarkable person and staying like that, and Kalin Vetsigian, for being my roommate the last two years and having more patience with me that I have ever witnessed in a person. My thanks too to Vivek Aji, Tae Kim, Patrick Chan and Nick Gutenberg.

I would also like to thank Prof. Bruce Fouke for his endless enthusiasm in his (and our) research and his permanently welcoming stance, George Bonheyo for helping us out immerse in Microbiology and keep a straight face in front of our..., let's say naive

[1] OK, this last one is a joke

questions. My thanks too to the rest of the Fouke group: Tom Schickel, Jorge Frias Lopez, Jim Kraus and Kelly Zimmerman.

Prof. Ceperley provided the Path Integral Monte Carlo (UPI) code which I used for the last part of my BEC work and I would like to thank him for that, and for always having an open door to my questions, intelligent or not. Greg Bauer was fundamental in setting up the UPI code for the first time and he can not imagine now how grateful I am. Bryan Clark has always been kind enough to share his knowledge of UPI. Ken Esler has also been really helpful with my computer problems.

I would like to thank Prof. Baym for the insightful conversations that I had with him regarding BECs. I would also like to thank Drew Gifford, who filled my ignorance gaps in this topic.

My acknowledgements to Swagatam Mukhopadhyay for being an honorable friend and Michelle Nahas for being as sweet as she is and the best example I know of a random stochastic process.

Greetings to all the Physics crowd including Bojan Tunguz, Kapil Rajaraman, Matt Gordon, Zigurts Majumdar, Dyutiman Das, Dean Eckhoff, Eun-ah Kim and Mike Lawler.

Special thanks to Uzuri Esteban Zarandona for I would not be who I am if it wasn't for her and thanks too to Wiebke Benthien for proving that people like her can exist.

I am indebted to Diana Arbaiza and Irune Del Rio for letting me stay at their places when much of this thesis was written. Greetings to Susana Vidal, Ana Vivancos, Antonio Rueda, Silvia, Jordi and the rest of the Spanish department.

My acknowledgements to Xavier Llora, Xavier Cartoixa, Javier Bareno, Neus, Marc, Diego and the rest of the Uiberia bunch.

Finally, my gratitude to all the people at "The Point" fencing club for starting me in a life-long vice: fencing. Thanks Rebecca, Mike and Dan.

Contents

Chapter 1

Introduction

The fact that the individual behavior of system components is strongly correlated leads to interesting consequences in both Physics and Biology. In Physics, this produces emergent properties such as superfluidity and the quantum hall effects. In Biology, it leads to self-organization patterns such as hierarchical social organization in ants or multicellular microbial communities.

This thesis focuses on two topics of great current interest: the statistical properties of macroscopic and microscopic ecological systems and the vortex structure in rotating Bose-Einstein condensates (BECs).

While it may be initially surprising that a theoretical physicist undertakes research in ecology, this topic has evolved naturally from the current physics research agenda, particularly the study of Complexity [1].

But, what is Complexity? The approach of science has for the most part been traditionally reductionistic. The underlaying idea is that Nature is governed by a small set of fundamental laws out of which all desired detailed knowledge could be derived: understand each of the parts of your system and you will understand the system. Centuries of reductionism have yielded successful and useful science and a wealth of detailed information. But the limits of reductionism are becoming increasingly apparent [2],[3].

With the advent of bioinformatics, this balance is beginning to change due to the widespread realization about the need to synthesise and integrate biological information.

One part of this is because of information overload. Molecular biology, for example, has generated enormous quantities of information on cell processes but has spent a comparatively little time integrating this knowledge [4]. The other part comes from the fact that understanding a system from the sum of its parts is not straightforward or sometimes even possible.

Take for example how, in spite of knowing the detailed behaviour of each transistor, wire and gate array with pinpoint accuracy, it becomes increasingly hard for engineers to design microprocessors. Chip designers need to rely on sophisticated modeling programs to be able to predict the collective behavior of their designs and prevent bugs in advance [4]. Another example is the genome project: considerable time and funding has been devoted to knowing every single nucleotide of human DNA (and others, too). We know the genes; and the function of many of them. Yet, prediction of the behaviour of gene networks is out of reach. Gene expression is governed not so much by the sheer amount of genes and their behavior, but rather by the combination and interactions between them. Otherwise, it would be impossible to understand the difference in sophistication between human beings and the tiny (959 cells) roundworm *C. elegans*: the number of genes for the former is less than double that for the latter [5].

Complexity studies precisely those kind of systems, which cannot be understood by simply adding the contributions of its parts. As such, their behaviour is not merely an extrapolation of the properties of underlying components. As the scale increases, entirely new properties appear and new laws and concepts are necessary [6].

The study of biocomplexity is particularly appealing to physicists because of the ubiquitous appearance of scaling laws of the type $Y \sim X^{\gamma}$. In Physics, scaling laws near phase transitions have been intensely studied because of their *universality*: very different systems can exhibit the same behaviour when in critical regimes [7]. Scaling laws in biological contexts seem to be more pervasive and hold for more decades than in Physics. Some examples are:

- Allometric laws: the metabolic rate of animals (e.g. number of heartbeats per second) scales with mass as $B \sim M^{3/4}$ for 20 decades [8] (but even this is not incontrovertible: despite scaling over so many decades, the value of the exponent is controversial [9],[10]).

- Cell scaling: the mean M and variance V in organ cell number of animals, plants, molds and algae show a power very close to $V \sim M^2$ for around 12 decades [11].

- Species Area Relationship: the number of different species S found in an area A scales as $S \sim A^z$ from hectares to continents [12].

General explanations have been proposed, such as Self Organized Criticality [13],[14] or Highly Optimized Tolerance [15] for the widespread appearance of scaling laws which, require the use of sophisticated mathematical machinery of the type commonly used by theoretical physicists [7].

These generic principles typically offer mechanisms for the occurrence of power laws on an abstract, non-specific level, which are rather unsatisfactory for biologists in general and even some physicists (especially our own group!). It's only recently that a large community physicists are learning to roll up their sleeves and get into the nitty gritty details. This is exactly what this thesis is about.

The first part of this thesis deals with the third example given above: the fractal species area rule. In particular, an explanation for its emergence and robustness will be proposed in chapter 4 and compared with real ecological data.

The second part presents work related to a project that studies how microbial ecosystems may be able to influence the formation of travertine terraces at geothermal carbonate springs (fig. 5.1), seven orders of magnitude larger in size. This is a multidisciplinary project involving geologists, microbiologists and physicists, who are trying to understand the formation of these terraces from a holistic approach. My main task has been so far to participate in field work and to create new ways to analyse data that quantify the organisms present. These analyses have proven very effective in characterizing the microbial ecosystem, and pave the way for the future studies on the impact of metabolic activity on biogeochemical cycles related to travertine precipitation.

The final part of this thesis presents a more conventional topic requiring statistical sampling: the study of rotating Bose Einstein condensates through Path Integral Monte Carlo simulations. Besides being a topic of great contemporary theoretical and experiment relevance, I am interested in how techniques derived from many-body physics can be used to study ecological and biological systems. After all, many-body physics has been dealing successfully with highly interacting systems for a long time. Specifically, Path Integral Monte Carlo techniques are rather promising, since they involve no uncontrolled approximations in dealing with interactions. Theoretically, the Gross-Pitaevskii equation is an excellent example of how to minimally include interactions in spatially-explicit models. My work shows that this equation holds for a greater range of densities than expected if renormalized properly. This is an interesting result in itself from a Physics point of view; and this work has provided me with the necessary familiarity with these methods for possible future use in biological contexts. For example, the statistics of ecological systems can be formulated in field theory terms [16], which might be addressable using PIMC.

1.1 Thesis outline

This thesis is roughly divided in three parts, corresponding to each of the different projects.

Chapters 2, 3 and 4 relate to the study of the fractal Species Area Relationship (SAR): chapter 2 presents a self-contained introduction to ecology for a Physics audience, chapter 3 explains the well-known model of self-similarity in ecology by Harte et al. [17] and my work on it, chapter 4 offers an explanation to the emergence and robustness of the fractal SAR in ecology along with ecological data to support it.

Chapter 5 contains an introduction to microbiology and work related to estimation of biodiversity while chapter 6 presents the results regarding microbial abundance.

Chapter 7 is an introduction to rotating Bose Einstein condensates that includes my own work regarding vortex structure and chapter 8 introduces the Path Integral Monte Carlo Method along with the results on the validity of the Gross-Pitaevskii equation.

The final chapter, 9, presents the conclusions.

Chapter 2

Introduction to ecology

In its broadest sense, ecology is defined as the“scientific study of the interactions that determine the distribution and abundance of organisms” [18]. The breadth of this definition leaves room for a large research field encompassing several subfields focused on the different levels of biological organization (fig. 2.1). Population biology, for instance, concentrates on interactions between a population and its environment. Community ecology studies populations of all species in an area at a particular time. Ecosystem ecology includes all the interactions between biotic and abiotic components of a system. Physiological ecology focuses on the adaptation of bodily processes to the physical environment, and genetic ecology analyzes the way an organism’s ecology affects its heredity [19]. A wide ranging science in scope then, my intention here is to provide an introduction to the specific parts to which my work is related. As will become apparent, the interactions between organisms and their environment, and the role of spatial degrees of freedom, make ecology a living example of pattern formation and statistical scaling in spatially-extended dynamical systems.

I will start with one of the main objects of study in ecology: the distribution of species abundances. This will be important to understand the emergence of power-law Species Area Relationships, as will be explained in section 4. Next, I will introduce the concept of cover (or occurrence) and its relation to species abundances, which

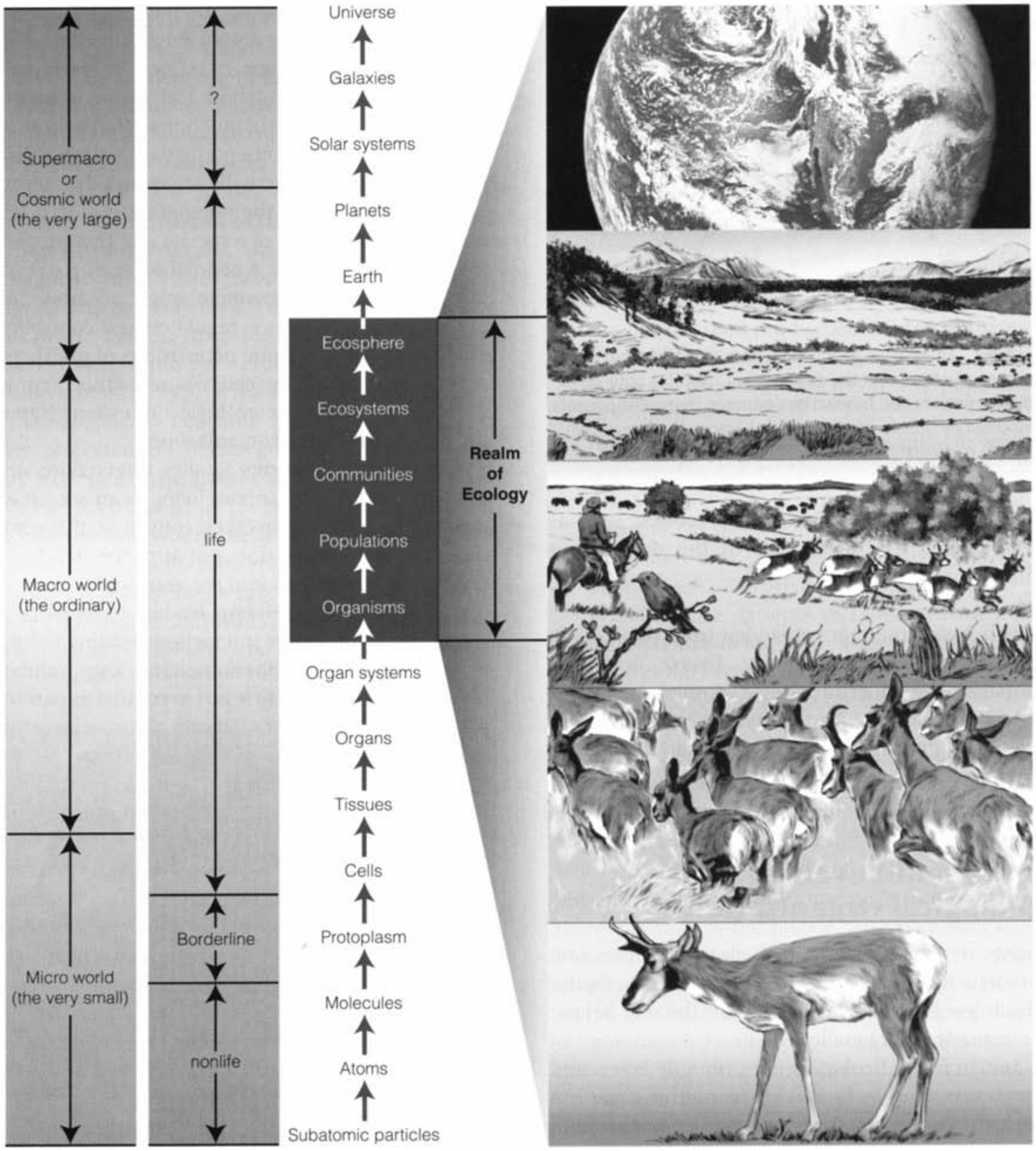


Figure 2.1: Ecology in context: different levels of self-organization of biological systems. The traditional realm of ecology has been to study biology at the level of populations, communities, ecosystems and the whole ecosphere. Populations are defined as groups of individuals of the same species which inhabit the same location. Communities are the populations of all species living in an area at a particular time while ecosystems comprehend communities plus the physical environment they are set in. Figure from ref. [19].

will be key in sections 4 and 6. The following section describes the Species Area Rule as an example of a well-known scaling law in ecology and one of the main topics of this thesis. Then, after a presentation of neutral models of ecology that have created a heated discussion in the field, I will conclude with an account of what I consider to be the key questions in ecology, from the physicist's perspective.

2.1 Abundance distributions

The species abundance distribution $P(n)$ is the fraction of species with n individuals and is one of the most intensely studied ecosystem characteristics in ecology [20],[21],[22],[23],[24]. Knowing how many species are abundant and how many are scarce is important not only to accurately describe an ecosystem, but also for practical applications such as biodiversity conservation and reserve design [12],[25].

Regarding abundance distributions the main questions we would like answered are: what are their functional form for a typical ecosystem? and what determines this form: how specifically must the details of the ecosystem be known in order to predict $P_(n)$ to a given level of accuracy?

The answer to the first question is that $P_(n)$ varies from ecosystem to ecosystem. Several distributions [26],[21] have been proposed and successfully fitted to data although Preston's lognormal [27] remains the most commonly reported. The following subsections will present this distribution along with others of historical and practical importance.

Regarding the second question, unfortunately, there is no clear consensus on what determines quantitatively the shape of the abundance distribution. Several models considering the internal dynamics of ecosystems have been presented. Most are extremely simplified pictures of the dynamics. Their focus is on how much information can be obtained from very general basic principles of ecosystem function rather than

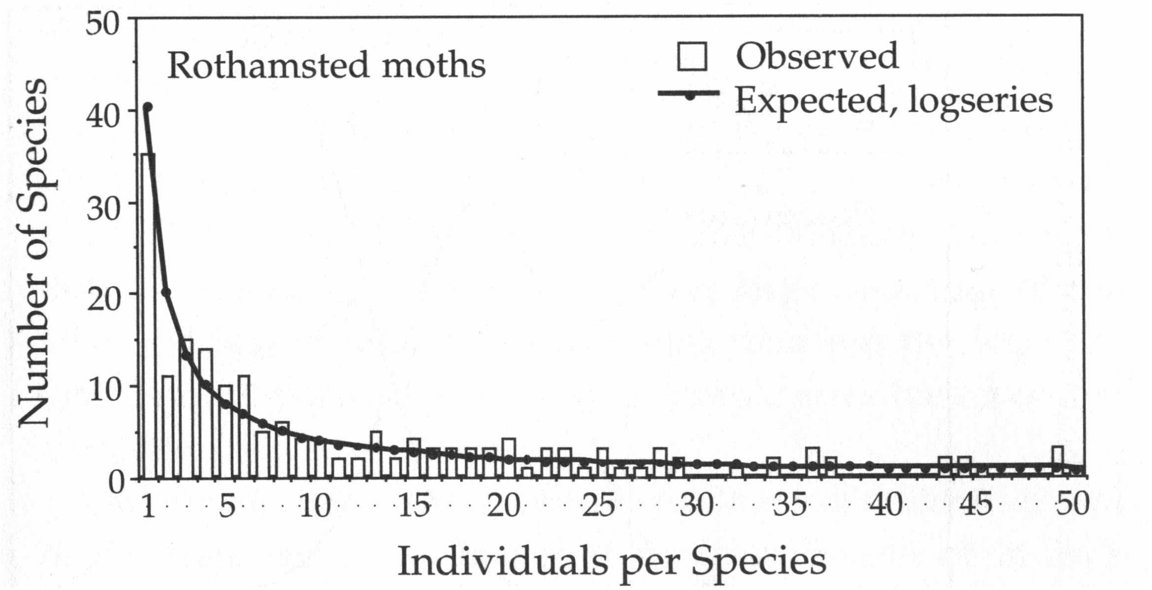


Figure 2.2: Fisher's logseries fit to data on species abundance in collections of moths at a light trap over a 4-year period at Rothamsted Field Station, U. K. Notice the lack of a mode, in contrast with the best documented abundance distribution function, the log normal (fig. 2.19). Taken from ref [20].

accurate descriptions. An example of this is MacArthur's broken stick model [28],[29], explained below in section 2.1.3. Recently though, more realistic models have been proposed, taking into account the dynamics of extinction, speciation and competition [20],[30],[31]. Among them, Hubbell's neutral model, which assumes all organisms in the community to be identical, has caused the greatest impact and controversy [32] in the field. This model will be described in section 2.5.

2.1.1 Fisher's logseries

One of the first published abundance distributions is due to Fisher, Corbet and Williams [33],[34]. Corbet and Williams had separately collected random samples of butterflies and moths respectively and took their data to the famous statistician Ronald Fisher for analysis. Noticing that the number of singletons (one individual), doubletons (two individuals), tripletons, .., etc decreased monotonically, he fit the data to a smooth hyperbolic progression which he called the logarithmic series. In ensembles obeying this distribution, the number of species $S(n)$ with n individuals is given by:

$$S(n) = \frac{\alpha x^n}{n} \tag{2.1}$$

where x $(0 < x < 1)$ is a positive constant that depends on the size of the sample [12]. The parameter α, known as Fisher's alpha, characterizes the diversity. By summing over all n one obtains the total species richness, S_T:

$$S_T = -\alpha \log(1 - x) \tag{2.2}$$

This model has been fit to several ensembles rich in species, particularly insects [35],[36]. Figure 2.2 shows a collection of moths fit to this distribution.

2.1.2 Preston's lognormal

A few years later, Preston [27] proposed a different abundance distribution based on his own bird data set. His data showed a mode that was not present in Fisher's logseries. To account for that he proposed his celebrated lognormal distribution. Abundance was binned in doubling categories called octaves (1,2,3-4,5-8,9-16,17-32,..etc). Species with abundance in one of the edges (1,2,4,8,16,32..) were divided equally between the adjacent categories. Under this binning, the abundance for each octave $R = 0, 1, 2, 3, 4..$ could be fit to a normal distribution:

$$S(R) = S_0 \, e^{-(R-R_0)^2/2\sigma^2} \tag{2.3}$$

where S_0 is a normalization constant.

Since the octave binning basically amounts to a $\log_2$ change of scale [26], this distribution is called the lognormal distribution. Preston also argued that the reason previous assemblages did not show a mode was due to small sample size, and predicted correctly (fig. 2.3) that as sampling continued samples well described by Fisher's logseries would become better fitted to the lognormal distribution.

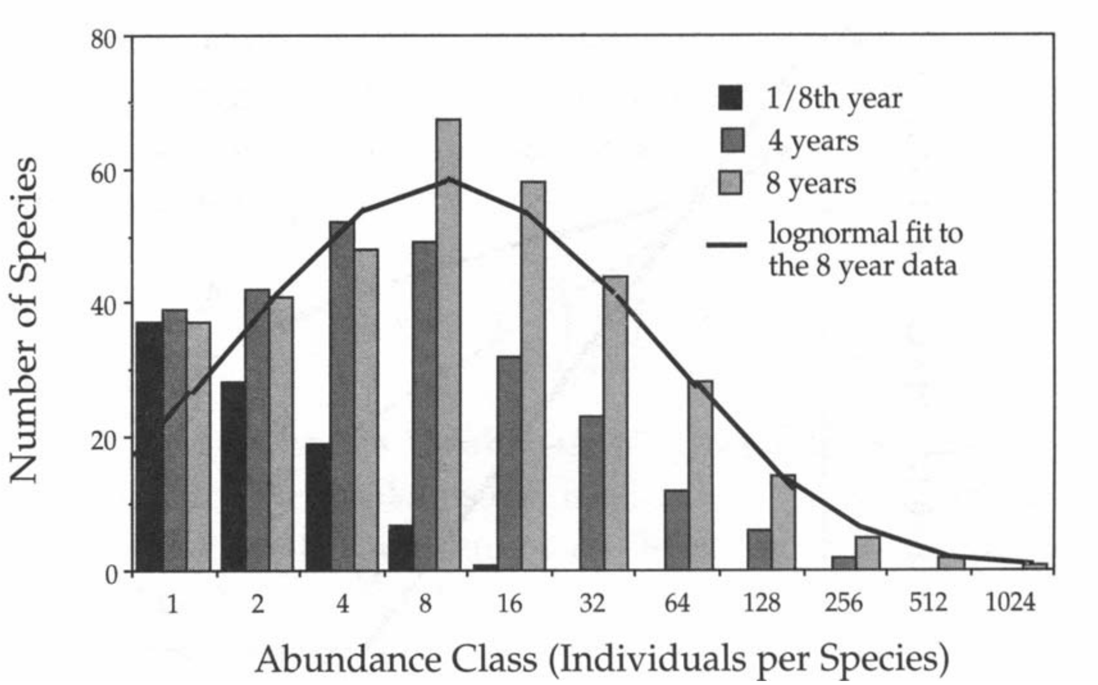


Figure 2.3: When the survey of moths at light traps at Rothamsted Field station was extended over more years, the distribution of individuals became lognormal, as predicted by Preston. Figure from ref. [20] and data from ref. [37].

The lognormal has since been found to provide a satisfactory description of a wide amount of abundance data describing birds, intertidal organisms, insects and plants [26],[27],[38],[39],[37],[40],[41],[42],[43],[41]. In fact, its apparent ubiquity led MacArthur to look for a general mechanism and to propose the niche partition model explained in the next section.

Perhaps the most important characteristic about the lognormal is that it is not just a descriptive pattern, but that it can be related to the internal ecosystem dynamics. It has been shown to be derivable from multiplicative processes and the law of large numbers [26] and from niche partitioning models [44]. It is not clear, though, how close these models are to real ecosystem dynamics.

Finally, it must be mentioned that the lognormal is not the definitive abundance distribution. Even though it fits many data sets, it does not fit them all. In particular, its main handicap is that it underestimates the amount of rare species [20],[17] (see fig. 2.4). This is particularly important for biodiversity estimation, since rare species

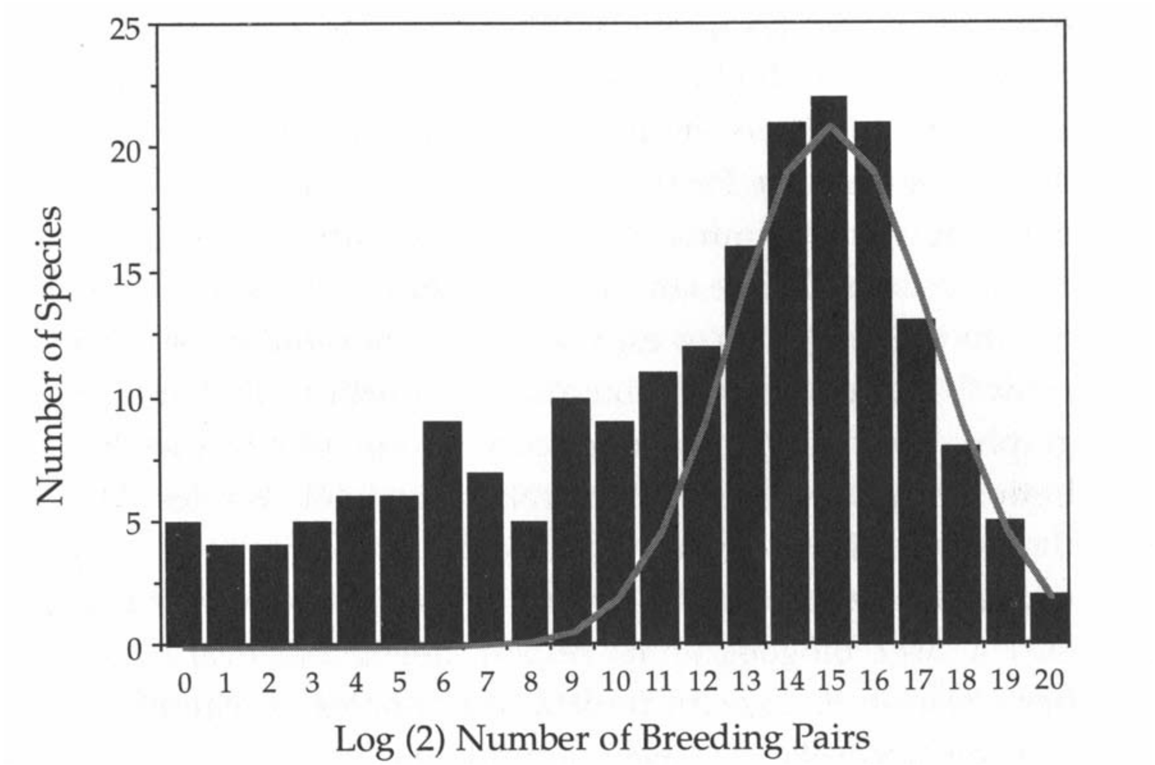


Figure 2.4: Non-lognormal distribution of relative species abundance for all British breeding birds. Notice that the lognormal (gray curve) predicts far less abundance of rare species. Figure taken form ref. [20]. Data from ref [45].

contribute in great measure.

2.1.3 Niche partitioning models

As mentioned above, the pervasiveness of the lognormal distribution led MacArthur [28], [29],[20] to propose that abundance distributions were consequence of very general principles of ecosystem functioning.

He proposed the "broken-stick" model, in which trophically similar species apportion the resources of their niche randomly among them. This is equivalent to taking a stick of given size and dividing it into S random parts, hence the whimsical nomenclature. The abundance would then be proportional to the amount of resources and the expected relative abundance of the ith least abundant species, P_i is given by:

$$P_i = \frac{1}{S}\sum_{x=i}^{S}\left(\frac{1}{x}\right) \tag{2.4}$$

which is definitely not a lognormal. Although, it has been fitted reasonably well to assemblages of closely related species of birds [29],[46],[47], snails [48] and microcrustaceans in lake-bed sediments [48],[49],[50],[51], it is not as commonly observed as the lognormal.

Enhancements have been made to this model [29], including a non-random division of resources [40],[52]. Under this modification, the most abundant species takes a fraction k of the available resources, the second most abundant takes k of the remaining $1 - k$ fraction and so on. Fits to data are available, but the consensus is that this model and its enhancements does not rival the lognormal in terms of its descriptive ability [26],[40],[43],[41].

More interestingly, a reinterpretation of this model in which the random breaks are made sequentially does yield a lognormal distribution because it describes a multiplicative cascade process [44]. In this sequential version of the model, the first break is random. Then one of the two resulting pieces is randomly chosen and broken. Next, one of the three pieces is picked and broken randomly, and so on.

While appealing, niche partitioning theories suffer from being unable to put forward a biologically-based mechanism for the division of niche resources [20]. At the same time, it is very hard to check the validity of its underlying assumptions about resource partitioning.

The neutral model explained in section 2.4, on the other hand, focuses on the population dynamics processes of birth, death, immigration and speciation. These are more accessible to experimental approaches.

2.2 Cover/Range/Occurrence

The concept of cover (or range, or occurrence as it is also known) as explained in figure 2.5 is often used to study the distribution of individuals. The cover of a species

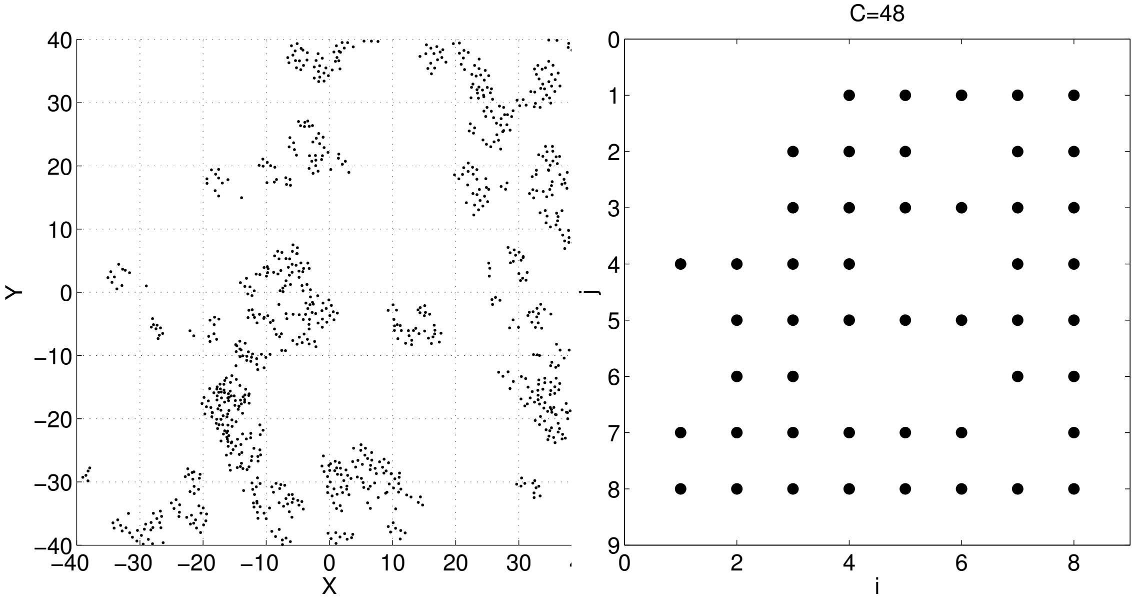


Figure 2.5: The cover, range or occurrence C of a species in a given area is determined by setting a grid of minimum size A_0 and counting the number of grid cells in which the species is present. In the case that $\sqrt{(A_0)}$ is of the order of the mean separation of individuals, the cover and the abundance are basically the same. If A_0 is bigger than the mean separation of individuals it is still possible to extrapolate the abundance from this information.

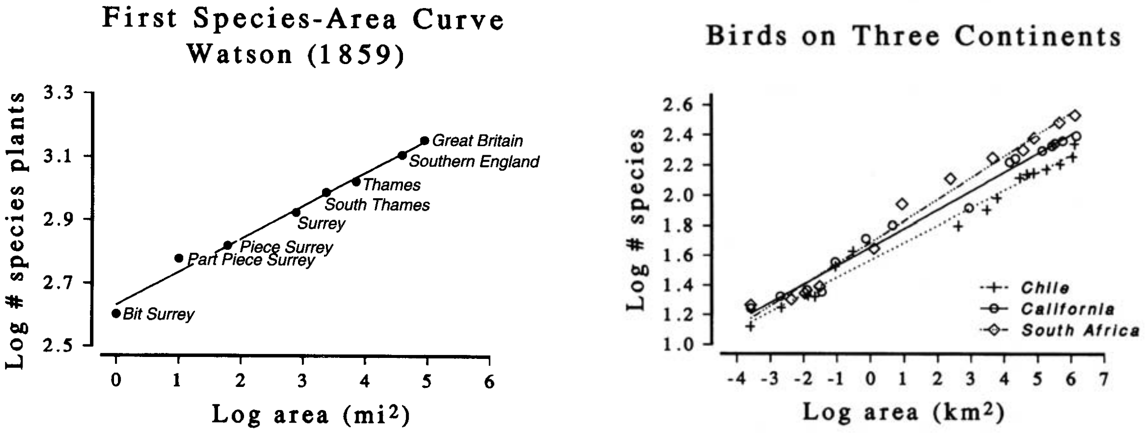
is strongly linked to its abundance. In fact, there are several methods [53],[54],[55] to derive the abundance of a species from its cover. The idea is to be able to scale down from a given cover to the minimum mapping unit and take that as the abundance. Interestingly, fractal extrapolations give overestimates and corrections to fractality must be added to get the right abundance. This means that the distribution of individuals is not fractal at all length scales, or that the fractal exponent changes with scale.

2.3 The Species Area Relationship

There are somewhere between 30 to 100 million species on earth [12]. How do we know this fact? There are definitely not enough ecologists on the planet to count them all! Our knowledge of biodiversity comes mainly from extrapolation of one of the oldest known ecological patterns: the Species Area Relationship (SAR). As its name implies, it relates an area with the number of different species it harbors.

Spatial patterns in general, regarding the location and abundance of species, are of central importance to ecology. Reliable and well tested rules of species richness, occupation and population sizes are useful tools for clarifying biologically relevant phenomena including disturbance [59], competition [60], division of niche space [44] and determination of community minimum areas.

From a more practical point of view, this knowledge is used in the design and management of reserves [18],[63],[64] and the estimation of extinction due to habitat loss. Our present knowledge of spatial patterns, for example, suggests that humanity's takeover of most of Earth's surface has provoked a mass extinction [65],[66], by the end of which only 5% of the actual biodiversity will remain. The time scale to reach this final equilibrium state is of the order of tens of thousands of years. We can, nonetheless, expect to have lost about half the plant and animal species on the planet



(a) The first instance of the SAR. This species area rule starts in Surrey county (England) and builds up to the whole Great Britain. It is the first empirical example of an ecological pattern. Taken from [12].

(b) Species Area curves for birds on three different continents. All of them have a similar value of z: Chile, 0.116; California, 0.125; South Africa, 0.143. Figure from ref. [12] after data from [56].

Figure 2.6: SARs for plants and birds.

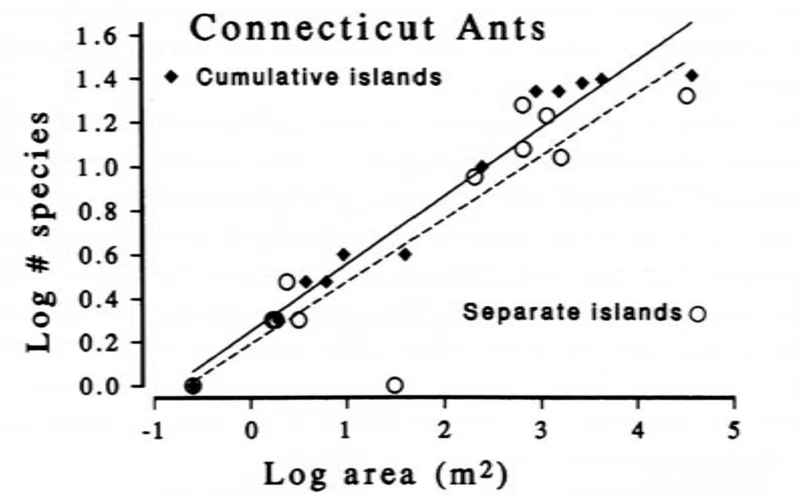


Figure 2.7: Species area curves for ants on Connecticut islands. SARs for islands are usually plotted island by island: on the vertical axis are the species found on an island of the area corresponding to the x coordinate. Plotting the curve cumulatively like for mainland patterns (including in each point all the previous species and area for the previous points), leads to a slight curvature, but the value of z does not change dramatically. Figure taken from ref. [12], assembled from data from ref. [57].

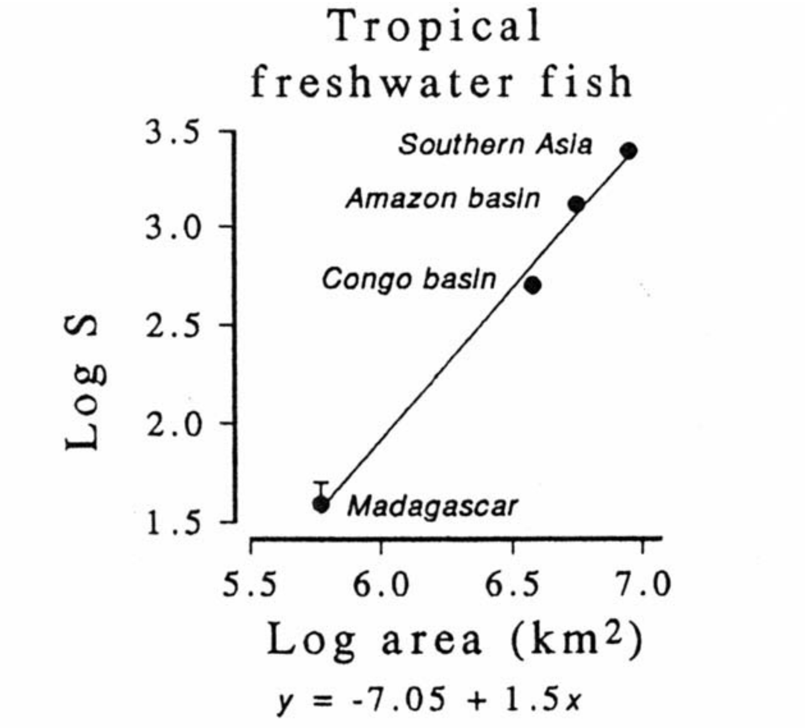


Figure 2.8: Tropical freshwater fish species plotted against area. The z value (1.5) is larger than for typical mainland plots of this scale (≈ 0.9) [58].

in a time scale of centuries. The Species Area Relationship and other spatial patterns are vital tools in understanding and ameliorating this phenomena that will certainly have an impact on human existence.

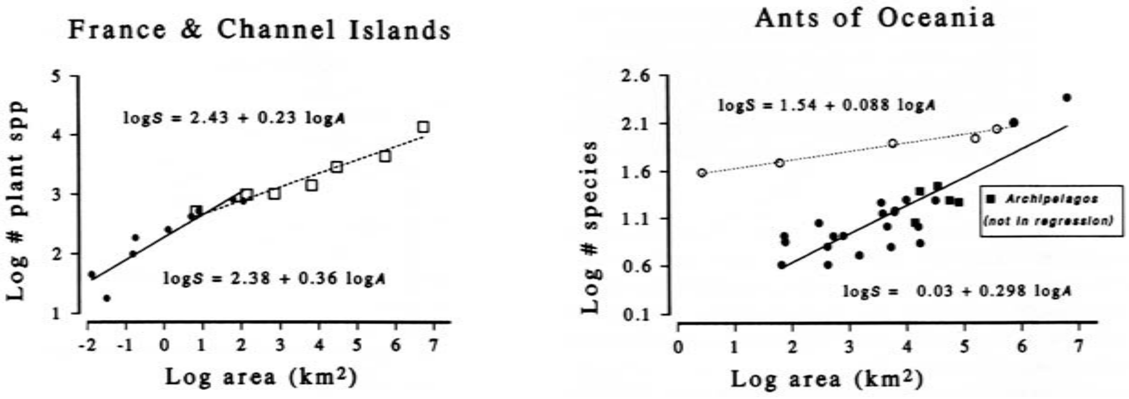
2.3.1 The fractal Species Area Relationship

Several functional relationships have been reported for the SAR (See fig. 2.14 and ref. [70] for a review) but the best documented are the exponential relationship $S = a + b\log(A)$ and, especially, the power law (fractal) dependence:

$$S = S_0 A^z \tag{2.5}$$

where the coefficient z is usually quoted to be 0.25, and S_0, a and b are empirical coefficients.

The fractality of the SAR seems remarkably universal. It is observed in such different taxa as plants, birds, insects, mammals and fish, climates ranging from



(a) SAR for plants of the channel islands compared to that of mainland France. The curve is steeper for islands. Graph from ref. [12], data from ref. [37].

(b) Ponerine and cerapachyine ant species on New Guinea, all of southeast Asia and several archipelagos. Solid dots and line represent all New Guinea, the separate islands and southeast Asia. Circles are parts of New Guinea (mainland). Dashed line represents them and whole island. Archipelagos (squares) not included in either regression line. Note again the steeper values for island SARs. Graph from ref. [12], data from ref. [61].

Figure 2.9: Differences between mainland and island SARs.

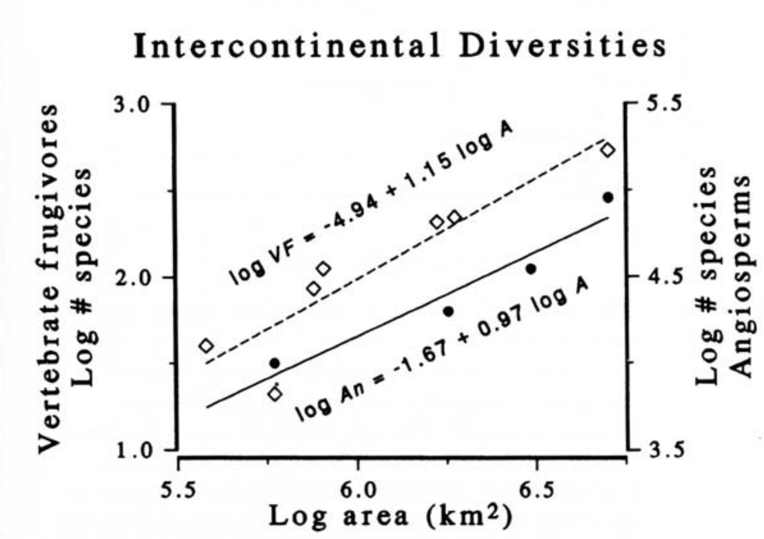


Figure 2.10: Interprovincial SAR for frugivores (fruit eaters) and angiosperms. The points include data from Africa, Malaysia, parts of Indonesia, New Guinea and the Neotropics. Frugivores include bats, birds and primates [12],[62]. Values of z close to unity are typical for interprovincial SARs, although there are exceptions (see next figure).

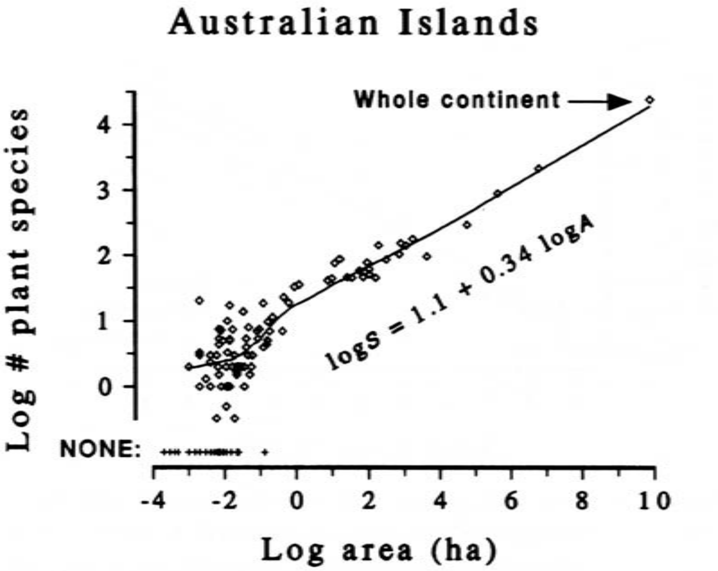


Figure 2.11: Plants species for Australian islands, including those close to Perth, Tasmania and the whole of Australia. The regression is locally weighted [12].

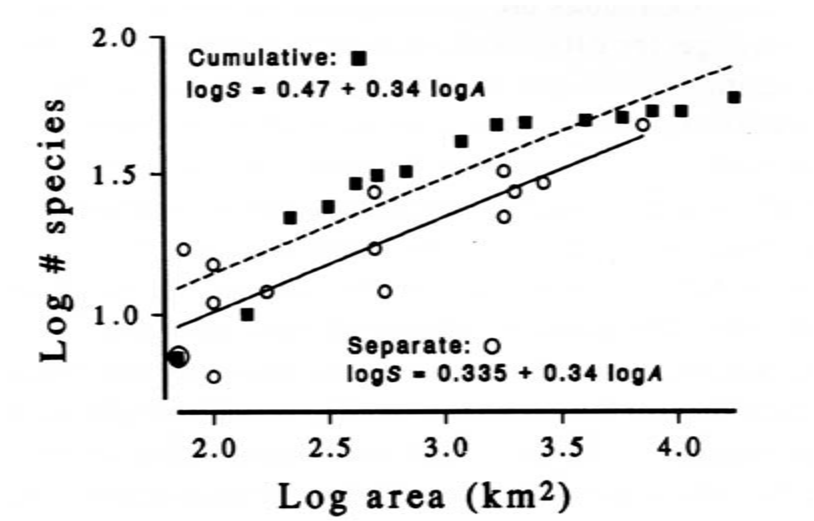
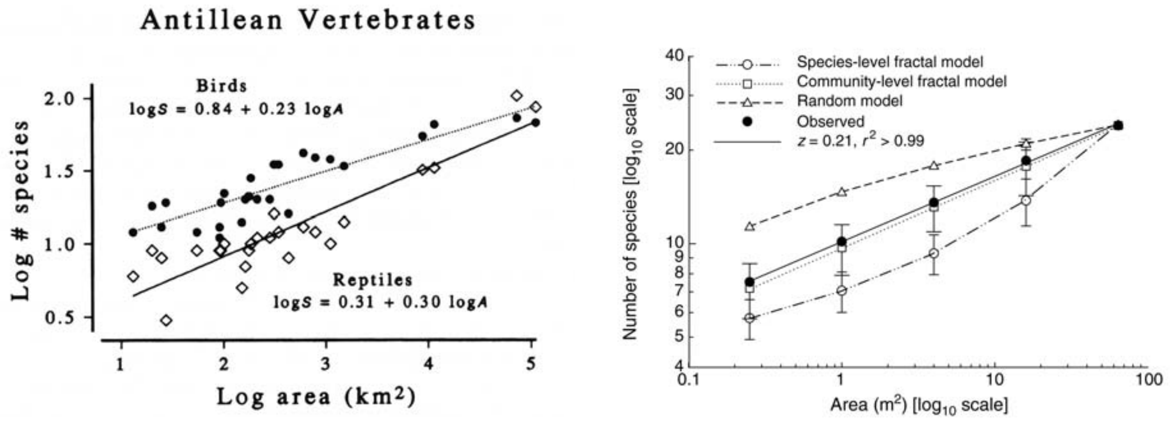


Figure 2.12: Birds of sky islands in the Northern Andes (paramos). Plotting the SAR cumulatively makes it curvilinear, but has little effect on the value of the slope. Figure taken from ref. [12], data from ref. [67].



(a) Bird (circles) and reptile (diamonds) species from the Caribbean versus area. Figure from ref.[12], data from ref. [68].

(b) Grassland plant community diversity in northern Napa and southern Lake counties, CA (120 km. north of San Francisco). Sampling site was only $64m^2$ in area. This allowed a systematic study of the site [69].

Figure 2.13: SARs for Antillean vertebrates and a grassland plant community.

Curve name	Model	Source
Power	$S = aA^b$	[71],[38]
Exponential	$S = a + b\log(A)$	[72],[73],[34]
Monod	$S = a(A/(b+A))$	[74],[75],[76]
Negative exponential	$S = a(1-\exp(-bA))$	[77],[78],[79]
Asymptotic regression	$S = a - bc^{-A}$	[80]
Rational function	$S = (a+bA)/(1+cA)$	[79]

Figure 2.14: Several functional relationships proposed for the SAR. By far, the most common and best documented is the power law, followed by the exponential [70].

temperate France to tropical (in the Antilles) and even data from the Pleistocene (figures 2.6 to 2.13). Such a widespread occurrence independent of the organism interactions seems to indicate that very general principles underlie its existence.

2.3.2 Spatial scale determines z

The value of the exponent z is not so universal. Frequently quoted as being close to $z = 1/4$, it typically varies between 0.15 and 0.4 [81]. There are, nonetheless, reported values as high as 1.5 (fig. 2.8) and as low as 0.086 [82].

The same exponent z does not apply to all scales (see fig. 2.15). In fact, it is usually acknowledged that the power law SAR is not one single pattern but actually four different patterns bundled into one because of their same functional relationship [12],[20]:

- SARs for small pieces of single biotas.
- SARs for larger pieces of single biotas:
  - Mainland SARs.
  - SARs among islands in an archipelago.

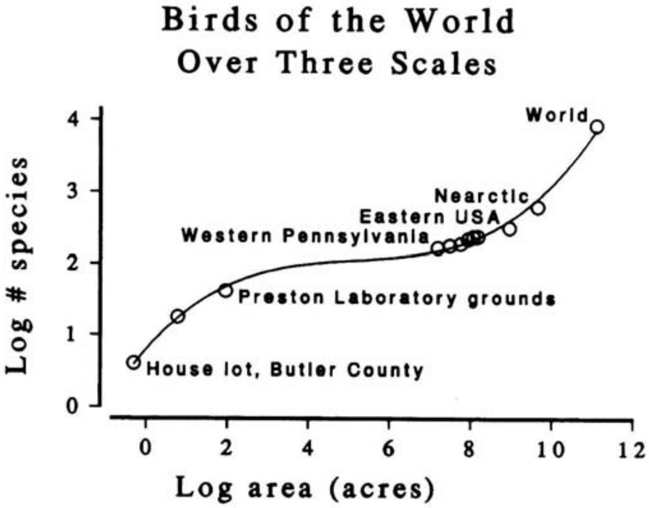


Figure 2.15: Frank Preston's species-area curve for land birds in western Pennsylvania. SARs typically display different slopes for large, medium and small scales [12],[83].

- SARs for large areas that had separate evolutionary history.

For small areas ($\approx$ 0.1ha) the SAR tends to show some curvature on a log-log plot. While for small scales it may be linear (see figs. 2.13(b) and 2.17.), the exponent is typically larger than for larger scales. In order to match lower exponents for larger scales, it must curve downwards.

As the sampled area increases, we enter regional scales comparable to the scales of individual dispersion. A sizeable portion of the individuals belongs to species already detected and z diminishes. Typical values of the scaling exponent are $z = 0.15 - 0.2$. According to Rosenzweig [12] most of the biodiversity increase at this scale is related to encountering different habitats as sampled area grows. Species will specialize to the different habitats and area will simply be a surrogate variable for habitat diversity. This is a good example of a niche oriented explanation that is at odds with neutral theories of diversity (see section 2.4). Evidence that, in some cases, biodiversity fits better to habitat diversity than area is shown in figure 2.16. Ignoring the technical difficulty of deciding when two habitats are different enough to

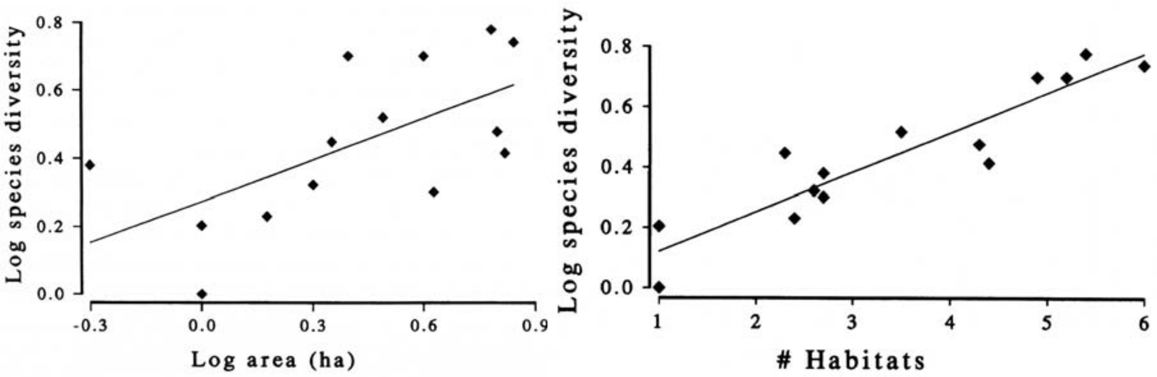


Figure 2.16: The diversity of small mammals in parts of Australia correlates better with the number of habitats than the area, thus supporting the view that the main origin of biodiversity is habitat variety. Figure from ref. [12]. Data from ref. [85].

be considered separately, this approach translates the question into: why are habitats distributed self-similarly?. This question seems to be largely unanswered, but the well known self-affinity of the earth landscape [84] has been suggested as a possibility [81].

Islands SARs are considered separately on several grounds: firstly, their biodiversity is heavily influenced by immigration, since islands are defined to be small enough that speciation is negligible. Secondly, exponents tend to have distinctly higher values: $z = 0.25 - 0.45$. Thirdly, the SAR plots refer to the diversity of islands as related to its area (fig. 2.7,2.9), not the nested design considered for mainland patterns (see fig. 2.17).

Biodiversity in an island seems to be determined by two factors: the number of habitats (as in the mainland case) *and* the balance of the processes of immigration and extinction on the island, as explained by the classical theory of island biogeography by MacArthur and Wilson [12],[86]. The exponent is higher because of the lower rate of immigration compared to the mainland. The mainland gets constantly replenished with immigrant individuals and this means that it can maintain some sink species, which have a negative birth minus death rate. In the case of an island, isolation means that no sink species will survive. Only source species (positive birth minus death rate) will remain. A mainland patch and an island of the same area will have roughly the same number of habitats and source species, but the island will lack the sink species that survive in the mainland patch only because of a vigorous immigration.

To understand qualitatively the influence of these considerations on the value of the exponent z, imagine taking half of the mainland patch and half of this island and comparing it with a second island of the same size. They all will have the same source species, but the second island will lack the sink species; the half of the first island will have some sink species due to the other half and the mainland plot will have sink species due to the other half and some more due the connection with the rest of the mainland. Therefore, for the same change in area, the change in diversity is more pronounced in the case of islands and leads to a higher z (steeper slope).

Finally, large regions of the earth have had separate evolutionary stories and are biologically independent. There is little correlation between species present in these different biological provinces and that gives rise to higher values of z (close to unity, typically $z = 0.8 - 1$). Regions are considered biological independent provinces if most of the species they harbor have been created through speciation within the region, as opposed to immigration from outside (such as islands, for example).

2.3.3 SAR methodology

The method of SAR data recollection deserves comment because not all data sets lead to a power law SAR, even if sampled from the same biome! Areas on which the number of species are to be counted must be contiguous when grouped to form a larger area. This is called the nested design. If one doesn't proceed in this way, the species-area curve will not be linear in a log-log plot, but instead forms a convex curve (see fig. 2.17). The reason is that because species cluster, two contiguous plots are more likely to have similar species. Separate plots will have more different species *in total* that if they were contiguous when added together to form the next area. Since the fraction of species in common between these two plots will decrease with

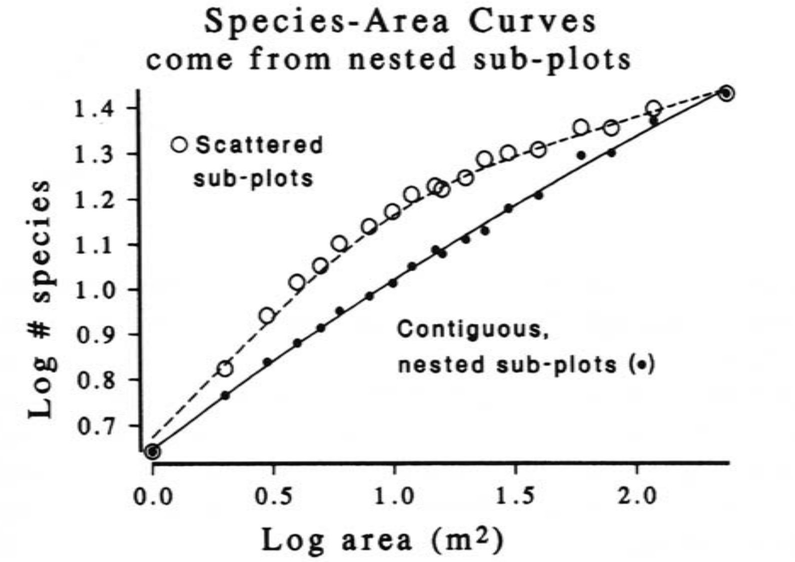


Figure 2.17: Species-area study on Michigan plants. By definition, a proper species-area curve must be assembled by adding adjacent areas to an initial plot. If we take additional plots scattered randomly around the final large area, this will lead to a convex curve. The reason is that, since individual are typically clustered, non-contiguous plots will have less species in common than adjacent one. This means that, when added together to form a bigger area, they will have higher total number of different species. Since the number of species in the smallest and largest area is independent of the sampling protocol, this necessarily leads to a convex curve as shown here. Figure taken from ref. [12]. Data from ref. [72].

distance, the most standardizable way of proceeding in assembling species-area curves is by nesting contiguous plots to create bigger areas [12].

Errorbars for the SAR data are usually not considered. In general the distribution around the mean will not be gaussian, and there is not necessarily a $\sqrt{N}$ dependence. This distribution is unknown: if it were known, we wouldn't be taking the data in the first place. The scatter of the data can give a measure of the error (look for example the small scatter for the thoroughly sampled site SAR in figure 2.13(b)).

2.3.4 A brief history of the Species Area Relationship

The earliest suggestion that a power law SAR is a good representation of data is due to Watson [87] in 1835:

> "On the average, a single county appears to contain nearly one half the whole number of species found in Britain; and it would, perhaps, not be a very erroneous guess to say that a single mile may contain half the species of a county"

Taking the area of Britain to be 89000 square miles and the area of a county to be around 300 square miles, his estimation implies a value of $z = 0.12$, which seems consistent with $z = 0.19$ for the British flora found in later work [81].

Nonetheless, the first to plot the data in a log-log format and suggest the equation that we now use was Arrhenius in 1921 [71]. Shortly afterwards Gleason [72] proposed the alternative exponential model which, however, does not seem to be as common a pattern as the power law dependence [81].

By 1943 Fisher had connected the problem of species diversity to that of species abundance [34]. Not long after, Preston (1948) introduced the lognormal relation for the relative species abundance (see section 2.1.2). Assuming random distributions of individuals, he showed that a common type of lognormal (canonical lognormal) would produce a species area relationship very close to a power law with exponent $z = 0.26$ [27]. Further work by May (1975) showed that this result was robust with respect to the choice of the lognormal distribution [26].

Unfortunately, species area rules with $z \neq 0.26$ are quite common and Preston's theory provided no reason to expect them. At the same time, the underlying assumptions of this derivation (scale invariant lognormal distributions and lack of correlation between ranges and abundances) were shown to be both unrealistic and critical for the derivation [88].

Recently (1999) Harte et al.[17] decided to turn the problem upside down. Since the fractal SAR was such a well-known and widespread pattern, they decided to *assume* it and see what consequences can be derived. There turned out to be plenty: an abundance distribution that resembles a lognormal but with more rare species (as found in real data); a connection to the Endemics Area Rule (similar to the SAR, but involving endemic species[1]); and a prediction for the fraction of species in common between two patches separated by a given distance among others (see section 3.3). It also provided a practical way to generate fractal species distributions [89].

Work by Harte's group [17],[90],[91],[92],[93],[69] has spurred great interest in the consequences of self-similarity [94],[95],[96],[97],[98]. This has added to the long standing interest in this well-known ecological pattern, producing an assortment of papers studying the consequences on the SAR of finite areas [99], clustering [100],[101], habitat heterogeneity and metapopulation dynamics [102] and its relationship with the scaling of trophic links [103], to give a few examples. Different approaches have been used to study the SAR, theoretically and experimentally, ranging from molecular phylogeny techniques (see section 5)[104] to multifractal analysis [105], random-placement models [106], extreme value functions [107], simple spatial contact models [108] or population dynamics oriented models [109],[110]. The neutral model by Hubbell,

[1] Endemic species are those restricted to a particular site/environment

which will be explained in the next section, deserves special mention.

In addition to this, recent efforts by the Center for Tropical Forest Science have yielded large, well sampled tropical forest plots in four different world sites. The data obtained from these sites are invaluable in the study of the SAR and other ecological patterns [111], [101].

Nevertheless, in spite of these recent efforts, there is still no consensus in regard to which processes produce power-law SARs and how the exponents can be predicted.

2.4 Hubbell's neutral model

The introduction of neutral models [20],[112],[113] has created a great deal of controversy in the field of ecology[32]. Neutral models assume that all organisms are on equal footing and consider dynamics involving solely death, birth, speciation and immigration. Considered irrelevant are niche exploitation, predation, parasitism, mutualism and many of the processes that ecologists traditionally invoke in order to explain ecosystem functioning. However, many of the most regular patterns observed in ecology (e.g. lognormal abundance distribution, power law species area relationships) are readily observed in these simplistic models. Hence the controversy. In the words of Jonathan Levine [2] [32]:

> "Neutrality starts with assumptions that are clearly wrong, but produces patterns that match what we see in nature."

The neutrality point of view has important consequences: downplaying the importance of the niche means that communities can easily be invaded and reserves must be larger in order to protect the rarer species. Experimental evidence in favor [20],[114] and against [115],[116] neutral theories has been provided and the debate has just

[2] UK Natural Environment Research Council's Center for Population Biology at Silwood park, London

begun.

In the rest of this chapter, I will briefly present Hubbell's unified neutral theory of biodiversity and biogeography.

2.4.1 Local community dynamics

Hubbell's neutral theory considers space only in an indirect way by describing the ecosystem at two scales: fast local community dynamics at small spatial scales and slower metacomunity dynamics on larger spatial scales. The metacomunity is supposed to be composed of all the local communities and to be large enough that speciation events are possible, whereas for the local community all new species come from immigration. The metacommunity is supposed to be an autonomous entity and receive no immigration.

The local community dynamics are depicted in figure 2.18. Assume there are J individuals at the local community. At the beginning of each cycle, species present (two in this case) occupy all resources. At each time step D individuals are killed by a disturbance. The vacant sites have a probability m of being replaced by an immigrant form the metapopulation. The remnant vacant sites not filled with immigrant are filled with the surviving species in a proportion equal to its local relative abundance in the local community after the disturbance. The immigrant species are chosen randomly among the metapopulation with a probability equal to its metapopulation relative abundance. Notice that this model is *not* spatial.

For each species, it can be shown that the local community markovian process will be attracted to one of the following absorbing states in the case of no immigration ($m = 0$): either extinction or monodominance (no other species are present). The times to reach these absorbing states are rather long (and decrease with growing disturbances D). Therefore, immigration from the metacommunity plays a very important role even if it is very limited. Additionally, since immigration is proportional to the metacommunity relative abundance distribution, this distribution will have an important influence on the local community relative abundance.

An analytical solution for the number of species with n individuals resulting from this process is available [117],[30]:

$$\langle \phi_n \rangle = \theta \frac{J!}{n!(J-n)!} \frac{\Gamma(\gamma)}{\Gamma(J+\gamma)} \int_0^{\gamma} \frac{\Gamma(n+y)}{\Gamma(1+y)} \frac{\Gamma(J-n+\gamma-y)}{\Gamma(\gamma-y)} e^{-y\theta/\gamma} dy \qquad (2.6)$$

where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$, $\gamma = \frac{m(J-1)}{1-m}$ and θ is the fundamental biodiversity number, introduced in the next section and proportional to the speciation probability.

This result has been shown to properly fit species abundances in tropical forest plots (fig. 2.19) and, in general, displays more rarity than the lognormal.

2.4.2 Metacommunity dynamics

The dynamics for the metacommunity are the same as for the local community except that there is no immigration and there are speciation events. In each birth, there is a speciation probability ν. Calculations become more complicated then, but it is possible to arrive at an expression for the number of species with n individuals:

$$\phi_n^M = \frac{\theta x^n}{n} \qquad (x = b/d) \qquad (2.7)$$

where b and d are the ratio of births and deaths per individual respectively. θ is the fundamental biodiversity number $\theta = S_M P_0 \nu / b$. S_M is the number of species in the metacommunity and P_0 is the probability that a species is not present.

Equation 2.7 is just the Fisher log series explained in section 2.1.1, with Fisher's α replaced by θ, the biodiversity parameter. Hubbell then concludes that, in a sense, both Fisher and Preston had the right abundance distribution, but for different time and space scales: Fisher for the metacommunity and Preston for the local community.

One might be surprised to see how similar dynamics yield such different relative abundances. It is fundamental to remember that speciation is very different to

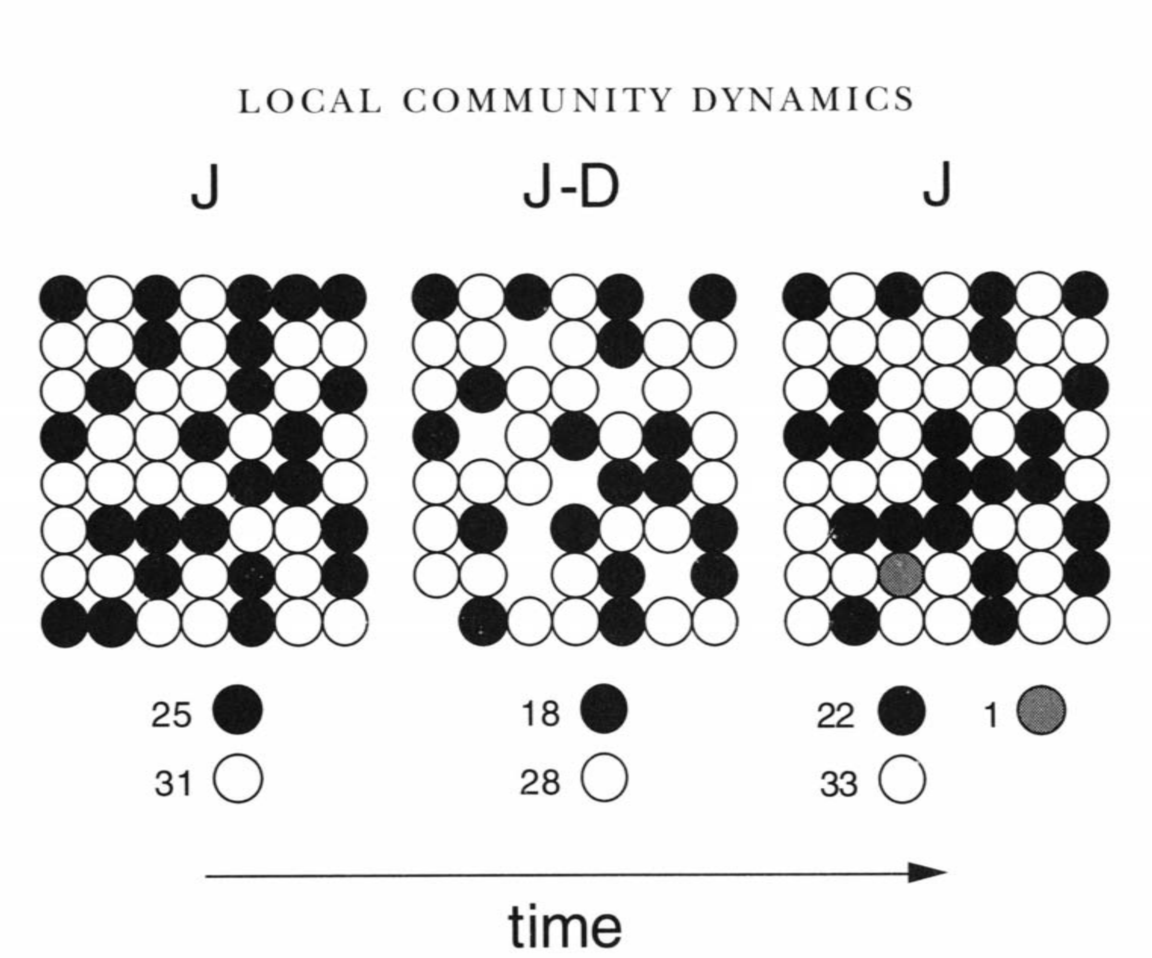


Figure 2.18: Local comunity dynamics for Hubbell's neutral model composed of competition and immigration. Assume there are J individuals at the local community. At the beginning of each cycle, species present (two in this case) occupy all resources. At each time step D individuals are killed by a disturbance. The vacant sites have a probability m of being replaced by an immigrant form the metapopulation. The remnant vacant sites not filled with immigrant are filled with the surviving species in a proportion equal to its local relative abundance in the local community after the disturbance. The immigrant species are chosen randomly among the metapopulation with a probability equal to its metapopulation relative abundance. Notice that this model is *not* spatial. The individuals are drawn in a grid to provide a visual picture, but their relative positions are irrelevant for the dynamics. Taken from ref.[20].

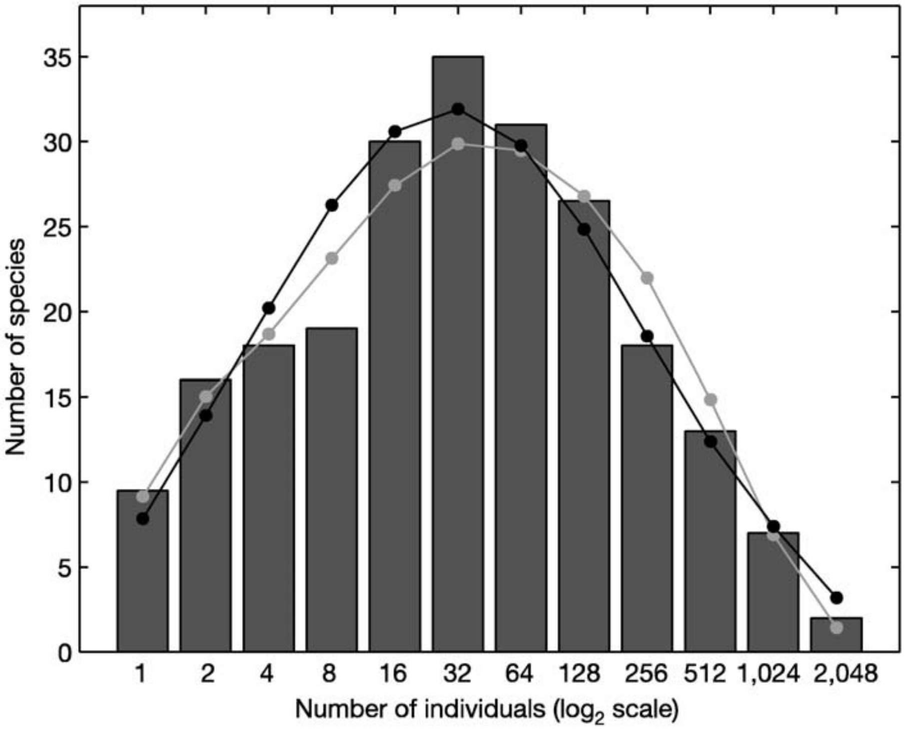


Figure 2.19: Tree species abundances in a 50 hectare plot of tropical forest in Barro Colorado Island, Panama [118]. Only trees with a diameter bigger than 10 cm were sampled. The data includes 225 species comprising 21457 individuals. The bars are the observed abundances binned into $\log_2$ abundance categories, following Preston's methods (see section 2.1.2). The first histogram bar represents $\langle\phi_1\rangle/2$, the second bar $\langle\phi_1\rangle/2 + \langle\phi_2\rangle/2$, the third bar $\langle\phi_2\rangle/2 + \langle\phi_3\rangle + \langle\phi_4\rangle/2$, the fourth bar $\langle\phi_4\rangle/2 + \langle\phi_5\rangle + \langle\phi_6\rangle + \langle\phi_7\rangle + \langle\phi_8\rangle/2$... etc. The black curve represents the best fit to a lognormal distribution $\langle\phi_n\rangle = \frac{N}{n}\exp(-(\log_2 n - \log_2 n_0)^2/2\sigma^2)$ $(N = 46.29, n_0 = 20.82$ and $\sigma = 2.98)$. The gray curve is the best fit to equation 2.6 ($m = 0.1$ and $\theta = 47.226$). Both represent a good fit to the BCI data. Figure taken from ref.[117].

immigration. The immigration into a local community is proportional to the metacommunity relative abundance, whereas speciation is completely random.

2.4.3 The species Area Relationship in Hubbell's model

Purporting to be a unified neutral theory of biodiversity and biogeography, Hubbell's theory must address the oldest ecological pattern known: the SAR. Unfortunately, this model treats space implicitly and that will prove to be quite constraining.

To start with, the unified neutral theory acknowledges with Rosenzweig [12] that the SAR is trifasic (fig. 2.15, section 2.3). For short scales the SAR is heavily dependant on the details of the ecosystem and, supposedly, not linear in a log-log plot (but see fig. 2.13(b)). Instead it is supposed to exhibit an upward convexity (negative lower convexity, see section 2.3). By taking the number of individuals to be proportional to the area: $n = kA$, the predictions above regarding species abundance can be recast in the form of SARs:

$$S(n) = S(A/k) \tag{2.8}$$

This produces a SAR curve with upward convexity that seems to fit well some invertebrate data, although needs some corrections to describe dispersally limited species.

For medium scales, it is impossible to continue with this approach, so Hubbell resorts to results from a voter model by Durrett and Levin [108] and introduces a spatially explicit version of the neutral model explained in the previous two sections. In this version, a grid of 201×201 local communities of $n = 4$ individuals each is used and immigration is limited to neighboring communities. Speciation is random and can appear anywhere in the lattice. By using this scheme, a power law SAR is found for intermediate scales, although only for 1.5 decades. The value of z is dependent only on the fundamental biodiversity number θ and the immigration rate m. For a significant range of reasonable values of these parameters, the SAR exponent is within the usually reported value of $z = 0.15 - 0.4$.

For larger scales, Hubbell, with Durrett and Levin, defines a correlation length L marked by the point at which the SAR deviates from a line in a log-log plot. This correlation length [20]:

> "measures and defines the natural length scale of a biogeographic process over which the metacommunity events are dynamically and evolutionarily connected."

L depends again only on m and θ, which are constrained by the value of z at medium ranges. For values of areas bigger than the square of this length, the communities are uncorrelated and exponents z close to 1 would be expected under neutral theory, as is the case for real data (section 2.3).

2.5 Perspectives in ecology

My personal interest in ecology resides in what it can teach us regarding pattern robustness and universality in highly interacting systems. The ubiquity of the lognormal abundance distribution and the power law Species Area Rule (see sections 2.1.2 and 2.3) tells us that many of the details of the interactions wash out and only a few are responsible for the regularities that we see.

I am therefore particularly startled that neutral theories have had so little consideration in ecology, in agreement with Bell [112]:

> "Indeed, few aspects of the history of ecology and evolutionary biology are more remarkable than the lack of development of an individual-based neutral theory of species diversity in community ecology during the entire 20 years when the neutralistic-selectionist debate over allelic variation was at its height in population genetics. ..."
>
> "... Perhaps ecologists find it difficult to accept that the differences they so

> clearly recognize among their study species have no functional significance, whereas geneticists, dealing with spots on a gel, are more inclined to neutralism."

I think the importance of neutral models is *precisely* that its dynamics are minimalistic. This is *prima facie* evidence of the remarkable robustness of the regular patterns observed in ecology. In fact, I think that one of the tasks of a complexity-oriented ecology would be to identify which dynamics lead necessarily and robustly to these patterns. By having this mapping, one would then be able to predict macroscopical properties from the knowledge of microscopical dynamics. Not only in ecology, but in other complex systems as well.

Consider the power law SAR. In chapter 4 it is demonstrated that it arises as a very general consequence of clustering and an abundance distribution close to the those found usually ecosystems. Only that. The details of the interactions in the ecosystem are irrelevant. Can we then expect this pattern to hold in systems characterized in a similar way? Languages, for example, arise in ways resembling the ecological allopatric speciation. Can we expect the SAR to hold for the number of languages spoken in an area?

Now consider abundance distributions. Neutral models involving as little as competition and selection (see section 2.4) exhibit regularly lognormal-like distributions of abundance. Free market economy relies on competition as a primary driving force. Maybe an equivalent pattern holds in e.g. the size of companies competing for market share [119]?

Granted, they are not the same systems and their dynamics may be quite different, but on the other hand, it doesn't take much for these laws to apply.

Unfortunately, experimental ecology is a tough science: ecosystems vary slowly, are typically large, time consuming to study properly and cannot be easily disturbed to test hypothesis. On the contrary, *microbial* ecosystems display much more accessible time and space scales, can be easily studied to test theoretical predictions and the experimental tools to study them have become available in the last decade (see chapter 5). Again, much of microbial ecosystem functioning will be different from that of higher taxa, but the robust patterns that we are interested in will likely apply to both worlds. I therefore believe that most of our ecological knowledge in the next century will likely come from microbial ecology.

Chapter 3

A self-similar model of ecology

This section is devoted to the self-similar model introduced by Harte et al.[17]. Whereas previous authors attempted to derive the power law SAR from known ecological patterns, Harte et al. take the opposite approach. They assume that the SAR is fractal, show that this is equivalent to self-similarity and see what can be obtained from this assumption. The results are quite remarkable as will be seen in section 3.3.

After explaining the model and its predictions, I will present my own work on this model. This work has been published as reference [98]. It involves a linear version of the recursion relation for the abundance distribution and an explanation for the scaling in this distribution that was assumed without proof by Banavar et al.[95].

In the final section, the shortcomings of this model will be presented.

3.1 Harte's model

In the model proposed by Harte et al.[17] an area A_0 with a number of species S_0 is considered. The number of individuals in each species is described by $P_0(n)$, where $S_0 P_0(n)$ is the expected number of species with n individuals. The area A_0 is chosen to be in a shape of a rectangle with its length being $\sqrt{2}$ times its width; such that by a bisection along the longer dimension it can be divided in two rectangles of shape similar to the original (see figure 3.1). $A_i = A_0/2^i$ is the area of a rectangle after the ith bisection. If a species is present in an area A_i, and nothing else is known about the species, there are three possibilities: it might be present *only* on the right subpartition of area A_{i-1} (probability $P(R'|L)$), *only* on the left one ($P(R|L')$) or in both ($P(R'|L')$). By symmetry $P(R'|L) = P(R|L')$; and a is defined as $P(R'|L) \equiv 1 - a$. The probability of finding a species on the right side, independently of what happens on the left side is:

$$\begin{aligned} P(R') &= P(R'|L) + P(R'|L') = 1 - a + 2a - 1 = a \\ &= P(L') \text{ by symmetry} \end{aligned} \tag{3.1}$$

Self-similarity is introduced by stating that a is independent of i, that is, scale.

The average number of species in an area A_i is then $S_i = a^i S_0$, assuming all the species share the same probability of presence a. To get the area dependence of S_i we write:

$$a^i = (2^{log_2(a)})^i = (2^i)^{log_2(a)} \equiv (2^i)^{-z} \tag{3.2}$$

and since $2^i = A_0/A_i$:

$$S_i = a^i S_0 = (A_i/A_0)^z S_0 \sim A_i^z \tag{3.3}$$

Dropping the i subindexes we get the traditional form of the SAR:

$$S \sim A^z \tag{3.4}$$

with $z \equiv -\log_2(a)$. Notice that this argument doesn't give any explanation as for the value of a and, consequently, z. It is just taken as a parameter.

A less straightforward conclusion can be obtained from self-similarity regarding $P_i(n)$ (expected fraction of species with n individuals for an area A_i). As demonstrated in figure 3.1) [17] the abundance distributions $P_i(n)$ are related for two contiguous levels by the nonlinear recursion relation:

$$P_i(n) = xP_{i+1}(n) + (1-x)\sum_{k=1}^{n-1} P_{i+1}(n-k)P_{i+1}(k) \tag{3.5}$$

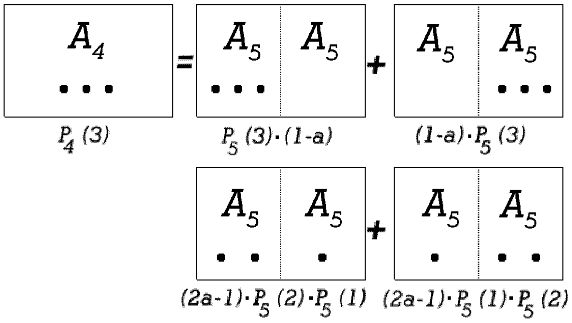


Figure 3.1: Explanation of Equation 3.5 . Let's consider the case $i = 4$ and $n = 3$. Circles correspond to individuals of a particular species found in a patch. On the left side there are three individuals in a patch A_4, on the right side all the possible ways in which those 3 individuals can be distributed in the two patches A_3. The probability of finding three individuals in a patch A_4 is then the addition of the probability that all the individuals are on one side (prob. $1-a$) times the probability that once all the individuals are on one side there are no individuals on one side and there are three individuals on the other side (prob. $1 * P_5(3)$) plus the probability that the species are present on both sides (prob. $2(1-a)$) times the probability that once the species are present on both sides there are two individuals on one side and 1 individual on the other one (prob. $P_5(2) * P_5(1)$). Taking $x = 2(1-a)$ and $1-x = 2a-1$ we find $P_4(3) = xP_5(3) + (1-x)2P_5(2)P_5(1)$. This can be generalized to obtain Equation 3.5. Figure taken from ref.[17].

where $x = 2(1 - a)$. This recursion relation requires an initial condition. It is supplied by defining a minimum patch $A_m = A_0/2^m$, such that it contains on average only one individual (see figure 3.3). Consequently, $P_m(n) = \delta_{n1}$. This also limits the maximum number of individuals that can be found in a patch A_i to 2^{m-i} so $P_i(n) = 0$ for $n > 2^{m-i}$.

The abundance distribution obtained from eq. 3.5 (fig. 3.2) is more similar, in general, to available data [111],[20] than the traditional lognormal (see section 2.1.2). In particular, the fraction of rare species is higher than for this distribution, a traditional handicap of the lognormal.

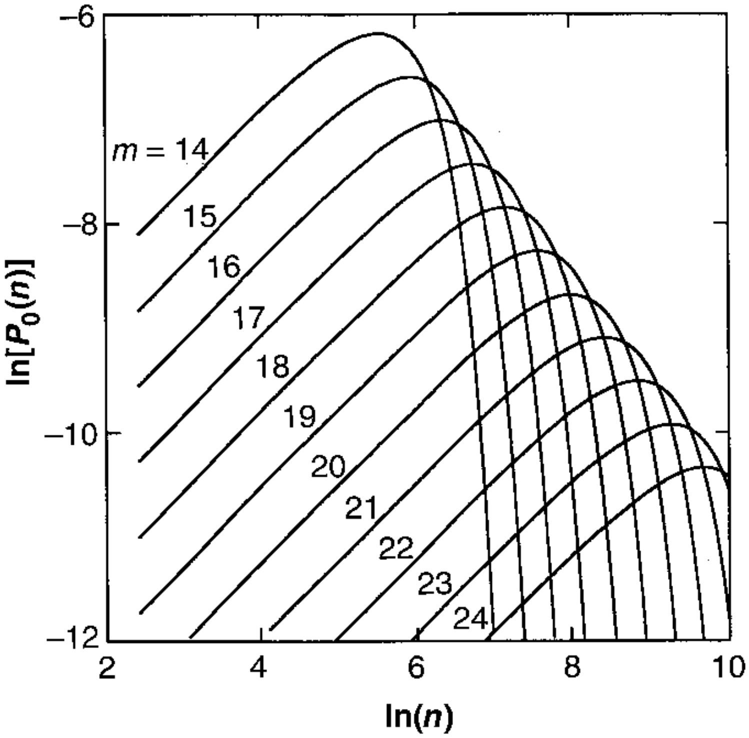


Figure 3.2: Species-abundance distributions in log-log axis obtained from equation 3.5 with $x = 0.484$. Taken from [17].

3.2 Community-level and species-level self-similarity

In the previous section, we assumed that all species distributed self-similarly with the same probability a. It is not necessary for this to be the case to get a power law SAR (community-level self-similarity). In fact, it looks quite unlikely that all species behave in the same way. One can define a counterpart of the probability a for each species j: α_j. In terms of these α_j, the ratio of average number of species at two different bisection levels (S_i and S_{i-1}) is [92],[120]:

$$a_i = \frac{S_i}{S_{i-1}} = \frac{\langle \prod_{j=1}^{i} \alpha_j \rangle}{\langle \prod_{j=1}^{i-1} \alpha_j \rangle} \tag{3.6}$$

If a_i is independent of i we will observe community-level self-similarity. This can happen if all α_j are the same and equal to a or if the α_j are different and have the proper scale dependence to make a_i constant with scale. Curiously, if all α_j are self-similar for each species, it can be proved that a_i won't be constant with scale [92],[120]:

community-level and species-level self-similarity are not compatible. The only ways to obtain community-level self similarity are that either all $\alpha_i = a$, or the α_i have an appropriate scale dependence as defined above.

3.3 Other consequences of self-similarity

The geniality of the self-similar model lies in its connection and unification of empirically supported ecological patterns with a minimal assumption. We have seen part of this in the previous section regarding the SAR and abundance distributions, but that is not all. By assuming that self-similarity holds at the community level ($a = 2^{-z}$) it is possible to show the applicability of the following patterns:

- The Endemics Area Rule (EAR) relates the amount of species found *only* in an area A with this area. By the same reasoning as in the previous section, the amount of species found only in a specified rectangle A_i is $E_i = (1-a)^i S_0$. By defining $z' = -\log_2(1-a)$ then $E(A_i)/E(A_j) = (A_i/A_j)^{z'}$ which necessarily means that:

$$E(A) = cA^{z'} \tag{3.7}$$

  For the typical value of $z = 0.25$, one obtains $z' = 2.65$ [90],[17].

- The fraction of species in common to two spatially separated patches of land separated by a distance d decreases with this distance as d^{-2z} [90][17].

- The amount of species found in an area depends on the shape of the latter: elongated areas will tend to hold more diversity that compact ones [17].

If we further assume that self-similarity holds not only for the whole ecosystem, but for each of the individual species i with probability α_i, then the following patterns are expected:

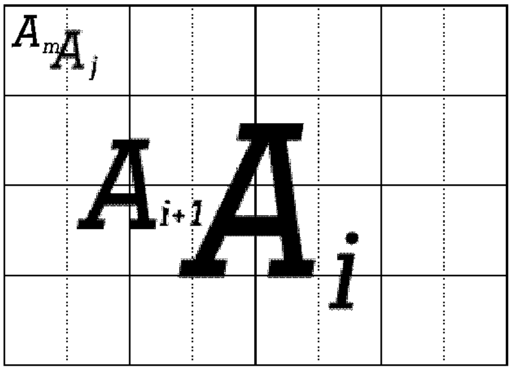


Figure 3.3: A_m is the minimum patch. A_j in this case comprises two minimum patches, but it can be of any size. In Equation 3.5 the contributions to $P_i(n)$ come from the two patches of size A_{i+1}, whereas in the case of the linear recursion relation they come from the 2^{j-1} patches of size A_j.

- The Range Abundance Relationship relates the range (or cover/occurrence as defined in section 2.2) of a species with its abundance. Species level self-similarity implies a power law dependence between the range R and both the species abundance n and the area of the census cell A. The exponent can be derived from the α_i [120].

- The correlation function ρ_i of individual of species i (or the relative neighborhood density by its ecological name) is also self-similar:

$$\rho_i(r) \propto r_i^w \tag{3.8}$$

  with $w_i = 2\log_2(\alpha_i)$ [89].

All of these patterns have been tested against real data as explained in the given references.

3.4 Linearization of the abundance distribution

Equation 3.5 is nonlinear, and thus inconvenient to handle efficiently. The purpose of this section is to derive the scaling relation for the probability distribution without making any assumptions about the existence of moments of the distribution. As we will see, this can be accomplished by deriving an equivalent linear relation for the probability distribution. This derivation sums up multiple patches at once, rather than proceeding strictly hierarchically as in the original derivation.

We consider that the contributions to $P_i(n)$ come from several (2^{j-i}) patches of area $A_j = A_i/2^j$ ("boxes") instead of from 2 patches of area $A_{i+1} = A_i/2$ as before (see figure 3.3). The probability of finding n individuals in A_i is then the sum over the probabilities of finding r of these "boxes" with the species present ($R^i_j(r)$), multiplied by the probability of finding n individuals in these r boxes ($Q^i_j(r,n)$):

$$P_i(n) = \sum_{r=1}^{2^{j-i}} R^i_j(r) Q^i_j(r,n) \tag{3.9}$$

Note that the index j is not summed over. It is arbitrary, indicating the size of the "box". For $j = i+1$ there are two boxes of area $A_i/2$ and the original result of Harte et al. is recovered, whereas for $j = m-1$ we will find a linear relation. But before establishing these results we calculate explicitly $R^i_j(r)$ and $Q^i_j(r,n)$:

- $Q^i_j(r,n)$ is the probability of finding n individuals in r boxes of size $A_i/2^j$ in a total area A_i:

$$Q^i_j(r,n) = \tag{3.10}$$
$$\begin{array}{ll} \sum\limits_{n_1 \ldots n_n=1}^{2^{m-j}} \quad (\prod_{l=1}^{r} P_j(n_l))\delta(n - \sum n_k) & r \leq 2^{j-i} \\ \qquad\qquad 0 & r > 2^{j-i} \end{array}$$

  This formula is the probability of finding n_1 individuals in the first box, n_2 in the second one, ... etc while the Kronecker delta limits the possibilities to those that add up to the total number of individuals n. 2^{j-i} is the maximum number of boxes and 2^{m-j} is the maximum number of individuals in each box.

- $R^i_j(r)$ is the probability of finding r boxes of size A_j in which the species is present, in a total area A_i. This is just:

$$R_j^i(r) = P_{m+i-j}(r) \tag{3.11}$$

This follows because the reasoning expressed in figure 3.1 can be applied to find the same recursion relation for $R_j^i(r)$ as for $P_i(n)$:

$$R_j^i(r) = xR_j^{i+1}(r) + (1-x)\sum_{k=1}^{r-1} R_j^{i+1}(k)R_j^{i+1}(r-k) \tag{3.12}$$

The initial conditions do not change either, with $R_j^j(r) = \delta_{r1}$. The only difference with the derivation for $P_i(n)$ is that r refers to the number of boxes (not individuals) and that the recursion has to be applied $j-i$ times instead of $m-i$ times.

We can now check that for $j = i+1$ we find the same result as before:

$$\begin{aligned} P_i(n) &= \sum_r R_1^i(r)Q_1^i(r,n) \\ &= R_1^i(1)Q_1^i(1,n) + R_1^i(2)Q_1^i(2,n) \end{aligned} \tag{3.13}$$

Reading off from Equation (4):

$$Q_1^i(2,n) = \sum_{k=1}^{n-1} P_{i+1}(k)P_{i+1}(n-k) \tag{3.14}$$

$$Q_1^i(1,n) = P_{i+1}(n) \tag{3.15}$$

$$R_1^i(1) = x \tag{3.16}$$

$$R_1^i(2) = 1-x \tag{3.17}$$

Hence we obtain:

$$P_i(n) = xP_{i+1}(n) + (1-x)\sum_{k=1}^{n-1} P_{i+1}(k)P_{i+1}(n-k) \tag{3.18}$$

as announced previously. To obtain a linear relation we set $j = m-1$ and obtain:

$$Q_{m-1}(r,n) = \sum_{n_1,\dots n_r=1}^{2^{m-j}} (\prod_{l=1}^{r} P_{m-1}(n_l))\delta(n - \sum_i n_i) \tag{3.19}$$

For $P_{m-1}(n)$ we only have the following possibilities:

$$P_{m-1}(n) = \begin{cases} x & n = 1 \\ 1-x & n = 2 \\ 0 & n \neq 1,2 \end{cases} \tag{3.20}$$

We find, denoting by $q = n - r$ the number of boxes with two individuals (factors of $P_{m-1}(2)$ in the equation above):

$$g(n,r) \equiv Q_{m-1}(r,n) = \frac{r!}{(r-q)!q!} x^{r-q}(1-x)^q \tag{3.21}$$

The first factor is the number of possible configurations in which there are q boxes with two individuals and $n - q$ with one individual. Finally we obtain:

$$P_i(n) = \sum_{r=1}^{2^{m-i-1}} P_{i+1}(r) g(n,r) \tag{3.22}$$

which is a *linear* relation involving $P_i(n)$ and $P_{i+1}(n)$.

3.5 Scaling in the abundance distribution

Equation (13) allows us to derive the scaling law that was assumed by Banavar et al. [95]:

$$P_i(n) = \frac{1}{n} f(\frac{n}{N_i^{\phi}}) \tag{3.23}$$

where N_i ($= 2^{m-i}$) is the maximum number of individuals in an area A_i and $\phi = 1-z$.

In order to achieve this, the following has to be done:

- First, find the **continuum limit** for $g(r,n)$. Since $g(r,n)$ is just a binomial distribution, it tends to a gaussian for large n:

$$\begin{aligned} g(n,r) &= \frac{r!}{(r-q)!q!}x^{r-q}(1-x)^q \\ \approx \frac{1}{\sqrt{2\pi r}}\frac{1}{\sqrt{x(1-x)}} &\exp\left(-\frac{1}{2}\frac{(q-rx)^2}{rx(1-x)}\right) \\ &= \frac{1}{\sqrt{\pi}\epsilon_{a,r}}\frac{1}{2a}\exp\left(-\frac{(r-n/2a)^2}{\epsilon_{a,r}^2}\right) \end{aligned} \tag{3.24}$$

$$\epsilon_{a,r} = \sqrt{2(2a-1)(1-a)r/(2a)^2} \tag{3.25}$$

$g(n,r)$ is the probability of finding n individuals in r boxes. This probability is highly peaked around $n = 2ar$, since $2a$ $(= 1(1-a) + 1(1-a) + 2(2a-1)$ $)$ is the average of individuals per box. The more boxes there are (bigger r) the sharper the peak. This means that for large r the only relevant values of n are those near $n = 2\mathbf{a}r$ and the expression given above for $g(n,r)$ is valid for large r (which implies large n).

- Second, rewrite everything in terms of a new variable x and a new probability density $\overline{P}_i(x)$. x replaces n and is the fraction of the total number of species: n/N_i (which varies from 0 to 1). $\overline{P}_i(n)$ is the density probability $P_i(x)/(1/2^{m-i})$, where $1/2^{m-i}$ is the distance between two points in the new variable x. In this way all $P_i(n)$ can be compared with each other in equal terms.

In terms of these new variables, the recursion relation can now be written as:

$$\overline{P}_i(x) = 2\sum_{y=1/2^{m-i-1}}^{1} g(2^{m-i}x, 2^{m-i-1}y)\overline{P}_{i+1}(y) \tag{3.26}$$

The continuum limit is found by taking m (and consequently the number of points $N_{i+1} = 2^{m-i-1}$) to an arbitrarily large value and using the continuum limit of $g(r,n)$ as defined above. The fact that the approximation for $g(r,n)$ is not a very good one for small values of n or r is of little importance in the limit of large m :

$$\begin{aligned}\overline{P}_i(x) &= 22^{m-i-1}\sum g(2^{m-i}x, 2^{m-i-1}y)\overline{P}_{i+1}(y)\underbrace{1/2^{m-i-1}}_{\Delta y} \\ &= \int_0^1 g^*(x,y)\overline{P}_{i+1}(y)dy \qquad (3.27)\end{aligned}$$

where $g^*(x,y) = 22^{m-i-1}g(2^{m-i}x, 2^{m-i-1}y)$ and is equal to $\frac{1}{a}\delta(y - x/a)$ in the limit of large m:

$$\begin{aligned}g^*(x,y) &= \frac{1}{\sqrt{\pi}}\frac{1}{2a}\frac{1}{\epsilon_{y,a}}\frac{1}{\underbrace{2^{(m-i-1)/2}}_{\delta}}\exp[\frac{(y-x/a)^2}{\epsilon_y, a^2}\underbrace{2^{m-i-1}}_{\delta^2}] \\ &= \frac{1}{\sqrt{\pi}}\frac{1}{2a}\frac{1}{\epsilon_{y,a}\delta}\exp\frac{(y-x/a)^2}{(\epsilon_{y,a}\delta)^2} \qquad (3.28)\end{aligned}$$

which is a standard representation of the Dirac delta function [121] in the x/a variable for $\epsilon_{y,a}\delta \to \infty$ (or $m \to \infty$):

$$\begin{aligned}&\lim_{m\to\infty}\int g^*(x,y)f(x)dx = f(ay) \\ &\Rightarrow \lim_{m\to\infty} g^*(x,y) = \frac{1}{a}\delta(y - x/a) \qquad (3.29)\end{aligned}$$

This implies that

$$\overline{P}_i(x) = \frac{1}{a}\overline{P}_{i+1}(x/a) \qquad (3.30)$$

which is, in terms of n and $P_i(n)$,

$$P_i(n) = \frac{1}{2a}P_{i+1}(n/2a) \qquad (3.31)$$

Since $a = 2^{-z}$ and $\phi = 1 - z$, multiplying the above equation by n and writing the explicit dependence of $P_i(n)$ on N_i as $P_i(n) = P(n, N_i)$:

$$\begin{aligned}&nP(n, N_i) = \frac{n}{2a}P(n/2a, N_{i+1}) = \frac{n}{2a}P(n/2a, N_i/2) \qquad (3.32) \\ &\Rightarrow f(n, N_i) \equiv nP(n, N_i) = \frac{n}{2^\phi}P(n/2^\phi, N_i/2)\end{aligned}$$

which is by definition $f(n/2^\phi, N_i/2)$. Since N_i is equal to a power of two this means that $nPi(n)$ is a function only of n/N_i^ϕ:

$$P_i(n) = \frac{1}{n}f(\frac{n}{N_i^\phi}) \qquad (3.33)$$

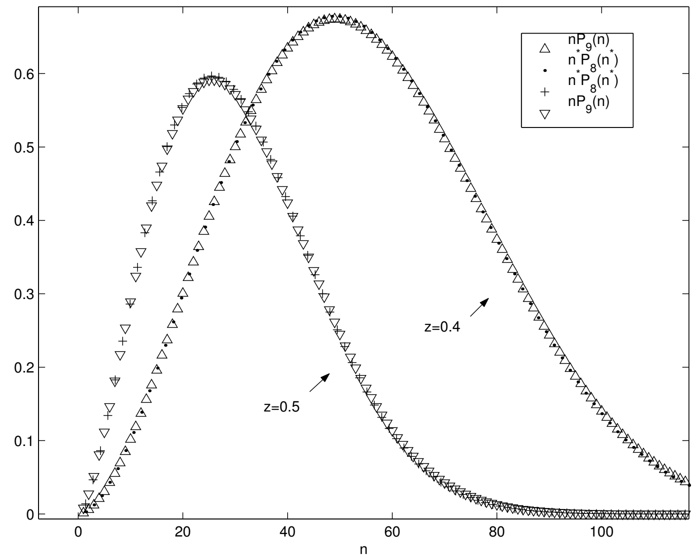


Figure 3.4: Scaling function $nP_0(n) = f(n/N_i^\phi)$ for $m = 8$ and $m = 9$, and $z = 0.4$ and $z = 0.5$. $n^* \equiv n/2^\phi = n/2a$.

As can be appreciated from the results above, a constant a (not dependent on i) is necessary to obtain the scaling law: otherwise ϕ would depend on i. In figure 3.4 we exhibit the scaling function for several z and demonstrate the scaling law.

3.6 Failures and successes of the self-similar model

The self-similar model is remarkable in the sense that it is able to connect a very simple assumption on self-similar placement with a host of related ecological patterns: a distribution of abundances that is more similar to those empirically found than the traditional lognormal, the Endemics Area Rule, species commonality, area shape dependence of biodiversity, Range Abundance Relationship and correlation functions. All of these are strongly supported by empirical evidence.

Despite its considerable successes there are some important limitations to this model. To start with, the grid description of the individuals distribution looks rather artificial when implemented. Figure 3.5 illustrates this. A more serious consequence of this grid dependence is that the SAR is not translational and rotationally invariant.

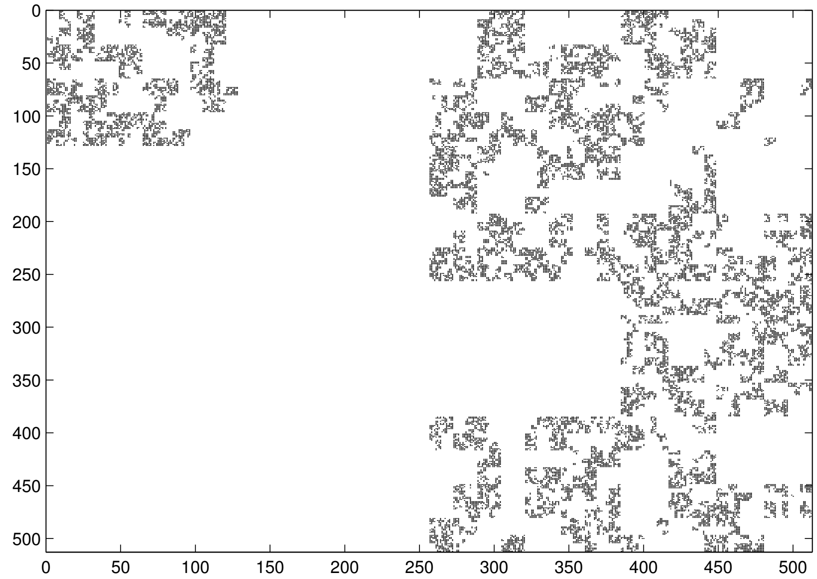


Figure 3.5: Implementation of a self-similar distribution produced with a probability of presence a constant with scale. At each level a random number $x \in [0,1]$ is generated: for $x < 1-a$ the species is present only on the left half, for $1-a < x < 2(1-a)$ the species is present only on the right half and for $x > 2(1-a)$ it is present in both. For each of the halves in which the species is present the process is repeated. The resulting distribution of individuals is shown above. As can be seen clustering is reproduced in this model, but in a way such that the grid used to lay the points generates artifacts like straight lines and rectangular shapes.

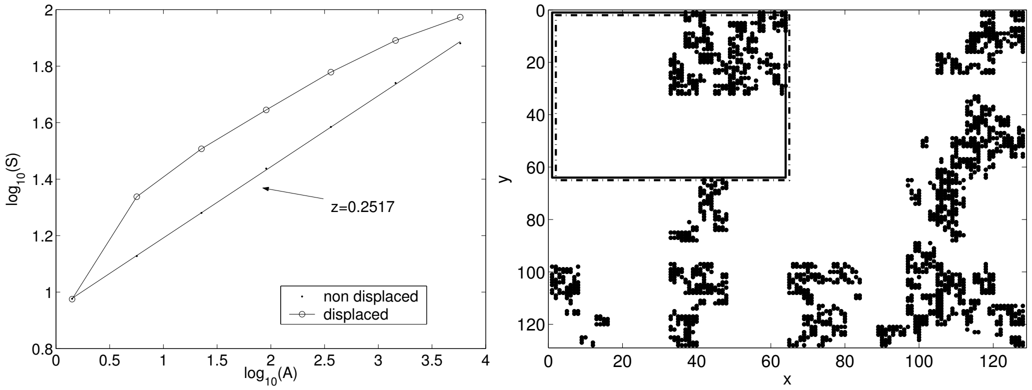


Figure 3.6: Evidence that individual distributions created by Harte's method (fig. 3.5) are unrealistically translationally non-invariant. An ensemble of 100 species ($z = 0.25$ for all of them) was created with this algorithm and then the SAR was measured for the area shown on the right (continuous line). A fractal SAR with exponent $z \simeq 0.25$ is found, as expected. If, for the same ensemble of species, we measure the SAR in an area that is displaced toward the right bottom corner by 1 space unit (discontinuous line), we find a convex, non fractal, SAR. This is unrealistic because we don't expect Nature to hold its patterns only for specific origin grid points.

Maddux [97] showed this in a rather convolute way. In my opinion, the clearest way to appreciate this fact is shown in figure 3.6.

Also, this construction describes the distribution of individuals for a very specific experimental setup. It automatically accounts for nested sampling. It can't readily answer why a non-fractal SAR would arise for a non-nested scheme (see fig. 2.17).

Finally, probably the biggest caveat is that, by construction, it leaves open the origin of the SAR and the factors that influence the exponent z.

In chapter 4, we will introduce a gridless, continuous, translationally and rotationally invariant scheme for creating individuals distributions and will pinpoint the factors that determine z.

Chapter 4

A geometrical explanation for the emergence of scale-invariant Species Area Relationships

4.1 Introduction

As explained in chapter 2, the relationship between the area occupied by a biome and the amount of biodiversity it supports is an intensively studied problem in ecology, of theoretical and practical importance [12],[18],[59],[60],[44],[18],[63],[64],[65]. Although other relationships are quoted and used [70], the most common representation for the Species Area Relationship (SAR) is a power law:

$$S = cA^z \tag{4.1}$$

The power law SAR seems to be a remarkably general pattern, observed in such different organisms as plants,birds,insects, mammals and fish and climates ranging from temperate in France to tropical in the Antilles (see section 2.3.1). Such widespread occurrence seems to indicate a very robust mechanism, independent of the details of the organism interactions [26]. Explanations for the ubiquity of the SAR have been proposed, some focusing on abundance distributions[27],[26], others focusing on the allocation of individuals,[99],[100],[106],[107] and still others focusing on population dynamics[20],[108],[109],[110]. Nonetheless, the consensus seems to be that the mechanisms for the appearance of the fractal SAR are not well known. In particular, it is hard to predict under which conditions the SAR will be power law and what the exponent will be.

In this chapter we will show that the scale-free SAR arises whenever two conditions are fulfilled: a) individuals of each species cluster and b) the abundance distribution for the species is similar to Preston's lognormal [27] but with a higher rarity, as is commonly observed in real ecosystems[20],[17] (see fig. 2.4). Our approach is purely geometrical and does not involve the internal dynamics of the system involving competition, selection, immigration, speciation.. etc. This doesn't mean that they are insignificant, but rather that they are relevant *only* in as much as they affect the two properties above. We will also show that the appearance of the power law SAR is robust to small changes in the details of the clustering or the abundance distribution.

Finally, we will test our theory with ecological data obtained by Jessica Green, John Harte and Annette Ostling.

4.2 A continuum definition of the Species Area Rule

The SAR is traditionally measured in experimental settings by imposing a grid on a given area and averaging the number of species found in each group of grid cells, as indicated in figure 4.1. We will refer to the number of species found in an area A averaged in this way as $S_G(A)$.

For the rest of this chapter we will use a continuum version of this concept: $S_C(A)$ = number of species found in a circular area around a given origin averaged

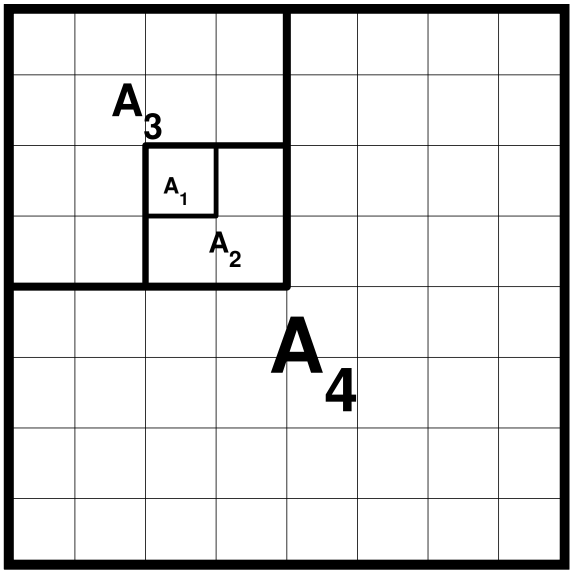


Figure 4.1: Traditional way of measuring the SAR for a given area. The area is divided in a grid and the number of species found is averaged over all possible nonoverlapping areas: $S_T(A1)$ is the average number of species found in nonoverlapping areas A_1, $S_T(A2)$ is the average number of species found in nonoverlapping areas A_2 and so on. One can see from this picture that this is similar to the procedure shown in figure 4.2. There will be, though, some differences for the areas not being centered on each other.

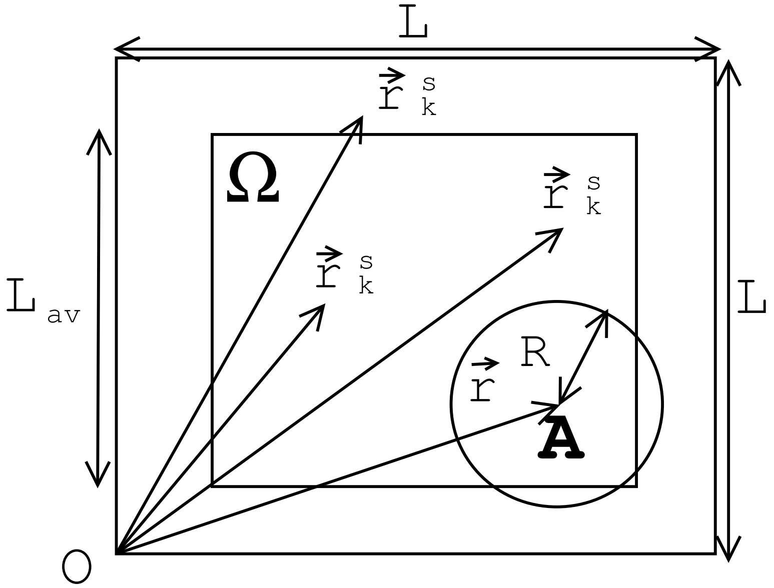


Figure 4.2: $S_C(A)$ is defined as the number of species found inside a circular area A of radius R around $\vec{r}$ and then averaged over $\vec{r}$ inside the area Ω. This is the content of equation 4.3.

over all possible origins. This is depicted in figure 4.2. We find this definition equally reasonable and more useful for a intuitive understanding of the SAR. The difference between these two definitions will in the end be minimal, and the fractal structure of the SAR will be preserved in both (see fig. 4.10).

The continuum version S_C has the advantage that it can be expressed in a more formal way as shown below in equation 4.3. This formal expression is convenient and will allow us to perform analysis without recourse to the use of intuition and graphical visualization.

Assume there are S species with n_s individuals ($s = 1..S$) in the area Ω (see figure 4.2) and the position vector for the kth individual belonging to species s is r_k^s ($k = 1..n_s$). Define the microscopic density of individuals to be:

$$\Gamma \equiv \sum_k \int_A \delta(\vec{r'} - \vec{r_k^s}) \tag{4.2}$$

Then the number of species within radius R is given by:

$$\begin{aligned} S_C(A) &= \frac{1}{L^2} \int_\Omega d\vec{r} \sum_s \Theta \overbrace{\left(\sum_k \int_A \delta(\vec{r'} - \vec{r_k^s}) d\vec{r'} \right)}^{\Gamma} \qquad (4.3) \\ & \Theta(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases} \\ & A = \{ \vec{z} \in \mathcal{R}^2 \ \mid \ |\vec{z} - \vec{r}| < R \} \\ & \Omega = \{ (x, y) \in \mathcal{R}^2 \ \mid \ 0 < x < L \, , \, 0 < y < L \} \end{aligned}$$

Note than in equation 4.3 the argument of the Θ function[1], will only be different from zero if any of the r_k^s are within the integration area A. In that case, the value of the Θ function will be one and the sum of these factors for all species is effectively

[1] The Θ or step function is defined as follows $\Theta(x) = 1$ for $x > 0$, zero otherwise. Related to the step function is Dirac's delta [121]. This function is defined such that $\int_A \delta(\vec{r'} - \vec{r}) d\vec{r} = 1$ if $\vec{r'}$ is inside A, zero otherwise.

the number of species that have at least one individual within a radius R of $\vec{r}$. The integration over Ω carries over the averaging over $\vec{r}$.

4.3 The Species Area Rule and extremal statistics

We will now link the SAR to the statistics of extremes [122],[123]. Intuitively, it is clear that for a given $\vec{r}$ (see fig. 4.2) the contribution of a single species to the total number of species in A will only depend on the closest individual to $\vec{r}$. Once a single individual is present in A, the location of the rest is irrelevant. For the purpose of the SAR, only whether *any* individual is present matters, not how many.

This can be proved rigorously by writing each of the factors in Γ as follows:

$$\int_A \delta(\vec{r'} - \vec{r_k^s}) d\vec{r'} = \Theta(|\vec{r_k^s} - r| - R) \tag{4.4}$$

since, as stated before, the integral will be unity if r_k^s is in the integration area A (within radius R of $\vec{r}$) and zero otherwise.

Now, for $R = 0$, all of the $\Theta(|\vec{r_k^s} - r| - R)$ are zero since none of the individuals are within a radius zero of $\vec{r}$ (assuming none of them is present at $\vec{r}$). As R increases $\Theta(|\vec{r_k^s} - r| - R)$ for the closest $\vec{r_k^s}$ will "switch on" while the rest are still zero. The precise value of R at which the rest will "switch on" is irrelevant for the purposes of equation 4.3, since Γ is the argument of a step function. Therefore the Θ function can be written:

$$\Theta\overbrace{\left(\sum_k \Theta(|\vec{r_k^s} - r| - R)\right)}^{\Gamma} = \Theta(r_m^s - R) \tag{4.5}$$

$$r_m^s = \min\{|\vec{r_k^s} - \vec{r}|, k = 1 : N_s\}$$

with r_m^s being the minimum distance to $\vec{r}$ for all individuals in species s. Equation 4.3 then becomes:

$$S_C(A) = \frac{1}{L^2}\int_\Omega d\vec{r} \sum_s \Theta(r_m^s - R) \tag{4.6}$$

Changing the order of sum and integration over the finite domain, we obtain:

$$S_C(A) = \sum_s \frac{1}{L^2} \int_\Omega d\vec{r}\, \Theta(r^s_m - R) = \sum_s F^s(R) \tag{4.7}$$

$$F^s(R) \equiv \frac{1}{L^2} \int_\Omega d\vec{r}\, \Theta(r^s_m - R) \tag{4.8}$$

Since $\Theta(r^s_m - R)$ is one for $r^s_m \leq R$ and zero otherwise, $F^s(R)$ is the fraction of the area Ω for which the closest individual of species s is within radius R; alternatively, this is the probability that selecting a random point from Ω the closest individual of species s is within radius R.

This result is exact and valid for any $S_C(R)$, not only fractal ones. The functions $F^s(R)$ can be calculated for a given distribution and their sum would exactly give $S_C(R)$.

Note the importance of the extremal character of the SAR. Bulk changes in the distribution of individuals may be irrelevant for the SAR as long as the minimum distance individuals stay the same. Conversely, changes in a small group of well selected individuals can change the SAR dramatically. This nonlinear characteristic will prove to be important when considering the robustness of the SAR in the following sections.

4.4 Modelling the distribution of individuals

In order to make further progress, and to calculate the functions $F^s(R)$, it is necessary to be able to model the distribution of individuals for each of the species present. This section presents a computer model that allows us to calculate $F^s(R)$.

Data from thoroughly sampled plots of several tropical forests [89] [124] show that individuals cluster in such a way that their correlation function $\rho_s(r)$ [2] (equivalent to

[2] The correlation function $\rho_s(r)$ is defined as the probability that two individuals of species s are at a distance $\vec{r}$ from each other.

the relative neighborhood density $\Omega_s(r)$ in references [89]) is close to a power law.

To properly mimic individual distributions, we will therefore use a bisecting tree algorithm that creates individual distributions whose correlation functions are power laws with exponents ω_s as shown in figure 4.3. The algorithm is depicted in figure 4.4, to which the reader is referred. An initial point r_i is chosen randomly within a distance d of the origin. Then, with probability α_s ($\equiv 2^{-\omega_s/2}$) only a left branch springs, with probability $1-\alpha_s$ only a right branch and with prob. $2\alpha_s - 1$ both branches spring. The process is repeated m times. The initial direction of the branches is chosen randomly and subsequently, after n splits, it is $\theta_n = \pi/2 + \delta_n$ radians with respect to the previous direction, where δ_n is randomly chosen between $\delta_n^{max} = \delta_0(\Delta\delta)^{2int(n/2)/m}$ and $-\delta_n^{max}$. Here, $int(n)$ represents the integer part of n. The length of the branch is halved in each branching: $l_n = l/2^n$. The final branch length is $l_m = 1$. This is a randomized, square-area version of the original construction by Harte et al. [17]. As such, it covers the whole area but because of the randomization of r_i and the angle variables it doesn't show artifacts due to the grid as the original version did (see fig. 3.6). For all fractal distributions in this paper $\delta_0 = 0.1$ rads, $\Delta\delta = 8$ and $m = 14$. The initial displacement was $d = 5$ and only the central 80×80 section was picked for analysis to avoid edge effects. Therefore, the lengths in fig. 4.2 are $L = 80$ and $L_{av} = 20$, giving an average over 400 sites. Our algorithm is a continuous and randomized version of the one proposed by Harte et al.[17], and avoids some of the problems pointed out by Maddux [97].

A pool of 6000 species (500 for each value of ω_s) was created. We will see in the following sections how to choose among this pool in order to get a power law SAR. Each individual distribution is characterized by its cover c (also known occurrence or range, see fig. 4.5) since this is the relevant quantity for the SAR: it doesn't matter how many individuals there are in the minimum area as long as there is at least one. In the end, we will be taking the length of the minimum displacement in the algorithm

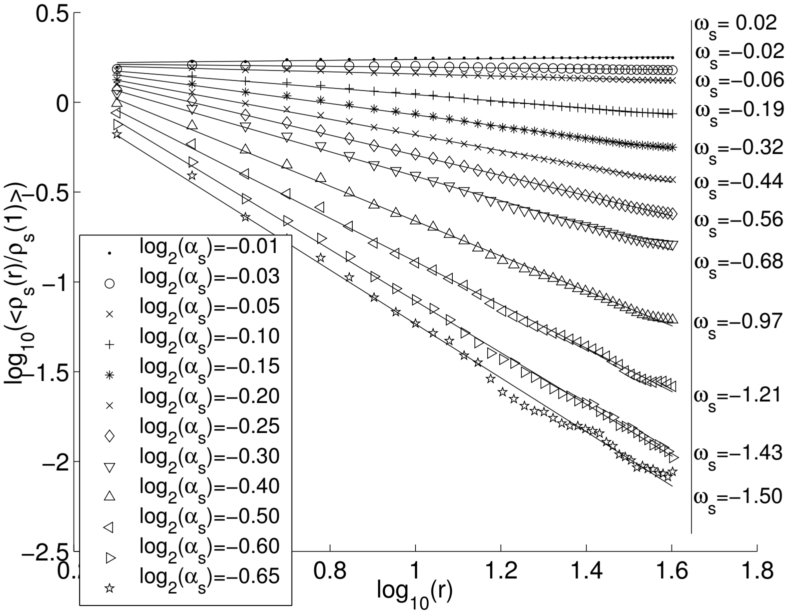


Figure 4.3: Correlation functions $\rho_s(r)$ for the fractal distribution of individuals. The average number of individuals of species s at a distance r from another individual of the same species is $2\pi r \rho_s(r)$. The correlation functions have been normalized by $\rho_s(1)$ and averaged over each of the 500 species for each α_s. In theory, ω_s should be equal to 2 $\log_2(\alpha_s)$ [89] but the randomization introduced in the individual distribution algorithm (fig. 4.4) and the finite size of the distribution change that value slightly (especially for low and high values of α_s). Plots have been offset for clarity.

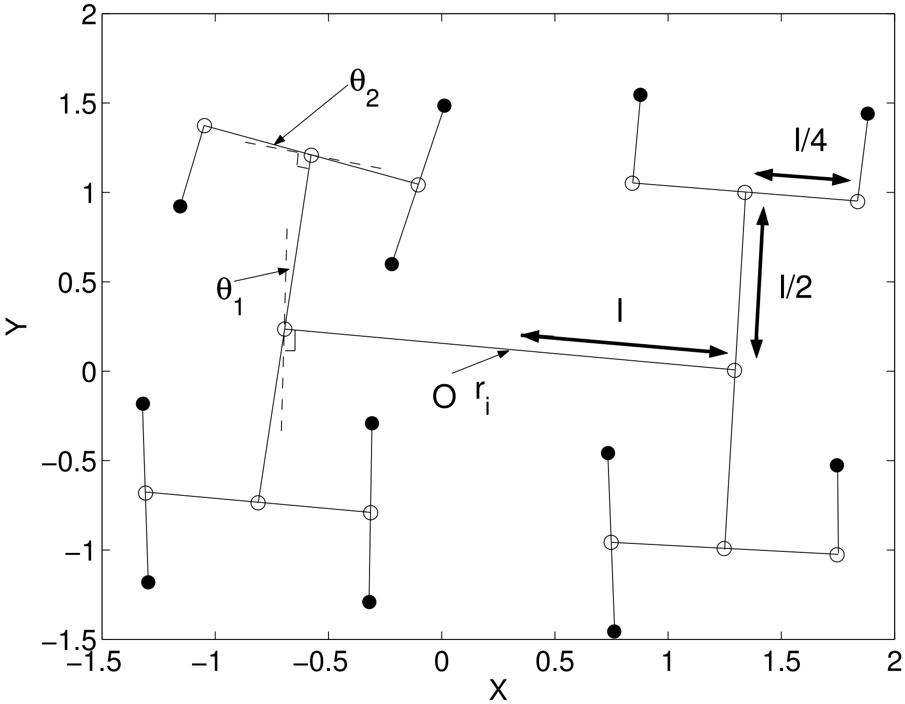


Figure 4.4: Graphical description of the algorithm to create individual distributions with fractal correlation functions $\rho_s(r)$. An initial point r_i is chosen randomly within a distance d of the origin. Then, with probability α_s ($\equiv 2^{-\omega_s/2}$) only a left branch springs, with probability $1-\alpha_s$ only a right branch and with prob. $2\alpha_s - 1$ both branches spring. The process is repeated m times. The initial direction of the branches is chosen randomly and subsequently, after n splits, it is $\theta_n = \pi/2 + \delta_n$ radians with respect to the previous direction, where δ_n is randomly chosen between $\delta_n^{max} = \delta_0(\Delta\delta)^{2int(n/2)/m}$ and $-\delta_n^{max}$. Here, $int(n)$ represents the integer part of n. The length of the branch is halved in each branching: $l_n = l/2^n$. The final branch length is $l_m = 1$. This is a randomized, square-area version of the original construction by Harte et al. [17]. As such, it covers the whole area but because of the randomization of r_i and the angle variables it doesn't show artifacts due to the grid as the original version did (see fig. 3.6). For all fractal distributions in this paper $\delta_0 = 0.1$ rads, $\Delta\delta = 8$ and $m = 14$. The initial displacement was $d = 5$ and only the central 80×80 section was picked for analysis to avoid edge effects. Therefore, the lengths in fig. 4.2 are $L = 80$ and $L_{av} = 20$, giving an average over 400 sites.

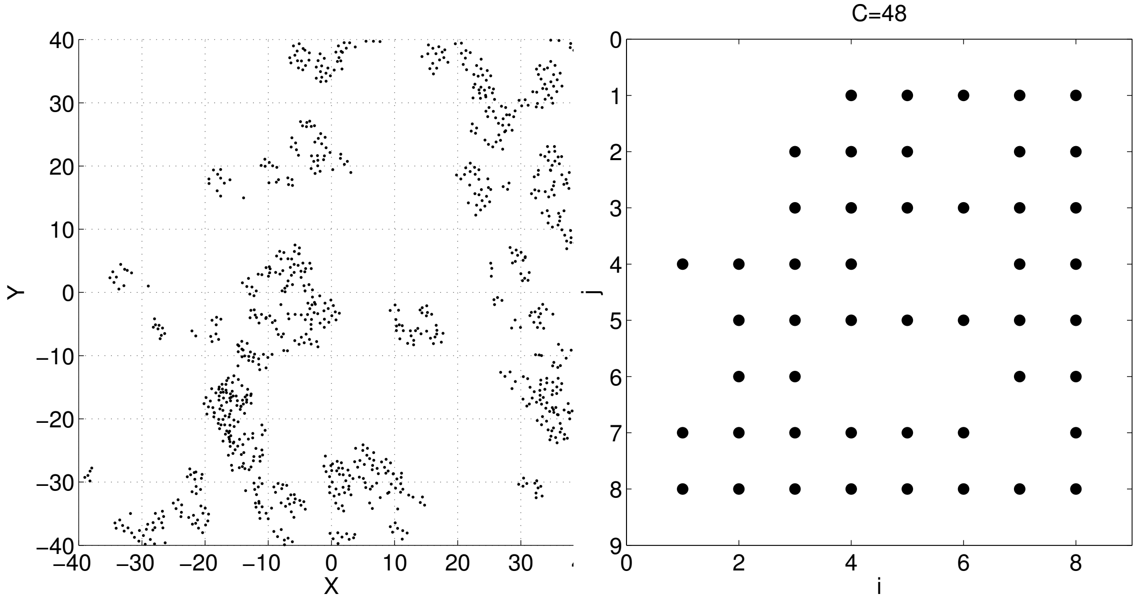


Figure 4.5: The left figure is an example of distribution generated through the algorithm explained in figure 4.4. The right figure is the presence matrix for this distribution. The matrix is found by imposing a grid on the distribution of individuals and making the value of the matrix one if any individual is present in the grid cell $\{i, j\}$, zero otherwise. The cover is the number of nonempty cells, in this case 48. Throughout this paper we will take the cell size to be the unity (here it was taken to be ten times that for demonstration purposes) and the cover and abundance will be equivalent.

explained in figure 4.4 as the size of the minimum area so cover and abundance will be the same. We will use them indistinguishably in the rest of the paper: abundance distributions and cover distributions will therefore be equivalent.

We are now in position to calculate $F_s(R)$ out of these allocations of individuals. $F_s(R)$ is the extremal distribution of the distances to the closest individuals from $\vec{r} \in \Omega$. Think of each $\vec{r}$ as a different draw of $n_s - 1$ distances to all other individuals of the same species, out of which only the minimum distance is kept. Repeating the procedure for each $\vec{r}$ yields a distribution of minimal distances.

These distributions have been heavily studied for the case of the distributions of extreme points (maximum or minimum) from n *independent, identically distributed* draws of an arbitrary distribution function $v(x)$. For this case, the extremal distribution is a universal function: one of the three classes of Fisher-Tippett distribution [125],[123],[122].

That result is not applicable here. The distances from a point in the area Ω to each of the individuals are not independent and identically distributed because they are clustered. If an individual is at a distance R from $\vec{r}$ it is very likely that other individuals are at a distance similar to R: they are strongly correlated. In the same way, the set of distances from a point $\vec{r}$ is correlated with the set of distances from a nearby point $\vec{r'}$.

It is, however, still possible to reduce all of them to a scaling function of the kind:

$$F^s(R) \simeq F\left(\frac{R - \langle R\rangle^s}{\sigma_s}\right) \tag{4.9}$$

where F is a universal function, $\langle R\rangle^s$ is the average of $f^s(r) \equiv \frac{dF^s(R)}{dR}\Big|_{r=R}$ and σ_s its variance (see figure 4.6):

$$\begin{aligned} \langle R\rangle^s &\equiv \int f^s(r) r dr \\ \sigma_s^2 &\equiv \int f^s(r)(r - \langle R\rangle^s) dr \end{aligned} \tag{4.10}$$

The function $f^s(r)$ is interpreted as the fraction of area for which the closest individuals are at a distance r. $\langle R\rangle^s$ is hence the average minimum distance to an individual of species s from points in Ω.

We have used a specific algorithm to obtain fractal correlation functions but our subsequent results seem to indicate that any other would have been acceptable as long as it reproduces aggregation, which is a common characteristic of populations [101]. In fact, we will see in section 4.8 that the specific shape of the correlation function will not be significant as long as it represents individuals clustering.

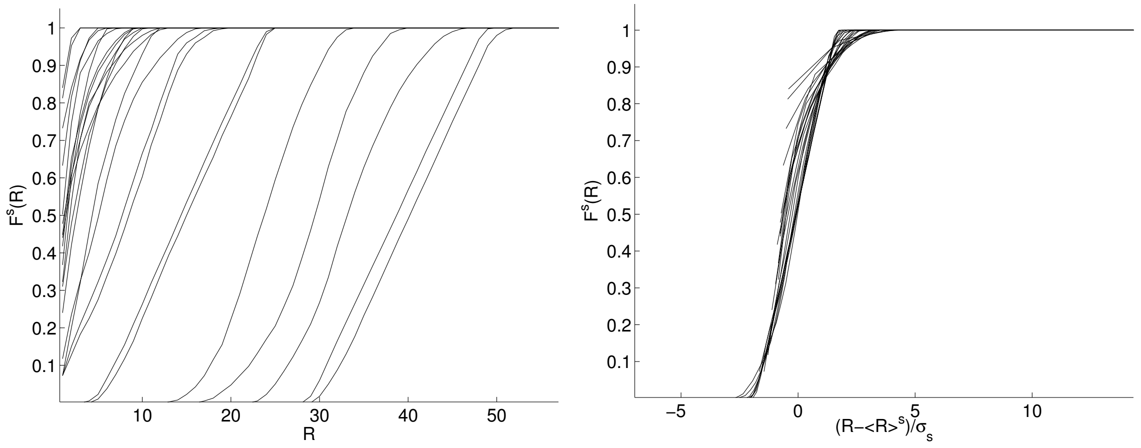


Figure 4.6: Data collapse of $F^s(R)$: the figure on the left is a plot of $F^s(R)$, the fraction of the area Ω for which the closest individual of species s is within radius R, without scaling. The figure on the right is the same plot, but the functions have been scaled by the variance σ_s and centered around the mean minimum distance $\langle R \rangle^s$. As can be seen they all collapse to the same function. Two species for each ω_s have been randomly chosen for this plot.

4.5 The scale-free SAR

In this section we will show that the distribution of $\langle R \rangle^s$ determines the SAR. We will also discuss the constraint distributions in order to obtain a power law for the SAR. From equation 4.7 and 4.9:

$$
\begin{aligned}
S_C(A) = S_C(R) &\simeq \sum_s F\left(\frac{R - \langle R \rangle^s}{\sigma_s}\right) \\
&\simeq \int_0^{R_m} D(r)\ F\left(\frac{R-r}{\sigma(r)}\right) dr \qquad (4.11)
\end{aligned}
$$

where $D(r)$ is the number of species with $\langle R \rangle^s \in (r, r+dr)$ and R_m is the maximum $\langle R \rangle^s$. If the variances $\sigma(r)$ were identically zero, $F((R-r)/\sigma(r))$ would become step functions and $S_C(R)$ would be the integral of $D(r)$. To see this, substitute $F\left(\frac{R-r}{\sigma(r)}\right) = \Theta(R-r)$ in the equation above and take the derivative with respect to R. The step function will become a Dirac delta and the integral will yield $D(r)$. Hence, obtaining a power law SAR of the kind $S_C(R) \propto R^{2z}$ would require $D(r) \propto r^\gamma$ $(z = 2(\gamma - 1))$.

Even if the variances were not zero, assuming a fractal $D(r) \propto r^\gamma$ and a linear $\sigma(r)(= ar)$ would similarly lead to a fractal SAR. For an arbitrary b ($b > 1$ without loss of generality):

$$\begin{aligned}
S_C(R/b) &\simeq \int_0^{R_m} D(r)F\left(\frac{R/b - r}{\sigma(r)}\right) dr \\
&= \int_0^{R_m} D(br/b)F\left(\frac{R - br}{\sigma(br)}\right) dbr/b \\
&\simeq b^{\gamma-1}\int_0^{bR_m} D(r')F\left(\frac{R - r'}{\sigma(r')}\right) dr' \\
&= b^{\gamma-1}\int_0^{R_m} D(r')F\left(\frac{R - r'}{\sigma(r')}\right) dr' \\
&+ b^{\gamma-1}\int_{R_m}^{bR_m} D(r')F\left(\frac{R - r'}{\sigma(r')}\right) dr' \\
&\simeq b^{\gamma-1}S_C(R) \qquad (4.12)
\end{aligned}$$

where the second integral in the line before the last is null because $D(r) = 0$ for $r > R_m$.

In reality, the $\sigma(r)$ function is neither negligible nor exactly linear. $\sigma(r)$ displays a more or less linear dependence until it saturates due to finite size effects (see fig. 4.7). It is better fitted to a logarithmic dependence $\sigma(r) \simeq a\log(r)$. This means that, even though $S_c(R)$ is determined by the distribution of average distances, a power law $D(r)$ is not equivalent to a fractal SAR. It does provide, however, an excellent starting point to obtain the correct $D(r)$. The next section focuses on which species abundance distributions naturally produce $D(r)$ distributions such that $S_C(R) \propto R^{2z}$.

4.6 Abundance distributions compatible with the fractal SAR

We will now show how the distribution of mean minimum distances $D(r)$ that produces a scale-free SAR will arise naturally from a typical abundance distribution.

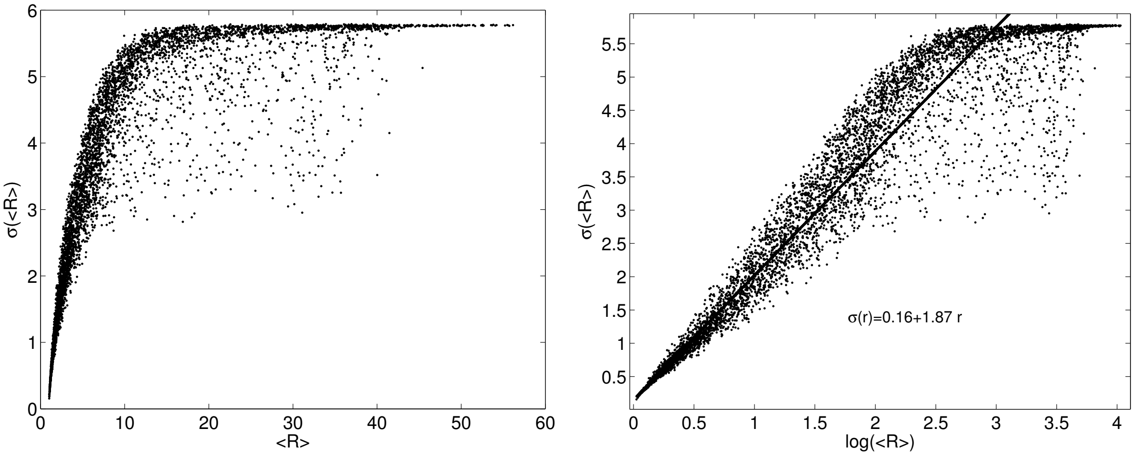


Figure 4.7: The variances of the mean minimum distances $\langle R\rangle^s$ initially grow close to linearly with $\langle R\rangle^s$ until they hit finite size effects. In fact, they fit better to a logarithmic dependence with $\langle R\rangle^s$. Line shown is not best fit, just a guide to the eye.

Intuitively, it is clear that the average minimum distance to an individual of species s, $\langle R\rangle^s$, should depend on the coverage c: for higher coverages the probability of finding an individual of that species within a radius R or $\vec{r}$ is greater than for smaller values of R. One could also imagine $\langle R\rangle^s$ being dependent on the correlation function exponent ω_s. In practice the only dependence on ω_s seems to be indirectly through the coverage (see fig. 4.8): higher $|\omega_s|$s (more tightly clustered) will tend to have lower coverages and therefore higher $\langle R\rangle^s$. Each species in the pool will thus be characterized exclusively by its cover c.

We define $g(r,c)$ as the fraction of species with coverage c that have $\langle R\rangle^s = r$. $D(r)$ is then related to $P(c)$ through the linear equation:

$$D(r) = \int_0^{Cmax} g(r,c)p(c)dc \tag{4.13}$$

Since $g(r,c)$ covers all space and is almost diagonal when properly binned (see fig. 4.9), by properly choosing $P(c)$ we can obtain any desired fractal $D(r)$ with exponent $\gamma < 0$.

For all practical purposes, a discretized version of equation 4.13 will be used from

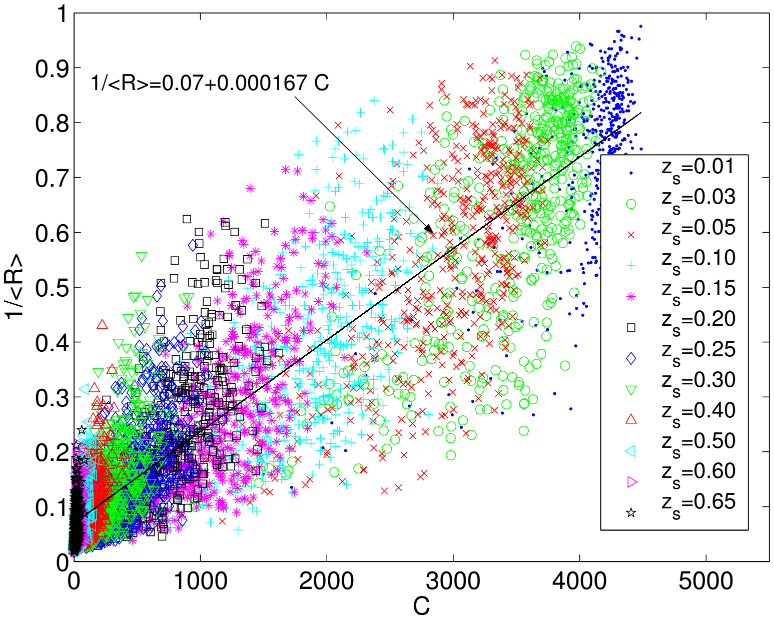


Figure 4.8: Dependence of average minimum distance $\langle R\rangle^s$ with the cover c. As expected it decreases with c. The fractal correlation function exponent seems to have little other effect other than determining the cover. We will therefore characterize each species by its cover.

now on. For the binning of both r and c we will use a logarithmic scale of doubling intervals: $[1,\ 2-3\ ;\ 4-7\ ;\ 8-15\ ;\ 16-31\ ;\ 32-63\ ;\ 64-127\ ...]$. Hence the discrete version of $g(r,c)$, $G_{i,j}$, represents the fraction of species with coverage $c \in [2^{j-1}, 2^j)$ and $\langle R\rangle^s \in [2^{i-1}, 2^i)$ (fig. 4.9). By definition $\sum_i G_{i,j} = 1\ \forall\ j$.

There are several good reasons for this base two logarithm scale. For the r variable this is justified because its variance grows linearly with r until finite size effects prevent it (fig. 4.7). Therefore a logarithmic scale keeps $G_{i,j}$ as close to diagonal as possible. Secondly, we are interested in forms of $D(r)$ which are scale free or close to it. Since this means $D(r)$ decreases fast, it is advisable to use increasing bins to obtain well converged averages. For the case of the cover variable c, if we assume that for this level of resolution (see figure 4.5) c and n are equivalent, this binning is roughly equivalent to the octave classification proposed by Preston and commonly used in ecology for abundance distributions [27]. Since populations tend to increase geometrically, the natural variable to be considered is the logarithm of the abundance [26].

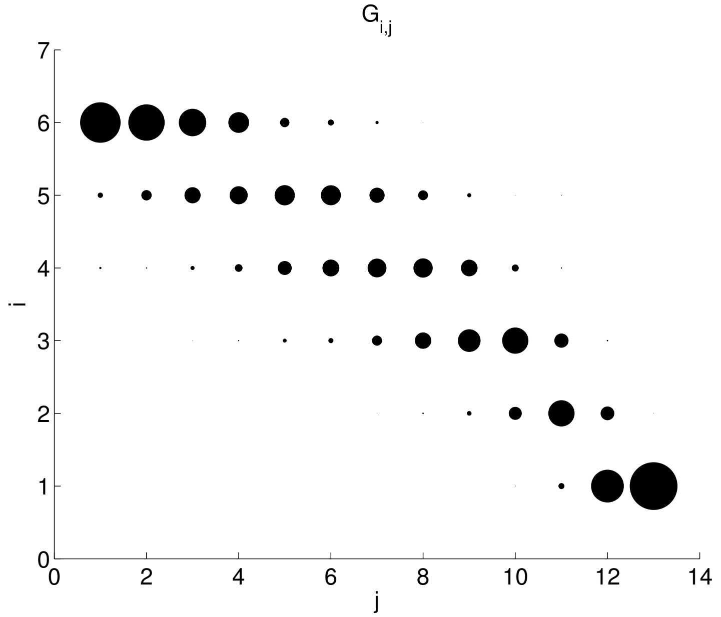


Figure 4.9: Matrix **G** binned in a logarithmic series of doubling intervals. $G_{i,j} \equiv$ fraction of species with coverage $c \in (2^{j-1}, 2^j]$ and $\langle R \rangle^s \in (2^{i-1}, 2^i]$. This logarithmic scale is justified for the r variable because its variance grows linearly with r until it encounters finite size effects (fig. 4.7) and for c, because is equivalent to the abundance n at this level of resolution (see fig. 4.5) and the geometrical increase of populations make the logarithm of the abundance the natural variable to be considered. By definition $\sum_i G_{i,j} = 1 \; \forall \; j$.

In terms of these doubling intervals equation 4.13 becomes:

$$D_i = \sum_{j=1}^{13} G_{i,j} P_j \quad i = 1, ..., 6 \tag{4.14}$$

Since $G_{i,j}$ is a 6×13 matrix and is close to diagonal, its rank is 6. Equation 4.14 is, hence, an underdetermined linear system [126]. This means that there is always a solution P_j^* to equation 4.14 such that:

$$D_i^* \equiv q(r_i)^\gamma = \sum_{j=1}^{13} G_{i,j} P_j^* \quad i = 1, ..., 6 \tag{4.15}$$

where $r_i \equiv 2^i$ and q is a normalization constant such that $\sum_{i=1}^{6} D_i = 1$. Furthermore, this solution is not unique. There exist $13 - 6 = 7$ vectors $\bar{P}_j^\alpha$ ($\alpha = 1, ..., 7,\ j = 1, ..., 13$) such that (in matrix notation) $\mathbf{G}\bar{\mathbf{P}}^\alpha = \mathbf{0}$ and therefore:

$$\mathbf{P}' \equiv \mathbf{P}^* + \sum_{\alpha=1}^{7} B^\alpha \bar{\mathbf{P}}^\alpha \tag{4.16}$$

is a solution to equation 4.15 for *any* choice of B^α.

This is not an artifact of the binning we have chosen. Had we chosen a less coarse description, the matrix would have considerably larger diagonal blocks and it would be possible to choose vectors among these subspaces that would give the same result when acted upon by the matrix $\mathbf{G}$. This wide range of solutions is a consequence of the variances growing linearly with $\langle R \rangle^s$, in turn a repercussion of the extremal nature of the SAR and the dimensionality.

This wide range of solutions is not a consequence of the specific size of system chosen, either. For different system sizes L the maximum value of $\langle R \rangle^s$ will scale as L and the maximum value of the coverage will scale as L^2. Hence, the linear systems will be underdetermined for any system size.

However, not all solutions to equation 4.15 are ecologically. Since P_j' represents probabilities we must demand positivity and unitarity:

$$P_j' \geq 0 \tag{4.17}$$

$$\sum_j P_j' = 1 \tag{4.18}$$

Additionally, one would expect the P'_j to have a smooth dependence on j in a stable ecosystem, so we would require that they do not vary wildly with j by choosing a tolerance t and impose:

$$\sum_{j=1}^{12}(P'_j - P'_{j-1})^2 < t \tag{4.19}$$

These conditions make the difference regarding the measure of probability distributions compatible with each power law exponent z, as will be explained in section 4.7. The continuum equivalent of equations 4.17-4.19 would be:

$$p'(c) \geq 0 \quad \forall\ c \tag{4.20}$$

$$\int_0^{c_{max}} p'(c)dc = 1 \tag{4.21}$$

$$\int_0^{c_{max}} \left(\frac{dp'(c)}{dc}\right)^2 dc < t \tag{4.22}$$

In practice, to choose a vector P^*_j that gives rise to a SAR of the type $S \propto A^z$ a selection of $\langle R\rangle^s$ according to the distribution D^*_i is taken from the total pool, in an iterative process, described as follows. The coverage distribution of these selection is taken as an initial guess of P^*_j. This SAR is usually not quite linear in a log-log plot, but it is a good start. This guess is then refined by small changes in the P^*_j using the information in figure 4.9 to obtain the right distribution of $\langle R\rangle^s$ for an acceptable scale-free SAR[3].

The results can be seen in figure 4.10. The corresponding abundance distributions $P^\dagger_j$ are shown in figure 4.11. Notice the change in mode location and the spread. For small z large covers are preferred and the distribution spread over all possible covers. Conversely, for high exponents z, the mode is centered around low coverage and the spread of the distribution is much smaller. The distributions fit rather well to lognormals, except for the fact that they exhibit more rarity, which is a well known

[3] By acceptable, I take the criteria that the linear correlation coefficient $r > 0.99$. The linear correlation coefficient for a pair of quantities (x_i, y_i) is [127]:
$r = \sum_i \frac{(x_i-\hat{x})(y_i-\hat{y})}{\sqrt{\sum_i(x_i-\hat{x})^2}\sqrt{\sum_i(y_i-\hat{y})^2}}$

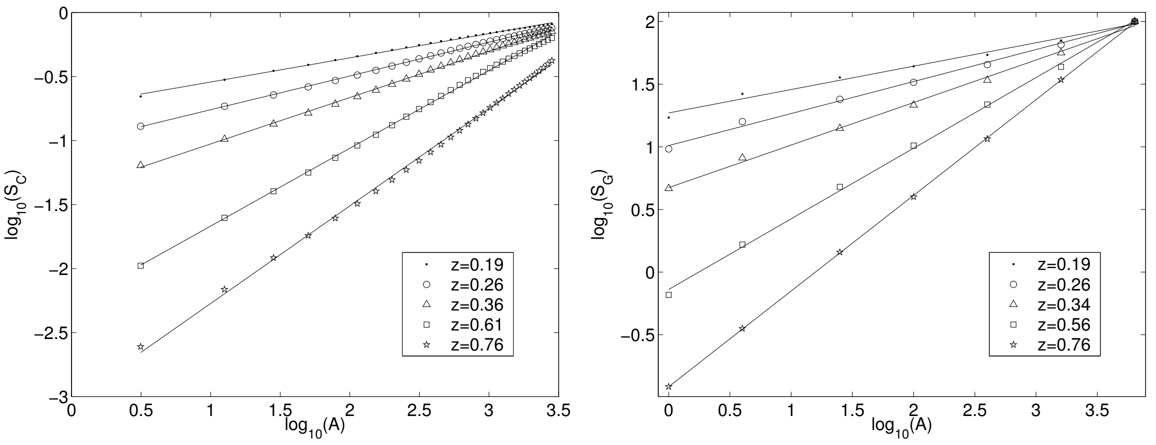


Figure 4.10: $S_C(A)$ and $S_G(A)$ for the distributions shown in fig. 4.11. As mentioned before, both share the fractal pattern. The exponents are mostly the same or very similar. Only in the case of $z = 0.19$ can some systematic deviation be appreciated.

characteristic of realistic abundance distributions [20] [17] (remember that we are considering the coverage and abundance distributions to be the same because of the small area of the censusing window, as explained in fig 4.5.).

4.7 The robustness of scale-free SARs

Everything stated above for the distribution of mean distances $D(r)$ can be applied to $S_C(R)$. It can also be written as a linear function of $p(c)$:

$$\begin{aligned} S_C(R) &= \sum_s F\left(\frac{R - \langle R\rangle^s}{\sigma_s}\right) \\ &= \int_0^{R_m}\int_0^{C_{max}} g(r,c)p(c)F\left(\frac{R-r}{\sigma(r)}\right)drdc \\ &= \int_0^{C_{max}}\int_0^{R_m} g(r,c)F\left(\frac{R-r}{\sigma(r)}\right)p(c)drdc \\ &= \int_0^{C_{max}} h(R,c)p(c)dc \end{aligned} \tag{4.23}$$

where:

$$h(R,c) \equiv \int_0^{R_m} G(r,c)F\left(\frac{R-r}{\sigma(r)}\right)dr \tag{4.24}$$

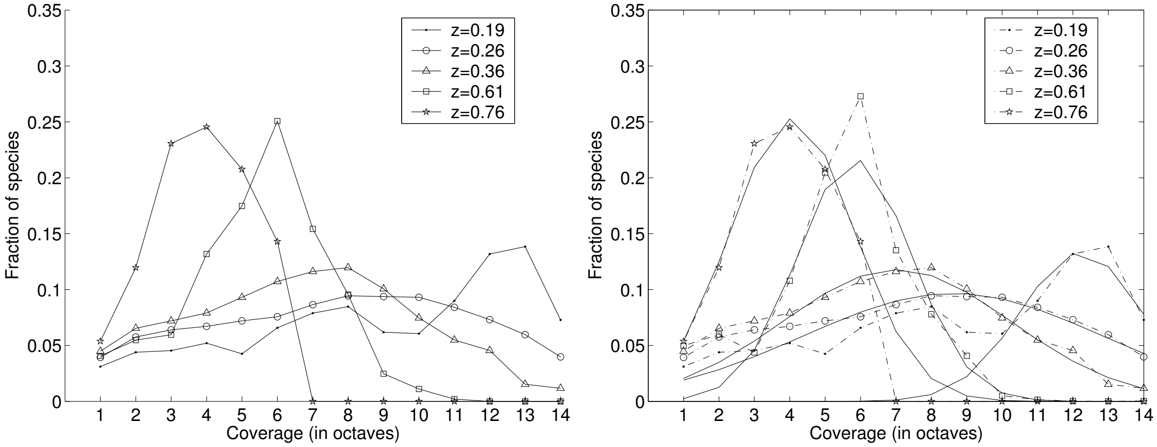


Figure 4.11: Abundance distributions for the different fractal SARs plotted following Preston's octave system. This is equivalent to the binning explained before, but species on the bin edges are divided equally between neighbors. Notice how the mode location and the spread change with z. For $z = 0.2$ we need a large fraction of the species with small $\langle R \rangle^s$. The way to assure that is to have a big fraction of species with high coverage and, therefore, low $\langle R \rangle^s$ (see fig. 4.8). Conversely, high zs $(= 0.6, 0.8)$ need higher $\langle R \rangle^s$, which means low coverage. Notice how the values $z = 0.2 - 0.4$ spread over all the possible coverages, making this values of z more likely (see fig. 4.12 and section 4.7). Right figure shows lognormal fits to the coverage distributions that produce power law SARs. The fit is rather good except for the low coverages, where the lognormal underestimates the fraction of species, as is typical of observed abundance distributions [20] [17].

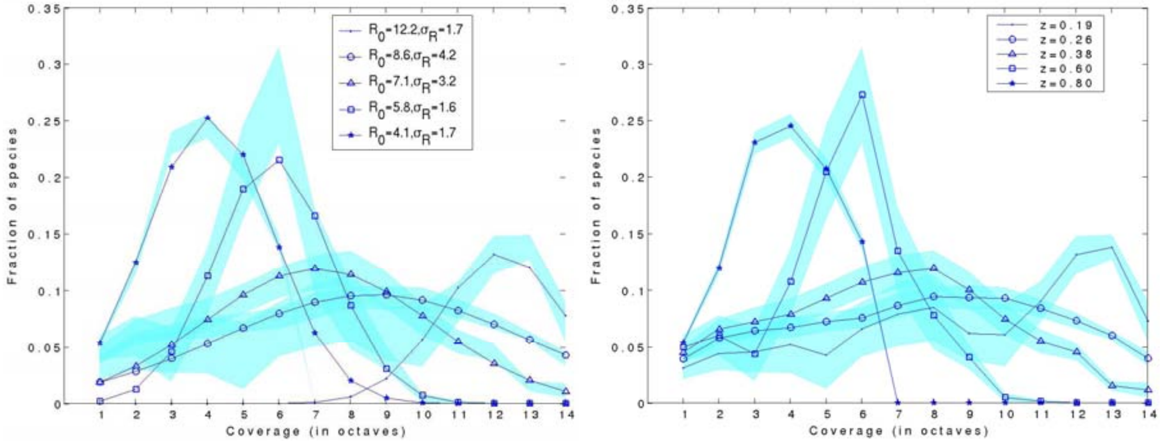


Figure 4.12: Plot of the relative abundances showing the robustness of the solutions. The semitransparent stripes indicate the variance of the abundance distributions that have the same scaling exponent z. The value of t in eq. 4.19 was chosen to be twice that of the original (blue line) distribution. The figure on the left shows the lognormal fits against the domain of abundance distributions that produce power laws.

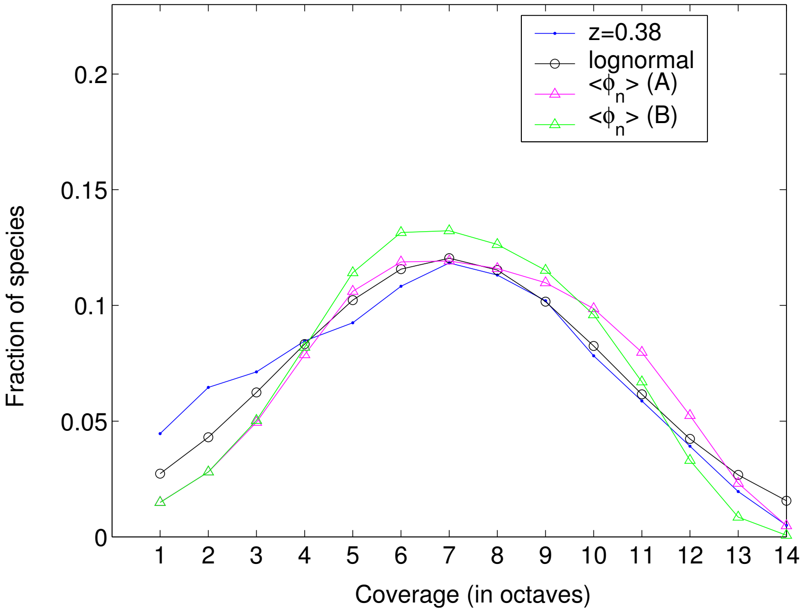


Figure 4.13: Comparison of the fit with a lognormal and with Hubbell's abundance distributions [20],[117]. Hubbell's distribution seems to decay too fast. For the lognormal, $R_0 = 7.0$ and $\sigma_R = 3.5$. For Hubbell's distributions: $\Theta = 10$, $m = 0.15$, $N = 10.000$ (A) and $\Theta = 7$, $m = 0.17$, $N = 12.000$ (B).

or in terms of the discrete binning:

$$\hat{S}_{Ci} = \sum_j H_{i,j} p_j \tag{4.25}$$

For the same reasons as mentioned above, equation 4.25 has at least one solution $\mathbf{P}^\dagger$ for a fractal SAR and there are again 7 vectors $\hat{\mathbf{P}}^\alpha$ that can be added to this solution to find *exactly* the same S_{Ci}. Obviously, these solutions must to obey conditions 4.17-4.19, too.

As announced previously, the appearance of the power law SAR is quite robust. There are several reasons for this. First, the number of abundance distributions that give rise to any fractal SAR seems to cover a large part of the probability phase space (figure 4.11). Therefore the occurrence of a fractal SAR is not something unavoidable, but is very likely to happen with reasonable abundance distributions. Second, any linear combination of the kernel vectors $\hat{\mathbf{P}}^\alpha$ gives exactly the same SAR. That increases the measure of probability distributions compatible with a fractal

SAR. In order to show the range of these equivalent solutions we have added random linear combinations of the vectors $\hat{\mathbf{P}}^{\alpha}$ to $\mathbf{P}^{\dagger}$ and selected those combinations that obey conditions 4.17-4.19. The results can be seen in fig. 4.12. Third, even the addition of vectors that do not belong to the kernel gives rise to distributions that would pass as power laws given the amount of noise that these plots usually have [12] (see section 2.3).

Interestingly, one can see why fractal power laws with exponents $z = 0.2 - 0.4$ are more common. Not only do their compatible distributions seem to cover more space, but the range of equivalent solutions through linear combinations of kernel vectors is larger. This happens because these distributions have non-null contributions from all coverages and, when adding kernel vectors, that tend to have non zero elements for all coverages, they are less likely to violate condition 4.17.

4.8 The importance of clustering

So far we have shown that the appearance of the fractal SAR is quite robust to changes in the abundance distribution: as far as it resembles a lognormal with more rarity, everything seems to be in qualitative and even semi-quantitative agreement with observations. We haven't, however, checked the robustness with regard to changes in the distribution of individuals. Which characteristic of the individual distributions is important for the power law SAR to appear? The answer is: clustering. As long as individuals cluster and their species relative abundance is as shown in the previous section we will find a fractal SAR, at least for the more stable distributions. It doesn't particularly matter whether this clustering is self similar, as for the distributions shown in section 4.4, or not.

To see why this is so, consider what would happen to $G_{i,j}$ if we took a fraction χ of the individuals for each species and redistributed them randomly. As emphasized in section 4.3, only the closest individuals matter when finding the average of the minimum distances $\langle R\rangle^s$. When individuals are redistributed randomly, the effect is that they become more evenly distributed over the area and for every position $\vec{r}$ (fig. 4.2) the closest individual tends, on average, to be closer. Hence, the minimum distances $\langle R\rangle^s$ decrease for a given cover as can be seen in figure 4.15. As χ increases, the elements $G_{i,j}$ corresponding to high $\langle R\rangle^s$ will be depleted and reasonable abundance distributions will produce an excess of $\langle R\rangle^s$ in the low and middle end. This will generate convex log-log plots (negative second derivative) as reported for random distributions [81].

All this seems to indicate that as long as individuals remain clustered it does not matter the exact way in which they are distributed, due to the insensitive nature of extremal statistics. One could then imagine that if we just divided the total number of individuals for each species into n_G groups and scatter them randomly, a similar $G_{i,j}$ would be found. This is exactly the case. We took the same distribution of coverages $p(c)$ and for each species divided the available cover into n_G groups. The number of groups n_G was randomly chosen between a minimum m_g and the maximum possible amount of groups M_G given the species coverage c and the group minimum size m_a ($M_G = c/m_a$). The leftover coverage units were clustered in groups of size equal to m_a and distributed randomly (fig. 4.14).

The results are very similar (fig. 4.16) to those obtained with self-similar distributions (figs. 4.6, 4.7, 4.8 and 4.9). The only difference seems to be that variances tend to be somewhat smaller for the clustered case. For $z = 0.2 - 0.4$, it is not hard to modify slightly the solutions for the self-similar distributions $\mathbf{P}^\dagger$ to obtain good fits to power-law SARs, but this is not the case for $z = 0.61, 0.76$.

The reason can be understood by inspecting the matrix $h(r)_j$ (fig. 4.18), which is the equivalent of $h(r,c)$ with binned coverages. Here, $S_C(R) = \sum_j h(R)_j P_j$ for a binned coverage distribution P_j. It can be seen that the fractal and cluster versions of

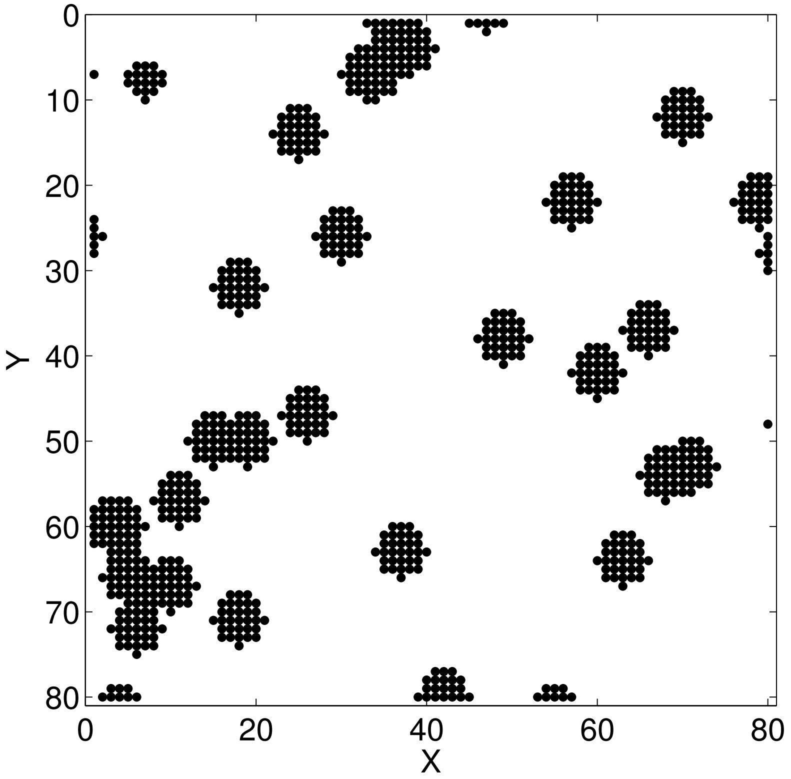


Figure 4.14: Cluster distribution. Given a cover c for a species, this was divided into an arbitrary number of groups n_G. This groups were aggregated into compact groups and spread randomly. The results regarding the SAR are very similar to those obtained with self-similar distributions (fig. 4.5). For all clustered distributions $m_a = 4$ and $m_g = 3$.

$h(r)_j$ are quite similar. The differences lie in two characteristics: firstly, for high j the cluster $h(r)_j$ trails the fractal one whereas for low j, the opposite occurs. Secondly, the cluster $h(r)_j$ rises much faster than the fractal ones. Since low values of z ($= 0.2 - 0.4$) have contributions from all j, these differences can compensate each other and can be counteracted by a slight change in the proportions P_j. On the other hand, higher values of z rely only on $h(r)_j$ using low values of j and in the cluster version of $h(r)_j$ these rise much faster (higher derivative) than in the fractal one, generating the curvature seen in figure 4.19. Changes in the parameters m_g, M_G, m_a do not seem to change any of this observations, although we do not disregard the possibility that there may exist combinations that are more successful. This will be object of a future analytical approach.

Here lies, hence, another reason for the commonness of values of z in the range $z = 0.2 - 0.4$: the power law SAR for these assemblages of species are more robust to differences in clustering.

4.9 Field data

We will now check the validity of our approach with real ecological data. This will also provide an example with low number of species where all the procedures explained above can be better followed intuitively.

The data set we use was collected and analyzed by Green et al.[69] at the Donald and Sylvia MacLaughlin University of California Natural Reserve (latitude $38^o51'N$, longitude $123^o34'W$) in northern Napa and southern Lake Counties, 120 km north of San Francisco, CA, USA. The particular site (known as Blue Ridge site henceforth) was 64 m^2 in area and included 37182 individuals and 24 plant species. The plot was divided into a square 16×16 grid and the sampling was performed from early May to late July 1998. This data was chosen because it is one of the most well-characterized

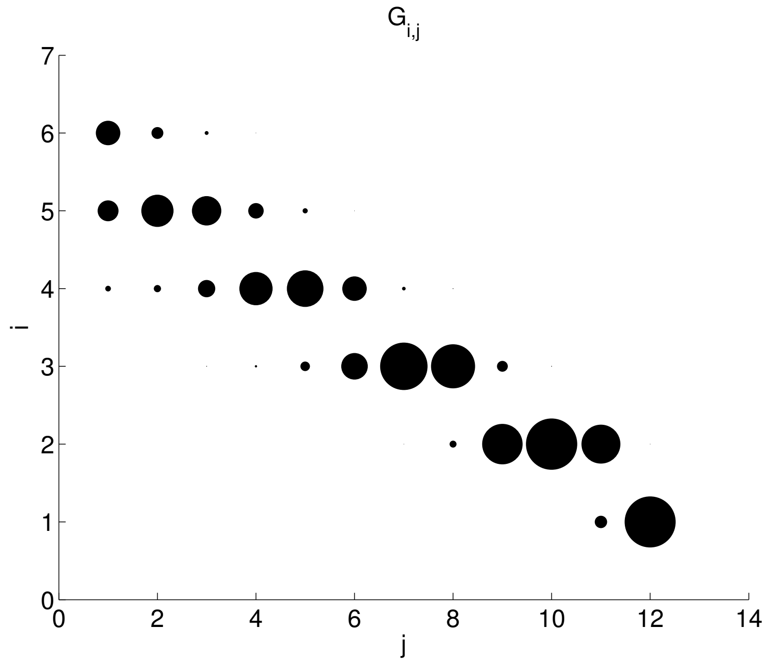


Figure 4.15: $G_{i,j}$ matrix for the fractal distributions with added noise. For each species a fraction $\chi = 0.5$ of the individuals has been randomly relocated. The average minimum distances $\langle R \rangle^s$ therefore decrease as can be readily observed from the comparison with figure 4.9.

in the literature: the site was completely sampled and a very good fractal SAR was reported (see fig. 4.20).

For this data set, the coverage distribution is divided between high and low cover, with few species in between (see fig. 4.21 and 4.23). In this case, the smallest sampling unit contains much more than one individual and abundance and cover are not equivalent, as can be seen in figure 4.22.

The scaling of $F^s(R)$ works well as expected (fig. 4.25), but the distribution of $\langle R \rangle^s$ is not fractal (fig. 4.24). This is a good example to show that $D(r)$ does not have to be fractal to obtain a power law SAR, and shows the importance of the nonzero variances to provide robustness to changes in $D(r)$. Let us comment on this.

Assume without loss of generality that the species number s has been chosen as the rank of $\langle R \rangle^s$ (i.e. the highest $\langle R \rangle^s$ corresponds to $s = 1$, the second one to $s = 2$... and so on). If the variances σ_s were identically zero then:

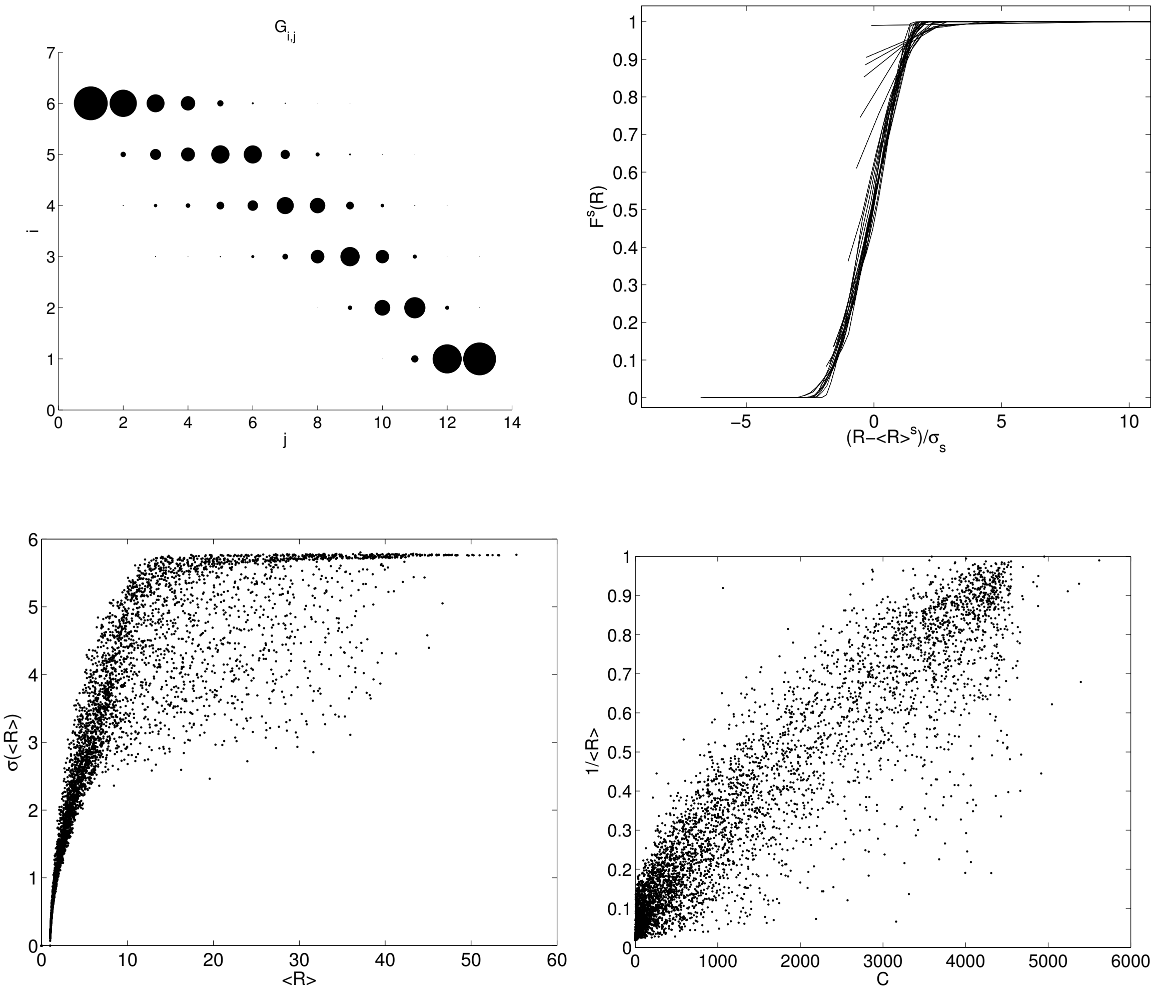


Figure 4.16: Results for the case in which individuals for each species are formed into clusters and spread randomly (fig. 4.14). Compare with figures 4.6,4.7,4.8,4.9 from the self-similar distributions explained in section 4.4. The results are very similar, except for the variances, which tend to be smaller for the clusters case. This produces the main difference in the matrix $h(r)_j$ (fig. 4.18).

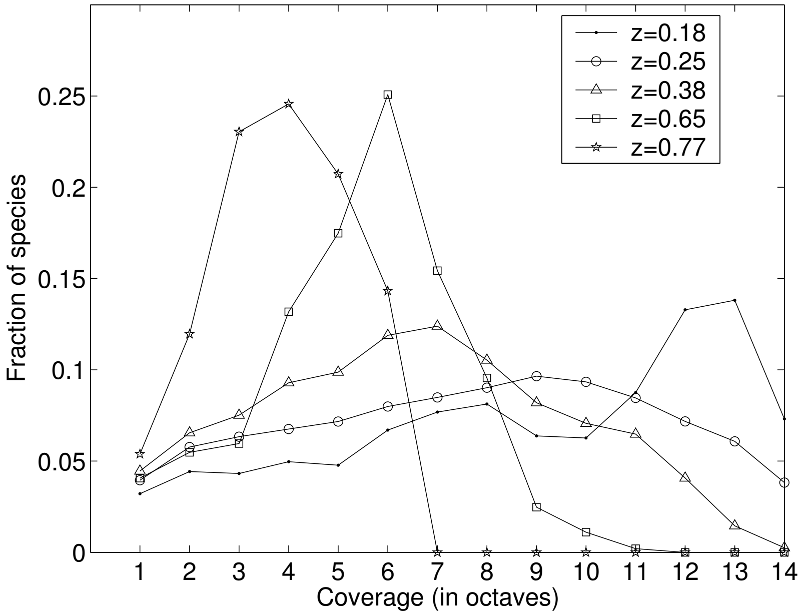


Figure 4.17: Abundances that produce the power law SARs in figure 4.19. They are basically identical to the case of self-similar distributions in fig 4.11.

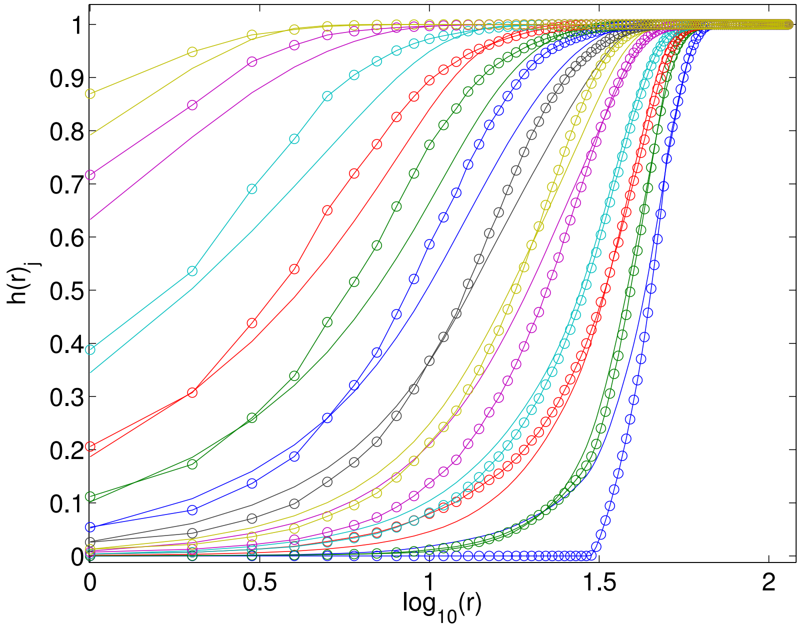


Figure 4.18: Matrix $h(r)_j$ for the self-similar $(-)$ and clustered (o) distribution of individuals. The index j decreases from left to right (high j indicates high coverage and $h(r)_j$ saturates earlier). Notice that for the clustered case $h(r)_j$ rises faster for low values of j (center and right part of the plot). This will produce the concave curvature in the power laws for $z = 0.65, 0.77$ in figure 4.19.

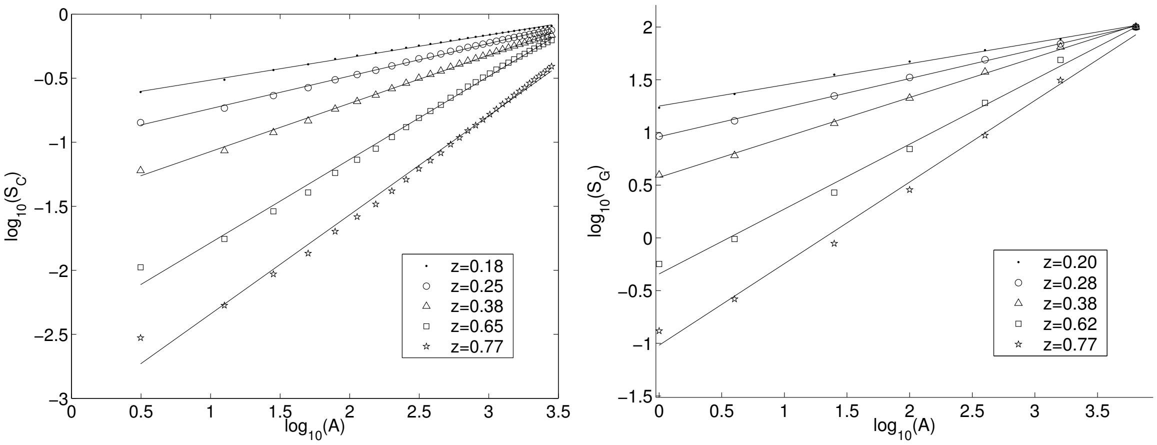


Figure 4.19: SARs for the clustered case. For $z = 0.2 - 0.4$ it is still possible to find coverage distributions to produce linear log-log plots, but the smaller variances (see fig. 4.16 and 4.18) produce concavity for higher values of z.

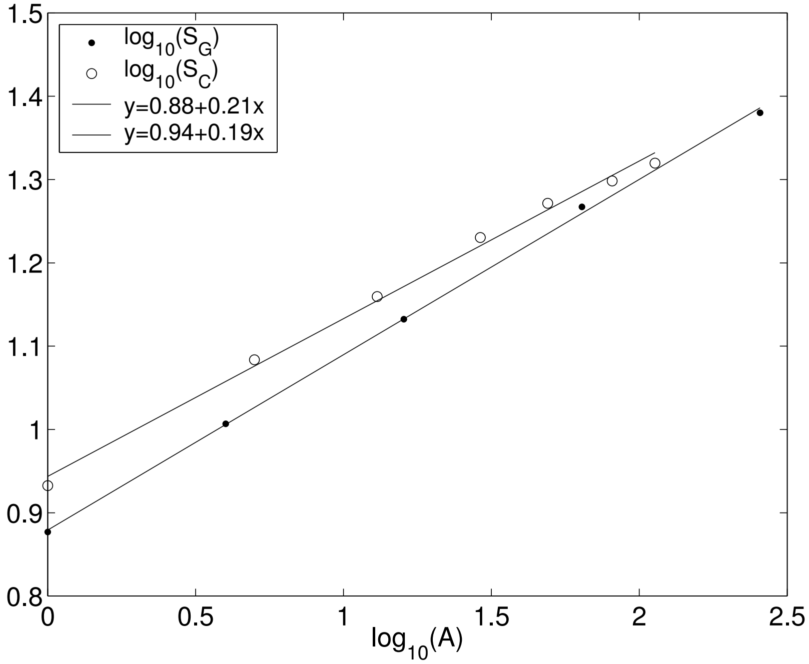


Figure 4.20: Comparison between both ways of measuring the SAR for the Blue Ridge data. Even though they differ slightly in quantitative terms, both still maintain a fractal nature. The variation in z is within the typical range when dealing with power law exponents.

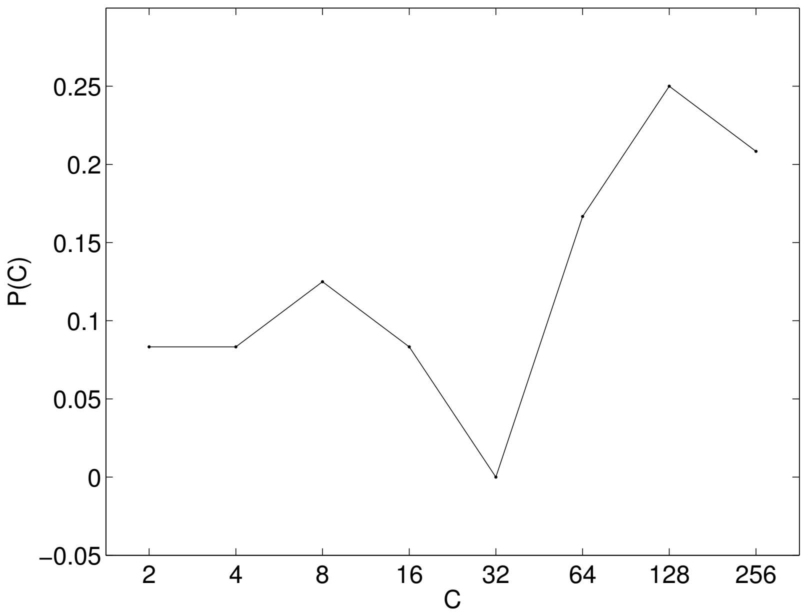


Figure 4.21: Distribution of coverages (C) for the Blue Ridge data. In spite of lacking species with medium C, we still obtain an excellent fractal SAR (fig. 4.20). Because of the high variances for $\langle R \rangle^s$ (fig. 4.26), the SAR is very robust to changes in $P(C)$.

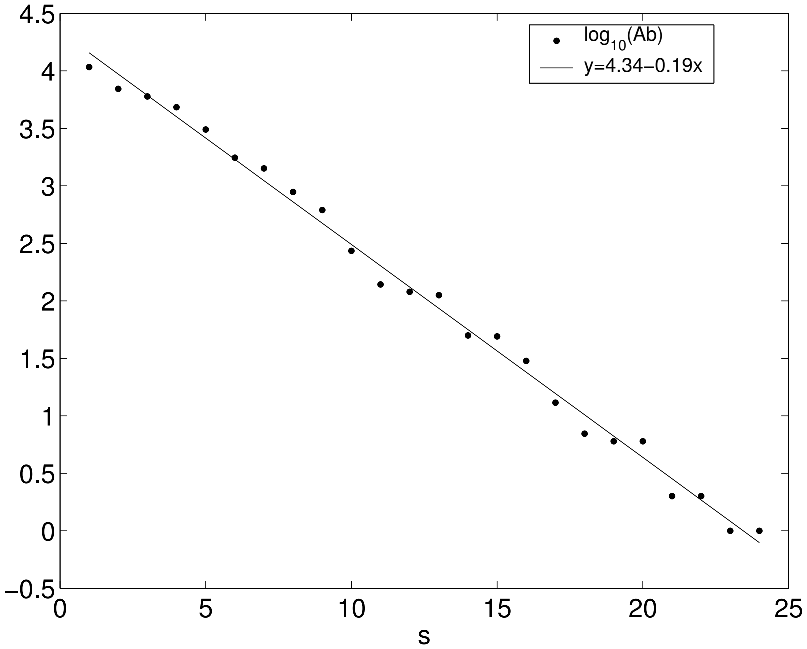


Figure 4.22: Ranked abundance for the field data set. In this case, abundance and cover are not equivalent, since the minimum sampling area is much bigger than the individual size.

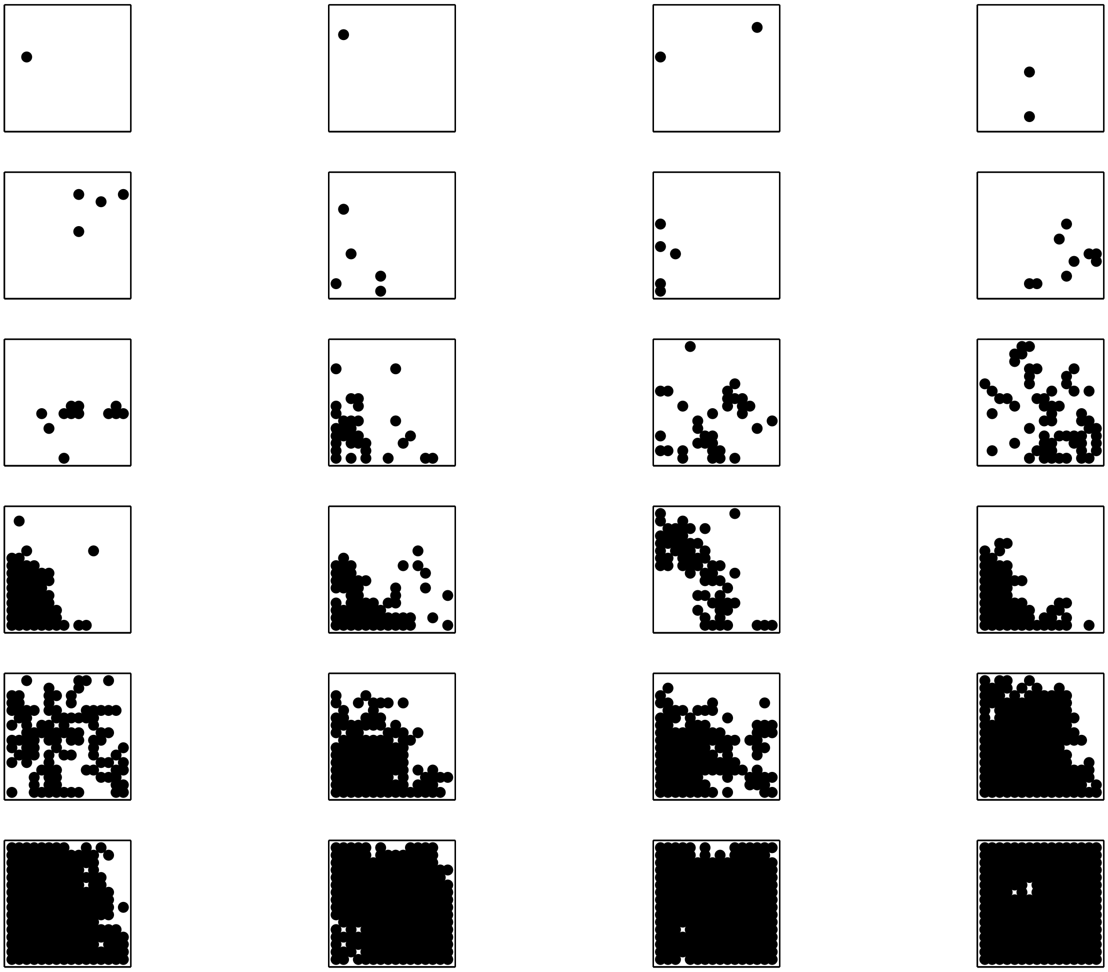


Figure 4.23: Spatial location for each of the species in the Blue Ridge data set in decreasing order of coverage. A matrix element is colored if an individual of that species is present in that position.

$$F^s(\langle R \rangle^m) = \begin{cases} 1 & m < s \\ 0 & m > s \end{cases} \tag{4.26}$$

so that $S(\langle R \rangle^s)$, according to equation 4.7, is the number of species with $\langle R \rangle^m < \langle R \rangle^s$ $(= S - s)$:

$$S(\langle R \rangle^s) = S + 1 - s \tag{4.27}$$

For a Species Area Rule that is fractal:

$$S(R) = S_0 \pi^z R^{2z} \tag{4.28}$$

since $A = \pi R^2$. Defining:

$$y \equiv \log_{10}(\langle R \rangle^s) \tag{4.29}$$

$$x \equiv \log_{10}(S) \tag{4.30}$$

$$x_0 \equiv \log_{10}(S_0 \pi^z) \tag{4.31}$$

we should obtain the following linear relation:

$$y = \frac{x - x_0}{2z} \tag{4.32}$$

which is not quite the case as can be seen in figure 4.24.

The high variances σ_s for small $\langle R \rangle^s$ (see figure 4.26), bring the value of S_0 (average number of species for the smallest area) to its actual value and smear the contribution of the functions $F^s(R)$ to compensate for the lack of middle values of $\langle R \rangle^s$ as seen in figures 4.21, 4.23 and 4.24.

In conclusion, field data shares the same characteristics as for the computer generated data: the functions $F^s(R)$, scaled by the mean and variance, collapse to a single function; the dependence of the variance σ_s with the mean $\langle R \rangle_s$ is similar as shown for computer data; and both S_C and S_G are power law, although with slightly different values of z.

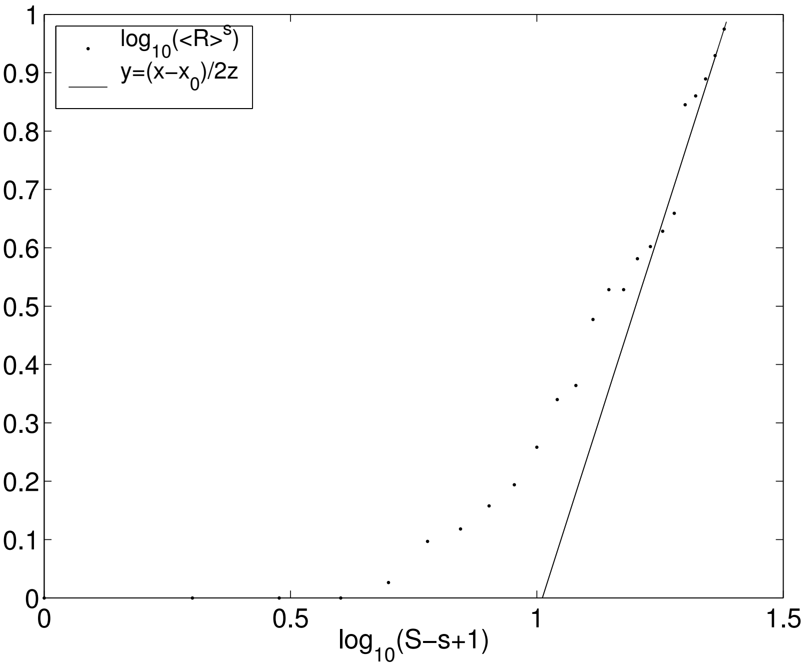


Figure 4.24: Log-log plot of the ordered values of $\langle R\rangle^s$ for the Blue Ridge data set. If variances σ_s were zero we would need a scale-free distribution of $\langle R\rangle^s$ to get a fractal SAR. High variances σ_s (fig. 4.26) smear the contribution of $F^s(R)$ over a fragment of size $\simeq 2\sigma_s$ (fig. 4.25) to produce a fractal SAR (fig 4.20) in spite of a lack of $\langle R\rangle^s$ in the middle range.

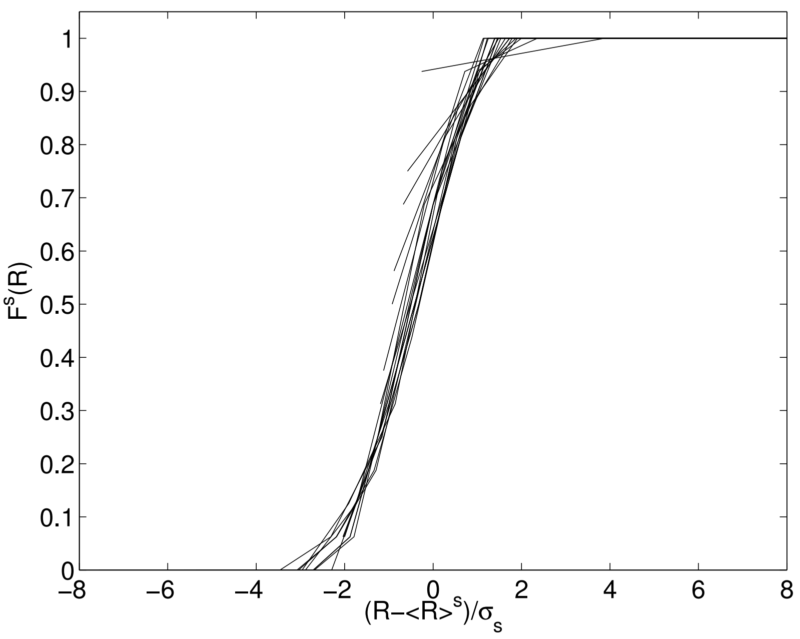


Figure 4.25: Scaling for the functions $F^s(R) \simeq F((R-\langle R\rangle^s)/\sigma_s)$ for the Blue Ridge data. The contribution from each $F^s(R)$ to $S_C(R)$ amounts to a total of 1 smeared over the interval: $[\langle R\rangle^s - 2\sigma_s, \langle R\rangle^s + 2\sigma_s]$, although the greatest change is inside the two centermost variances.

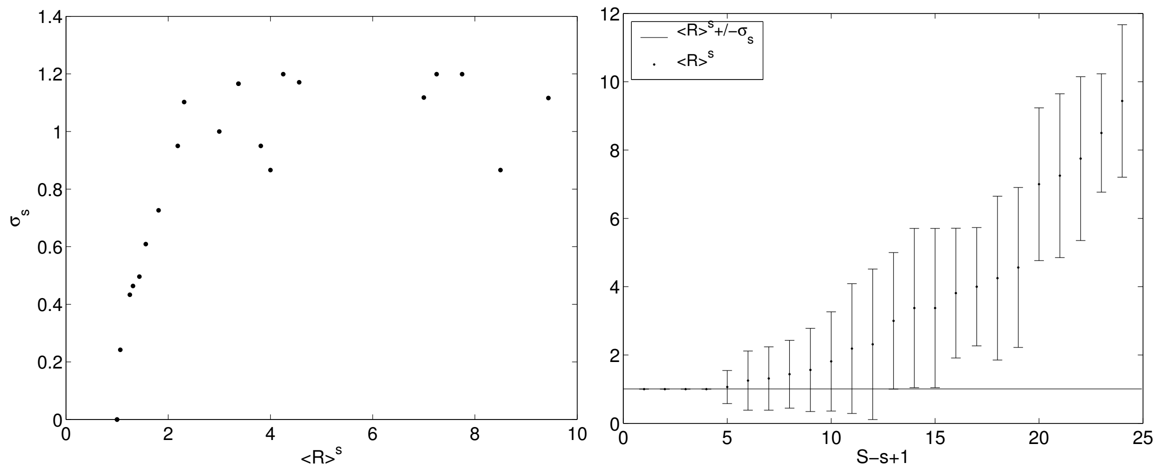


Figure 4.26: The plot of the left shows variances of $\langle R\rangle^s$ (σ_s) vs $\langle R\rangle^s$ for the Blue Ridge data. As usual, the relation is linear until it hits a maximum due to the finite size of the patch. The right plot presents the ordered values of $\langle R\rangle^s$ versus the rank s for the Blue Ridge data. Two variances range is also plotted. Each species contributes to the SAR $S_C(R)$ over approximately two variances up to a maximum of 1. Up to 12 species contribute to the initial value of S_0, which explains the starting point of the line in figure 4.24.

4.10 Conclusions

In this chapter, we have presented a continuum description of the Species Area Relationship and an intuitive decomposition into contributions by each species. In essence it transforms a 2D problem into a 1D problem.

With this construction we have shown that power law SARs arise from fractal clustering and abundance distributions that are of the type usually found in real ecosystems. The exponent z is mostly determined by the shape of the abundance distribution.

Furthermore, we have shown that the appearance of these power law SARs is somewhat independent (robust) of the specific details of either the abundance distribution or the clustering mechanism, especially for exponents $z = 0.2 - 0.4$, which explains its predominance in gathered data [12]. One should therefore expect to find fractal SARs whenever clustering and lognormal abundance distribution with extra rarity are observed.

This work suggests a new and robust mechanism for the emergence of power law SARs that is not obviously related to previous proposals [7],[128],[129], and which can be compared with field ecological data with satisfactory results.

Chapter 5

Microbial ecology in the context of terrace formation in Yellowstone National Park

5.1 Introduction

The present and following chapters will present my work regarding the study of microbial ecosystems in the carbonate hot springs at Yellowstone National Park (YNP). It is part of a larger multidisciplinary project involving geomicrobiologists (led by Prof. Bruce Fouke, Dept. of Geology UIUC), microbial ecologists (led by Prof. Alison Murray, Desert Research Institute) and physicists (led by Prof. Nigel Goldenfeld, Dept. of Physics UIUC). The goal of this project is to study the possible effect of microorganisms on the formation of geological structures, in particular the travertine terraces at the hot springs in Yellowstone National Park (see fig. 5.1). Do microbes play an important role in the formation of these characteristic structures? Indications that this may indeed be the case are presented in section 5.4.

My work has involved characterization of the microbial ecosystems found in the hot springs in terms of their biodiversity (section 5.7) and relative abundance of

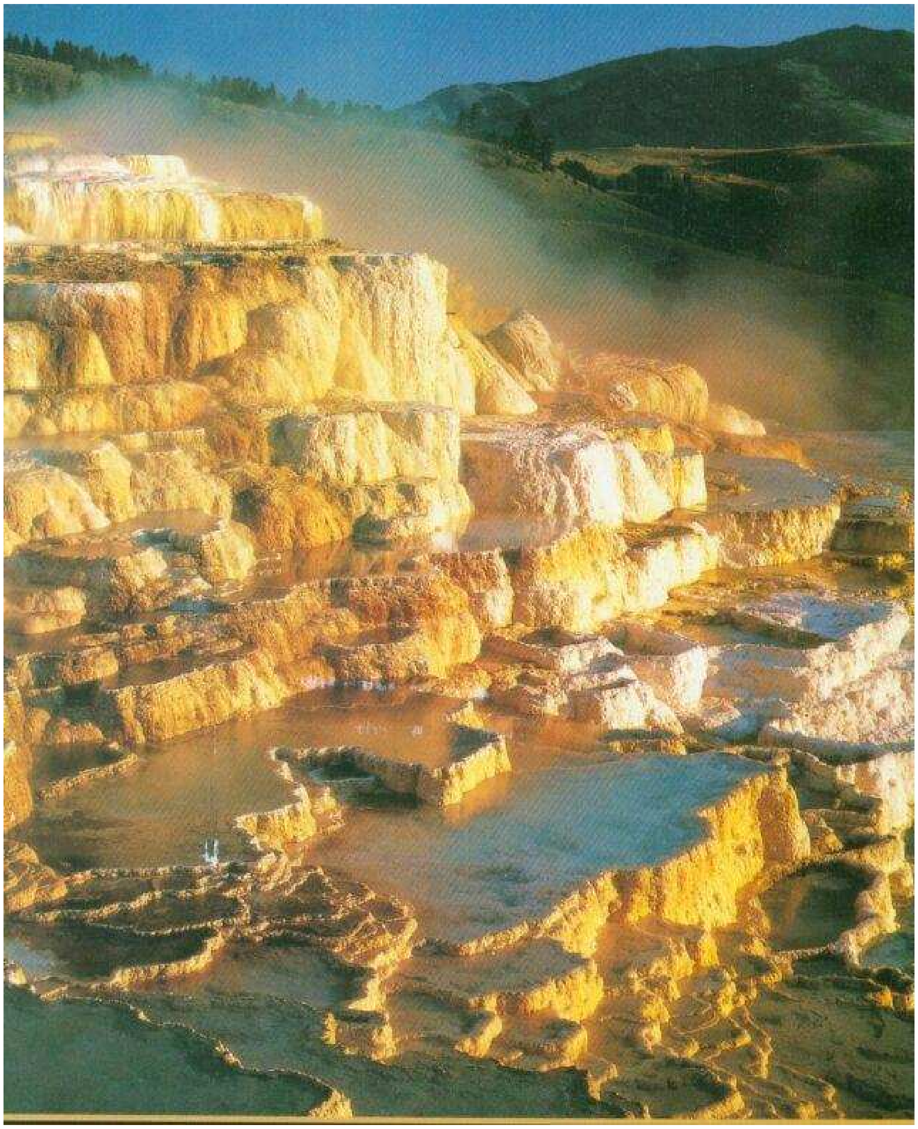


Figure 5.1: Typical travertine terrace formation at Yellowstone National Park hot springs. Water erupts from the vent at the top and deposits $CaCO_3$ as it flows downhill. Under such conditions, one would intuitively expect a simple structure, such as a mould, or a canyon, to form. Instead, these very characteristic terraces develop. Notice how terraces come in different sizes, but basically the same shape. The picture suggests the formation may be self-similar: with no explicit landmark to guide the eye, the observer finds it difficult to know the size of the structures shown. In fact, the largest terraces shown here are about 10 meters across. The study of the generation of these structures is an interesting pattern formation problem in itself (one on which our group is working on) but what makes it even more interesting is that microbes may have some influence on this process. Microbial life is known to accelerate and produce precipitation at microscopical scales (see section 5.4). Could they, too, influence the large scale structure of these formations, seven orders in magnitude larger? This is the question that our group is set to answer. Photo by Larry Ulrich.

microbial species (chapter 6). This work has produced two papers to be published under references [130] and [131].

5.2 What is Geomicrobiology?

Geomicrobiology studies phenomena arising from the interplay of microbiological and purely physicochemical processes across multiple scales in space (microns for microbiological scales to thousands of kilometers for geological scales) and time (nanoseconds for microbiological scales to eons for geological scales).

One of its primary interests is to be able to extrapolate the conclusions of studies in modern settings to inaccessible environments. A good example of this is posed by the controversy regarding the martian meteorite ALH84001 found in Antarctica in 1984 [132]. This meteorite had been ejected from Mars, supposedly after an impact from an asteroid or a comet, and later fell to Earth. Several conspicuous marks on it (see fig. 5.2) have been claimed to be indicative of a biological origin [133]. Controversy ensued [134], making it clear that it is necessary to expand our knowledge regarding what kind of structures attest to biological genesis. By studying, here on earth, accessible microbial ecosystems, it should be possible to assess which formations require microbial interaction to appear. Eventually, this information could be extrapolated to (e.g.) Martian rocks and the detection of extraterrestrial life; and with more foundation, to the fossil evidence of the earliest forms of life (stromatolites) whose origin is still debated.

The reason we focus on Geo*micro*biology rather than just Geobiology is that microorganisms are the main actors on the biological influence on geological features. Not only do they account for more than 50% of the living protoplasm on Earth and have been present for twice as long as multicellular organisms, but they also accumulate ten times more nutrients than the uprunners in this category: plants [136],[137].

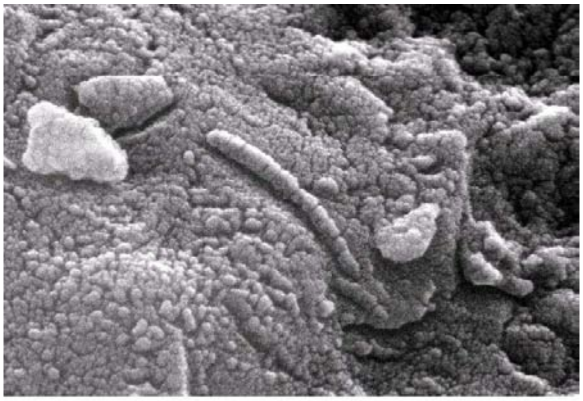


Figure 5.2: High-resolution scanning electron microscopy image of the martian meteorite ALH84001. The ovoid structure was suggested as evidence compatible with the existence of life on Mars, since they resemble the shape of microfossils found on Earth. Unfortunately, there is little knowledge about which structures denote biological origins. Hence the interest of this project. Picture taken from NASA website [135].

It is no surprise then, that they are considered the most important biological geochemical agents on earth [138].

So, if microbes are so abundant, so active and have been affecting the planet for so long why are their effects not conspicuous? The answer is that their effects are all around us! For example, the atmosphere, as we know it, is a consequence of microbial activity: the high levels of O_2 (that allow the proper functioning of all multicellular life) are entirely due to photosynthesis that was initiated by cyanobacteria 2.3 billions of years ago. In fact, it was the rising levels of O_2 due to the advent of photosynthesis that altered the levels of the methane gas CH_4, which kept the archean earth warm, ultimately giving rise to the first well documented glaciation 2 to 3 billions years ago [139]. Nowadays, most of the O_2 in the atmosphere is produced by multicellular plants, but microorganisms still produce 99% of it in the oceans.

Another good example of the geological effects of microbes is the composition of the soil, heavily determined (especially regarding nutrients: Nitrogen and Phosphorus) by the labor of decomposition of organic material carried over by bacteria [140],[141]. Regarding lower depths, the demonstrated accelerated precipitation

Figure 5.3: Carlsbad Caves in New Mexico. These caves have been created by microbes secreting sulfuric acid and dissolving the cave walls to its actual size: the "big room" can be as high as 100 m (≈ 30 storey building) and has an area of around $30,000$ m^2 (≈ 6 American football fields). Figure taken from ref. [144].

of minerals (e.g. ZnS, Au and Fe) also suggests that, over geological time, microorganisms may be the source of formations of ore deposits [138].

To finish with, possibly the most visually striking instance of microbially originated geological formations are Carlsbad Caves in New Mexico (see fig. 5.3). These caves have been formed by bacteria secreting sulfuric acid and slowly dissolving the walls of the caves up to its actual remarkable size [142],[143].

This list is just a personal selection of interesting examples. The list of geomicrobiological effects increases continuously as our knowledge of microbial ecology develops [137].

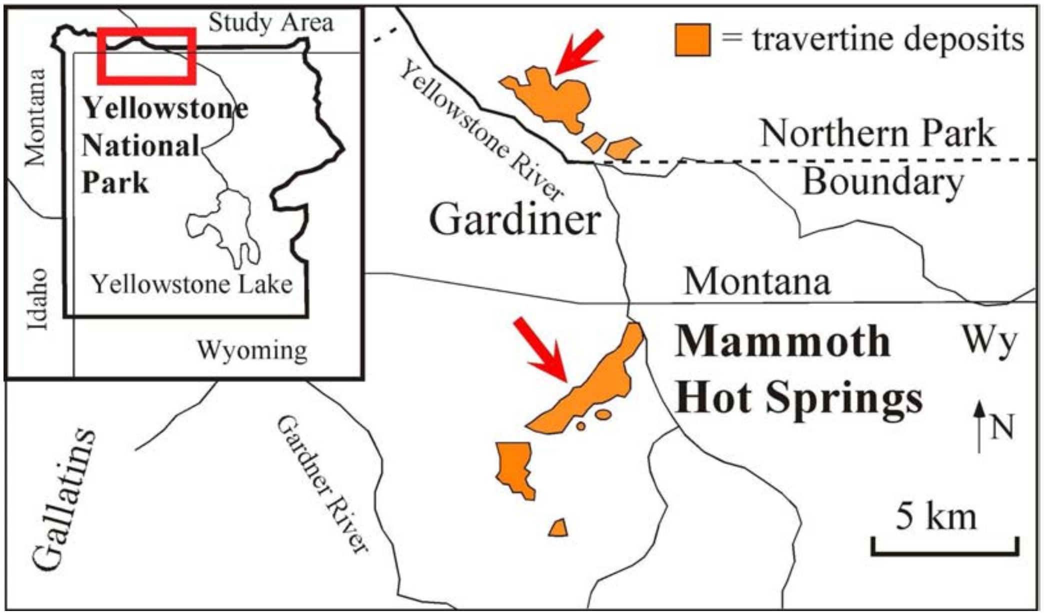


Figure 5.4: Location of Mammoth Hot Springs in Yellowstone National Park.

5.3 The system under study: Yellowstone Hot Springs

The travertine[1] terraces at Yellowstone National Park (figs. 5.1 and 5.8.) are an optimal choice for studying the putative effect of microbial life on geology for a variety of reasons. From a general point of view, carbonates are an interesting compound because they precipitate at life-permitting temperatures, are sensitive to the conditions of the environment and are the most common sedimentary rocks at the Earth's surface [146], [147], [148], [149] [150]. The particular travertine terraces at Yellowstone National Park are appealing because they are among the few protected from human contamination.

The specific hot spring where we have carried out our research is Spring AT-1 located on the Angel Terrace, in the upper terrace region of the Mammoth Hot Springs Complex (see fig. 5.4). This hot spring is closed to the public and has been the object of previous research by members of our group [145], [149].

[1] Travertine is a broad term used to refer to all nonmarine carbonate precipitates formed in or near terrestrial springs, rivers, lakes and caves [145]

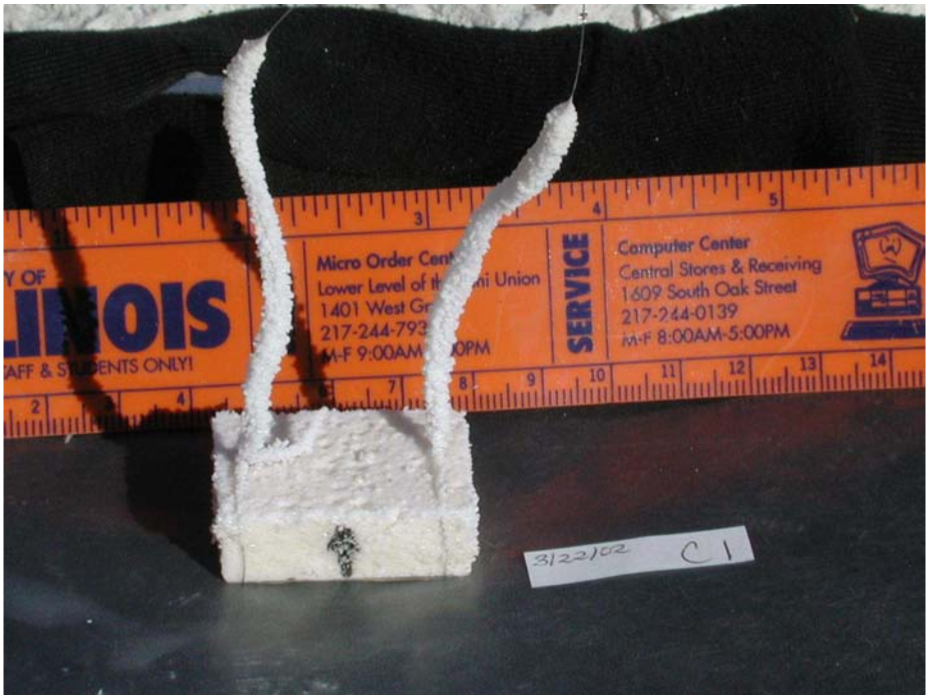


Figure 5.5: Travertine block left in the hot spring for 24 hours. Notice the large amount of deposition on top of the travertine and on the wires. This high deposition creates a very hostile environment for microbial life, which must somehow avoid being encrusted (fig. 5.6). Picture courtesy of George Bonheyo and Bruce Fouke.

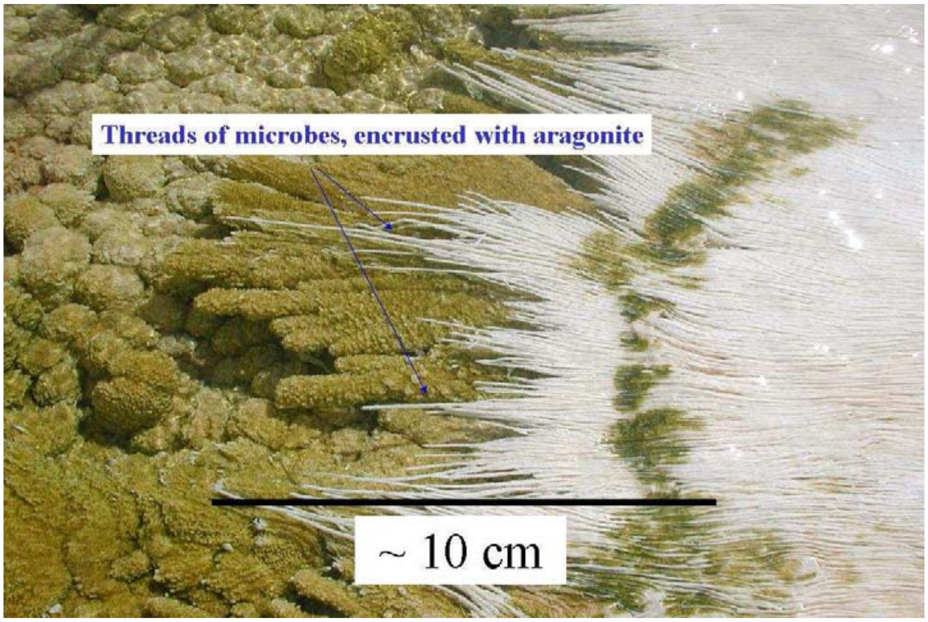


Figure 5.6: The high deposition rate creates a very hostile environment to bacteria, which are easily encrusted. Here are shown threads of bacteria encrusted in aragonite (one of the two crystallographic forms of $CaCO_3$) at the Apron and Channel facies (see section 5.3). Picture courtesy of Bruce Fouke and George Bonheyo.

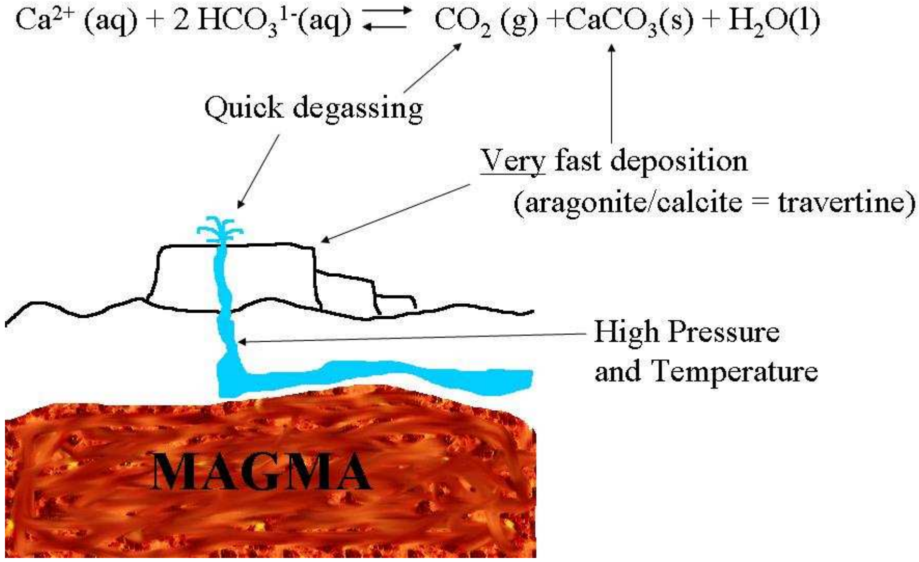

Ca^{2+} (aq) + 2 HCO_3^{1-}(aq) $\rightleftarrows$ CO_2 (g) +$CaCO_3$(s) + H_2O(l)
Quick degassing
Very fast deposition
(aragonite/calcite = travertine)
High Pressure
and Temperature
MAGMA

Figure 5.7: Cartoon picture of the process of terrace formation. Under Yellowstone National Park lies one of the largest volcanoes on earth [151] (inactive at the time). The magma lying not far from the surface (as close as 2 miles) heats up the groundwater, charged in Ca^{2+} ions. As water erupts and flows downhill, it degasses and cools down, precipitating calcium carbonate in two crystallographic forms: aragonite and calcite. The equilibrium reaction above is given to offer an intuitive first understanding of the process. Because of the downflow and the turbulence, water never comes into equilibrium. Nonequilibrium chemistry is fundamental to understand the precipitation process.

The travertine terraces form as a consequence of subsurface waters (heavily charged with Ca^{2+} ions) erupting from a vent and depositing $CaCO_3$ as it flows downhill, lowering its temperature and losing CO_2 (see fig. 5.7). This degassing is very quick and therefore prompts very fast deposition rates, typically between 1 and 5 mm. per day [150],[145],[152] as can be seen in figure. 5.5. In Geology, a science used to measure time in scales of millions of years, this is extremely fast! An immediate consequence of the rapid growth rate is that the system changes greatly between visits. It is therefore necessary to be able to divide it into subsystems that can be identified independently of location and size. These subsystems are called **facies** and have their own chemical (temperature, pH, alkalinity...) and geological characteristics (crystallographic structure and crystal morphology) that are consistent in time. Each of these facies is linked to each other by the water flow. The five facies model for this system [145] is schematically shown in figure 5.9 and in the real system in figure 5.8. The five facies are:

- Vent (V): water erupts from the subsurface and starts degassing, cooling down and increasing pH.

- Apron and Channel (AC): the water flows down without accumulating. Threads of bacteria encrusted in aragonite form by the high precipitation rates (see fig. 5.6). Aragonite is one of the forms in which calcium carbonate crystallizes. The other one is calcite.

- Pond (P): the flowing water accumulates in a pond doing most of the deposition and creating the largest terraces.

- Proximal Slope (PS): as the water overflows the Pond, it runs into the Proximal Slope where deposition is slower and hence the terraces are smaller.

- Distal Slope (DS): finally, the water runs into the rest of the surrounding area, mixing with soil and organic material.

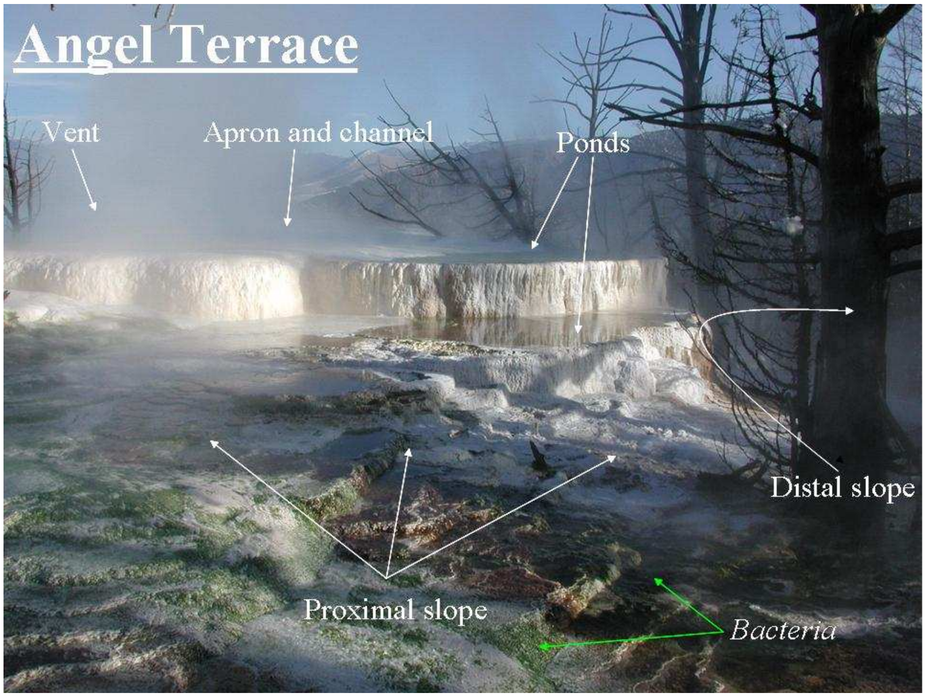
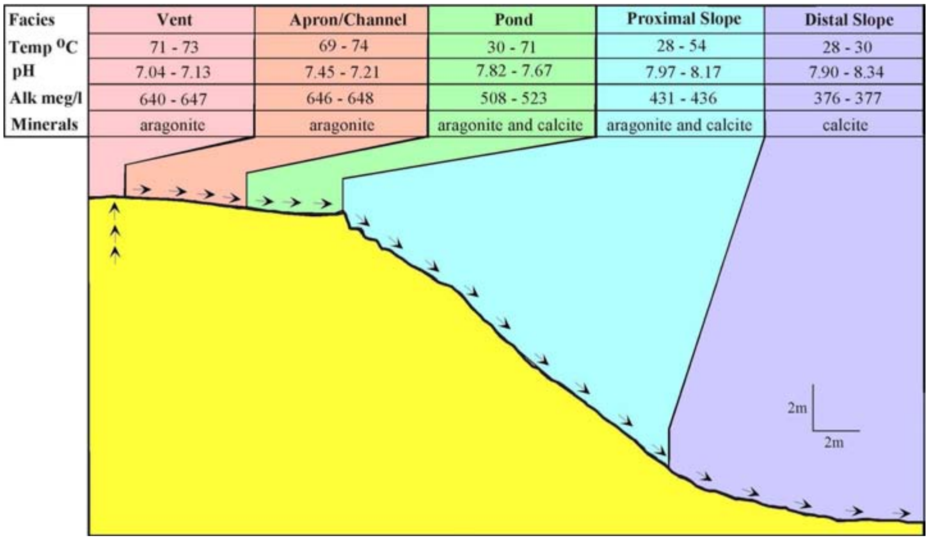


Figure 5.8: Angel Terrace in Mammoth Hot Springs, Yellowstone National Park, our study site. The five facies are indicated: the water erupts at the Vent at high temperature and flows down the Apron and Channel, degasssing and cooling, to accumulate at the Pond. Most deposition is done here. Water then overflows the Pond to run into the Proximal Slope, where terraces are smaller because of a lower rate of deposition. Finally, the water mixes with organic material and soil in the Distal Slope. The green and brown colors on the travertine are thick mats of bacteria. Microbes are present in the water, on top of the travertine and encrusted inside the travertine.



Facies	Vent	Apron/Channel	Pond	Proximal Slope	Distal Slope
Temp ^{0}C	71 - 73	69 - 74	30 - 71	28 - 54	28 - 30
pH	7.04 - 7.13	7.45 - 7.21	7.82 - 7.67	7.97 - 8.17	7.90 - 8.34
Alk meg/l	640 - 647	646 - 648	508 - 523	431 - 436	376 - 377
Minerals	aragonite	aragonite	aragonite and calcite	aragonite and calcite	calcite

Figure 5.9: Schematics of the facies. As water flows down temperature falls continuously from 70^oC to 30^oC, while pH increases continuously from 7 to 8 because of the degassing. Crystal morphology is characteristic of each facies. Figure taken from ref. [149].

The facies model provides an overarching spatial and temporal framework to contextualize geochemical and biological measurements, and be able to link them to system scale environmental processes [149].

5.4 Indications of microbial influence

The explanation given above regarding the creation of terraces doesn't seem to require the involvement of biological processes. Yet, bacteria are known to play important roles in precipitation for other systems, and these mechanisms are expected to be applicable to our case.

To start with, microbes act as passive nucleation sites for the deposition. This is very relevant for the sedimentation since turbulent water like that found in the YNP hot springs is always out of equilibrium. Under these conditions the lack of availability of nucleation sites can severely limit the rate of precipitation [153]. Plain visual evidence that this can be an important effect in the AT-1 can be seen in figure 5.6, where aquificales colonies are encrusted by aragonite.

Another, more interesting way in which microbes can influence the deposition of carbonates is by altering the levels of CO_2 or pH in the water. This may happen in several ways. One of them is through autotrophic reactions (such as photosynthesis), which take CO_2 out of the water. This is equivalent to enhancing the degassing (see fig. 5.7), which in turn increases precipitation. Another possibility is through heterotrophic reactions, which use carbon compounds as a source of energy. Several heterotrophic metabolic pathways have, as an end result, an increase in pH, which in turn shifts carbonate-bicarbonate equilibrium towards the production of CO_3^{2-}. Excess availability of Ca^{2+} then produces more $CaCO_3$. [154].

All these effects are "passive", in the sense that cells change the environmental characteristics to make precipitation more likely but do not directly create $CaCO_3$ on their own. Active precipitation, in which carbonate particles are produced by ionic exchanges through the cell membrane, is also hypothesized to be an important general mechanism that can accelerate the precipitation of carbonates [154]. Furthermore, it has been proposed that active precipitation chemically favours bacterial survival and proliferation in naturally deposition prone environments [155]. Part of the precipitation observed in fig. 5.6 may be caused by these active effects.

Specifically how each of these mechanisms affect our system is the object of ongoing and future research but, all things considered, microbial-induced carbonate precipitation is a well known effect, observed in lab and field experiments that is neither restricted to particular taxonomic groups nor to specific environments [154],[156],[157]. It would be natural and not surprising if it were to occur in our system.

Evidence of microbial activity has been reported in the levels of carbon fractionation: for energetic reasons microbes prefer to use C^{12} to C^{13} in their metabolism and a departure from expected fractionation levels can indicate biological influence [158]. In the case of Mammoth Hot Springs (fig. 5.10), the Pond, Proximal Slope and Dis-

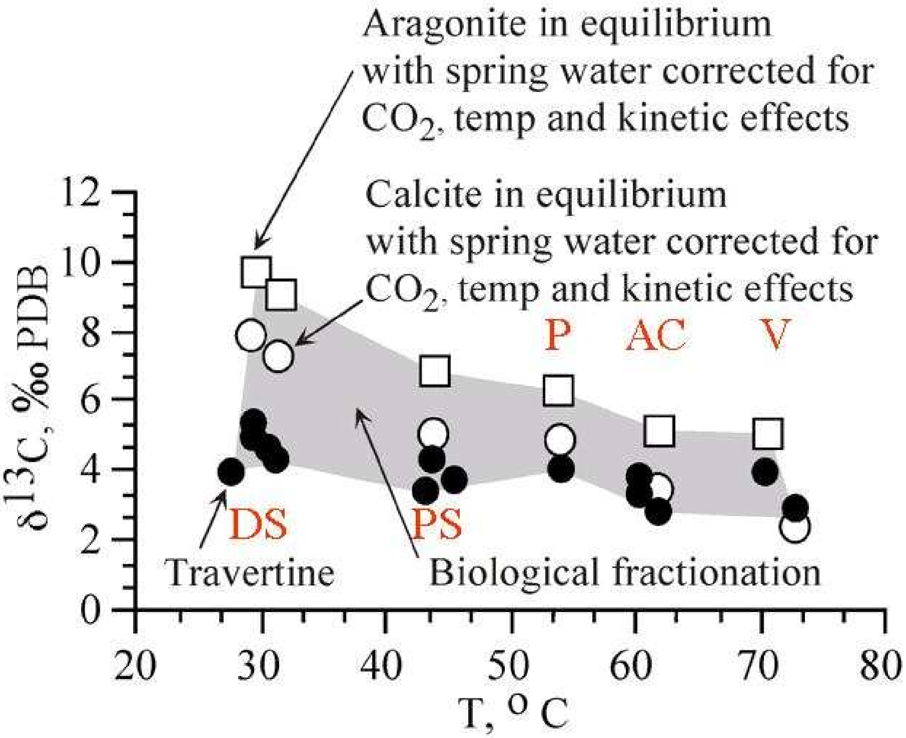


Figure 5.10: Carbon fractionation for each facies. Squares and white circles are the values expected without biological intervention, i.e. equilibrium values with corrections for nonequilibrium consequences: CO_2 degassing, temperature and kinetic effects. Black dots are values measured. The Pond, Proximal Slope and Distal Slope facies display lower than expected fractionation, indicating possible biotic respiration (release of CO_2). Fractionation is defined as: $\delta^{13} = 1000 \times [C^{13/12}/C^{13/12}_{std} - 1]$, where $C^{13/12}$ is the isotope ratio and $C^{13/12}_{std}$ is the PDB international standard. PDB refers to the Cretaceous belemnite formation at Peedee in South Carolina, USA. Figure courtesy of Bruce Fouke.

tal Slope facies display lower than expected fractionation, indicating possible biotic respiration (release of CO_2).

Microbial induced precipitation is therefore, on one hand, expected as a general phenomena, and on the other, suggested by carbon fractionation levels in our system. The next step is to determine if there is enough of it to affect terrace formation. In order to do that, it is necessary to know which microbes are present along with their metabolism. Additionally, it is important to have a grasp of which are the most abundant, and hence more able to influence the water chemistry.

The detection of microbial species has been accomplished through the 16S rRNA method (see next section) by our collaborators. I assisted in the analysis of the large quantities of raw data used for the study in section 5.6. In order to know which fraction of the total amount of species has been detected so far, I have used an array of methods (section 5.7) to estimate the total biodiversity. Furthermore, I have used this dataset to produce a rough estimate of relative abundances (see chapter 6), using a method that is novel in the context of the analysis of clone libraries.

5.5 The 16S rRNA gene method

All of the detection of microorganisms in this project has been done though the 16S rRNA method, which I will explain in this section.

The 16S rRNA method for classification and detection of microorganism has revolutionized microbial ecology. Before its inception, cultivation in order to assess metabolism was necessary for microbial identification, since morphological characteristics alone are not sufficient for this purpose [159]. The main problem with this approach is that only an small fraction (0.1-10%) of the bacteria that can be typically visualized in a sample have been successfully cultivated [160]. Only a meager part of the complete microbial world can be accessed.

16S rRNA analysis put an end to this situation by classifying microorganisms by a single gene: the 16S rRNA gene. This method works because, generally speaking, this gene is present in all microorganisms. Its expression is ribosomal RNA (rRNA) which is involved in ribosome creation and these, in turn, in protein folding: a fundamental process for cell functioning. All cells must have this gene (or an equivalent) in order to function properly. The gene's structure is extremely similar for virtually all cells since this vital function was created early in the evolution of life and its functioning is essential for life. It is nonetheless not identical and this is why this method works for classification: although the coding part of the gene must be the same for cell machinery to operate properly, the non-coding or "junk" DNA part of the gene is still subject to random mutations [161] as time progresses. If two lineages from a single species diverge and speciate the accumulation of random mutations in the non-coding part of the gene (which does not affect the species rate of survival because it is not involved in the functioning of the cell) will be larger the more time elapsed since the speciation event. The percentage difference of these non-coding parts would then offer a natural classification for microorganisms. This idea was first used by Carl Woese and coworkers [162] to create the phylogenetic tree of all life on Earth, and lead to the discovery of an unexpected branch of life: archaea, previously assumed to be part of the bacteria branch (see fig. 5.11).

Norman Pace and coworkers [164],[165] had the seminal idea to apply this method to environmental DNA, thus bypassing the need for cultivation. This revolutionized microbial ecology and opened up possibilities that have enormously enlarged our knowledge of the microbial world [159],[160], including the viability of the present project.

I will now give a semi-detailed account of the steps involved in identifying the microbial diversity present in an ecosystem as was used in this project (more specific details can be found in ref. [130][2]):

- Environmental sample collection: Samples are collected from the system under study. Usually these are taken from diverse environments to get the most possible diversity. One of the key differentiating factors of our project has been the extent to which the environmental context has driven the choice of microbial samples; in our work several samples were taken from all five facies and

[2] An excellent account for the uninitiated can be found on the "Virtual Lab Series" CD-ROM (Bacterial ID lab) distributed free by the Howard Hughes Institute through its webpage: http://www.biointeractive.org

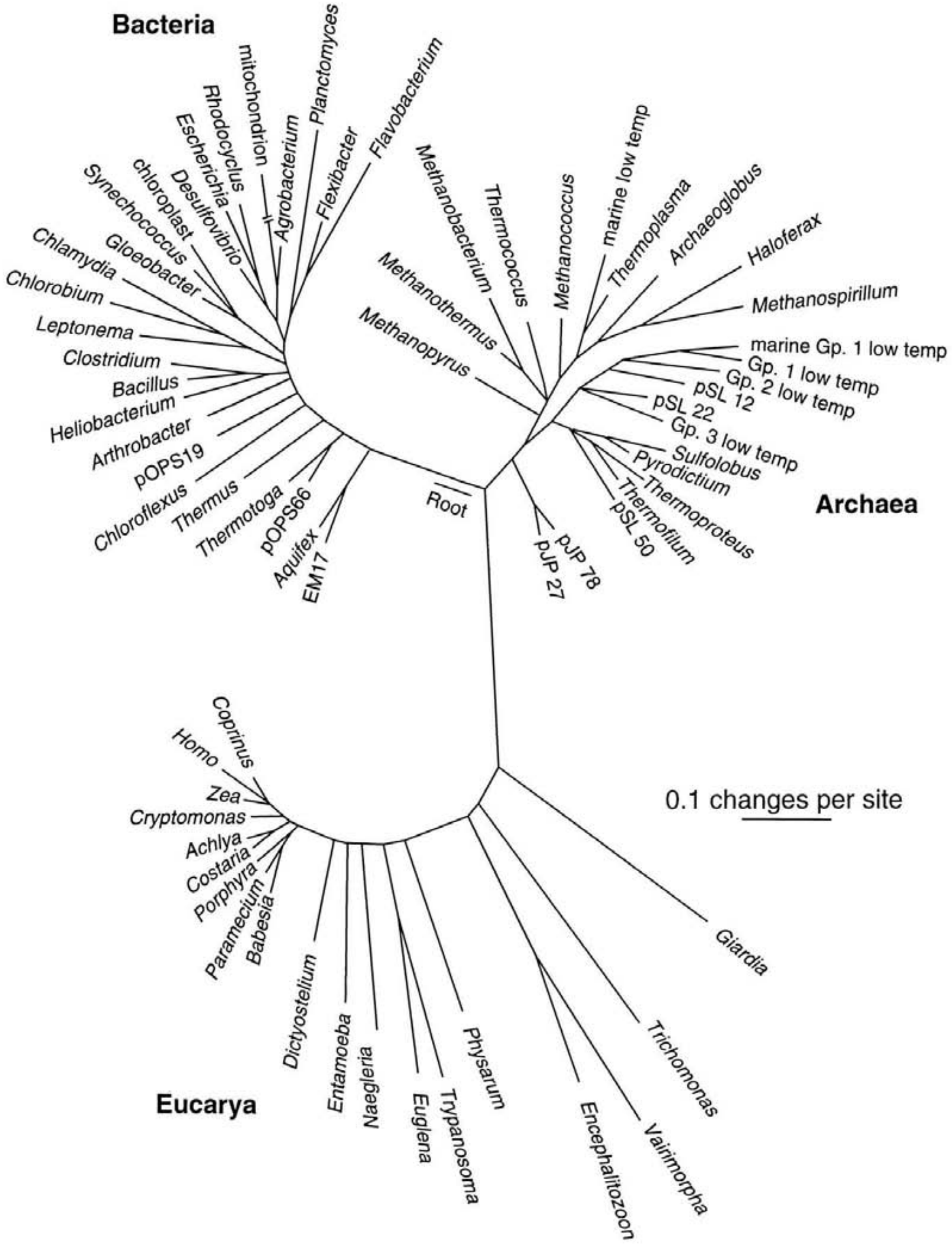


Figure 5.11: Phylogenetic tree of all life on earth based on small subunit rRNA sequences (16S for bacteria and archaea, 18S for Eukarya). The tree length represents nucleotide difference between rRNA sequences and is roughly equivalent to the time elapsed since the organisms became different species. Notice the clustering around the three main domains of life: Eukarya, Bacteria and Archaea. The root of the tree represents a common ancestor for all existing life on Earth. The Archaea domain branched off first, which suggests that they are the most primitive organisms. On the other end, Eukarya, which comprises all multicellular organisms, branched off the last. Archaea were grouped with bacteria prior to the advent of phylogenetic techniques, which revealed its independent nature as one of the three domains of life on Earth. Figure taken from ref. [163].

the different mediums: water (filtered), the thick mat of bacteria on top of the travertine, and the sediment itself. Samples are preserved frozen.

- DNA extraction: For the purpose of 16S rRNA detection, we only need this gene. The rest of the cell is superfluous. The DNA is inside the cell walls and it is necessary to break them (this is called lysis) in order to obtain it. There are several methods to achieve this; either physically (e.g. by including solid beads smaller than the size of a cell and beating the solution vigorously to pierce the walls), chemically (e.g. alkaline lysis) or thermally (e.g freeze and thaw in cycles). After this it is necessary to purify the solution to get only the DNA.

- PCR amplification of 16S rRNA genes: Out of the whole DNA we only want the 16S rRNA part. We would like to amplify this gene (signal) out of the rest of the DNA (noise). PCR achieves just this. By using the complementarity of base pairs and temperature driven naturation-denaturation cycles, it is possible to *exponentially* amplify a selected gene [161]. It must be mentioned, however, that different bacterial species 16S rRNA genes are amplified differently. Hence, the final amplified relative ratio of 16S rRNA genes is not a faithful indicator of the initial relative abundance of microbial species: it is severely biased.

- PCR product purification: The sample needs to be purified in order to get only the amplified 16S rRNA genes and not the rest of the DNA or the PCR byproducts.

- Ligation of PCR amplified 16S rDNA: We are now left with a solution of 16S rRNA genes (16S rDNA refers to the DNA genes that produces ribosomal RNA=rRNA). How do we separate them? In order to do that we introduce them into the DNA of a host cell (one per cell) by taking advantage of a mechanism called lateral gene transfer. This refers to the quite shocking concept (for multicellular organisms like us) of exchanging pieces of DNA between cells. To put it bluntly, for a human being this would be equivalent to (e.g.) exchanging a blue eyes gene with somebody and inserting it into your own DNA to display blue eyes. It is called lateral gene transfer and is one of several vehicles of gene dispersal in microorganisms [166]. This procedure is carried out through plasmids (basically, pieces of DNA) and the act of including the 16S rRNA genes into the plasmids is called ligation.

- Transformation of ligation mixture into competent *E. coli* cells: Plasmids with the ligated 16S rRNA genes are introduced by the *E. coli* cells (the standard tool of molecular biology) into their own DNA. This is called transformation and only one plasmid per cell is introduced.

- Isolation of individual cells: Now that there is only one 16S rRNA gene per cell the solution is spread thin over a petri dish so that individual cells are separated. Enough growth substrate is added so the cells multiply and form colonies of clones that are visible with the naked eye (remember a single cell is around 1 μm in size). To eliminate the cells that didn't take up any plasmids, an antibiotic is added to the plate. The corresponding antibiotic resistant gene has already been included in the plasmid so that only clones of cells with plasmid inserts survive. Finally, the insertion spot for the 16S rRNA gene has been arranged to be within the Beta glucuronidase gene, which confers a blue color to the colonies. If the plasmid is within a cell, and the cell took the 16S rRNA gene into its DNA, then the colony will disrupt the Beta glucuronidase gene and look white, otherwise it will display a blue color. Therefore the surviving colonies of clones which are white are the clones with the plasmid and the 16S rRNA gene insert. Each individual colony corresponds to a single initial cell and the white colonies are selected and taken to the sequencing facility for sequencing of the 16S rRNA gene.

- Clone screening for uniqueness: In the case of our project, over 14000 clones were produced [130]. The current cost of sequencing makes it prohibitively expensive to sequence them all, so only unique clones are sequenced. To select the unique clones the RFLP (Restriction Fragment Length Polymorphism) method is used [161].

- DNA sequencing: The sequence of the 16S rRNA gene is introduced in a data bank of genes called GenBank [167] from which information regarding whether it is a known bacterial species, previous location detections, and associated metabolism can be obtained.

5.5.1 Species vs OTUs

So far, I have continuously referred to the concept of microbial species and assumed they can be identified through their 16S rRNA gene sequence. This point is controversial. In fact, the consensus is that it is necessary to take into account genomic AND phenotypic[3] information to define species [168]. Furthermore, even if prokaryotic species could be defined based only on genetic information, the 16S rRNA is but a small part of the whole genome and doesn't always reflect the whole genome difference. There are known examples of different species with identical or very similar 16S rRNA sequences [168].

For all these reasons, instead of referring to a 16S rRNA sequence as identifying different species we will refer to them as identifying different Operational Taxonomic Units (OTU). To be considered different OTUs, their 16S rRNA sequences must differ by more than a given percentage. The agreed upon value of percentage difference is customarily 3% [168], but this is purely heuristic and there is no reason other than

[3] Phenotype refers to the way in which the genotype is expressed: the visible or otherwise measurable physical and biochemical characteristics of an organism, as a result of the interaction of genotype and environment [168].

previous experience to select this value. Thus, three different OTU defintions were used in the Yellowstone National Park study: 0.5%, 1% and 3% difference. Our conclusions do not depend much on this choice, as expected.

5.6 Microbial ecosystem partitioning in YNP hot-springs

By comparing the 16S rRNA sequences found with Genbank data (see the last paragraph of section 5.5) it is possible to see whether the corresponding bacteria have been previously found anywhere else on Earth and whether they have been cultivated and their metabolism is known.

Several of the 16S rRNA sequences found in this project (14 out of 193 for the 3% definition) had never been detected before, but most had, and their metabolic characteristics are known or could be inferred to be similar to a closely related sequence. Bacterial communities in our system changed from being primarily chemolithotropic (use inorganic chemicals as energy sources; typical of extreme environments at high temperatures) in the vent, to photoautotrophic (extract energy from photosynthesis) at the pond and finally to heterotrophic (incapable of photosynthesis, and therefore requiring carbon compounds as carbon source) in the distal slope [130].

Most interestingly, very strong partitioning of the ecosystems can be observed in the system (fig. 5.12): around 90% of the OTUs detected are present in only one of the facies and only 2 (out of 200 to 300 depending on the OTU definition) were detected in all five facies. This is not because of lack of sampling. We believe we have detected the most abundant OTUs as shown in the next section. Neither is it an artifact of having too narrow a definition of OTU, as can be seen in the tables from fig. 5.13. As the OTU definition decreases the number of OTUs increases as expected, but the number of OTUs in common between facies stays constant.

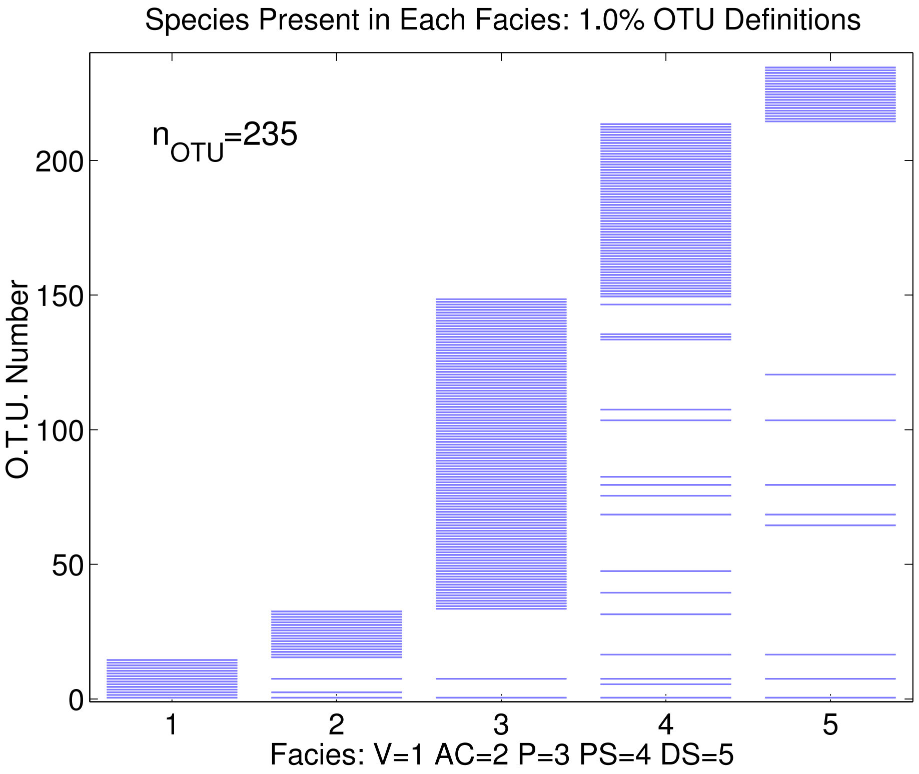


Figure 5.12: Microbial ecology partition by facies. In this graph the X axis represents facies and the Y axis represents OTUs. If an OTU is present in a facies a bar is drawn. Microbial communities partition tightly to facies, even though they were defined on purely physical and geological terms. This is not an effect of a coarse OTU definition as can be seen in figure 5.13.

OTU commonality between facies for each OTU definition

0.5%	V	AC	P	PS	DS
V	**23**	2	3	5	2
AC		**24**	2	3	2
P			**167**	17	10
PS				**114**	8
DS					**40**

1.0%	V	AC	P	PS	DS
V	**32**	3	2	3	2
AC		**24**	2	4	3
P			**167**	14	7
PS				**114**	6
DS					**114**

3.0%	V	AC	P	PS	DS
V	**6**	3	2	2	2
AC		**20**	5	6	4
P			**99**	14	8
PS				**71**	6
DS					**28**

Figure 5.13: Number of OTUs in common for each facies, for OTU definitions 0.5%,1% and 3%. As can be readily observed, even though the number of OTUs increases with a tighter OTU definition (as expected) the number of OTUs in common between two given facies changes only slightly. This indicates that the partitioning is real, and not an artifact of the OTU definition.

In conclusion, microbes seem to track the facies that were defined from a purely geological point of view. Microbial ecosystems are therefore faithful indicators of depositional environments. It is not clear at the moment whether this is because the microbial ecosystems create these environments or because they track the water chemistry of the host environment.

5.7 Biodiversity estimates

The first practical question that arises when surveying a new environment is: How do we know when we have taken enough samples to detect most microbial OTUs?

This section will explain the methods I have used to answer it: accumulation curves (section 5.7.1) and total biodiversity estimators.

Fortunately, biodiversity estimation is a standard problem in traditional ecology [169] and there are several methods available ([170],[171],[172],[173],[174],[175],[176],[177],[178],[179],[180],[181],[182]). They can be classified into three types:

- Non-parametric methods: These methods do not use any fitted parameters in modeling the sampling. They include the jackknife, bootstrap and Chao estimators [176],[177],[178],[175],[182],[171],[172].

- Extrapolation of accumulation curves: the sampling process is modeled and a accumulation curve is derived (e.g.: $S(n) = S_{max}(1 - e^{-Kn})$, $n \equiv$ number of samples). Fitting the parameters of the analytical expression gives an estimate for the total number of species (S_{max}). This category includes fits to exponential and hyperbolic curves [173],[174],[171],[172].

- Extrapolation of ecological patterns: the abundance data available is fit to known widely observed ecological patterns [170],[179]. For example, the abundance distribution $P(n)$ (fraction of total number of species with n individuals) is assumed to be lognormal; the parameters are obtained from a fit to available abundance data, and the integration of $P(n)$ gives the total number of species. Another option is to use the Species Area Rule which states that the number of different species S in a given area A is $S \equiv A^z$.

I will only be using the first two types of methods because it is not known whether the lognormal abundance distributions and Species Area Rule are applicable to microbial ecosystems, and the data collected were neither sufficient nor appropriate for methods of the third type.

In the following sections I will expand on each of these methods and introduce accumulation curves.

5.7.1 Accumulation curves

Accumulations curves are a basic tool of macroscopic ecology [169]. They plot the number of total OTUs found as the number of samples n increases. Since the plot should not depend on how the samples were numbered the graph is averaged over all the possible permutations of the samples[4]. For the case of a well-sampled ecosystem, these graphs should flatten out for large n, because no new OTUs are detected as more samples are taken. The results for the YNP data are shown in figures 5.14, 5.15 and 5.16. They show very little curvature, if any, indicating that we are far from the saturating regime. Their shape, though, gives us information on the total number of OTUs as explained in sections 5.7.4 and 5.7.5.

5.7.2 Bootstrap method

The main idea in bootstrap methods is to sample with replacement from the available data in order to obtain accurate estimates of desired statistical quantities. It is a

[4] In cases where the number of permutations was too large to calculate within a convenient amount of computer time, a total of 1000 permutations was randomly chosen and averaged over.

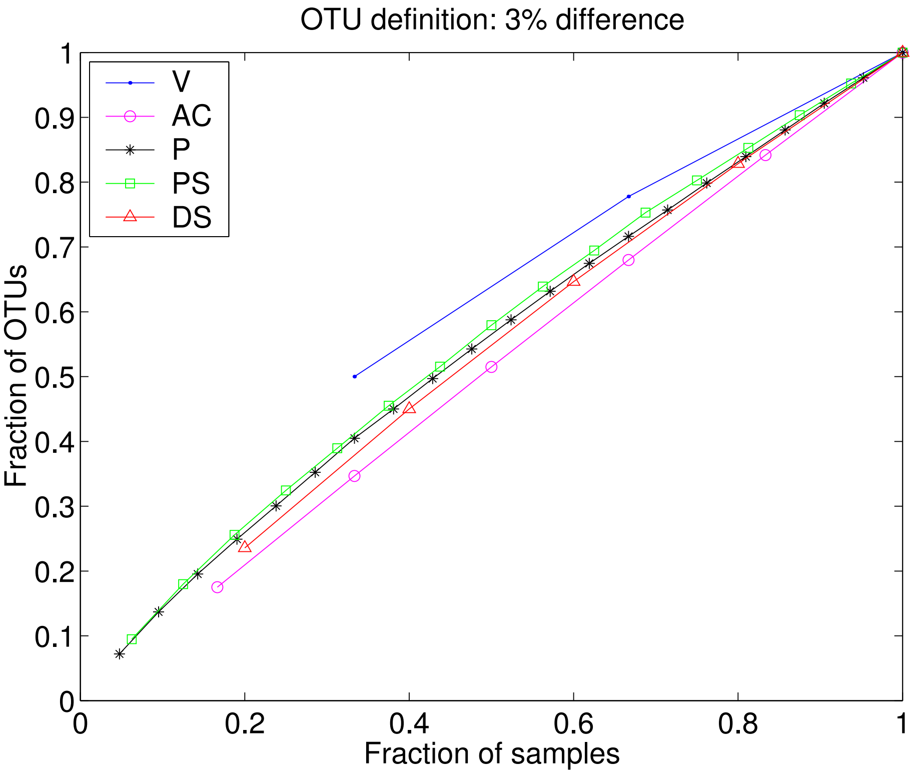


Figure 5.14: Accumulation curves for species as function of samples (see section 5.7.1). Only the vent, proximal slope and pond facies show some curvature indicating that they are the better sampled than the rest. The number of samples for each facies is: Vent (V): 3, Apron and channel (AC): 4, Pond (P): 25, Proximal slope (PS): 14, Distal slope (DS): 6.

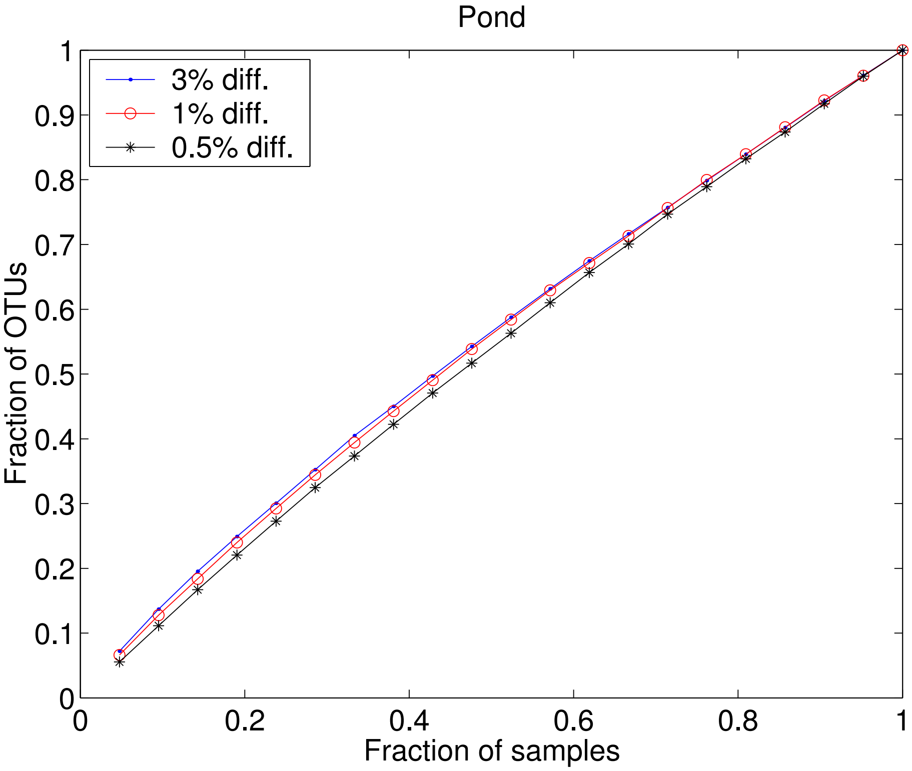


Figure 5.15: Comparison between different OTU definitions for the pond facies. As expected, the accumulation curves becomes more linear with tighter definitions.

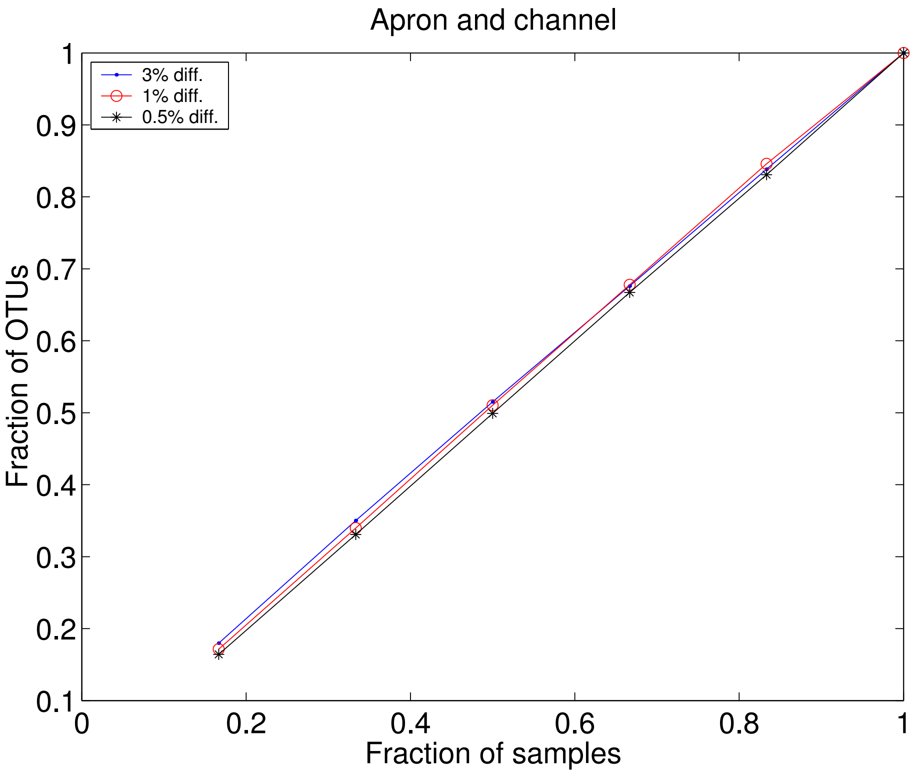


Figure 5.16: Accumulation curves for the apron and channel. Accumulation curves are almost linear, indicating a very poor sampling. In the case of 0.5% the curvature is so meager that it will be impossible to extract information from it.

general method with wide applicability [183],[184], that has over the last 20 years or so become an accepted technique for computer-intensive analysis of data. It's particular advantage is that the special assumption of classical (analytical) statistical tests may not be known to apply to the dataset in question.

The bootstrap method works as follows [185] [169]: assume that there are n independent identically distributed measurements (x_i) of an unknown cumulative distribution function F. The bootstrap method [186] prescribes the following steps in order to obtain an estimate of the desired statistic θ:

1. Construct an empirical probability distribution $\hat{F}$, by putting a weight $1/n$ at each of the measurements x_i.

2. Draw a sample of size n from the n data points with replacement. This means that each time you draw a data point it is not eliminated from the data points list. This sample is called the "bootstrap" sample.

3. Calculate the estimate of the desired statistic θ based on the bootstrap sample.

4. Repeat steps 2 and 3 N times to obtain N estimates of θ, denoted $\hat{\theta}_i$ ($i = 1, ..., N$). N should be large enough that the result converges (usually $50 \leq N \leq 200$).

The bootstrap estimate, $B_n(\theta)$ is given by:

$$B_n(\theta) = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i \tag{5.1}$$

with a variance:

$$var_{est}[B_n(\theta)] = \frac{1}{N-1}\sum_{i=1}^{N}[\hat{\theta}_i - B_n(\theta)]^2 \tag{5.2}$$

This general method can be applied to our problem of estimating the total OTU richness of an ecosystem out of a partial census given by several samples of OTU diversity. In particular, we will consider what information can be extracted from the OTUs observed in multiple samples from the same spatial location.

Let us assume that we have n samples with S_i OTUs each. The total number of different OTUs in all samples will be denoted S_0 and it will be the initial bootstrap estimate for the total OTU richness S_T.

We are interested in finding the bias of the initial estimate S_0: $bias = S_0 - S_T$ ($\equiv \theta$ in the notation above). The bootstrap estimation of the bias is ([169] page 570, [185]) $bias_{est} = \tilde{S}_0 - S_0$, where $\tilde{S}_0$ is the bootstrap estimate of the number of total different species in the n samples and S_0 is the bootstrap estimate of the total OTU richness.

This bootstrap estimate $\tilde{S}_0$ is obtained following the procedure above: N bootstrap samples are generated and the total number of OTUs is determined for each of them. The average of these gives $\tilde{S}_0$. Normally it would be necessary to perform a simulation to obtain the bootstrap samples and the average, but in this case the required statistic $\tilde{S}_0$ is simple enough that we can use the analytic expression ([185] [187]):

$$\tilde{S}_0 = \sum_{j=1}^{S_0} (1 - (1 - Y_j/n)^n) \tag{5.3}$$

where Y_j is the number of samples in which OTU j is present and $(1 - Y_j/n)^n$ is the probability that OTU j is not included in the bootstrap sample.

By equating the *bias* and $bias_{est}$ one obtains [185]:

$$S_T = 2S_0 - \tilde{S}_0 = S_0 + \sum_{j=1}^{S_0} (1 - Y_j/n)^n \tag{5.4}$$

The estimate of the variance can be similarly obtained:

$$\begin{aligned} var(\tilde{S}_0) = \sum_{j=1}^{S_0} (1 - Y_j/n)^n - [1 - (1 - Y_j/n)^n] \qquad (5.5) \\ + \sum \sum\nolimits_{i \neq j} [(Z_{jk}/n)^n - (1 - Y_j/n)^n (1 - Y_k/n)^n] \end{aligned}$$

where Z_{jk} is the number of samples in which *both* OTUs j and k are absent.

3% difference

	V	AC	P	PS	DS
Exponential	9.3^*	131 ± 31	246 ± 21	190 ± 19	86 ± 5
Hyp. (ML)	11.8 ± 0.85	316 ± 29	279 ± 29	224 ± 20	147 ± 0.6
Hyp. (LF)	12 ± 1	315 ± 17	294 ± 15	233 ± 12	147 ± 3
Lee-Chao ($\hat{N}$)	15 ± 11	238 ± 90	651 ± 1000	389 ± 510	139 ± 89
Lee-Chao ($\tilde{N}$)	11 ± 6.4	$175^\dagger$	638 ± 997	379 ± 490	104 ± 48
Bootstrap	7.2 ± 1.4	26 ± 16	129 ± 111	92 ± 78	36 ± 18
Jackknife	8.7 ± 2.4	36 ± 49	176 ± 295	125 ± 195	47 ± 86
OTUs found	6	20	99	71	28

Figure 5.17: Species estimates for each of the facies calculated using the methods explained in the text. (V = vent, AC = apron and channel, P = pond, PS = proximal slope, DS = distal slope). In the case of the Lee-Chao estimator, the $\tilde{N}$ version is a bias corrected for the case when the true value of the coefficient of variation is large. Hyp. (ML) stands for hyperbolic curve, maximum likelihood and Hyp. (LF) stands for hyperbolic, linear fit. * The exponential fit for the Vent doesn't have a variance since there are only three samples, the derivative has two points and only one line can fit through two points. † The coefficient of variation $\tilde{\gamma}$, and hence the variance, comes negative out of equation 5.23. The reason is that $f_1 = 19$, $f_2 = 1$ and $f_k = 0 \; \forall \; k > 2$.

1% difference

	V	AC	P	PS	DS
Exponential	35.3*	171 ± 53	308 ± 23	265 ± 24	92 ± 5
Hyp. (ML)	47 ± 3	519 ± 681	401 ± 26	317 ± 33	157 ± 1
Hyp. (LF)	48 ± 4	517 ± 85	414 ± 16	328 ± 19	157 ± 3
Lee-Chao ($\hat{N}$)	63 ± 53	263 ± 100	815 ± 1221	611 ± 823	149 ± 97
Lee-Chao ($\tilde{N}$)	40 ± 22	193†	788 ± 1166	584 ± 772	112 ± 52
Bootstrap	19 ± 13	28 ± 16	154 ± 171	111 ± 86	37 ± 19
Jackknife	23 ± 26	38 ± 55	210 ± 426	152 ± 301	49 ± 94
OTUs found	15	21	118	84	29

Figure 5.18: Abbreviations are the same as in figure 5.17. *† Variances are missing for the same reasons as in figure 5.17.

5.7.3 Jackknife method

The jackknife method is a precursor of bootstrap. Instead of resampling, one sample at a time is eliminated in order to reduce the bias [183],[184].

Consider n independent identically distributed observations x_i and compute the statistic $\theta = f(x_1, ..., x_n)$. The jackknife estimate is given by following the procedure [185]:

1. Remove one of the observations, x_i for example.

2. Compute θ with the rest of the observations: $(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$ and call it $\hat{\theta}_{-i}$.

3. Compute the pseudo value $\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$.

4. Repeat 1 through 3 n times for $i = 1, ..., n$.

0.5% difference

	V	AC	P	PS	DS
Exponential	58^*	$0^§$	611 ± 53	446 ± 422	147 ± 3
Hyp. (ML)	85 ± 3	$0^§$	998 ± 92	696 ± 16	303 ± 1
Hyp. (LF)	86 ± 5	$0^§$	1009 ± 31	701 ± 18	302 ± 6
Lee-Chao ($\hat{N}$)	110 ± 88	$0^‡$	1066 ± 1001	989 ± 1217	208 ± 70
Lee-Chao ($\tilde{N}$)	64 ± 32	$0^‡$	1006 ± 971	950 ± 1148	$150^†$
Bootstrap	29 ± 32	32 ± 20	221 ± 381	151 ± 161	53 ± 44
Jackknife	36 ± 67	44 ± 80	304 ± 941	210 ± 603	70 ± 198
OTUs found	23	24	167	114	41

Figure 5.19: Abbreviations are the same as in figure 5.17. * Variance is missing for the same reasons as in figure 5.17. † As in the case for 3%, the coefficient of variation $\tilde{\gamma}$ comes negative out of equation 5.23 because most of the OTUs are detected in only one sample. ‡ For the apron and channel, all OTUs are detected in only one sample and the method understandably breaks down (see eq. 5.21).§ As can be seen in figure 5.16, the accumulation curve shows no curvature at all and therefore the estimation method breaks down.

The jackknife estimate (first order) is:

$$J_n^1(\theta) = 1/n \sum \hat{\theta}_i \tag{5.6}$$

and the variance estimate is:

$$var_{est}[J_n^1(\theta)] = \sum_{i=1}^{n} (J_n^1(\theta) - \hat{\theta}_i)^2)/(n(n-1)) \tag{5.7}$$

For our case, these can be rewritten to yield and estimate of the total number of OTUs and its variance (first order only):

$$J_n^1(S_T) = S_0 + [r_{1(1)}(n-1)]/n \tag{5.8}$$

$$var_{est}[J_n^1(S_T)] = \frac{1}{n} \sum_{i=1}^{n} [r_{1i} - (1/n) r_{1(1)}]^2 \tag{5.9}$$

where r_{1i} is the number of OTUs found only in sample i and $r_{1(1)}$ is the number of OTUs found in exactly 1 sample.

5.7.4 Exponential curve

This and the following section describe parametric methods for biodiversity estimations. The sampling process is modeled and an accumulation curved derived. The parameters in this expression are obtained by fitting the data.

The exponential accumulation curve proposed in [173] and first used in [77] is:

$$S(n) = S_{max}(1 - e^{-Kn}) \tag{5.10}$$

Where $S(n)$ is the number of OTUs found in n samples, S_{max} is the total number of OTUs and K is a fitted constant. The fit was done by defining $x \equiv e^{-n}$ and taking the derivative:

$$\frac{dS}{dn} = \frac{ds}{dx}\frac{dx}{dn} = S_{max}(-k)x^k(-x) = S_{max}(-x^k) \tag{5.11}$$

Plotting $\log(\frac{dS}{dn})$ versus $\log(x)$ allows us to find k and $S_{max}k$. The variance of S_{max} is given by the standard error of the linear fit slope. The results are given in the table in fig. 5.17 and the linear fits in figures 5.20, 5.21 and 5.22.

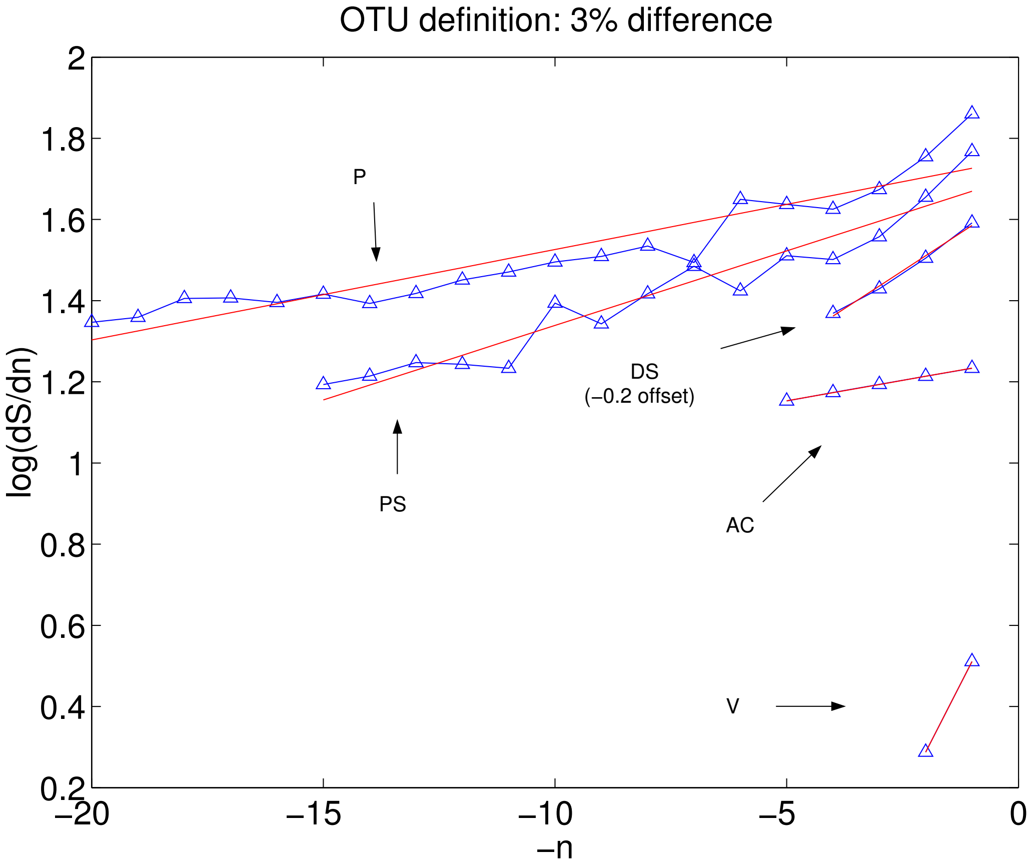


Figure 5.20: Linear fits to $log(\frac{dS}{dn})$ versus $log(x)$ for the exponential curve. The little variance in these estimates (see fig. 5.17) is due to the rather good fit of these plots. Numbers in parenthesis indicate the offset applied to the graphs for clarity.

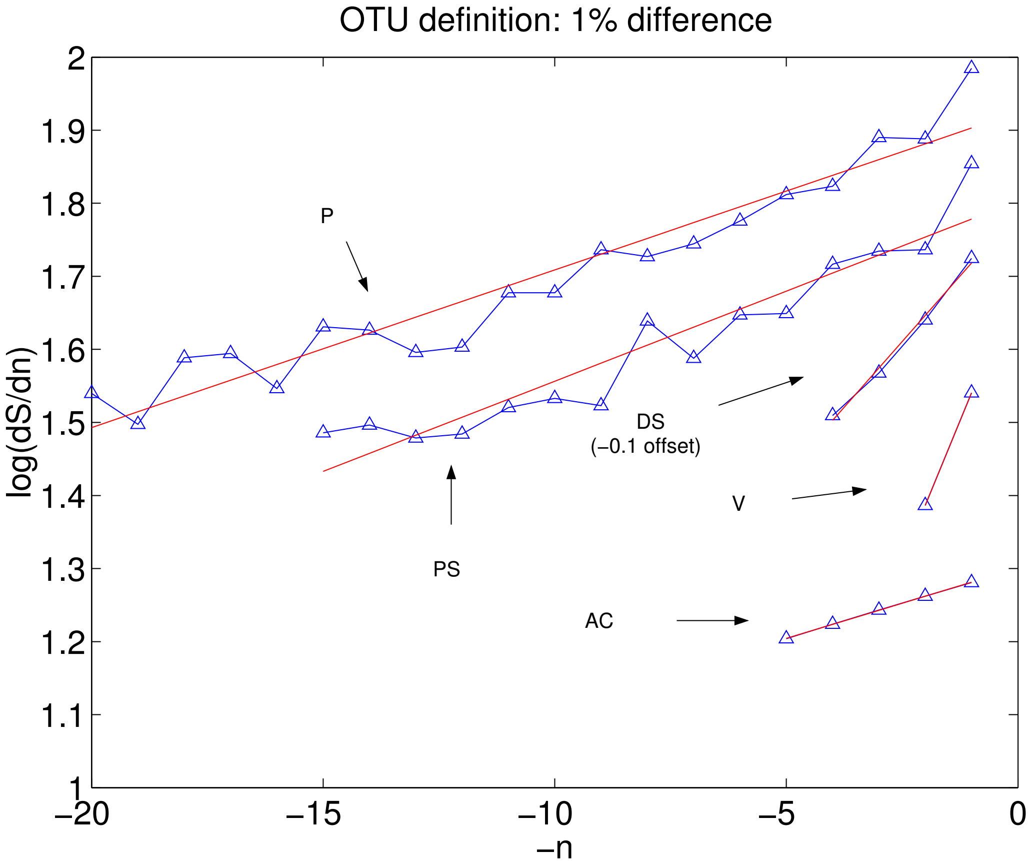


Figure 5.21: Same as figure 5.20 but for a 1% OTU definition.

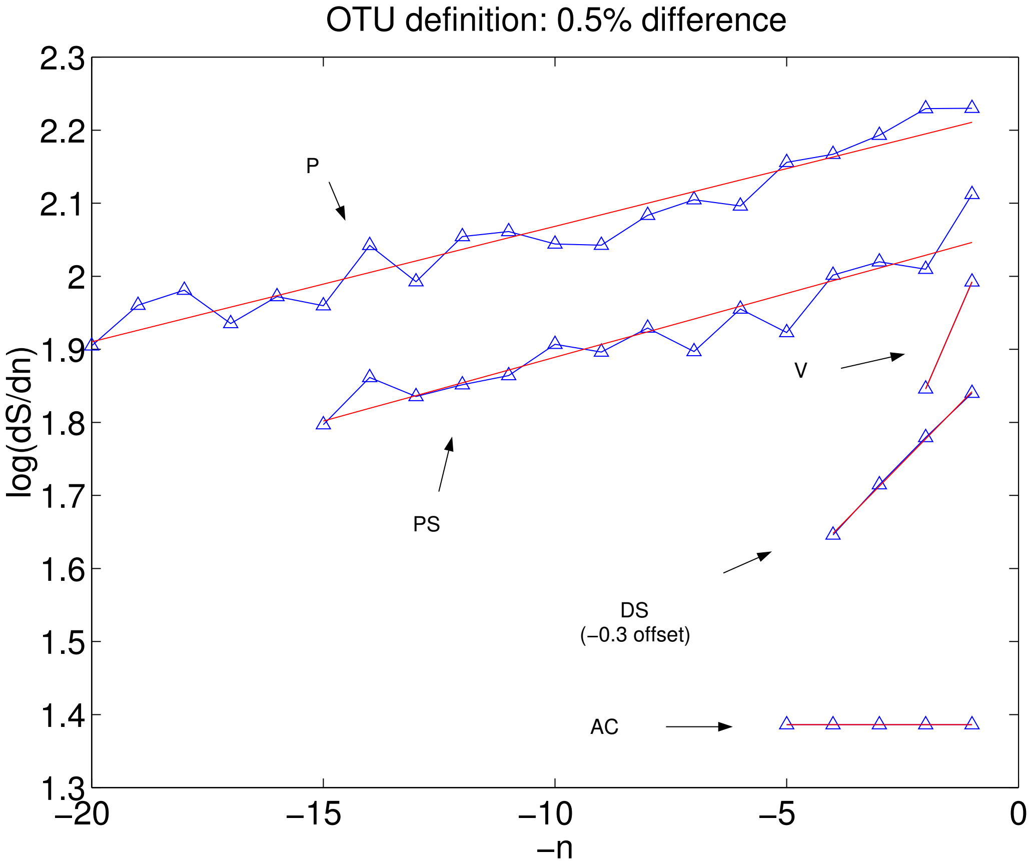


Figure 5.22: Same as figure 5.20 but for a 0.5% OTU definition.

5.7.5 Two parameter hyperbola

Another commonly used parametrization for the accumulation curve is ([173],[75]):

$$S(n) = \frac{S_{max} n}{B + n} \tag{5.12}$$

B being a fitted constant and the rest of symbols being the same as above.

It would be possible to fit this curve through a linear fit, but apparently this yields different results depending on the transformation of 5.12 that is used ([173],[174]). Having reviewed several possibilities Raaijmakers ([174]) leans toward using a maximum likelihood method for:

$$S(n) = S_{max} - \frac{BS(n)}{n} \tag{5.13}$$

By making the transformation:

$$X_i = \frac{S(n)}{n} \tag{5.14}$$

$$Y_i = S(n) \tag{5.15}$$

expressions for the constants can be derived:

$$B = \frac{\overline{X} S_{yy} - \overline{Y} S_{xy}}{\overline{Y} S_{xx} - \overline{X} S_{xy}} \tag{5.16}$$

$$S_{max} = \overline{Y} + B\overline{X} \tag{5.17}$$

where S_{yy},S_{xx} and S_{xy} are the sums of squares and cross products of the deviations $Y_i - \overline{Y}$ and $X_i - \overline{X}$.

The estimates are given in the table in fig. 5.17. Along with them, the results for a linear fit of the expression:

$$\frac{n}{S(n)} = \frac{B}{S_{max}} + \frac{n}{S_{max}} \tag{5.18}$$

where, for the purposes of the linear fit, $y \equiv \frac{n}{S(n)}$ and $x \equiv n$. The plots can be seen in fig. 5.23. The accumulation curve function for this method is the same but

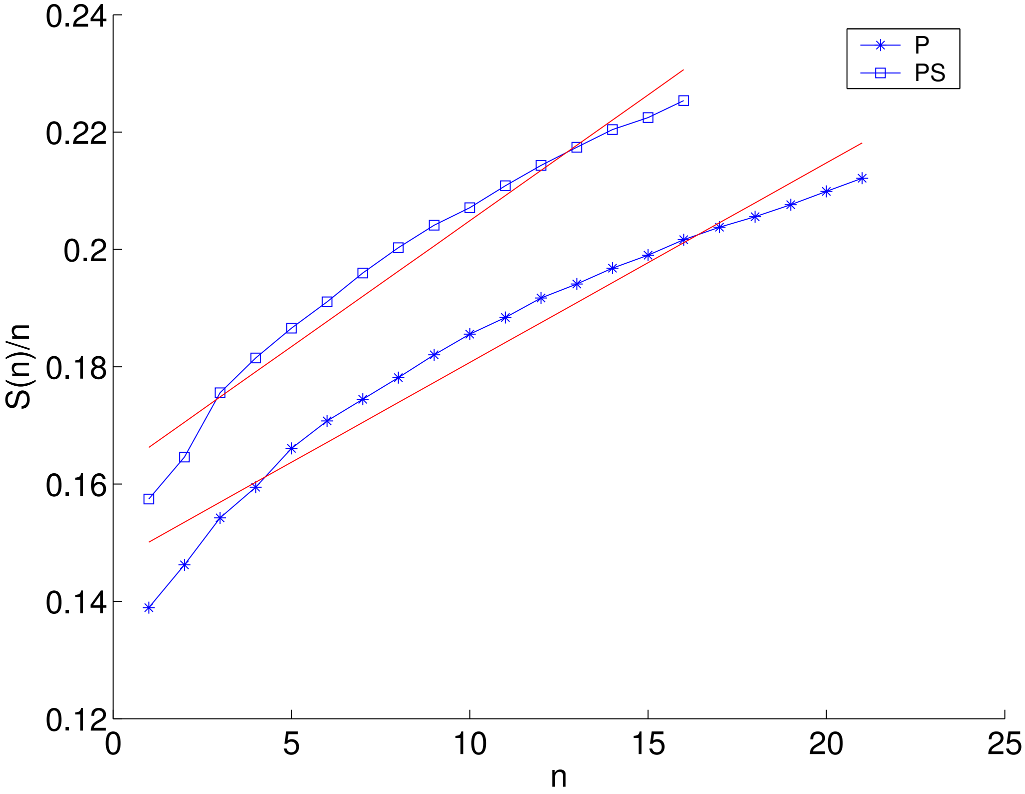


Figure 5.23: Linear fits to $n/S(n)$ versus n for the hyperbolic curve.

the fitting procedure is different. This additional procedure was carried out because the variance for the maximum likelihood method seemed too small and Colwell [173] already warned that the expressions for the variance could make little sense in the method by Raaijmakers. This gives more reasonable values for the variances and the estimates are not that different.

5.7.6 Lee and Chao estimator

Lee and Chao ([172],[171]) model sampling by assuming that each species is drawn with a probability p_i (i labeling species) in each sample. The number of individuals of each species (X_i in their notation) is then multinomially distributed. For p_i being different for each species and constant in time the following estimators for the total number of species are proposed:

$$\hat{N} = \frac{D}{\hat{C}} + \frac{f_1}{\hat{C}}\hat{\gamma}^2 \tag{5.19}$$

$$\tilde{N} = \frac{D}{\tilde{C}} + \frac{f_1}{\tilde{C}}\tilde{\gamma}^2 \tag{5.20}$$

where

$$\hat{C} = 1 - \frac{f_1}{\sum_{k=1}^{t} k f_k} \tag{5.21}$$

$$\tilde{C} = 1 - \frac{f_1 - 2f_2/(t-1)}{\sum_{k=1}^{t} k f_k} \tag{5.22}$$

and

$$\hat{\gamma}^2 = \frac{D/\hat{C} \sum k(k-1)f_k}{(t-1)(\sum k f_k)^2} - 1 \tag{5.23}$$

$\tilde{\gamma}^2$ being the same as $\hat{\gamma}^2$ with $\hat{C}$ replaced by $\tilde{C}$.

In the formulas above D is the total number of species collected, f_k is the number of species captured in exactly k samples, t is the number of samples and γ is the coefficient of variation (standard error / mean). The second estimate ($\tilde{N}$) is a bias corrected version of $\hat{N}$ for the case when the true value of the coefficient of variation is large.

5.7.7 Conclusions

Tables 5.17, 5.18, 5.19 and figures 5.24 and 5.25 show the results for all of the methods explained above. Jackknife and Bootstrap offer the lowest estimates, the Lee and Chao estimator produce the highest and the hyperbolic and exponential fits yield a middle ground. Obviously, as one can see in the accumulation curves (figs. 5.14, 5.15 and 5.16), we have just began to probe the biodiversity of the system and one can only expect to get a rough idea of the total OTU diversity. Even taking this into account, the differences between the estimates are substantial, up to a maximum of more that 400% difference depending on the facies. The likely reason for this is that most of these methods are applicable for a larger number of samples than we have (a minimum of 3 at the vent to a maximum of 25 in the pond).

Which estimate should we believe? To start with, the bootstrap and Jackknife estimators are known not be useful for sparsely sampled communities or for those with a large number of rare species [169]. This is evident from equations 5.3 and 5.8, since their maximum value is twice the number of observed species or OTUs. This explains their low values.

The Lee and Chao estimators produce preposterously large variances. They are based on a series expansion up to third order from the case that all p_i ($\equiv$ probability OTU i is drawn) are the same. It is not clear that this is applicable to the present data where the values of p_i are very different (see section 6). More importantly, the accuracy of this estimator seems very vulnerable to a low amount sample number and/or poor sampling. This can be seen in equations 5.19 and 5.21. In the case of poor sampling most OTUs will be detected in only one sample and equation 5.21 can become negative or even singular. Even when this is not the case, the low number of OTUs present in more than one sample makes equation 5.21 and hence equation 5.19 unreliable.

Finally, the estimates from the hyperbolic and exponential fit are very similar, but as can be seen in figures 5.20, 5.21, 5.22 and 5.23 the fit for the exponential is much better: the deviation from the line is random, whereas in the case of the hyperbolic fit, there is an obvious trend.

We will therefore use the exponential fit accumulation curve for our estimates. This means that present sampling has uncovered from 15% to 40% of the present biodiversity, depending on the facies. In spite of this, it is likely that we have detected the most abundant OTUs. Since the exponential curve can be derived assuming random sampling of individuals [130],[188], the fact that it gives a good fit suggests that random sampling is a good description of the process. Therefore, one can expect to find the most abundant OTUs first.

It is also worth noticing that the greatest biodiversity can be found in the Pond, even though it is not the largest facies. We suggest that this is because it is the facies with most fluctuations in pH, temperature and water flow rate, therefore promoting biodiversity. Obviously, one should take into account the size of the system when comparing relative biodiversities, since the relation is nonlinear as explained in chapter 2.3 when referring to the Species Area Rule ($S \propto A^z$). An intensive parameter (indpendent of size) such as z would be best to characterize the biodiversity, but we lack the appropriate data to measure it. In any event, this approach has provided important information on the level of sampling achieved and the expected biodiversity in the system.

Now that we have some reason to believe that the most abundant OTUs have been likely detected, the next step is to estimate relative abundance, since the impact on water chemistry of biogeochemically-active organisms will be weighted by their abundance.

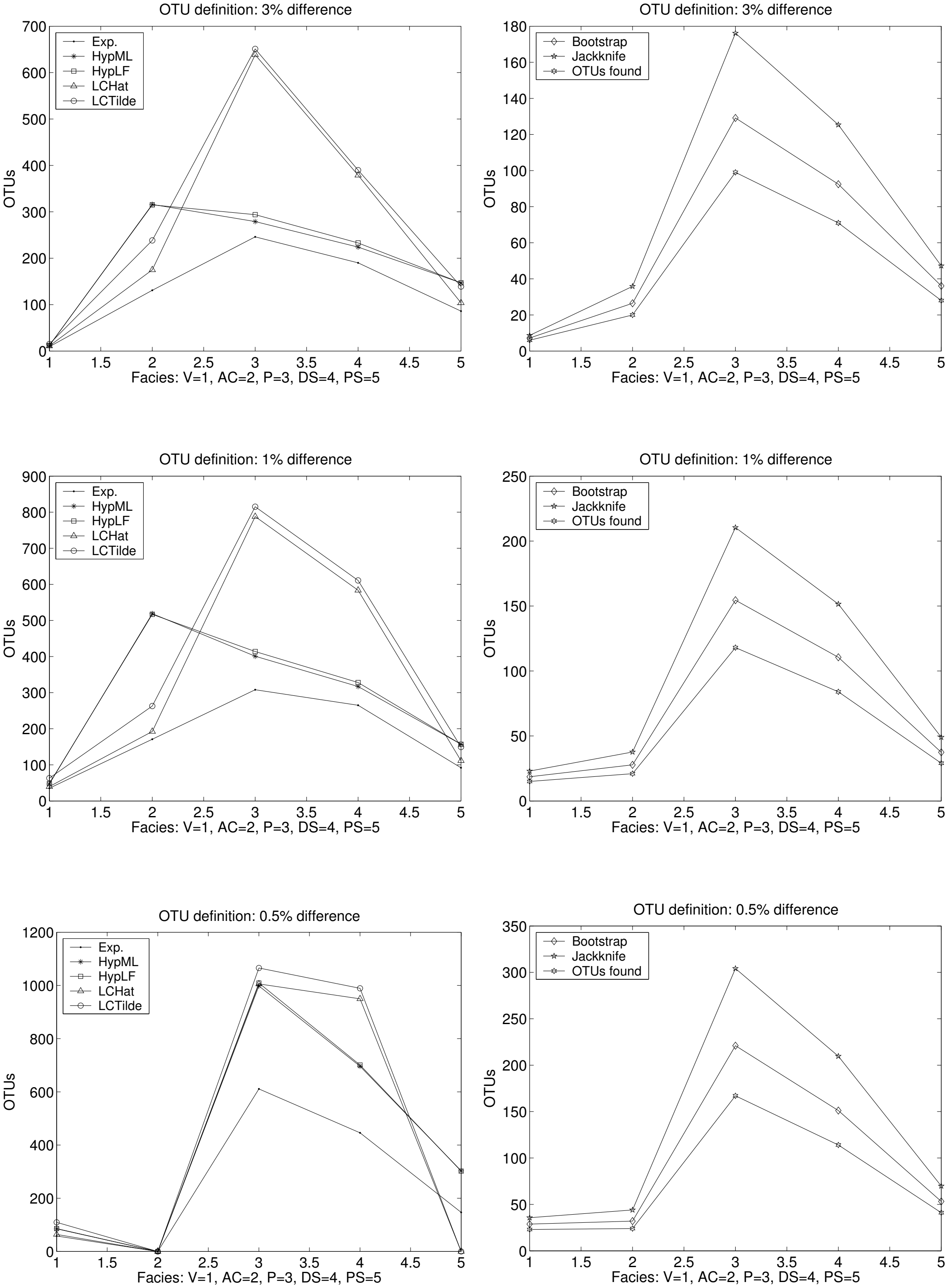


Figure 5.24: OTU estimation for each facies and each method explained in the text: exponential accumulation curve fit, hyperbolic accumulation curve fit using the maximum likelihood (ML) method and the linear fit (LF) method and the the Lee-Chao estimators for $\hat{N}$ (1) and $\tilde{N}$ (2).

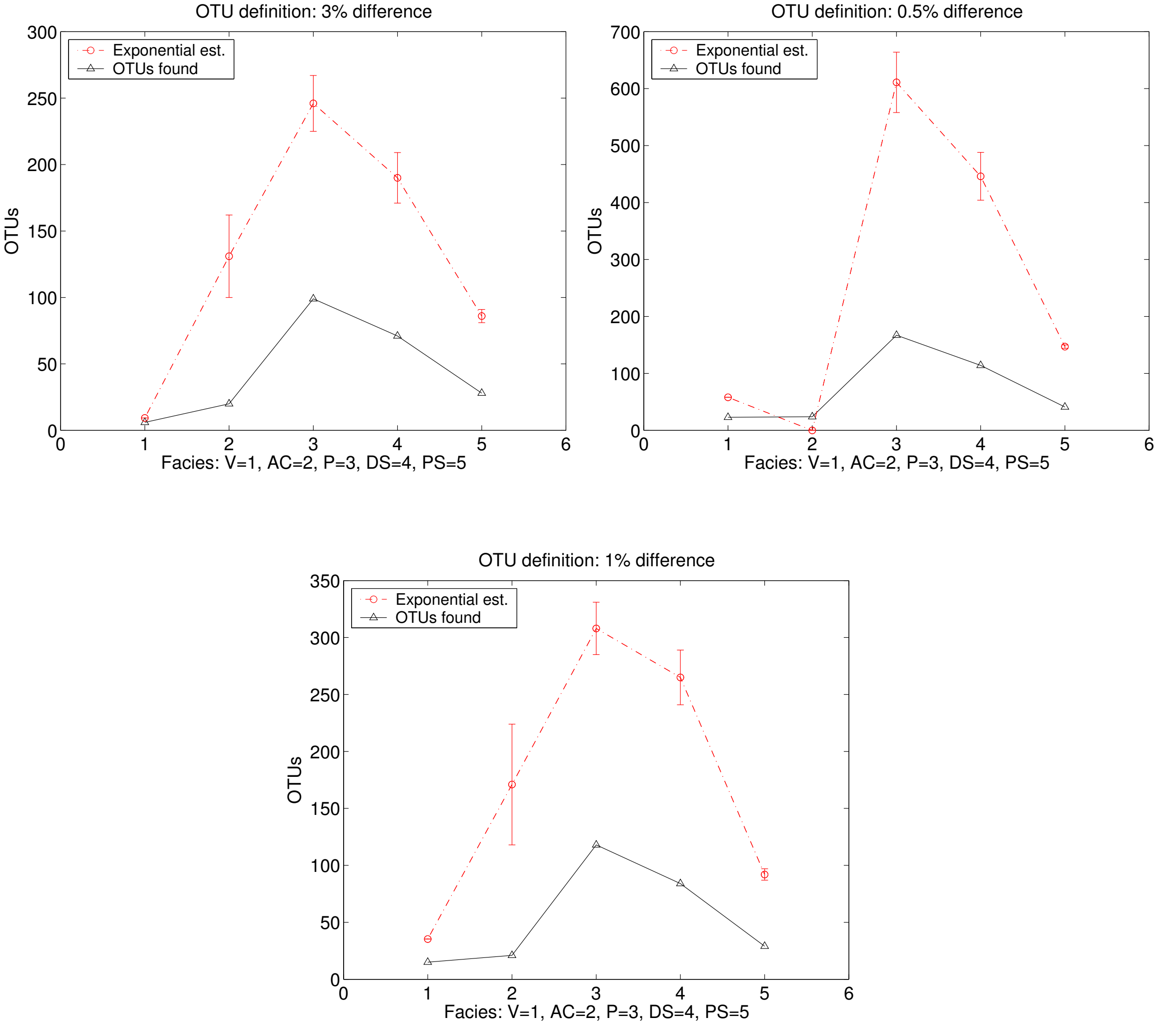


Figure 5.25: OTU estimates and OTUs found for each definition for the preferred method: exponential curve fit. The estimate for the Apron and Channel in the case of a 0.5% gives a null result since the accumulation curve is linear (see fig. 5.16).

Chapter 6

Microbial abundance distribution at Yellowstone's Mammoth Hot Springs

6.1 Introduction

Until recently, the study of microbial ecology was narrowly constrained by the difficulty of identifying bacteria outside of cultures. Sequencing of small subunit RNA genes [164], [165] put an end to this situation by permitting the classification and comparison of microbial species directly from an environmental sample. Alas, inherent biases in PCR amplification and cloning [189] prevent clone libraries obtained in this way to be reliable accounts of relative abundances. Several methods have subsequently been developed to provide quantification of abundances: Quantitative PCR, Most Probable Number PCR, competition PCR and dot-blot hybridization among others ([190], [191], [159], [160], [192]), each of them with their own advantages and disadvantages.

In this chapter I propose a complementary approach to determining relative abundances by using just clone library sequence information. Its implementation is fast, cheap and only requires the use of a computer. It is designed to be independent of the extraction and PCR amplification biases that make clone abundances unreliable, and gives a large scale, system-wide estimate of abundances, as opposed to estimates for a small volume from a specific point in the system. The technique is intended to be a quick, convenient first step in assessing the relative abundances of a microbial ecosystem, to be followed by a more in-depth accurate study based on the other tools mentioned above.

We illustrate this method with data from a hot spring in Yellowstone National Park. Its application yields information on the most abundant OTUs and the ones most capable of driving the ecosystem metabolism. Our analysis pinpoints OTUs to focus our work on, allowing us to concentrate future efforts on the 10 or 15 most abundant instead of the full 300 OTUs detected in previous work [130].

6.2 Relative abundance estimation

We define the relative abundance ρ_i as the fraction of total individuals in the system belonging to OTU i : $\rho_i = n_i/n$. Here n_i is the number of individuals belonging to OTU i and n is the total number of individuals. $i = 1..S$, S being the total number of OTUs in the system. Samples are assumed to be collected at each facies (see chapter 5) and to be processed by DNA extraction, 16S rRNA gene PCR amplification, cloning and clone screening to create a clone library as explained in ref. [130] and chapter 5.

Ideally, one would count every individual in the system and assign it to an OTU to calculate ρ_i. This is unfeasible first and most obviously because of the impossibility of sampling the whole system and, secondly, this procedure would give a very biased estimate of relative abundance. The reason is that clone libraries exhibit biases due to differential PCR amplification. A small preference in primer binding for a certain OTU type is exponentially amplified and distorts abundances greatly. Other biases are introduced by the DNA extraction, ligation and transformation, but they lack the exponential growth inherent to PCR DNA amplification. We will therefore assume that the only information which is trustworthy is the presence or absence of each OTU. Even this is biased because a species can be present but not detected in a particular sample. However, as we show below, repeated independent sampling of the environment largely overcomes this problem. In particular, the technique presented here is not affected strongly by PCR bias, because it uses only binary (presence or absence) information about each OTU.

In this situation, if the system were homogeneous and there were no detection errors (see appendix 6.8), every sample would give the same result. This is patently not the case for the data shown in section 6.5.

Why does the variation from sample to sample arise? There are three main contributions: spatial and temporal variation and detection errors. As explained in section 6.5, samples were taken in different spatial locations within the same facies and at different times of the year and day. Microbial species show preferred ranges of temperature and pH ranges and have been shown to partition fairly tightly to given facies, as described in chapter 5. It is therefore not surprising that variations of pH, temperature and other facies characteristics give rise to distinctive patterns in the location of OTUs.

Detection errors come mainly from two sources: firstly, the processes of extraction, amplification, ligation and transformation have an intrinsic variability in their success rate (see appendix 6.8). Secondly, some of the clones submitted for sequencing (after screening) to the sequencing facility were lost in failed reactions, and hence never identified. Note that the probability of failed sequencing is independent of the OTU abundance, so, in principle, OTUs with high relative abundance might not be detected. With the amount of available data it is hard to tell how much variance is due to spatial and temporal variation, and how much is due to detection errors.

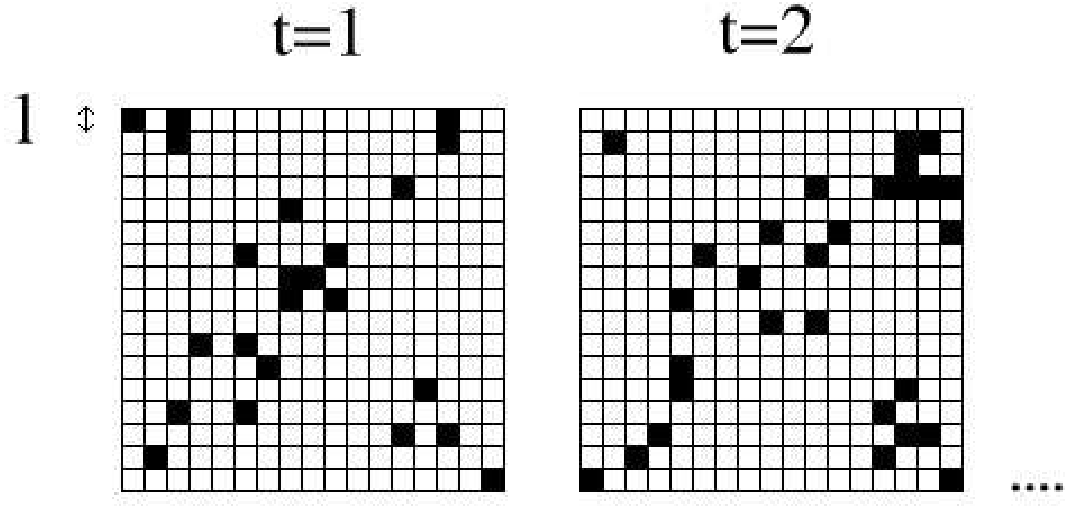


Figure 6.1: Cover is a "coarse grained" abundance in the sense that each cell will contribute if the species is present inside it, independently of its abundance. The cover is the number of cells in which a given OTU is present divided by the total number of cells (equation 6.1), for each of the possible times t ($t = 1..T$). For example, for $t = 1$ the cover is $C_i^1 = 23/289$, and for $t = 2$, $C_i^2 = 29/289$. Each of the squares is a diagram representing one of the facies in the system. This is the same concept as used in macroecology (see section 2.2).

In spite of all this, if we assume that most of this variance comes from spatial and temporal variations, we can use the collected data to obtain estimates of "coarse grained" abundances as explained in figure 6.1.

Assume that the square represents one of the facies in the system, properly divided into smaller cells of the size l. The correlation length l is defined to be small enough so that sampling within the boundaries of a cell always yields the same result. One can define the *cover* of OTU i to be the fraction of subcells in which OTU i is present over the total number of subcells:

$$C_i^t = \frac{X_i^t}{X} \tag{6.1}$$

and the time averaged coverage is:

$$C_i = \sum_{t=1}^{T} \frac{C_i^t}{T} \tag{6.2}$$

As a starting point, and lacking any better hypothesis, we will assume that the

		OTU number							
Sample		1	2	3	4	5	6	7	
A		●	●		●	●			
B		●	●	●		●			
C		●	●		●			●	
D		●					●	●	
$\hat{C}_i$	(	4	3	1	2	2	1	2	)/4
$\hat{\rho}_i$		26.7%	20%	6.7%	13.3%	13.3%	6.7%	13.3%	

Figure 6.2: Example of how to calculate the estimates of the coverage $\hat{C}_i$ and abundance $\hat{\rho}_i$ for a given number of samples according to equations 6.4 and 6.5. Only information of presence or absence of a given OTU is used.

coverage is proportional to the abundance ($C_i \propto \rho_i$) and as a result:

$$\rho_i = \frac{C_i}{\sum_i C_i} \tag{6.3}$$

Sampling the whole facies to find the true coverage C_i is out of reach. Random sampling from each facies yields estimates (denoted by the hat) that should converge quickly:

$$\hat{C}_i = \frac{N_i}{N} \tag{6.4}$$

$$\hat{\rho}_i = \frac{\hat{C}_i}{\sum_i \hat{C}_i} \tag{6.5}$$

where N_i is the number of samples in which OTU i is present and N is the total number of samples (see fig. 6.2).

Equation 6.5 provides a useful estimate of relative coverages and abundance. The geobiochemical impact of a given organism's metabolic activity is weighted by its abundance, so this information is helpful in assessing the role of micro-organisms in mineralization.

Finally, equation 6.5 provides an estimate for the relative abundance but not its variance, critical to ascertaining the variability of ρ_i across the facies and in time.

The bootstrap method presented in the next section gives us an estimate not only of relative abundance, but also of the variance.

6.3 The Bootstrap method

The Bootstrap method is a data resampling method of wide applicability introduced by Efron in 1979 to assess the accuracy of statistical estimates and provide bias corrections [186]. A broad exposition can be found in references [193], [183] and [194]. Here we limit ourselves to explaining its use for the case at hand.

How can the bootstrap method be used to obtain a variance for the relative abundance estimate in 6.5? Traditionally, one would divide the N samples in M groups of N/M samples and obtained the estimates of $\hat{\rho}_i^s$ $(s = 1..N/M)$ for each of these groups as per equation 6.5. The variance would be obtained from the usual formula: $var(\hat{\rho}_i) = \sum_s (\hat{\rho}_i^s - \hat{\rho}_i^{av})^2$, where $\hat{\rho}_i^{av}$ denotes the average of $\hat{\rho}_i^s$. For large enough N this would converge to the desired variance. Nonetheless, this procedure means having less samples for each estimate $\hat{\rho}_i^s$ $(s = 1..M)$ and a worse estimation. For example, for $N = 8$ (as for data in section 6.5) having 4 groups would lead to a meager 2 samples per group.

The bootstrap explores the variance by forming groups in which some samples are left out but which have the same amount of samples per group as the total initial number of samples. This is achieved by choosing these groups through sampling with replacement as explained in figure 6.3: R bootstrap groups are created and a relative abundance estimate $\hat{\rho}_i^s$ is calculated for each. The bootstrap theorems state that, for large enough R, the behaviour of the $\hat{\rho}_i^s$ around $\hat{\rho}_i$ mimics the behaviour of $\hat{\rho}_i$ around ρ_i (see appendix 6.7 and reference [194]). One can therefore obtain an improved estimate and its variance by treating the bootstrap groups as independent

Initial samples		1st BS resampled group			sth BS resampled group			Rth BS resampled group	
$s=0$		$s=1$			s			$s=R$	
A		C			D			D	
B	$\Rightarrow \hat{\rho}i$	A	$\Rightarrow \hat{\rho}_i^1$	.....	B	$\Rightarrow \hat{\rho}_i^s$	.....	A	$\Rightarrow \hat{\rho}_j^R$
C		A			A			B	
D		D			D			C	

Figure 6.3: The bootstrap method applied to estimate the bias and variance of $\hat{\rho}_i$. R groups of four samples are generated by sampling with replacement from the original samples. This means that the samples in each group are chosen randomly among the original samples and each time a sample is selected for the group it is returned to the original set, so it can be chosen again. Therefore each group is not just a permutation of the initial samples. For each group, $\hat{\rho}_j^s$ is generated using equation 6.5 and the estimate of the bias and the variance are given by equations 6.6 and 6.7.

measurements:

$$\rho_i^{BS} = \frac{1}{R}\sum_{s=1}^{R}\hat{\rho}_i^s \tag{6.6}$$

$$var(\hat{\rho}_i^{BS}) = \frac{1}{R}\sum_{s=1}^{R}(\hat{\rho}_i^s - \hat{\rho}_i^{BS})^2 \tag{6.7}$$

The bootstrap principle as stated above is not always applicable (e.g. extremal statistics [194]) and the convergence to the right distribution must be proved for each estimator [193]. Equations 6.6 and 6.7 refer to functions of sample means for which the probability distribution of the bootstrap resampling has been proved to converge to the probability distribution of the estimates in the limit of large N (See appendix 6.7).

Incidentally, it is worth saying that in this case, because of the linear character of the statistic $\hat{\rho}_i$, the bootstrap estimate is the same as the non-bootstrapped case.

6.4 Test distributions

The theorem in appendix 6.7 proves the consistency of the bootstrap estimator in the asymptotic limit, which is a necessary condition for the validity of the bootstrap method. The real interest of the bootstrap are its fixed sample properties. The variance eventually converges to the true variance for $N \to \infty$, but for a finite N it also gives a measure of the variability of relative abundances by omitting the use of certain samples, therefore yielding an account of its reliability. There is no general theory for fixed sample properties of the bootstrap. Its performance is usually examined through empirical simulations [193], [183].

In this section we assess the performance of the bootstrap by generating a series of samples from a known abundance distribution ρ_i and checking how close the bootstrap estimation using N of these samples is to the original distribution.

We suppose that each of the S OTUs in the system are present in each sample with probability $\rho(i) \propto i^{-0.65}$ $(i = 1..S)$. The total number of OTUs S is chosen to be be $S = 200$ since the number of OTUs in, for example, the water filters of the Pond is 43 and previous results [130] indicate that $20\% - 25\%$ of the total diversity has been sampled.

Figures 6.4 and 6.5 show the results of the bootstrap estimates $\hat{\rho}_i(N)$, for sample numbers $N = 10, 100$ and $R = 10000$ as compared to the original relative abundance ρ_i. The results are satisfactory, with the target abundance within the variance of the estimate. As expected, estimates improve with increasing N. For low N the estimates overshoot slightly since not all S OTUs have been detected and therefore the detected OTUs are given a higher relative abundance than the real one.

R is in practice chosen large enough so that further increases don't change the estimate appreciably.

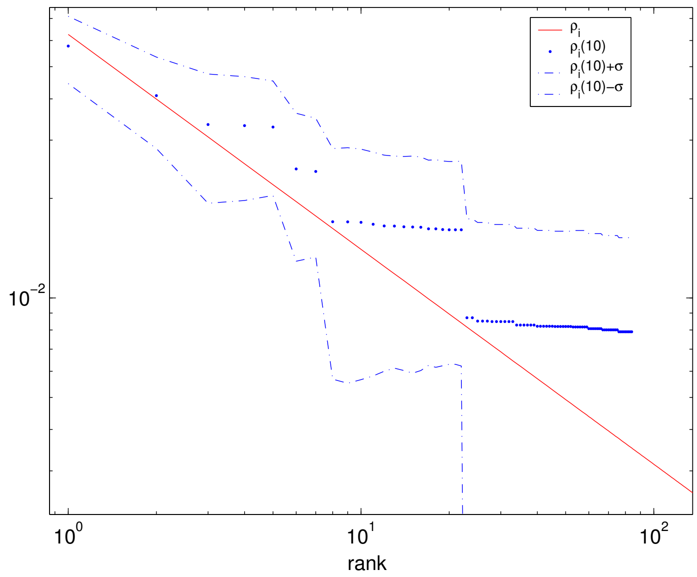


Figure 6.4: Bootstrap estimate for $N = 10$ samples. In spite of this low number of samples it is possible to get a hint of the underlying distribution. The estimate overshoots because for such low amount of samples not all OTUs have been detected and therefore the detected OTUs are assumed a higher relative abundance so they all add up to 1.

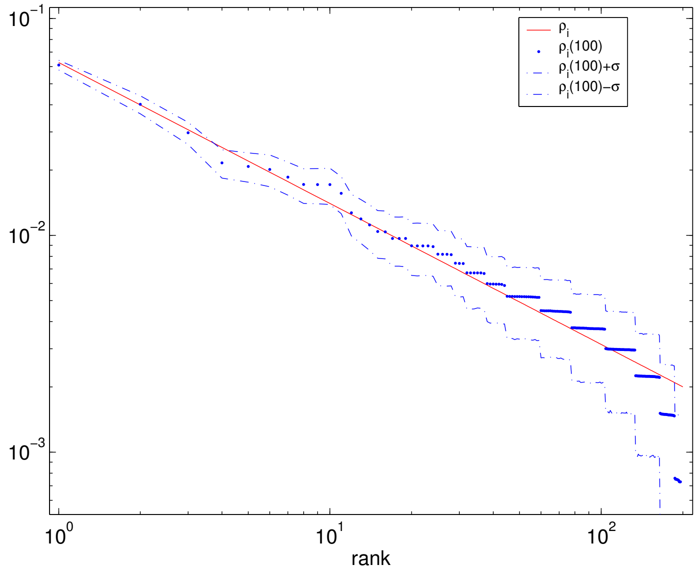


Figure 6.5: The bootstrap estimate improves for $N = 100$ as expected and becomes quite close to the target distribution.

6.5 Yellowstone National Park data

6.5.1 Study site

Up to 50 samples were taken during an interval of 4 years at Spring AT-1, located on Angel Terrace, in the upper terrace region of the Mammoth Hot Springs complex at Yellowstone National Park (see section 6.5). This spring presents the typical characteristics of springs at the complex: hot waters erupt from the vent and flow downhill cooling down, quickly degassing CO_2, increasing in pH and precipitating travertine at extremely fast rates ($\sim$ 1.5 m per year). This produces its characteristic terraces formations. Samples were taken for all the five facies: vent, apron and channel, pond, proximal slope and distal slope and two different mediums: filtered water from the flow and pieces of travertine substrate up to 2 cm. deep, including the thick mat of bacteria on top of it (See chapter 5 for details of facies definitions and more specific information on the site).

Bacteria were identified through 16S rRNA gene identification as explained in [130] and chapter 5. Three difference sets of OTU definitions were used, based on sequence differences of 0.5%,1% and 3%.

6.5.2 Bootstrap estimates

The procedure for obtaining the abundance ρ_k^{BS} is the same as explained above with a total of $R = 10000$ bootstrap samples being used. The results are given in the form of rank abundance plots in figures 6.9, 6.13 and 6.17, and in the form of rank tables for all facies and mediums in figures 6.8, 6.12 and 6.16, along with the number of samples for each case. In all these graphs, mat or substrate refer to pieces of travertine substrate up to 2 cm. deep, including the thick mat of bacteria on top of it.

As can be seen, only the Pond and Proximal Slope facies have enough samples for the resulting abundances to be meaningful. Nonetheless, even for 3 or 4 samples the results give a qualitative idea of trends in relative abundances.

In the case of the Pond and Proximal Slope the abundances seem to fit a power law for the water samples, in contrast with the mat and substrate.

It can also be noticed that among the highest ranking OTUs there is a certain degree of commonality in the case of the water samples, but not so in the rest. This could be due to the downstream flush of cell being only feasible in the water. Also, from the top ranking OTUs in the water, very few are found in the top ranks of the substrate. It can therefore be concluded that encrustment is not random, i.e. that some species are more able to avoid it than others.

Finally, OTU 5 (using the 3% definition) seems to be the most abundant in all facies. This OTU is an unknown beta proteobacterium and corresponds to OTU 8 in the 1% definition. Under the 0.5% definition, it is divided between OTU 39 and several others not shown in the tables because they are not among the most abundant. This seems to suggest that OTUs defined up to 1% occupy the same ecological niche, whereas an OTU definition based on a genetic difference of less than 0.5% is not ecologically useful. In accord with this, high variances for abundance estimations are noticed in the case of the 0.5% definition, hinting that this may be too narrow a distinction for OTU definitions, at least with such a small number of samples.

6.6 Conclusion

We have presented a computational method that uses clone library information to provide an estimate of relative abundance. This can be used to give a first hint about which bacterial OTUs are more relevant and supply possible candidates for later quantitative work involving (e.g.) hybridization probes.

This method has been demonstrated using data from hot springs at the Yellowstone National Park to provide estimations of relative abundances for different facies and mediums. The most abundant OTUs have the greatest potential for influencing the degassing of CO_2 which, in turn, produces calcium carbonate precipitation and ultimately gives rise to the formation of the travertine terraces.

The data for abundances seem to fit well a power law for the water samples, similar in functional form to other power laws involving size and rank, such as the Pareto law in economical systems [195]. The abundance patterns are interesting in themselves as a characteristic of the microbial ecosystem and are useful in planning future sampling.

Commonality of most abundant species is high in the water and scarce in the substrate or within mediums in the same facies. This fact can be attributed to the water downflush of bacteria. In any case, it would be limited to the most abundant bacteria, since there is very little commonality of OTUs between facies [130] (see section 6.5).

The little commonality of OTUs between the water and the substrate seems to indicate that encrustment is not random: some OTUs are more able to avoid it than others. This, in turn, suggests a non-passive role of microorganisms in carbonate precipitation.

Finally, the use of 3 different sets of OTU definitions permits us to explore the issue of the proper definition of OTUs/species. We conclude that differentiating OTUs by 0.5% may be excessive and advocate the 1% difference definition.

6.7 Appendix A

This appendix states the result that proves the applicability of the bootstrap procedure for the estimator $\hat{\rho}_i$.

We define X_j^i such that $X_j^i = 1$ if species i is present in sample j and $X_j^i = 0$

otherwise. In this case, $\hat{\rho}_i(N) = \sum_j X^i_j / \sum_{i,j} X^i_j$ (as per equation 6.5). Then, if X^i_j are random with finite second moments, it can be proved that the distribution of $\sqrt{N}(\hat{\rho}^s_i(N) - \hat{\rho}_i(N))$ will converge in the asymptotic limit ($N \to \infty$) to the same distribution as $\sqrt{N}(\hat{\rho}_i(N) - \rho_i)$, namely a gaussian with variance σ_i (which depends on the average of X^i_j). More explicitly (see [193], example 3.3):

$$\mathrm{P}\{\sqrt{N}\,(\hat{\rho}^s_i(N) - \hat{\rho}_i(N)) < x \;\; \forall i, s\} \to_{a.s.} \Phi(x/\sigma_i) \tag{6.8}$$

where $\Phi(x)$ is the standard normal distribution:

$$\Phi(x) = 1/\sqrt{2\pi} \int_{-\infty}^{x} e^{-x^2/2} \tag{6.9}$$

Notation: $P\{A\}$ denotes probability that clause A is true and $\to_{a.s.}$ denotes almost surely convergence or convergence with probability $1: \; P\{X_n \to X\} = 1 \Rightarrow X_n \to_{a.s.} X$.

6.8 Appendix B

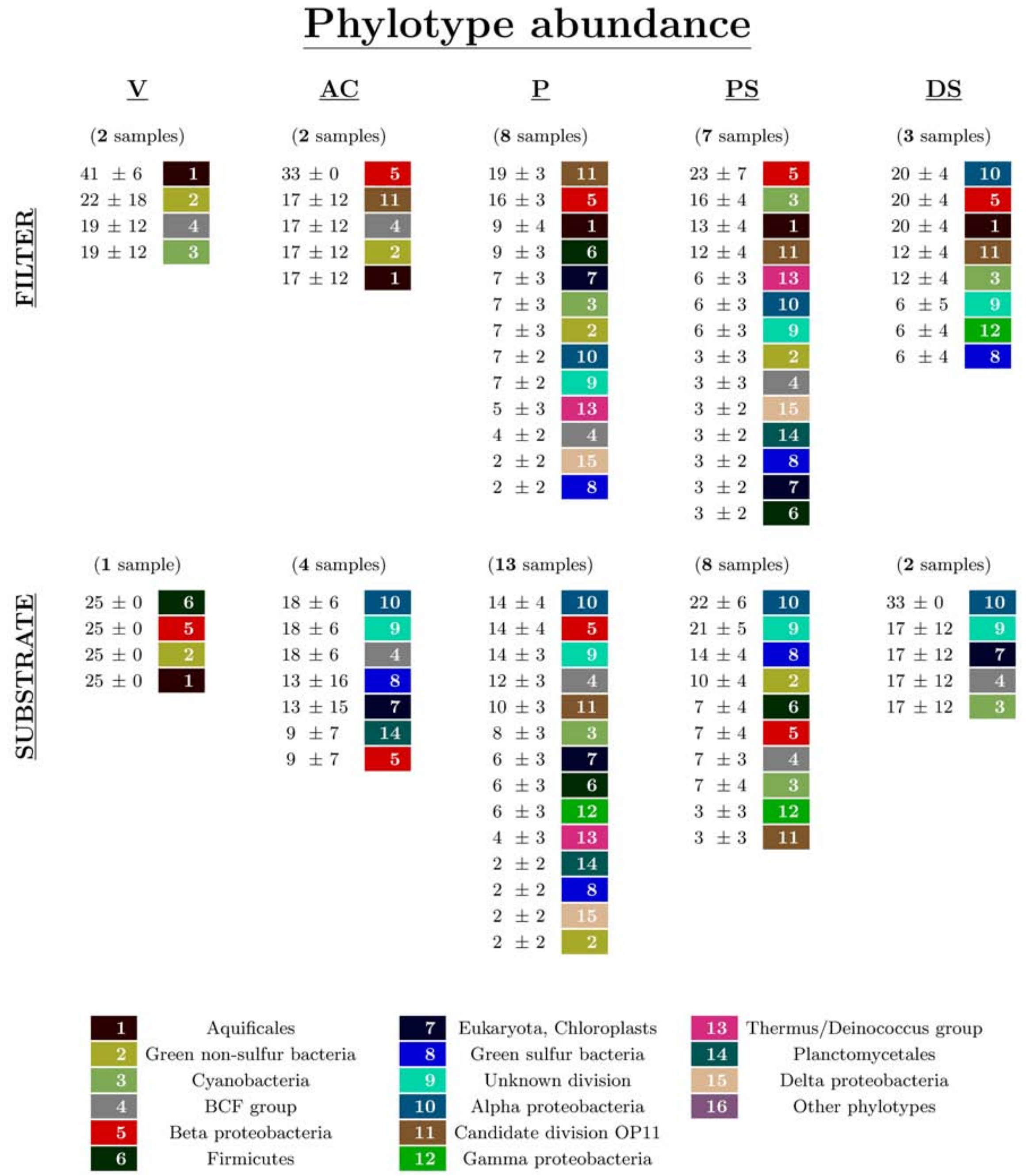
The purpose of this appendix is to show that, aside from detection errors, all relevant OTUs in a sample are detected.

Environmental samples go through the process of DNA extraction, PCR amplification, ligation and transformation. The result is a series of clone colonies, each with a different DNA insertion. Up to 200 of these colonies are randomly chosen for further analysis, which involves singling out unique OTUs through RFLP analysis. We assume that due to the biases of the previous processes, the relative fraction of clones belonging to OTU i is $\rho^*_i \neq \rho_i$ at this stage. The probability that OTU i is present in any of these 200 colonies is one minus the probability that it is not present: $P_i = 1 - (1 - \rho^*_i)^{200}$. For $\rho^*_i > 0.000005$, $P_i < 0.001$, so every OTU with a relative abundance higher that 0.0005% will be present with a probability bigger than 99.9%.

In some cases the first 18 clones had the same RFLP signature and examination of the rest was dropped. In this case the probability that at least one other OTU is present in the first 18 clones is $1 - (\rho_i^*)^{18}$, so proceeding as above, it can be shown that if $\rho_i^* > 0.99994\%$ then, with a probability greater than 99.9%, no other OTUs appear in these 18 clones. The rest of the OTUs then amount to less than 0.00006%. In some other instances, after RFLP analysis of the first 18 clones, the next 18 offered no new results and the rest of the analysis was dropped. In this case, the probability that OTU i is not present in the first 36 clones is $(1-\rho_i^*)^{36}$. It can then be concluded that if $\rho_i^* < 0.002779\%$, then OTU i is not present with more than 99.9% probability.

The conclusion from this analysis is that any OTU with a relative abundance $\rho_i^* > \rho_c = 0.0028\%$ should have been detected, barring detection errors. Certainly our result ρ_i^* is only an approximation for the actual value ρ_i. Whatever biases in the extraction, amplification, ligation and transformation processes are present are, we believe, not strong enough as to be able to drive the relative abundance ρ_i of any relevant OTU ($\rho_i > 1\%$) below the detection threshold ρ_c.

Finally, it should be mentioned that the procedures described above are ideal. In many cases, less than 200 clones with inserted DNA are obtained, sometimes even only one. This may be due to many reasons: cell lysis problems, primer annealing difficulties in PCR reactions, variations in ligation success rates, variations in receptivity of transformation host cells, purification problems after PCR reactions... etc. These effects are hard to control and contribute to our detection errors.



Phylotype abundance

FILTER

V		AC		P		PS		DS	
(2 samples)		(2 samples)		(8 samples)		(7 samples)		(3 samples)	
41 ± 6	1	33 ± 0	5	19 ± 3	11	23 ± 7	5	20 ± 4	10
22 ± 18	2	17 ± 12	11	16 ± 3	5	16 ± 4	3	20 ± 4	5
19 ± 12	4	17 ± 12	4	9 ± 4	1	13 ± 4	1	20 ± 4	1
19 ± 12	3	17 ± 12	2	9 ± 3	6	12 ± 4	11	12 ± 4	11
		17 ± 12	1	7 ± 3	7	6 ± 3	13	12 ± 4	3
				7 ± 3	3	6 ± 3	10	6 ± 5	9
				7 ± 3	2	6 ± 3	9	6 ± 4	12
				7 ± 2	10	3 ± 3	2	6 ± 4	8
				7 ± 2	9	3 ± 3	4		
				5 ± 3	13	3 ± 2	15		
				4 ± 2	4	3 ± 2	14		
				2 ± 2	15	3 ± 2	8		
				2 ± 2	8	3 ± 2	7		
						3 ± 2	6		

SUBSTRATE

V		AC		P		PS		DS	
(1 sample)		(4 samples)		(13 samples)		(8 samples)		(2 samples)	
25 ± 0	6	18 ± 6	10	14 ± 4	10	22 ± 6	10	33 ± 0	10
25 ± 0	5	18 ± 6	9	14 ± 4	5	21 ± 5	9	17 ± 12	9
25 ± 0	2	18 ± 6	4	14 ± 3	9	14 ± 4	8	17 ± 12	7
25 ± 0	1	13 ± 16	8	12 ± 3	4	10 ± 4	2	17 ± 12	4
		13 ± 15	7	10 ± 3	11	7 ± 4	6	17 ± 12	3
		9 ± 7	14	8 ± 3	3	7 ± 4	5		
		9 ± 7	5	6 ± 3	7	7 ± 3	4		
				6 ± 3	6	7 ± 4	3		
				6 ± 3	12	3 ± 3	12		
				4 ± 3	13	3 ± 3	11		
				2 ± 2	14				
				2 ± 2	8				
				2 ± 2	15				
				2 ± 2	2				

1	Aquificales	7	Eukaryota, Chloroplasts	13	Thermus/Deinococcus group
2	Green non-sulfur bacteria	8	Green sulfur bacteria	14	Planctomycetales
3	Cyanobacteria	9	Unknown division	15	Delta proteobacteria
4	BCF group	10	Alpha proteobacteria	16	Other phylotypes
5	Beta proteobacteria	11	Candidate division OP11		
6	Firmicutes	12	Gamma proteobacteria		

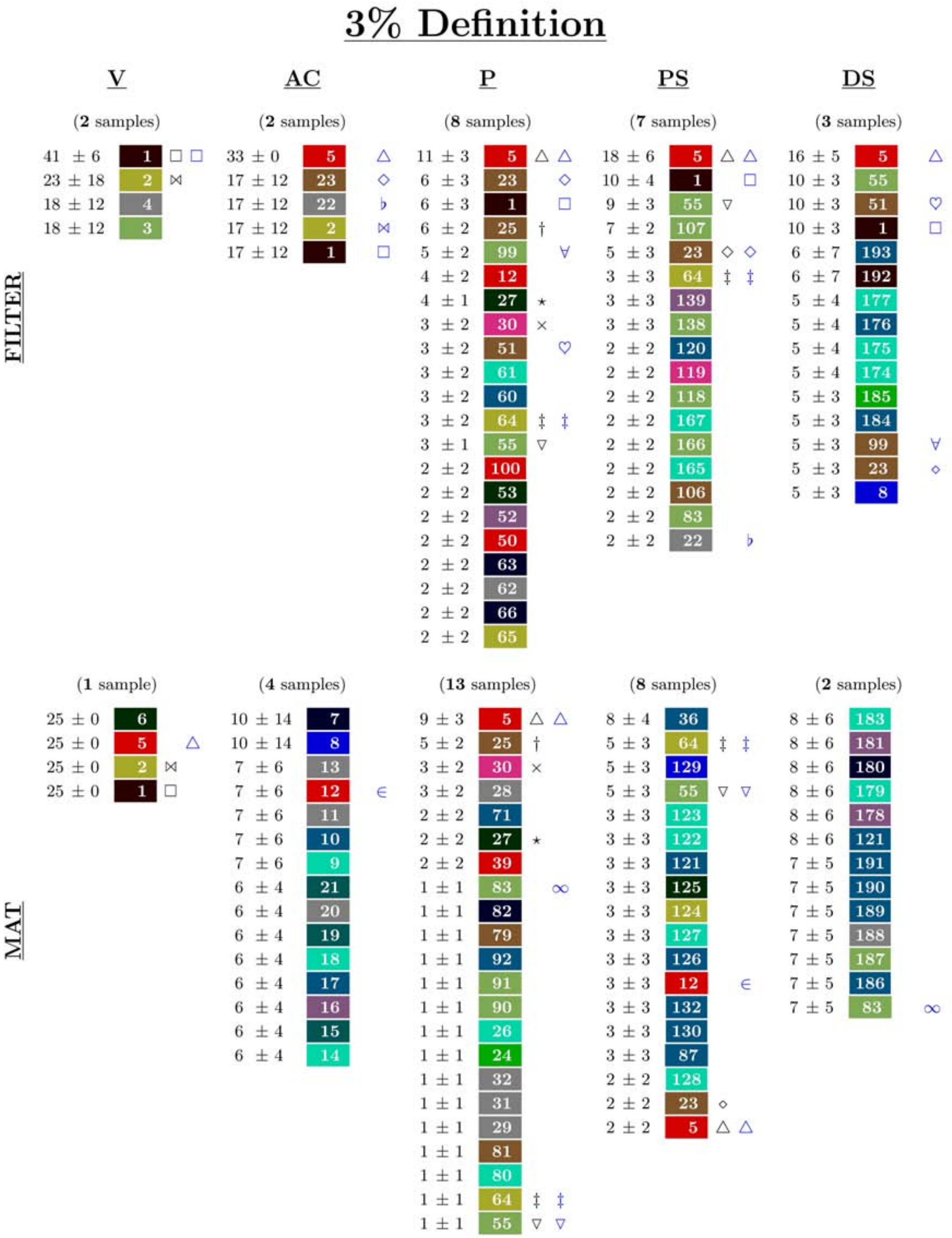
Figure 6.6: Phylotype relative abundances and variances for each facies and medium. Each present phylotype is identified by a color throughout the whole paper. Numbers change for each grouping (phylotypes and 3%, 1%, 0.5% differences).

3% Definition

1	SM2-G04	Aquificales
2	SM2-G02	Green non-sulfur bacteria
3	6-2-99-6 # 10	Cyanobacteria
4	6-2-99-6 # 12	BCF group
5	SM1-E12	Beta proteobacteria
6	FL1-F09	Firmicutes
7	FL13-E05	Eukaryota, Chloroplasts
8	FL10-G07	Green sulfur bacteria
9	FL13-A10	Unknown division
10	FL13-A11	Alpha proteobacteria
11	FL13-A12	BCF group
12	SM1-G05	Beta proteobacteria
13	FL13-B02	BCF group
14	FL13-A01	Unknown division
15	FL13-A02	Planctomycetales
16	FL13-A03	Verrucomicrobium group
17	FL13-A04	Alpha proteobacteria
18	FL13-A05	Unknown division
19	FL13-A08	Planctomycetales
20	FL13-A07	BCF group
21	FL13-A09	Planctomycetales
22	SM1-D01	BCF group
23	6-2-99-10 # 13	Candidate division OP11
24	SM2-E10	Gamma proteobacteria
25	SM1-E10	Candidate division OP11
26	FL8-F12	Unknown division
27	SM2-D03	Firmicutes
28	FL8-H07	BCF group
29	FL8-B07	BCF group
30	SM1-A01	Thermus/Deinococcus group
31	FL8-H02	BCF group
32	SM2-D04	BCF group
36	SM2-B05	Alpha proteobacteria
39	SM1-E01	Beta proteobacteria
50	FL14-A01	Beta proteobacteria
51	FL14-A03	Candidate division OP11
52	FL14-A05	Fibrobacteria/Acidobacteria
53	FL14-A08	Firmicutes
55	SM2-D09	Cyanobacteria
60	FL14-D05	Alpha proteobacteria
61	FL14-C12	Unknown division
62	FL14-H07	BCF group
63	FL14-D06	Eukaryota, Chloroplasts
64	SM1-F10	Green non-sulfur bacteria
65	FL14-E06	Green non-sulfur bacteria
66	FL14-E08	Eukaryota, Chloroplasts
71	FL14-F11	Alpha proteobacteria
79	FL13-E04	Candidate division OP11
80	FL13-D02	Unknown division
81	FL13-E01	Candidate division OP11
82	FL14-G05	Eukaryota, Chloroplasts
83	SM1-D12	Cyanobacteria
87	FL13-B08	Alpha proteobacteria
90	FL6-F06	Cyanobacteria
91	FL6-F11	Cyanobacteria
92	SM2-D12	Alpha proteobacteria
99	6-2-99-9 # 10	Candidate division OP11
100	6-2-99-9 # 13	Cyanobacteria
106	SM1-F02	Candidate division OP11
107	SM1-C10	Cyanobacteria
118	FL6-H08	Cyanobacteria
119	FL6-H09	Thermus/Deinococcus group
120	FL6-H11	Alpha proteobacteria
121	SM2-A11	Alpha proteobacteria
122	SM2-A12	Unknown division
123	FL7-E02	Unknown division
124	FL7-A06	Green non-sulfur bacteria
125	FL7-A05	Firmicutes
126	FL7-E05	Alpha proteobacteria
127	FL7-H09	Unknown division
128	FL7-C04	Unknown division
129	SM2-A03	Green sulfur bacteria
130	SM2-B06	Alpha proteobacteria
132	SM2-B07	Alpha proteobacteria
138	6-2-99-10 # 15	Cyanobacteria
139	SM2-H05	Epsilon proteobacteria
165	SM1-C09	Unknown division
166	SM1-D11	Cyanobacteria
167	SM2-G06	Unknown division
174	FL12-G06	Unknown division
175	FL12-G08	Unknown division
176	FL13-H07	Alpha proteobacteria
177	SM2-F06	Unknown division
184	FL10-H03	Alpha proteobacteria
185	FL12-A07	Gamma proteobacteria
192	6-2-99-11 # 7	Aquificales
193	6-2-99-11 # 8	Alpha proteobacteria

Figure 6.7: OTU numbers with their corresponding defining sequence and division for 3% difference definition.

3% Definition



FILTER

V

(**2** samples)

41 ± 6	1	□	□
23 ± 18	2	⋈	
18 ± 12	4		
18 ± 12	3		

AC

(**2** samples)

33 ± 0	5		△
17 ± 12	23		◇
17 ± 12	22		♭
17 ± 12	2		⋈
17 ± 12	1		□

P

(**8** samples)

11 ± 3	5	△	△
6 ± 3	23		◇
6 ± 3	1		□
6 ± 2	25	†	
5 ± 2	99		∀
4 ± 2	12		
4 ± 1	27	⋆	
3 ± 2	30	×	
3 ± 2	51		♡
3 ± 2	61		
3 ± 2	60		
3 ± 2	64	‡	‡
3 ± 1	55	▽	
2 ± 2	100		
2 ± 2	53		
2 ± 2	52		
2 ± 2	50		
2 ± 2	63		
2 ± 2	62		
2 ± 2	66		
2 ± 2	65		

PS

(**7** samples)

18 ± 6	5	△	△
10 ± 4	1		□
9 ± 3	55	▽	
7 ± 2	107		
5 ± 3	23	◇	◇
3 ± 3	64	‡	‡
3 ± 3	139		
3 ± 3	138		
2 ± 2	120		
2 ± 2	119		
2 ± 2	118		
2 ± 2	167		
2 ± 2	166		
2 ± 2	165		
2 ± 2	106		
2 ± 2	83		
2 ± 2	22		♭

DS

(**3** samples)

16 ± 5	5		△
10 ± 3	55		
10 ± 3	51		♡
10 ± 3	1		□
6 ± 7	193		
6 ± 7	192		
5 ± 4	177		
5 ± 4	176		
5 ± 4	175		
5 ± 4	174		
5 ± 3	185		
5 ± 3	184		
5 ± 3	99		∀
5 ± 3	23		◇
5 ± 3	8		

MAT

V

(**1** sample)

25 ± 0	6		
25 ± 0	5		△
25 ± 0	2	⋈	
25 ± 0	1	□	

AC

(**4** samples)

10 ± 14	7		
10 ± 14	8		
7 ± 6	13		
7 ± 6	12		∈
7 ± 6	11		
7 ± 6	10		
7 ± 6	9		
6 ± 4	21		
6 ± 4	20		
6 ± 4	19		
6 ± 4	18		
6 ± 4	17		
6 ± 4	16		
6 ± 4	15		
6 ± 4	14		

P

(**13** samples)

9 ± 3	5	△	△
5 ± 2	25	†	
3 ± 2	30	×	
3 ± 2	28		
2 ± 2	71		
2 ± 2	27	⋆	
2 ± 2	39		
1 ± 1	83		∞
1 ± 1	82		
1 ± 1	79		
1 ± 1	92		
1 ± 1	91		
1 ± 1	90		
1 ± 1	26		
1 ± 1	24		
1 ± 1	32		
1 ± 1	31		
1 ± 1	29		
1 ± 1	81		
1 ± 1	80		
1 ± 1	64	‡	‡
1 ± 1	55	▽	▽

PS

(**8** samples)

8 ± 4	36		
5 ± 3	64	‡	‡
5 ± 3	129		
5 ± 3	55	▽	▽
3 ± 3	123		
3 ± 3	122		
3 ± 3	121		
3 ± 3	125		
3 ± 3	124		
3 ± 3	127		
3 ± 3	126		
3 ± 3	12		∈
3 ± 3	132		
3 ± 3	130		
3 ± 3	87		
2 ± 2	128		
2 ± 2	23	◇	
2 ± 2	5	△	△

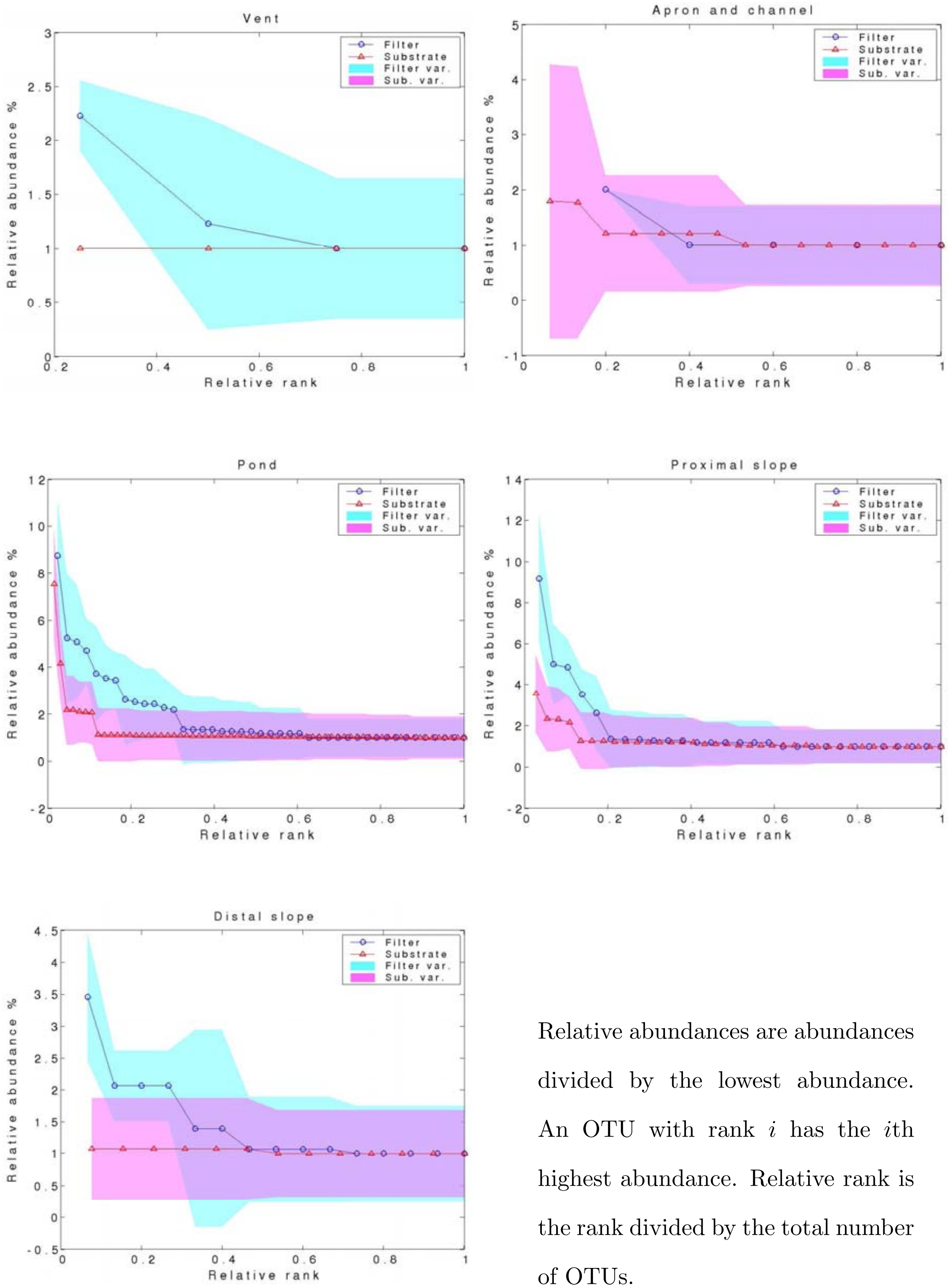
DS

(**2** samples)

8 ± 6	183		
8 ± 6	181		
8 ± 6	180		
8 ± 6	179		
8 ± 6	178		
8 ± 6	121		
7 ± 5	191		
7 ± 5	190		
7 ± 5	189		
7 ± 5	188		
7 ± 5	187		
7 ± 5	186		
7 ± 5	83		∞

Figure 6.8: Most abundant OTUs for the 3% difference definition. Figures are relative abundances with their variances. Numbers are identification OTU numbers given in figure 6.7. Black symbols mark OTUs that are present in another medium in the same facies. Blue symbols mark OTUs that are present in another facies in the same medium. Colors indicate phylotypes according to the code in figure 6.6.

3% Definition



Relative abundances are abundances divided by the lowest abundance. An OTU with rank i has the ith highest abundance. Relative rank is the rank divided by the total number of OTUs.

Figure 6.9: Plots of relative abundances versus relative rank for the 3% difference definition.

3% Definition

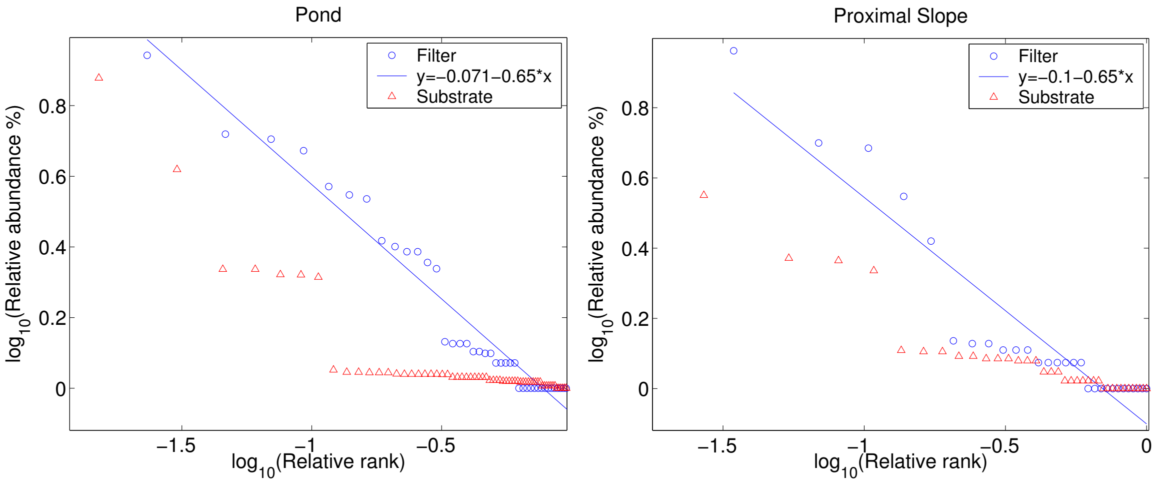


Figure 6.10: Plots of the logarithm of relative abundances versus logarithm of relative rank for the 3% difference definition. Only the filter samples from the pond and proximal slope facies seem to adjust well to a power law, within the limits imposed by the small amount of samples used (i.e. steps in the lower right end). Substrate sample plots from both facies curve upwards.

1% Definition

1	SM2-G04	Aquificales	100	FL13-E04	Candidate division OP11
2	6-2-99-7 # 12	Aquificales	101	FL13-D02	Unknown division
3	SM2-G02	Green non-sulfur bacteria	102	FL13-E01	Candidate division OP11
4	6-2-99-6 # 10	Cyanobacteria	103	FL14-G05	Eukaryota, Chloroplasts
5	6-2-99-6 # 12	BCF group	104	SM1-D12	Cyanobacteria
6	SM2-G01	Aquificales	108	FL13-B08	Alpha proteobacteria
7	6-2-99-6 # 8	Aquificales	112	FL6-F06	Cyanobacteria
8	SM1-E12	Beta proteobacteria	113	FL6-F11	Cyanobacteria
9	FL1-D10	Aquificales	114	SM2-D12	Alpha proteobacteria
10	FL1-D11	Aquificales	121	6-2-99-9 # 10	Candidate division OP11
11	FL1-E07	Aquificales	122	6-2-99-9 # 13	Cyanobacteria
12	FL1-F09	Firmicutes	132	FL5-E10	Aquificales
13	FL1-H03	Aquificales	135	SM1-C10	Cyanobacteria
14	FL1-H10	Aquificales	136	SM1-F10	Green non-sulfur bacteria
15	FL1-H12	Aquificales	150	FL6-E05	Aquificales
16	FL13-E05	Eukaryota, Chloroplasts	151	FL6-H05	Beta proteobacteria
17	FL10-G07	Green sulfur bacteria	152	FL6-H07	Candidate division OP11
18	FL13-A10	Unknown division	153	FL6-H08	Cyanobacteria
19	FL13-A11	Alpha proteobacteria	154	FL6-H09	Thermus/Deinococcus group
20	FL13-A12	BCF group	155	FL6-H11	Alpha proteobacteria
21	FL13-B01	Beta proteobacteria	156	SM2-A11	Alpha proteobacteria
22	FL13-B02	BCF group	157	SM2-A12	Unknown division
23	FL13-A01	Unknown division	158	FL7-E02	Unknown division
24	FL13-A02	Planctomycetales	159	FL7-A06	Green non-sulfur bacteria
25	FL13-A03	Verrucomicrobium group	160	FL7-A05	Firmicutes
26	FL13-A04	Alpha proteobacteria	161	FL7-E05	Alpha proteobacteria
27	FL13-A05	Unknown division	162	FL7-E09	Alpha proteobacteria
28	FL13-A08	Planctomycetales	163	FL7-E10	Alpha proteobacteria
29	FL13-A07	BCF group	164	FL7-E11	Beta proteobacteria
30	FL13-A09	Planctomycetales	165	FL7-H09	Unknown division
31	6-2-99-8 # 7	Aquificales	168	SM2-A03	Green sulfur bacteria
32	SM1-D01	BCF group	169	SM2-B06	Alpha proteobacteria
33	FL5-E03	Candidate division OP11	171	SM2-B07	Alpha proteobacteria
35	SM1-E10	Candidate division OP11	178	6-2-99-10 # 15	Cyanobacteria
37	SM2-D03	Firmicutes	179	SM2-H05	Epsilon proteobacteria
38	FL8-H07	BCF group	210	SM2-G06	Unknown division
48	SM2-B05	Alpha proteobacteria	222	FL7-G01	Eukaryota, Mitochondria
52	SM1-E01	Beta proteobacteria	223	FL7-F09	Unknown division
64	FL14-A01	Beta proteobacteria	224	SM2-B11	Eukaryota, Chloroplasts
65	FL14-A03	Candidate division OP11	225	FL7-F11	Eukaryota, Mitochondria
66	FL14-A05	Fibrobacteria/Acidobacteria	226	FL7-G02	Alpha proteobacteria
67	FL14-A08	Firmicutes	227	SM2-B12	Unknown division
69	SM2-D09	Cyanobacteria	230	FL7-G08	Alpha proteobacteria
72	SM1-G05	Beta proteobacteria	231	SM2-C12	Cyanobacteria
75	FL14-C04	Candidate division OP11	232	FL7-G11	BCF group
77	FL14-C12	Unknown division	233	FL7-G12	Alpha proteobacteria
80	6-2-99-10 # 13	Candidate division OP11	234	FL7-H12	Alpha proteobacteria
83	FL13-F03	Unknown division	235	SM2-C02	Alpha proteobacteria

Figure 6.11: OTU numbers with their corresponding defining sequence and division for 1% difference definition.

1% Definition

FILTER

V			AC			P			PS			DS		
(**2** samples)			(**2** samples)			(**8** samples)			(**7** samples)			(**3** samples)		
26 ± 5	1	□ □	29 ± 3	8	†	9 ± 3	8	† †	13 ± 3	8	†	15 ± 5	8	†
14 ± 12	3	× ×	16 ± 12	33		6 ± 3	80	∂	8 ± 3	1	□	9 ± 2	69	
14 ± 12	2		16 ± 12	32		5 ± 2	1	□	8 ± 2	69	⊙ ⊙	9 ± 2	65	⋄
11 ± 7	7		13 ± 9	31		4 ± 2	121	⊢	6 ± 2	135		9 ± 2	1	□
11 ± 7	6	⊗ ⊗	13 ± 9	3	×	4 ± 2	72		3 ± 3	151		6 ± 7	237	
11 ± 7	5		13 ± 9	1	□	4 ± 1	75		2 ± 2	150		6 ± 7	236	
11 ± 7	4					4 ± 1	37	▽	2 ± 2	6		5 ± 3	221	
						3 ± 2	65	⋄	2 ± 2	179		5 ± 3	220	
						3 ± 2	77		2 ± 2	178		5 ± 3	219	
						2 ± 1	132		2 ± 2	80	∂	5 ± 3	218	
						2 ± 1	69	⊙ ⊙	2 ± 2	83		5 ± 3	217	
						1 ± 2	122		2 ± 2	155		5 ± 3	229	
						1 ± 2	67		2 ± 2	154		5 ± 3	228	
						1 ± 2	66		2 ± 2	153		5 ± 3	121	⊢
						1 ± 2	64		2 ± 2	152		5 ± 3	80	∂
												5 ± 3	17	

SUBSTRATE

V			AC			P			PS			DS		
(**1** sample)			(**4** samples)			(**13** samples)			(**8** samples)			(**2** samples)		
9 ± 0	15		10 ± 14	16		9 ± 3	8	† †	5 ± 4	48		8 ± 6	227	
9 ± 0	14		10 ± 14	17		5 ± 2	35		5 ± 3	168		8 ± 6	226	
9 ± 0	13		7 ± 6	22		2 ± 2	38		4 ± 3	69	⊙ ⊙	8 ± 6	225	
9 ± 0	12		7 ± 6	21		2 ± 1	52		3 ± 3	158		8 ± 6	224	
9 ± 0	11		7 ± 6	20		2 ± 1	37	▽	3 ± 3	157		8 ± 6	223	
9 ± 0	10		7 ± 6	19		1 ± 1	104	∩	3 ± 3	156		8 ± 6	222	
9 ± 0	9		7 ± 6	18		1 ± 1	103		2 ± 3	160		7 ± 5	235	
9 ± 0	8	†	6 ± 4	30		1 ± 1	100		2 ± 3	159		7 ± 5	234	
9 ± 0	6	⊗	6 ± 4	29		1 ± 1	114		2 ± 3	136		7 ± 5	233	
9 ± 0	3	×	6 ± 4	28		1 ± 1	113		2 ± 2	171		7 ± 5	232	
9 ± 0	1	□	6 ± 4	27		1 ± 1	112		2 ± 2	169		7 ± 5	231	
			6 ± 4	26		1 ± 1	102		2 ± 2	108		7 ± 5	230	
			6 ± 4	25		1 ± 1	101		2 ± 2	165		7 ± 5	104	∩
			6 ± 4	24		1 ± 1	83		2 ± 2	164				
			6 ± 4	23		1 ± 1	69	⊙ ⊙	2 ± 2	163				
									2 ± 2	162				
									2 ± 2	161				

Figure 6.12: Most abundant OTUs for the 1% difference definition.

1% Definition

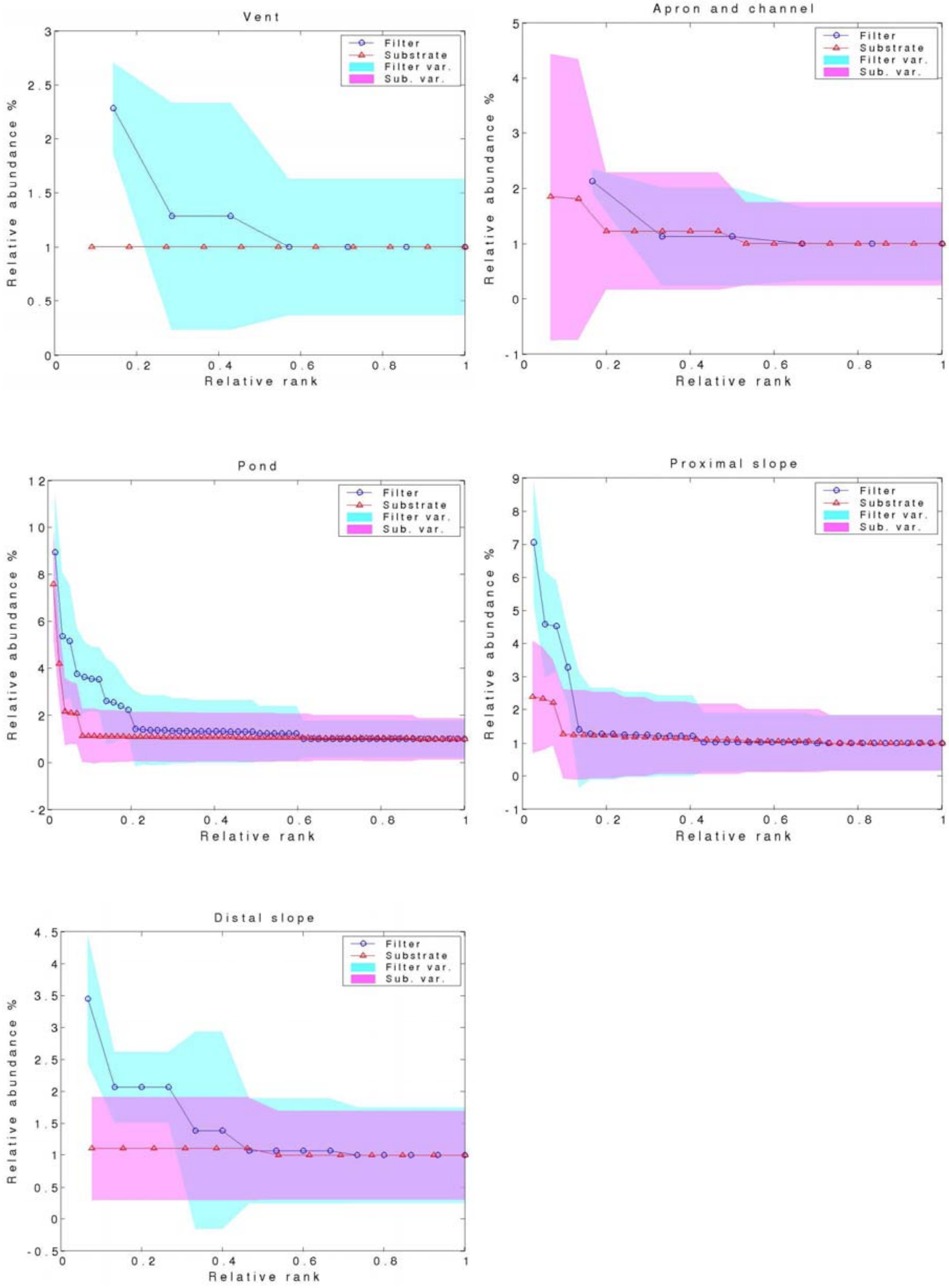


Figure 6.13: Plots of relative abundances versus relative rank for the 1% difference definition

1% Definition

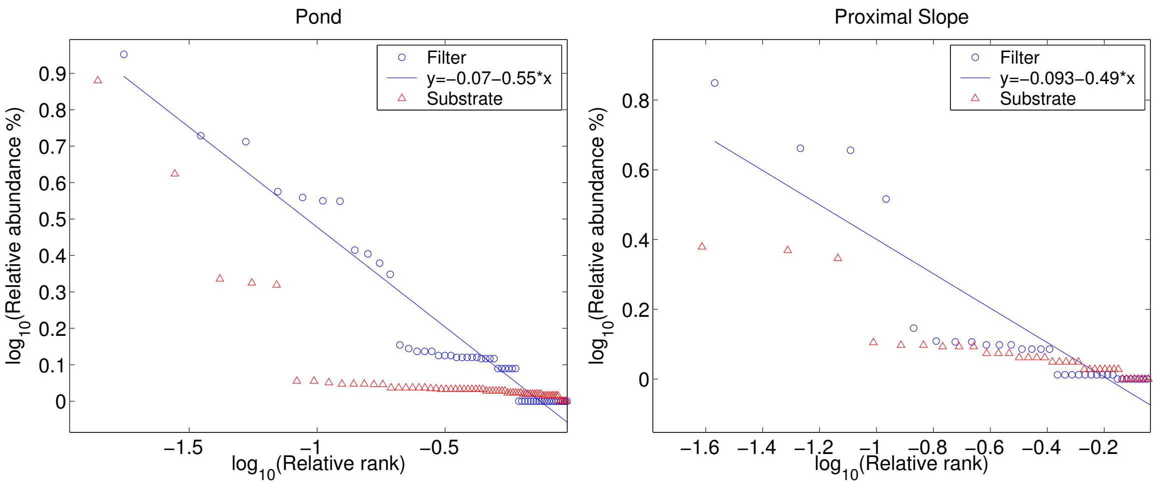


Figure 6.14: Plots of the logarithm of relative abundances versus logarithm of relative rank for the 1% difference definition.

0.5% Definition

1	SM1-G03	Aquificales
2	6-2-99-7 # 12	Aquificales
3	6-2-99-8 # 13	Aquificales
4	6-2-99-7 # 17	Aquificales
5	SM2-G04	Aquificales
6	SM2-G02	Green non-sulfur bacteria
7	6-2-99-6 # 10	Cyanobacteria
8	6-2-99-6 # 12	BCF group
9	6-2-99-6 # 17	Aquificales
10	6-2-99-6 # 8	Aquificales
11	6-2-99-11 # 13	Beta proteobacteria
12	FL1-D09	Aquificales
13	FL1-D10	Aquificales
14	FL6-E07	Aquificales
15	FL1-E02	Aquificales
16	FL1-E07	Aquificales
17	FL1-F09	Firmicutes
18	FL1-G05	Aquificales
19	FL1-H03	Aquificales
20	FL6-E06	Aquificales
21	FL1-H10	Aquificales
22	FL1-H12	Aquificales
23	SM2-G01	Aquificales
24	FL13-E05	Eukaryota, Chloroplasts
25	FL10-G07	Green sulfur bacteria
26	FL13-A10	Unknown division
27	FL13-A11	Alpha proteobacteria
28	FL13-A12	BCF group
29	FL13-B01	Beta proteobacteria
30	FL13-B02	BCF group
31	FL13-A01	Unknown division
32	FL13-A02	Planctomycetales
33	FL13-A03	Verrucomicrobium group
34	FL13-A04	Alpha proteobacteria
35	FL13-A05	Unknown division
36	FL13-A08	Planctomycetales
37	FL13-A07	BCF group
38	FL13-A09	Planctomycetales
39	SM1-E12	Beta proteobacteria
49	SM1-E10	Candidate division OP11
53	FL8-H07	BCF group
69	SM1-E01	Beta proteobacteria
70	FL8-D03	Thermus/Deinococcus group
71	FL8-D08	Firmicutes
86	FL14-A03	Candidate division OP11
92	SM2-H09	Firmicutes
94	FL5-G09	Beta proteobacteria
99	FL14-C12	Unknown division
102	6-2-99-10 # 13	Candidate division OP11
105	FL14-D08	Candidate division OP11
106	SM2-G09	Beta proteobacteria
131	FL13-E04	Candidate division OP11
134	SM2-D09	Cyanobacteria
143	FL14-G05	Eukaryota, Chloroplasts
144	SM1-D12	Cyanobacteria
149	FL13-B08	Alpha proteobacteria
154	FL6-F06	Cyanobacteria
155	FL6-F11	Cyanobacteria
156	SM2-D12	Alpha proteobacteria
163	6-2-99-9 # 10	Candidate division OP11
164	6-2-99-9 # 13	Cyanobacteria
165	6-2-99-9 # 2	Candidate division OP11
169	6-2-99-10 # 2	Beta proteobacteria
181	FL5-E10	Aquificales
182	SM2-G07	Aquificales
190	SM1-F10	Green non-sulfur bacteria
212	FL6-H05	Beta proteobacteria
213	FL6-H07	Candidate division OP11
214	FL6-H08	Cyanobacteria
215	FL6-H09	Thermus/Deinococcus group
216	FL6-H11	Alpha proteobacteria
218	SM2-A11	Alpha proteobacteria
219	SM2-A12	Unknown division
220	FL7-E01	Alpha proteobacteria
221	FL7-E02	Unknown division
222	FL7-A06	Green non-sulfur bacteria
223	FL7-A04	Green non-sulfur bacteria
224	FL7-A05	Firmicutes
236	SM2-A03	Green sulfur bacteria
237	SM2-B06	Alpha proteobacteria
239	SM2-B07	Alpha proteobacteria
303	FL12-H02	Beta proteobacteria
304	FL12-G10	Beta proteobacteria
305	FL12-G06	Unknown division
306	FL12-G08	Unknown division
307	FL12-G09	Cyanobacteria
308	FL13-H06	Aquificales
309	FL13-H07	Alpha proteobacteria
310	SM2-F06	Unknown division
311	FL7-F08	Eukaryota, Mitochondria
312	FL7-F09	Unknown division
313	SM2-B11	Eukaryota, Chloroplasts
314	FL7-F11	Eukaryota, Mitochondria
315	FL7-G01	Eukaryota, Mitochondria
316	FL7-G02	Alpha proteobacteria
317	FL7-H10	Eukaryota, Mitochondria
318	SM2-B12	Unknown division
324	FL7-G08	Alpha proteobacteria
325	SM2-C12	Cyanobacteria
326	FL7-G11	BCF group
327	FL7-G12	Alpha proteobacteria
328	FL7-H12	Alpha proteobacteria
329	SM2-C02	Alpha proteobacteria
330	6-2-99-11 # 7	Aquificales
331	6-2-99-11 # 8	Alpha proteobacteria

Figure 6.15: OTU numbers with their corresponding defining sequence and division for 0.5% difference definition.

0.5% Definition

FILTER

V			AC			P			PS			DS		
(**2** samples)			(**2** samples)			(**8** samples)			(**7** samples)			(**3** samples)		
17 ± 0	3	†	14 ± 12	45		4 ± 2	11	∀	7 ± 2	106	‡	8 ± 6	11	∀
17 ± 0	1	△ △	14 ± 12	44		3 ± 2	105		6 ± 2	134	⋄	6 ± 2	303	
8 ± 6	6	□	14 ± 12	43		3 ± 1	94		5 ± 2	1	△	6 ± 2	134	
8 ± 6	5	⊘	10 ± 6	42		2 ± 2	86	⊙	4 ± 2	169		6 ± 2	106	
8 ± 6	4		10 ± 6	41		2 ± 2	102		4 ± 2	11	∀	6 ± 2	86	⊙
8 ± 6	2		10 ± 6	40		2 ± 2	99		3 ± 2	182		5 ± 7	331	
8 ± 6	10		10 ± 6	39		2 ± 1	165		2 ± 1	39		5 ± 7	330	
8 ± 6	9		10 ± 6	5	⊘	2 ± 1	5	⊘	2 ± 2	212		3 ± 3	310	
8 ± 6	8		10 ± 6	3	†	2 ± 1	181		2 ± 2	216		3 ± 3	309	
8 ± 6	7					2 ± 1	106	‡	2 ± 2	215		3 ± 3	308	
						2 ± 1	92		2 ± 2	214		3 ± 3	307	
						2 ± 1	39	⋆	2 ± 2	213		3 ± 3	306	
						1 ± 1	164					3 ± 3	305	
						1 ± 1	163					3 ± 3	304	
												3 ± 3	1	△

MAT

V			AC			P			PS			DS		
(**1** sample)			(**4** samples)			(**13** samples)			(**8** samples)			(**2** samples)		
6 ± 0	23		10 ± 14	24		4 ± 2	11	∀ ∀	4 ± 3	236		7 ± 5	329	
6 ± 0	22		10 ± 14	25		3 ± 1	39	⋆	4 ± 2	134	⋄	7 ± 5	328	
6 ± 0	21		7 ± 6	30		2 ± 1	53		2 ± 3	219		7 ± 5	327	
6 ± 0	20		7 ± 6	29		2 ± 1	69		2 ± 3	218		7 ± 5	326	
6 ± 0	19		7 ± 6	28		2 ± 1	49		2 ± 3	221		7 ± 5	325	
6 ± 0	18		7 ± 6	27		1 ± 1	156		2 ± 3	220		7 ± 5	324	
6 ± 0	17		7 ± 6	26		1 ± 1	155		2 ± 2	224		7 ± 5	144	
6 ± 0	16		6 ± 4	38		1 ± 1	154		2 ± 2	223		6 ± 4	318	
6 ± 0	15		6 ± 4	37		1 ± 1	144		2 ± 2	222		6 ± 4	317	
6 ± 0	14		6 ± 4	36		1 ± 1	143		2 ± 2	190		6 ± 4	316	
6 ± 0	13		6 ± 4	35		1 ± 1	131		2 ± 2	239		6 ± 4	315	
6 ± 0	12		6 ± 4	34		1 ± 1	71		2 ± 2	237		6 ± 4	314	
6 ± 0	11	∀	6 ± 4	33		1 ± 1	70		2 ± 2	149		6 ± 4	313	
6 ± 0	6	□	6 ± 4	32								6 ± 4	312	
6 ± 0	5	⊘	6 ± 4	31								6 ± 4	311	
6 ± 0	1	△												

Figure 6.16: Most abundant OTUs for the 0.5% difference definition.

0.5% Definition

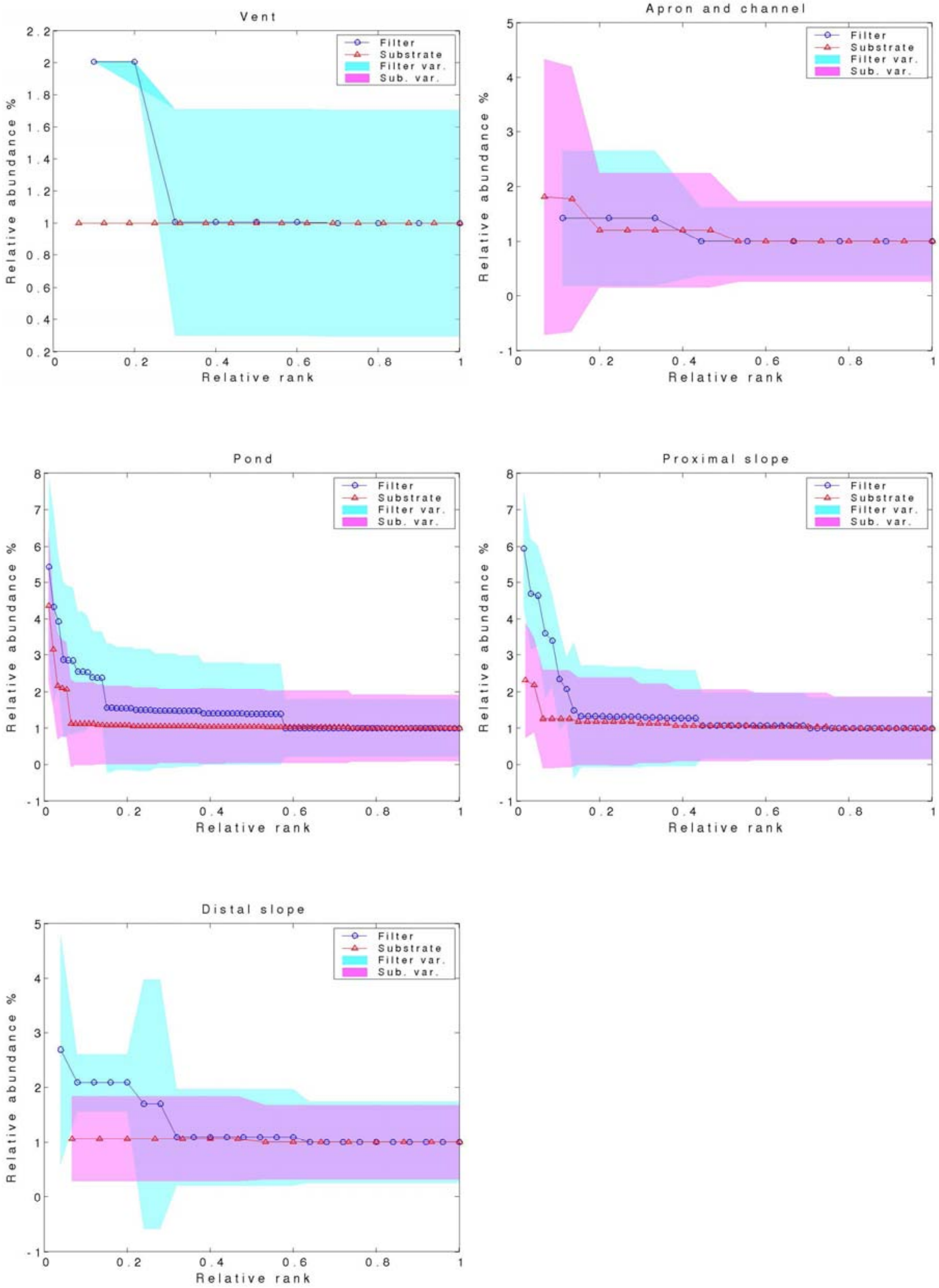


Figure 6.17: Plots of relative abundances versus relative rank for the 0.5% difference definition

0.5% Definition

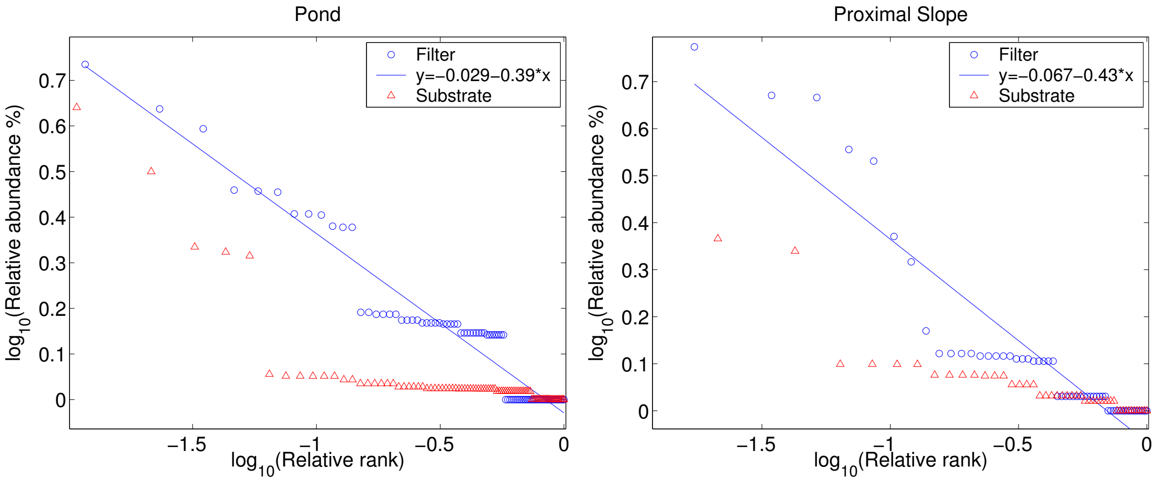


Figure 6.18: Plots of the logarithm of relative abundances versus logarithm of relative rank for the 0.5% difference definition.

Chapter 7

Rotating Bose Einstein Condensates

The world we live in is a quantum world. At microscopic scales, reality can simply not be understood without recourse to quantum phenomena such as position/momentum indetermination and energy quantization. For macroscopic observers, this quantum nature is certainly not obvious. Otherwise, quantum mechanics would not have taken until the 20th century to be developed [196].

Bose-Einstein Condensates (BECs) are one of the few examples in which this quantum nature of reality becomes conspicuous at macroscopic scales. The superfluid nature of these gases is particularly noticeable in the case of fast rotation. Under these conditions a normal fluid, such as water for example, develops a single hole in the center due to centripetal forces. Superfluids instead form an array of vortices, similar to the case of superconductors under a magnetic field [197]. Unlike superconductors though, the destruction of the superfluid state is not due to the overlap of vortex cores and destruction of Cooper pairs. In fact, mean field calculations based on the Gross-Pitaevskii equation [198] suggest that, for an increasing rotation rate, the core size reduces in such a way that cores never overlap. This description must break down for high rotation rates when core sizes become so small that mean field theory is no longer applicable. Understanding how this happens has been the goal of this project. Path Integral Monte Carlo (PIMC) simulations involve no uncontrolled approximations in dealing with interacting particles and are ideally suited for this task [199].

In this chapter, I present my own work regarding the self-consistent calculation of the chemical potential (section 7.1.3), a self-similar solution for the vortex structure in a lattice (sections 7.3 and 7.4) and I show the importance of considering the higher density corrections to the Gross-Pitaevskii equation (section 7.5).

In the next chapter, I will present the results involving PIMC calculations relevant to anharmonic traps.

7.1 Bose-Einstein Condensates (BECs)

Bose-Einstein Condensation (BEC) refers to macroscopic occupation of the ground quantum state [200]. Predicted by Einstein and Bose in 1924 [201],[202], the first connection to observed natural phenomena was made by London in 1938 for superfluid ^{4}He [203]. Superfluid ^{4}He has been intensively studied since then, as the epitome of a quantum fluid. Unfortunately, the strong interactions between Helium atoms means that, typically, less than 10% of its particles are in the condensed state.

The advent of laser cooling and magnetic trapping among other techniques, eventually produced in 1995 the first experimental realization of BECs in alkali gases: rubidium, sodium and lithium[204],[205],[206],[207]. These gases provide nowadays the most studied implementations of Bose-Einstein Condensates. Their densities are very low $\approx 10^{13} - 10^{15}\text{cm}^{-3}$, allowing for high superfluid fractions and very good modelling in terms of mean field theory. Typical temperatures are of the order of 10^{-5}K to 10^{-8}K, which means that the only way to enclose these gases without undesired heat transfer is through magnetic trapping [200]. The most commonly used trap is the harmonic trap, described by the potential:

$$V(r_\perp, z) = \frac{1}{2} m(\omega_\perp^2 r_\perp^2 + \omega_z^2 z^2) \tag{7.1}$$

where the symbol $r_\perp$ corresponds to the distance from the center of the trap perpendicular to the rotation axis.

Non-harmonic traps, in which the trapping transverse to the rotation axis grows faster than harmonic:

$$V_\perp(r_\perp, z) = \frac{1}{2} m[\omega_\perp^2 r_\perp^2 (1 + \lambda r_\perp^2) + \omega_z^2 z^2] \tag{7.2}$$

are also used [208].

In the following sections, I will introduce the Gross-Pitaevskii equation that is used to model the behavior of alkali BECs. Along with it, I present a self-consistent calculation of the applicable chemical potential and the modification of the Gross-Pitaevskii equation for high densities.

7.1.1 The Gross-Pitaevskii equation

The Gross-Pitaevskii equation is one of the main tools for the prediction of behavior in alkali gases BECs [200],[198],[209]. It assumes a mean field theory approach in which the N-particle wave function is a symmetrized product of single particle wavefunctions. In the case of a BEC, these wavefunctions are all the single particle ground state:

$$\Psi(\vec{r}_1, \vec{r}_2, ..., \vec{r}_N) = \prod_{i=1}^{N} \Phi(\vec{r}_i) \tag{7.3}$$

The density of particles is assumed to be small enough that the effective interacting potential can be assumed to be a Dirac delta [210] of strength U_0, and therefore the effective Hamiltonian reads:

$$\mathcal{H} = \sum_{i=1}^{N} \left[\frac{p_i^2}{2m} + V(\vec{r}_i) \right] + U_0 \sum_{i<j} \delta(\vec{r}_i - \vec{r}_j) \tag{7.4}$$

The energy is then:

$$E = N \int d\vec{r} \left[\frac{\hbar^2}{2m} |\vec{\nabla}\Phi(\vec{r})|^2 + V(\vec{r})|\Phi(\vec{r})|^2 + \frac{N-1}{2} U_0 |\Phi(\vec{r})|^4 \right] \tag{7.5}$$

Intuitively, this is the energy for N independent particles plus a term that adds an interaction energy proportional to the overlap of wavefunctions. A more rigorous derivation in terms of pseudopotentials can be found in reference [211]. This functional for the energy only applies when the space variation is bigger than the characteristic length scale given by the coherence length:

$$\frac{1}{\epsilon_0^2} = \frac{gn}{\lambda} \tag{7.6}$$

The wave function of the condensed state is defined as $\psi(\vec{r}) = N^{1/2}\Phi(\vec{r})$ such that the particle density $n(\vec{r}) = |\psi(\vec{r})|^2$. In terms of $\psi(\vec{r})$, the energy is:

$$E \;\; = \;\; \int d\vec{r} \left[\lambda |\psi(\vec{r})|^2 + V(\vec{r})|\psi(\vec{r})|^2 + \frac{1}{2} g |\psi(\vec{r})|^4 \right] \tag{7.7}$$

where $g = \frac{N-1}{N} U_0/2 \simeq U_0$ and $\lambda \equiv \hbar^2/2m$.

The ground state will, by definition, have the smallest energy possible. The ground state wavefunction will minimize the energy in equation 7.7 subject to the constraint that the total number of particles is constant: $N = \int d\vec{r}\psi(\vec{r})$. This is equivalent to minimizing the quantity $E - \mu N$, where the chemical potential μ is a Lagrange multiplier (see section 7.1.3). The minimization is achieved through functional differentiation [212]:

$$\begin{aligned} &\frac{\delta(E - \mu N)}{\delta \psi(\vec{r})} = 0 \quad \Rightarrow \\ &\quad -\frac{\hbar^2}{2m} \nabla^2 \psi(\vec{r}) + V(\vec{r})\psi(\vec{r}) + g|\psi(\vec{r})|^2 \psi(\vec{r}) = \mu \psi(\vec{r}) \end{aligned} \tag{7.8}$$

This has the form of the Schrödinger equation plus an interaction term and with μ (instead of the energy) being the eigenvalue.

For a uniform condensate equation 7.8 turns into:

$$\mu = U_0 |\psi(\vec{r})|^2 = U_0 n \tag{7.9}$$

where n is the particle density.

This result can also be obtained through the thermodynamic relation $\mu = \partial E/\partial N$ [200].

The interaction factor g can be obtained from the knowledge of the atomic interaction. For the case of hard spheres of diameter σ (=scattering length) used in the PIMC simulations[200]:

$$g = 8\pi\frac{\hbar^2}{2m}\sigma \tag{7.10}$$

7.1.2 Corrections to the Gross Pitaevskii for high densities

As stated above, in order for the Gross-Pitaevskii functional to be valid the density must be low enough. For higher densities, corrections apply to the energy of a gas of hard spheres [213]:

$$\frac{E}{V} = 4\pi\frac{\hbar^2}{2m}\sigma\, n^2\left[1+\frac{128}{15}\left(\frac{n\sigma^3}{\pi}\right)^{1/2}+8\left(\frac{4}{3}\pi-\sqrt{3}\right)n\sigma^3\log\left(n\sigma^3\right)+O(n\sigma^3)\right] \tag{7.11}$$

The prefactor is the interaction term in equation 7.7 for a homogeneous condensate. The first two corrections are independent of the actual potential [214] and depend only on the scattering length σ. We will only be using the first correction, since the addition of the second extra factor (log term) lowers the energy excessively if not compensated by further terms. In general, the first correction alone performs very well without further adjustments [213].

The corrected Gross-Pitaevskii functional is then:

$$\begin{aligned} E &= \int d\vec{r}\left[\frac{\hbar^2}{2m}|\vec{\nabla}\psi(\vec{r})|^2+V(\vec{r})|\psi(\vec{r})|^2\right.\\ &\quad\left.+\frac{1}{2}g|\psi(\vec{r})|^4(1+4.81\sigma^{3/2}|\psi(\vec{r})|+19.65\sigma^3|\psi(\vec{r})|^2\log\left(\sigma^3|\psi(\vec{r})|^2\right)\right]\\ &= \int d\vec{r}\left[\lambda|\vec{\nabla}\psi(\vec{r})|^2+V(\vec{r})|\psi(\vec{r})|^2+\mathcal{U}(\psi(\vec{r}))\right] \end{aligned} \tag{7.12}$$

where:

$$\mathcal{U}(\psi(\vec{r})) \equiv \frac{1}{2}g|\psi(\vec{r})|^4(1+a\sigma^{3/2}|\psi(\vec{r})|+d\sigma^3|\psi(\vec{r})|^2\log\left(\sigma^3|\psi(\vec{r})|^2\right) \tag{7.13}$$

and $a = 4.81 = 128/15\sqrt{\pi}$ and $d = 19.65 = 8(4\pi/3-\sqrt{3})$.

7.1.3 The chemical potential

The section presents a self-consistent derivation of the chemical potential. This is not a small order correction, but will turn out to be very important in order to explain the scaling properties of the vortices.

The chemical potential μ is the Lagrange multiplier corresponding to minimizing the energy:

$$E = \int \lambda |\vec{\nabla}\psi(\vec{r})|^2 + \hat{\mathcal{U}}(\psi(\vec{r}))d\vec{r} \tag{7.14}$$

where

$$\hat{\mathcal{U}}(\psi(\vec{r})) \equiv \mathcal{U}(\psi(\vec{r})) + V(\vec{r}) \tag{7.15}$$

under the constraint:

$$N = \int |\psi(\vec{r})|^2 d\vec{r} \tag{7.16}$$

We will show that the thermodynamic relation $\mu = \partial E/\partial N$ holds for non-uniform condensates. In what follows, we will take the wavefunction to be real, since in all following manipulations, we will assume a given "frozen" phase that will become an addition to the potential (see section 8.1.7). An analogous procedure can be used in the case of a full imaginary wave function, by doing independent variations for the real and imaginary part.

Let's start by splitting $\psi(\vec{r})$ as $\psi(\vec{r}) = \sqrt{n}f(\vec{r})$, where $\sqrt{n}$ is a normalizing factor. The equations above are rewritten then as:

$$\begin{aligned} E &= \lambda n \int \left[|\vec{\nabla} f(\vec{r})|^2 + \frac{1}{\lambda n}\hat{\mathcal{U}}(\sqrt{n}f(\vec{r})) \right] d\vec{r} \\ &= \lambda n \int \left[|\vec{\nabla} f(\vec{r})|^2 + \bar{\mathcal{U}}(f(\vec{r}), \sqrt{n}) \right] d\vec{r} \end{aligned} \tag{7.17}$$

where

$$\bar{\mathcal{U}}(\chi(fr), \sqrt{n}) \equiv \frac{1}{\lambda n}\hat{\mathcal{U}}(\sqrt{n}f(\vec{r})) \tag{7.18}$$

and

$$N = n \int f^2(\vec{r}) d\vec{r} \tag{7.19}$$

Let's assume that $\psi(\vec{r}) = \sqrt{n} f(\vec{r})$ is the minimum energy configuration and make a small variation $\eta(\vec{r})$, controlled by the parameter ϵ:

$$\begin{aligned} f(\vec{r}) &\rightarrow f(\vec{r})(1+\epsilon\eta(\vec{r})) & (7.20) \\ \sqrt{n} &\rightarrow \sqrt{n}(1-\epsilon k) & (7.21) \end{aligned}$$

where ϵ is a small parameter, and $\eta(\vec{r})$ and k are related by the normalization condition:

$$\begin{aligned} N &= n\int f^2(\vec{r})(1-2\epsilon k)(1+2\epsilon\eta(\vec{r})) \\ &= N + 2\epsilon n \int f^2(\vec{r})(\eta(\vec{r})) - k)d\vec{r} \\ &\Rightarrow\ k = \frac{\int f^2(\vec{r})\eta(\vec{r})d\vec{r}}{\int f^2(\vec{r})d\vec{r}} = \frac{\int f^2(\vec{r})\eta(\vec{r})d\vec{r}}{q f_m^2 V} \end{aligned} \qquad (7.22)$$

where

$$q \equiv \int f^2(\vec{r})d\vec{r}/f_m^2 V \qquad (7.23)$$

and f_m is the maximum of $f(\vec{r})$.

The energy under this variation $\eta(\vec{r})$ is to first order in ϵ:

$$\begin{aligned} \frac{E(\epsilon)}{\lambda} &= n(1-2\epsilon k)\int \left[|\vec{\nabla} f(1+\epsilon\eta)|^2 + \bar{\mathcal{U}}(f(1+\epsilon\eta), \sqrt{n}(1-\epsilon k))\right] d\vec{r} \\ &= n(1-2\epsilon k)\int \left[|\vec{\nabla} f|^2 + 2\epsilon\vec{\nabla} f \vec{\nabla}(f\eta) + \bar{\mathcal{U}}(f,\sqrt{n}) + \eta\chi\epsilon\frac{\partial\bar{\mathcal{U}}}{\partial f} - \epsilon\sqrt{n}k\frac{\partial\bar{\mathcal{U}}}{\partial\sqrt{n}}\right] d\vec{r} \\ &= n(1-2\epsilon k)\left[\frac{E(0)}{\lambda n} + \int -2\epsilon\eta f\nabla^2 f + 2\eta f\epsilon\frac{1}{2}\frac{\partial\bar{\mathcal{U}}}{\partial f} - 2\epsilon\sqrt{n}k\frac{1}{2}\frac{\partial\bar{\mathcal{U}}}{\partial\sqrt{n}}\right] \end{aligned} \qquad (7.24)$$

where we have used Green's theorem:

$$\begin{aligned} \int_V \vec{\nabla}(f\eta)\vec{\nabla} f d\vec{r} &= \int_S \eta f\vec{\nabla} f\cdot d\vec{\sigma} - \int_V \eta f\nabla^2 f d\vec{r} & (7.25) \\ &= -\int_V \eta f\nabla^2 f d\vec{r} & (7.26) \end{aligned}$$

The surface integral is zero because $\vec{\nabla} f(\vec{r})$ must be null at the edge of the cell. In a stationary state no net current should be entering or leaving the system.

Hence, to first order in ϵ

$$\begin{aligned}\frac{E(\epsilon)}{\lambda} = & \\ & n\left[\frac{E(0)}{n} + \int\left(-2\epsilon\eta f\nabla^2 f + 2\eta f\epsilon\frac{1}{2}\frac{\partial\bar{\mathcal{U}}}{\partial f} - 2\epsilon k\frac{E(0)}{\lambda n} - 2\epsilon\sqrt{n}k\frac{1}{2}\frac{\partial\bar{\mathcal{U}}}{\partial\sqrt{n}}\right)d\vec{r}\right] \\ = \; & n\left[\frac{E(0)}{n} + 2\epsilon\int\eta f\left(-\nabla^2 f + \frac{1}{2}\frac{\partial\bar{\mathcal{U}}}{\partial f} - f\left[\frac{E(0)}{\lambda n}\frac{1}{qVf_m^2} + \int\frac{\sqrt{n}}{2}\frac{\partial\bar{\mathcal{U}}}{\partial\sqrt{n}}\frac{d\vec{r}'}{qVf_m^2}\right]\right)d\vec{r}\right]\end{aligned} \tag{7.27}$$

where we used equation 7.22. The minimization of the energy with respect to the variation then entails:

$$-\nabla^2 f(\vec{r}) + \frac{1}{2}\frac{\partial\bar{\mathcal{U}}}{\partial f} - \frac{f(\vec{r})}{qVf_m^2}\left[\frac{E(0)}{\lambda n} + \int\frac{\sqrt{n}}{2}\frac{\partial\bar{\mathcal{U}}}{\partial\sqrt{n}}d\vec{r}\right] = 0 \tag{7.28}$$

which is the standard Gross-Pitaevskii equation with chemical potential:

$$\mu = \frac{\lambda}{qVf_m^2}\left[\frac{E(0)}{\lambda n} + \int\frac{\sqrt{n}}{2}\frac{\partial\bar{\mathcal{U}}}{\partial\sqrt{n}}d\vec{r}\right] \tag{7.29}$$

We could have obtained this result for μ in a different way, by using the thermodynamic relation:

$$\mu = \frac{\partial E}{\partial N} = \frac{1}{qAf_m^2}\frac{\partial E}{\partial n} = \frac{\lambda}{qVf_m^2}\left[\int\left[|\vec{\nabla}f(\vec{r})|^2 + \bar{\mathcal{U}}(f(\vec{r}),\sqrt{n}) + \frac{\sqrt{n}}{2}\frac{\partial\bar{\mathcal{U}}}{\partial\sqrt{n}}\right]d\vec{r}\right] \tag{7.30}$$

For the case of the unmodified Gross-Pitaevskii functional this becomes:

$$\mu = \frac{\lambda}{qVf_m^2}\int\left[|\vec{\nabla}f|^2 + f^2V(\vec{r}) + \frac{f^4}{\epsilon_0^2}\right]d\vec{r} \tag{7.31}$$

Notice the difference with the energy: there is no 1/2 factor in front of the interaction energy. The coherence length ϵ_0 is defined as:

$$\frac{1}{\epsilon_0^2} = \frac{gn}{\lambda} \tag{7.32}$$

For the modified Gross-Pitaevskii the result is:

$$\mu = \frac{\lambda}{qVf_m^2}\int\left[|\vec{\nabla}f|^2 + f^2V(\vec{r}) + \frac{f^4}{\epsilon_0^2}(1 + \frac{5}{4}\bar{a}f(r))\right]d\vec{r} \tag{7.33}$$

where $\bar{a} \equiv a(\sigma^3 n)^{1/2}$.

7.2 BECs under rotation

The fact that for a BEC all particles are in the same state, and their velocity is proportional to the gradient of the phase of the wavefunction imposes important constraints [200] on the response of superfluids to rotation:

$$\vec{v} = \frac{\hbar}{m}\vec{\nabla}\varphi \tag{7.34}$$

Since the phase must be single valued, this means that around a closed countour the change of the phase must be a multiple of 2π:

$$\Delta\varphi = \oint \vec{\nabla}\varphi \cdot d\vec{l} = 2\pi l \tag{7.35}$$

where l is a positive or negative integer, or zero.

From equation 7.34, it follows that the circulation is quantized:

$$\Lambda = \oint \vec{v}d\vec{l} = l\frac{h}{m} \tag{7.36}$$

For a purely azimuthal flow, assuming the velocity to be dependent only on the distance to the vortex core r:

$$\vec{v} = l\frac{\hbar}{mr}\hat{u}_\theta \tag{7.37}$$

where $\hat{u}_\theta$ is the unit vector in the azimuthal direction.

The angular momentum carried by a vortex is the integral, weighed by the square of the condensate wave function, of $l\hbar$:

$$L_z \propto l\hbar \tag{7.38}$$

while the energy is proportional to the square of the velocity:

$$E \propto \hbar^2 l^2 \tag{7.39}$$

This means that it is energetically more favourable to store e.g. two quantums of angular momentum $\hbar$ in two singly quantized vortices than in a double quantized

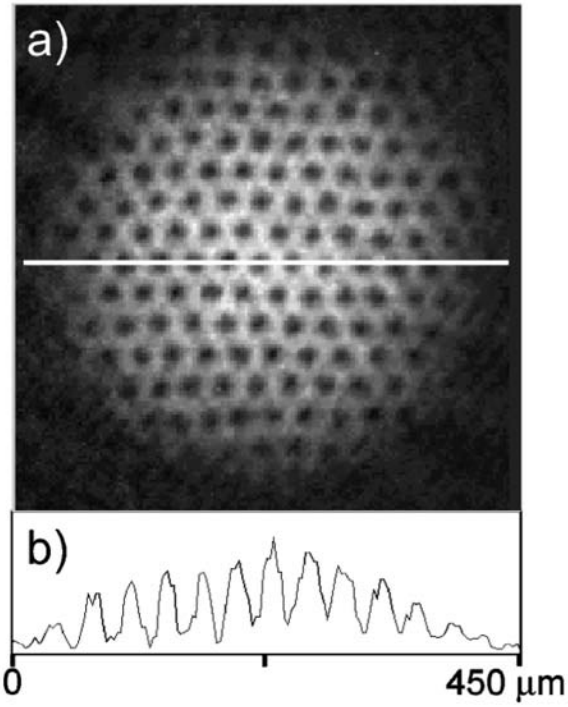


Figure 7.1: Example of vortex lattice seen along the rotation axis. The density profile shown below is taken along the white horizontal line. Figure from ref. [215].

one ($l = 2$). Hence, as the rotation rate increases and more angular momentum needs to be stored, the number of vortices increases and they adopt the least energetic configuration, which turns out to be a triangular lattice (see fig 7.1)

The area of each individual vortex also decreases with the rotation rate Ω. The reason for this is that, as shown by Feynman [216], the lowest energy for a superfluid (as for a normal fluid) is that in which the vorticity is constant throughout the liquid and equal to twice the rotation rate:

$$\vec{\nabla} \times \vec{v} = 2\vec{\Omega} \tag{7.40}$$

Nonetheless, the velocity field is irrotational (eq. 7.34) except for the contribution of the vortices (equation 7.37):

$$\vec{\nabla} \times \vec{v} = \hat{z}\frac{lh}{m}\delta^2(r) \tag{7.41}$$

where $\hat{z}$ is a unit vector in the direction of the rotation vector and $\delta^2(\vec{r})$ is a 2D delta function centered at the vortex core $\vec{r}$. Therefore the only way of meeting the above condition is by a proper configuration of the vortices so that the average vorticity in an area is:

$$\int_A |\vec{\nabla} \times \vec{v}| = 2\Omega A \tag{7.42}$$

If we take A to be the area of a single vortex:

$$\frac{h}{m} = 2\Omega A \tag{7.43}$$

and, therefore, the area of a vortex decreases with the rotation rate as:

$$\Omega = \frac{\pi\hbar}{mA} \tag{7.44}$$

So, as rotation increases the number of vortices grows and their separation diminishes.

Similar phenomena are observed in Type II superconductors, where the magnetic field plays the analog of the rotation rate. In this case, though, the vortex cores overlap above a critical field H_{c2} and the cooper pairs break up [217]: the normal, non-superconducting phase has a lower energy. In the case of rotating BECs, the bosons cannot disassemble, and hence the ultimate fate must be different.

What happens, then, to BECs at high rotation rates? The answer to this question is presently pursued very actively, both from a theoretical and experimental point of view [218]. The expected, theorized outcome is presented next for each of the two possible traps: harmonic or non-harmonic.

Harmonic traps

For harmonic traps there is no phase transition provoked by the core overlap for the simple reason that cores never touch, as explained in the the next section. The cores scale with vortex size [198], so there is never contact, as experimentally observed [219]. Additionally, harmonic traps have the same dependence on r as the centrifugal potential: $\frac{1}{2}m\omega_\perp^2$ vs. $-\frac{1}{2}m\Omega^2$. Thus, the maximum rotation rate that can be achieved is $\omega_\perp$ before the containment is precluded and the condensate disperses. As the rotation rate becomes closer and closer to this limit, the condensate will spread and the density will become lower and lower. In this limit, the interaction terms become less important compared to the Coriolis force. The system is formally analogous to a particle in a magnetic field [218]. An approach similar to that used for particles in a quantum Hall regime is applicable [209] and most particles condense into the lowest Landau level (LLL), with the interaction acting as a small perturbation.

At these regimes, very small distortions of the vortex lattice can produce large changes in density distributions [218]. Consequently, soft collective modes of the vortex lattice (Tkachenko waves) can lead to singular behavior for the order parameter phase correlations, destroying the superconducting state even at zero temperature [220],[221],[222]. This may happen even before the lattice melts and becomes a vortex liquid, a transition that has not yet been studied systematically [218].

At even higher rotation rates, the system enters a sequence of highly correlated incompressible fractional quantum Hall-like states, as seen in simulations [218].

Non-harmonic traps

In the case of non-harmonic traps, the core size reduces self-similarly with the size of the cell, as in the case of harmonic traps and as shown in the next section. The main difference is that the rotation rate can be increased to any level, since the potential will always trap the particles. At high enough rotation rates, a hole begins to form in the center of the trap as predicted theoretically and observed in simulations of the Gross-Pitaevskii equation [198], [223],[224],[225],[226]. At even higher rotation rates a single multiply quantized vortex is expected from theory and Gross-Pitaevskii equation simulations [198],[223],[224],[225],[226],[227].

7.2.1 A two scale description of vortex lattices

This section explains how to separate the behavior of a vortex lattice in two scales: a long wavelength coarse scale, where the details of the vortex structure have been integrated out, and a short wavelength one, where out of the specifics of the long wavelength density distribution of particles, only the average value of the density matters. This procedure was first used for BECs by Fischer and Baym [198], who used it to show that the vortex cores reduce self-similarly with the vortex size.

We start by writing the order parameter as:

$$\psi(\vec{r}) = e^{i\varphi(\vec{r})} f(\vec{r}) \sqrt{n(\vec{r})} \tag{7.45}$$

where $f(\vec{r})$ carries the short scale spatial dependence (vanishing at each core vortex), $\sqrt{n(\vec{r})}$ is a slowly varying real envelope function and $\varphi(\vec{r})$ is the phase. $f(\vec{r})$ is normalized to average to one over each unit cell. We will assume that the length scale of variation in the direction parallel to the rotation axis is much greater than than in the perpendicular direction, and we drop all dependences on z. We will also assume that the rotation of the container and the vortex lattice is the same, although, in reality, a slight discrepancy is expected [198].

In these terms, the Gross-Pitaevskii energy is:

$$E = \int \left[\lambda |\vec{\nabla}\psi(\vec{r})|^2 + \hat{\mathcal{U}}(\psi(\vec{r})) \right] d\vec{r} \tag{7.46}$$

We know that the kinetic energy in the above equation is the combination of a rigid body motion plus the vortex rotation around its core. For each vortex then, the kinetic energy can be written as the sum of of the kinetic energy of the vortex as if all its mass was concentrated in the center of mass plus the kinetic energy relative to the center of mass [228]. This is a property of the transformation of kinetic energies, applicable to quantum or classical systems. If we assume that the envelope function $\sqrt{n(\vec{r})}$ remains constant within a vortex, the kinetic energy in the moving frame of reference is [229],[198]:

$$\int |\vec{\nabla}\psi(\vec{r})|^2 d\vec{r} = \sum_j -\frac{1}{2} m R_j^2 N(\vec{R}_j)\Omega^2 + \sum_j n(\vec{R}_j) \int_j f^2(\vec{r}) \left(|\delta v(\vec{r})|^2 - \frac{1}{2} m\Omega^2 r^2 \right) d\vec{r} \tag{7.47}$$

The first term and last terms are the centrifugal potential as if all the vortex mass was concentrated in the center and for each vortex, respectively. The middle term corresponds to the velocity relative to the center of mass ($\delta v(\vec{r})$). The sum is over all cells j and the integral is over each cell j. The total number of particles in a cell is $N(\vec{R}_j) \equiv n(\vec{R}_j) \int_j f^2(\vec{r}) d\vec{r}$.

If we further assume that $f(\vec{r})$ depends only on the distance to the vortex core, and that the phase gradient in the moving frame $\vec{\nabla}\varphi'(\vec{r})$ only has components in the azimuthal direction ($\hat{u}_\theta$), the local velocity is:

$$\delta\vec{v} = \vec{\nabla}(f(\vec{r}) e^{i\varphi'(\vec{r})}) = e^{i\varphi'(\vec{r})} \vec{\nabla} f(\vec{r}) \hat{u}_r + i\, f(\vec{r}) e^{i\varphi'(\vec{r})} |\vec{\nabla}(\varphi'(\vec{r}))| \hat{u}_\theta \tag{7.48}$$

where $\hat{u}_r$ and $\hat{u}_\theta$ are unit vectors in the radial and azimuthal directions.

Adding the potential energies, independent of the frame of reference, one obtains the total energy in the moving frame [198]:

$$E' = -\frac{1}{2} I \Omega^2 + \sum_j E_j \tag{7.49}$$

$$E_j = n(\vec{R}_j) \int_j \left[\frac{1}{2m} \left(f(\vec{r})^2 |\vec{\nabla}\phi_j|^2 + |\vec{\nabla} f(\vec{r})|^2 \right) - \frac{1}{2} m\Omega^2 r^2 + \frac{1}{n(\vec{R}_j)} \hat{\mathcal{U}}(f(\vec{r}) \sqrt{n(\vec{R}_j)}) \right] d\vec{r}$$

where I is the moment of inertia: $I = \sum_j N(\vec{R}_j) m\Omega R_j^2$.

Similarly, the z component of the angular momentum $L_z$can be written as sum of contributions from center of mass movement and movement relative to the center of mass:

$$L_z = \sum_j m\Omega R_j^2 N(\vec{R}_j) + \sum_j n(\vec{R}_j) \int_j f^2(\vec{r})(\vec{r} \times \vec{\nabla}\phi_j)_z d\vec{r} \tag{7.50}$$

The equilibrium configuration can be found minizing E' or minimizing the energy in the lab frame $E = E' + \Omega L_z$ subject to the constraint that L_z is fixed. Both procedures are equivalent [230].

In the case that $n(\vec{r})$ can not be considered constant over a cell, the results are somewhat more involved [230]:

$$E' = -\frac{\Omega^2}{2}\left(\bar{I} + \frac{N}{2\Omega}\right) + \int \frac{(\vec{\nabla}\sqrt{n})^2}{2m} d\vec{r} - N\Omega + \sum_j E_j \tag{7.51}$$

$$E_j = n(\vec{R}_j)\int_j \left[\frac{1}{2m}\left(f(\vec{r})^2|\vec{\nabla}\phi_j|^2 + |\vec{\nabla}f(\vec{r})|^2\right) - \frac{1}{2}m\Omega^2 r^2 + \frac{1}{n(\vec{R}_j)}\hat{\mathcal{U}}(f(\vec{r})\sqrt{n(\vec{R}_j)})\right] d\vec{r}$$

Here $\bar{I}$ is the smoothed moment of inertia: $\bar{I} = \int mn(\vec{r})r^2 d\vec{r}$.

In any event, the idea is the same: it is possible to decouple the energy into two scales: small scale vortex structure and large scale vortex occupation.

Fischer and Baym[198] used this approach with a simple approximation to $f(r)$ and the phase $\phi(r)$:

$$\phi(\vec{r}) = \theta; \qquad f(r,z) = \begin{cases} A(r/r_0) & 0 \leq r \leq r_0 \\ \\ A & r_0 \leq r \leq l \end{cases} \tag{7.52}$$

where A is a normalization constant and θ is the azimuthal angle.

By doing this, they were able to calculate the energies E_j as a function of the core radius r_0, using as $n(\vec{R}_j)$ the average mean density. Plugging this result into equation 7.49, it was possible to find the core radius and the long wavelength density profile $n(\vec{r})$ that minimized the total energy. Two different confining potentials were considered: a harmonic trap and a cylindrical bucket.

For the harmonic trap, the confining potential $(mr^2\omega_\perp^2/2)$ is of the same kind as the centrifugal potential $(-mr^2\Omega^2/2)$ and for high enough rotation rates Ω they will cancel and preclude confinement. As the rotation rate reaches this critical point $(\Omega = \omega_\perp)$ the atom cloud increases in size in the direction perpendicular to the rotation and the average density tends to zero. Under these conditions the Gross-Pitaevskii equation is always applicable and it can be shown that the area of the vortex core is a fixed fraction of the total vortex area [198].

The cylindrical bucket potential ($V(r) = \infty$ for $r > R_b$, zero otherwise) was considered as a tractable approximation to an unharmonic potential. For this case, a hole develops in the center of the trap beyond a critical rotation rate Ω_h, with a radius $R_h = R\sqrt{1 - \Omega_h/\Omega}$, where R is the radius of the trap. The mean density is then $n = N/(A_v L_z)$, where A_v is the available area: $A_v = \pi(R^2 - R_h^2) = \pi R^2 \Omega_h/\Omega$ and L_z is the size of the system in the z direction. Since the rotation rate is proportional to the vortex area A(see section 7.2):

$$nA = \frac{N}{R^2}\frac{\hbar}{m\Omega_h} = \text{const} \tag{7.53}$$

Therefore, the number of particles in a vortex is constant. The vortex area, nonetheless, decreases with the rotation rate $a_v \propto 1/\Omega$ and the average density grows as Ω. At some point the Gross-Pitaevskii equation must fail and this regime is treated in the next section, devoted to the Path Integral Monte Carlo, which can treat arbitraritly strong interacting systems.

For this case, there is also a regime for which cores scale self-similarly with vortex size and it starts when the hole in the middle of the trap appears.

7.3 Self-similar solutions to the Gross-Pitaevskii equations

In the remainder of this chapter, we will try to provide a better description of the vortex structure than the simple model assumed by Fischer and Baym [198].

In this section, we will prove that the solutions to the GP-equation with no confining potential satisfy a scaling solution. In the next section, we will find this solution through finite difference techniques and in the last one, we will show the need for the extra length scale given by the scattering length σ.

Let's start with the energy for the lattice cell j:

$$E_j = \lambda n(\vec{R}_j) \int_j \left[|\vec{\nabla} f(\vec{r})|^2 + |\vec{\nabla}\phi_j(\vec{r})|^2 f^2(\vec{r}) - \frac{1}{2}\frac{m^2\Omega^2 r^2}{\hbar^2} f^2(\vec{r}) + \frac{gn(\vec{R}_j)}{2\lambda} f^4(\vec{r}) \right] d\vec{r} \tag{7.54}$$

As in ref. [200], we will define the coherence length as:

$$\frac{1}{\epsilon_0^2} = \frac{gn(\vec{R}_j)}{\lambda} \tag{7.55}$$

Defining $F(r, l_x) \equiv |\vec{\nabla}\phi_j(\vec{r})|^2$ (l_x = distance from vortex center to cell edge, see section 8.1.8) and introducing the dependence of the rotation rate on the vortex area $\Omega = \pi\hbar/mA$ ($A = 2\sqrt{3}l_x^2$) the energy becomes:

$$E_j = \lambda n(\vec{R}_j) \int_j \left[|\vec{\nabla} f(\vec{r})|^2 + F(r, l_x) f^2(\vec{r}) - \frac{1}{2}\frac{\pi^2}{12}\frac{r^2}{l_x^4} f^2(\vec{r}) + \frac{f^4(\vec{r})}{2\epsilon_0^2} \right] d\vec{r} \tag{7.56}$$

Since, as shown in section 8.1.8, $F(r, l_x)$ is of the type $F(r, l_x) = \bar{F}(r/l_x)/l_x^2$, the function $H(r, l_x) \equiv F(r, l_x) - \pi^2 r^2/3l_x^4$ is also of the type $H(r, l_x) = 1/l_x^2 \bar{H}(r/l_x)$, hence:

$$E_j = \lambda n(\vec{R}_j) \int_j \left[|\vec{\nabla} f(\vec{r})|^2 + \frac{1}{l_x^2}\bar{H}(r/l_x) f^2(\vec{r}) + \frac{f^4(\vec{r})}{2\epsilon_0^2} \right] d\vec{r} \tag{7.57}$$

Functional differentiation to minimize E_j, with chemical potential $\mu = \lambda k$ yields:

$$-\nabla^2 f(\vec{r}) + \frac{1}{l_x^2}\bar{H}(r/l_x) f(\vec{r}) + \frac{f^3(\vec{r})}{\epsilon_0^2} = k f(\vec{r}) \tag{7.58}$$

with the boundary condition that the gradient of the wave function is null at the cell edge. The phase gradient already obeys this condition because of the lattice symmetry, so for $f(\vec{r})$ this condition means:

$$\vec{\nabla} f(\vec{r}) = 0 \quad \text{at} \quad r\cos(\theta) = l_x \tag{7.59}$$

where θ is the angle to the closest line perpendicular to the cell edge going through the center (see fig. 7.2).

We can see in equation 7.58 that k has units of inverse length squared. There are therefore, only three length scales in the problem: L_x, ϵ_0 and $1/\sqrt{k}$.

Because of dimensional arguments, the solution must be a function involving ratios of these scales [231]. We will seek a solution of the type:

$$f(r, \epsilon_0, l_x, k) = \frac{\epsilon_0}{l_x}\bar{f}\left(r/l_x, kl_x^2\right) \tag{7.60}$$

Substituting in the above Gross-Pitaevskii equation this yields ($r' = r/l_x$):

$$\begin{aligned} -\frac{1}{l_x^2}\nabla^2\bar{f}(\vec{r'}) + \frac{1}{l_x^2}\bar{H}(\vec{r'})\bar{f}(\vec{r'}) + \frac{\bar{f}^3(\vec{r'})}{l_x^2} &= k\bar{f}(\vec{r'}) && (7.61)\\ \Rightarrow -\nabla^2\bar{f}(\vec{r'}) + \bar{H}(\vec{r'})\bar{f}(\vec{r'}) + \bar{f}^3(\vec{r'}) &= kl_x^2\bar{f}(\vec{r'}) && (7.62) \end{aligned}$$

and the chemical potential is:

$$\begin{aligned} k &= \frac{1}{qAf_m^2}\int_A \left[|\vec{\nabla} f(\vec{r})|^2 + \frac{f^2(\vec{r})}{L_x^2}\bar{H}(r/L_x) + \frac{f^4(\vec{r})}{\epsilon_0}\right] d^2\vec{r} \\ &= \int_A \frac{\epsilon_0^2}{L_x^2}\left[\frac{1}{L_x^2}|\vec{\nabla}\bar{f}(\vec{r'})|^2 + \frac{\bar{f}^2(\vec{r'})}{L_x^2}\bar{H}(\vec{r'}) + \frac{\bar{f}^4(\vec{r'})}{L_x^2}\right] d^2\vec{r} \Big/ \frac{\epsilon_0^2}{L_x^2}\int_A \bar{f}^2(\vec{r'})d^2\vec{r} \\ &= \frac{1}{L_x^2}\int_{A/L_x^2}\left[|\vec{\nabla} f(\vec{r'})|^2 + f^2(\vec{r})\bar{H}(\vec{r'}) + f^4(\vec{r'})\right] d^2\vec{r'} \Big/ \int_{A/L_x^2} \bar{f}^2(\vec{r'})d^2\vec{r'} \\ &= \frac{1}{L_x^2}\frac{\bar{Q}(kL_x^2)}{\bar{q}(kL_x^2)} \end{aligned} \tag{7.63}$$

where:

$$\begin{aligned} \bar{Q}(kL_x^2) &\equiv \int_{A/L_x^2}\left[|\vec{\nabla} f(\vec{r'})|^2 + f^2(\vec{r})\bar{H}(\vec{r'}) + f^4(\vec{r'})\right] d^2\vec{r'} && (7.64)\\ \bar{q}(kL_x^2) &\equiv \int_{A/L_x^2} \bar{f}^2(\vec{r'})d^2\vec{r'} && (7.65) \end{aligned}$$

The integrals are over a cell of side unity and depend only on the ratio $t \equiv kL_x^2$. Therefore,

$$t = \frac{Q(t)}{q(t)} \tag{7.66}$$

The sought solution will be the one that obeys this constraint and the boundary conditions, as shown in the next section.

7.4 Finite difference solution to the Gross-Pitaevskii equation

In this section, we will use a finite difference shooting method to solve the unmodified Gross-Pitaevskii for the scaling function $\bar{f}(\vec{r})$:

$$
\begin{aligned}
-\nabla^2 \bar{f}(\vec{r}) + \bar{H}(\vec{r})\bar{f}(\vec{r}) + \bar{f}^3(\vec{r}) = t\bar{f}(\vec{r}) \\
\vec{\nabla}\bar{f}(\vec{r}) = 0 \quad \text{at} \quad r\cos(\theta) = 1
\end{aligned}
\tag{7.67}
$$

Since the vortex array, vortex cell and phase gradient all have a six-fold symmetry, we will seek a six-fold symmetry. Hence, the involved functions can be expressed in terms of harmonics as follows:

$$\bar{f}(\vec{r}) = \sum_{p=-p_m}^{p_m} \bar{f}_p(r) e^{i6p\theta} \tag{7.68}$$

$$\bar{H}(\vec{r}) = \sum_{p=-p_m}^{p_m} \bar{H}_p(r) e^{i6p\theta} \tag{7.69}$$

Because the function $\bar{H}(\vec{r})$ has very small harmonic components for $p \neq 0$ (remember that the squared phase is mainly $1/r^2$ for a vortex), few harmonics are needed (see section 8.1.8). In practice three suffice, although we will use $p_m = 4$. Because of the sixfold symmetry we will only need to solve the equation for one sixth of the whole cell. Aditionally, each of these wedges is symmetric with respect to the center segment, so $\bar{f}_p(r) = \bar{f}_{-p}(r)$.

The divergence, interaction and potential term are then:

$$\nabla^2 \bar{f}(\vec{r}) \;=\; \sum_p \frac{1}{r}(r\bar{f}'_p(r))' e^{i6p\theta} + \frac{1}{r^2}\bar{f}_p(r)e^{i6p\theta} \tag{7.70}$$

$$\bar{f}^3(\vec{r}) \;=\; \sum_{p,m,n} \bar{f}_p(r)\bar{f}_m(r)\bar{f}_n(r)e^{i6(p+m+n)\theta} \tag{7.71}$$

$$\bar{H}(\vec{r})\bar{f}(\vec{r}) \;=\; \sum_{p,m} \bar{f}_p(r)\bar{H}_m(r)e^{i6(p+m)\theta} \tag{7.72}$$

Putting all the pieces together into the equation, multiplying by $e^{i6p\theta}$ and integrating over θ to separate the harmonics one gets:

$$\frac{1}{r}(\bar{f}_p'(r) + r\bar{f}_p''(r)) = \left(\frac{p^2}{r^2} - 1\right)\bar{f}_p(r) + \sum_n \bar{f}_n(r)\bar{H}_{p-n}(r) + \sum_{m,n}\bar{f}_m(r)\bar{f}_n(r)\bar{f}_{p-n-m}(r) \tag{7.73}$$

The derivatives can be approximated by [232]:

$$\begin{aligned} \frac{d\bar{f}_p(r)}{dr} &\simeq \frac{\bar{f}_p(r+dr) - \bar{f}_p(r-dr)}{2dr} & (7.74)\\ \frac{d^2\bar{f}_p(r)}{dr^2} &\simeq \frac{\bar{f}_p(r+dr) - 2\bar{f}_p(r) + \bar{f}_p(r+dr)}{dr^2} & (7.75)\end{aligned}$$

and substituting into the differential equation ($e \equiv dr/r$):

$$\begin{aligned}\bar{f}_p(r+dr) &= \frac{1}{1+e/2}\Bigg[(\frac{e}{2}-1)\bar{f}_p(r-dr) + 2\bar{f}_p(r) + e^2\Bigg(\bar{f}_p(r)\left(\frac{p^2}{r^2}-1\right)\\ &\quad + \sum_n \bar{f}_n(r)\bar{H}_{p-n}(r) + \sum_{n,m}\bar{f}_n(r)\bar{f}_m(r)\bar{f}_{p-n-m}(r)\Bigg)\Bigg] \qquad (7.76)\end{aligned}$$

To start the iterative solution,we use the starting point $r^* \simeq 0.1$ (to avoid the singularity), taking $\bar{f}_0(r^*) = cr^*$ and $\bar{f}_0(r^* - dr) = c(r^* - dr)$ for the lowest harmonic, and $\bar{f}_p(r^*) = 0$ for the rest it is possible to find $\bar{f}_p(r)$ for $r > r^*$ by using equation 7.76. The initial slope c is chosen by the shooting method so that the boundary conditions are met. The grid is formed by 10,000 and 100 points for the radial variable and the angular variable, respectively.

The result for the case without the centrifugal potential ($H(r, l_x) = F(r, l_x)$) can be seen in figures 7.2 and 7.3. It turns out that the only solution which matches the boundary conditions at $r/l_x = 1$ is that with $kl_x^2 = 3$. The ratio $Q(t)/q(t)$ for this solution is $Q(t)/q(t) = 2.9966 \simeq 3$, in accordance with equation 7.66.

The addition of the centrifugal potential changes the shape of the solution, but keeps the scaling properties. On the other hand, the addition of harmonic trap terms ($\frac{1}{2}\omega_\perp r^2$) or an extra term in the interaction energy (see section 7.1.2) breaks the self-similarity of the solution, since it adds a new lengthscale to the problem. This solution should provide a good starting point for perturbative analysis.

Unfortunately, this solution is not valid in its present form. As we will see in the next section, we need to take into account an extra length scale in the scaling function.

7.5 The need for a fourth length scale

Not all is well with the scaling solution presented above. To see why, let's apply the normalization condition to the scaling solution:

$$\frac{N}{Lz} = n(\vec{R}_j)\frac{\epsilon_0^2}{L_x^2}\int_A \bar{f}^2(\vec{r'}d\vec{r} = n(\vec{R}_j)\epsilon_0^2\bar{q}(t)$$

Then, from the definition of the coherence length ϵ_0 (eq. 7.55):

$$\begin{aligned} n(\vec{R}_j)\epsilon_0^2 &= \frac{\lambda}{g} = \frac{1}{8\pi\sigma} \\ \Rightarrow \frac{N}{L_z} &= \frac{\bar{q}(t)}{8\pi\sigma} \end{aligned} \tag{7.77}$$

where L_z is the size of the system in the z direction.

The function $\bar{q}(t)$ only depends on $t = \mu L_x^2/\lambda$, which must take the value 3 in order to match the boundary conditions, as explained in the previous section. Therefore, it appears that there is only one possible value for the linear density N/L_z that is allowed and it is equal to $\bar{q}(3)/8\pi\sigma$.

This is shocking in itself, but what is more surprising is that this "acceptable" density, depends on the hard sphere diameter σ, which is much smaller than the other lengths involved in the problem (cell size, coherence length... etc) and does not enter into the determination of the shape of the vortex (only in the normalization of $\bar{f}$).

The interpretation of this argument is that the length scale σ somehow determines the solution. Imagine that we used the modified Gross-Pitaevskii equation:

$$-\nabla^2\bar{f}(\vec{r}) + \bar{H}(\vec{r})\bar{f}(\vec{r}) + \bar{f}^3(\vec{r})\left(1 + \frac{a}{\sqrt{8\pi}}\frac{\sigma}{l_x}\bar{f}(\vec{r})\right) = t\bar{f}(\vec{r}) \tag{7.78}$$

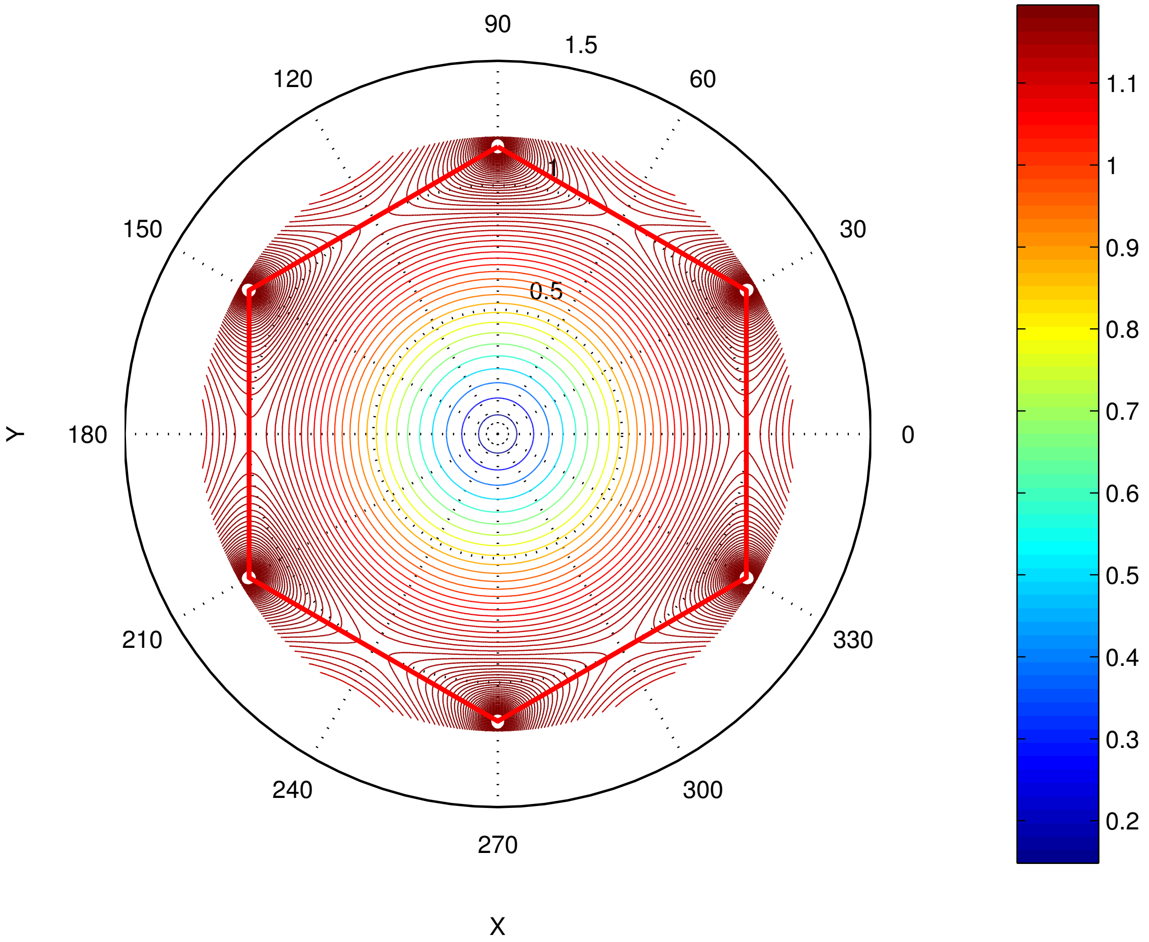


Figure 7.2: Finite difference solution $f(\vec{r})$ to equation 7.67. Only the solution with $kL_x^2 = 3$ matches the boundary conditions at $r' = r/l_x = 1$. For this solution $Q(t)/q(t) = 2.9966 \simeq 3$ as expected from self-consistency. The red line marks the boundary of the cell, across which the gradient must be zero. The angle θ for equation 7.67 is measured for each of the six symmetry-related equivalent wedges from the closest line that passes through the center and is perpendicular to one of the edges.

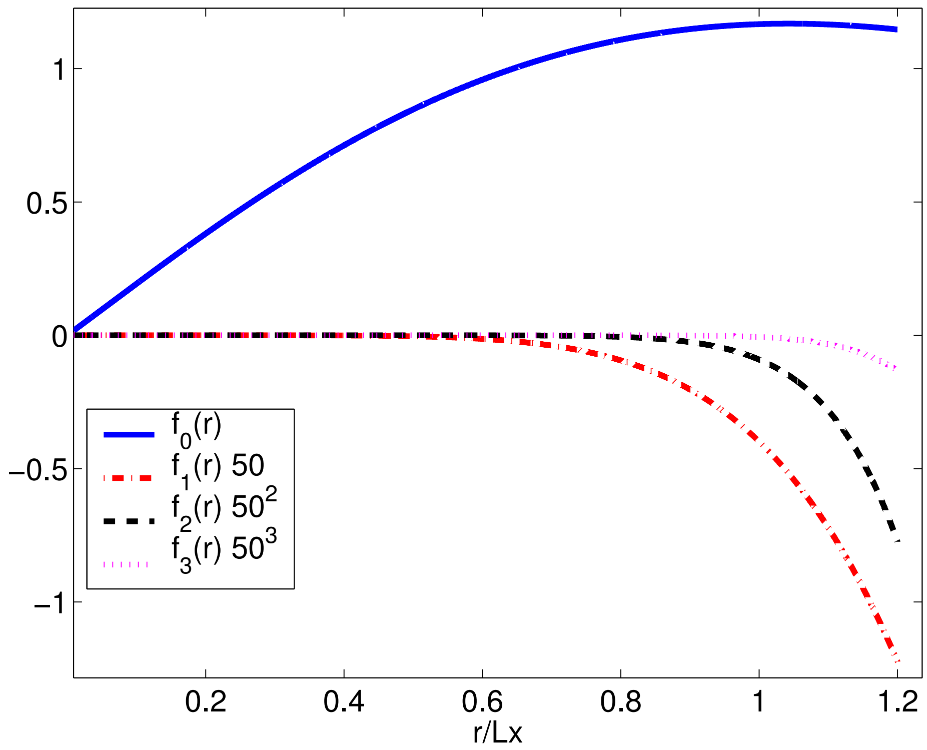


Figure 7.3: Modes $f_p(r)$ for the solution shown in figure 7.2. As expected $p = 0$ is the largest and the others are much smaller with larger values of p. Notice the scale difference for each mode (factor of 50).

even though it seems that the extra term is negligible for low enough densities. The extra length scale σ provides an exit to our paradox. From dimensional analysis the scaling function $\bar{f}$ depends on σ through the relation:

$$f(r, \epsilon_0, l_x, k, \sigma) = \frac{\epsilon_0}{l_x}\bar{f}(r/L_x, kl_x^2, \sigma/l_x) \tag{7.79}$$

Then, the normalization condition becomes:

$$\frac{N}{L_z} = \frac{\bar{q}(t,\sigma)}{8\pi\sigma} = \frac{\bar{q}(\mu l_x^2/\lambda, \sigma/l_x)}{8\pi\sigma} \tag{7.80}$$

Now, the addition of the extra term allows us to meet the boundary conditions so that $\mu l_x^2/\lambda = 3$ is not the only possible solution. Then, for a given σ and l_x, μ could be modified to accommodate the actual density N/L_z.

This is an example of the well-known mathematical phenomenon of intermediate asymptotics and incomplete similarity [231]. The extra factor $\frac{a}{\sqrt{8\pi}}\frac{\sigma}{l_x}\bar{f}$ is a singular perturbation and the $\sigma \to 0$ limit is singular. As we will see in section 8.2, the new term does not change the solution profile significantly, but it does change the scaling, allowing the boundary conditions to be met for a different value of $\mu l_x^2/\lambda$ that can adhere to the normalization condition in equation 7.80. The scaling solution presented in section 7.4 properly treated with perturbative techniques [233] provides and exact solution to the structure of the vortex in a single cell. Coupled with the knowledge of the long wavelength density $n(\vec{R}_j)$ from the work of Fischer and Baym [198], this gives a semi-analytical description of vortex lattices.

7.6 Conclusion

We have presented a mathematical self-similar solution to the Gross-Pitaevskii equation for a vortex cell that can be used in conjunction with perturbative methods to provide a semi-analytical physical description of vortex lattices.

Chapter 8

Path Integral Monte Carlo studies of rotating Bose Einstein Condensates

In this chapter I will present results regarding the structure of a single vortex cell in a lattice (see section 7.2.1) for the case of an unharmonic bucket-like trapping potential. In this case, the density grows as the rotation rate increases and the Gross-Pitaevskii description breaks down. A method like the PIMC, able to capture this behavior, becomes necessary. We will show that, surprisingly, the Gross-Pitaevskii equation remains valid for much higher densities than it is supposed to.

8.1 Path Integral Monte Carlo in a nutshell

The Path Integral Monte Carlo (PIMC) calculates static properties of quantum systems in thermal equilibrium without uncontrolled approximations. For the case of helium, for which the interaction potential is precisely known, it has been used to make remarkably accurate predictions of e.g. specific heat, pair correlation functions and superfluid fractions [199].

An in depth review of the PIMC can be found in reference [199]. The purpose of this section is not to present a detailed explanation of this method, but rather to offer a quick and intuitive overview. Hopefully, it will be enough for the casual reader to understand how it was used for the case of vortex lattices in BEC and grasp its virtues and limitations without dwelling on unnecessary details.

8.1.1 Feynman's path integral formalism

Assume we have a system of N particles in a volume V at temperature T and the exact eigenvalues and eigenfunctions of the hamiltonian $\mathcal{H}$ are ϕ_i and E_i. In thermal equilibrium the probability of the state i being occupied is proportional to e^{-E_i/K_bT} and the equilibrium value of an operator $\mathcal{O}$ is:

$$\langle \mathcal{O} \rangle = Z^{-1} \sum_i \langle \phi_i | \mathcal{O} | \phi_i \rangle \tag{8.1}$$

where the partition function is:

$$Z = \sum_i e^{-\beta E_i} \tag{8.2}$$

In the position representation, we write $\langle \mathcal{O} \rangle$ by inserting the realization of the identity $\mathcal{I} = \int dR |R\rangle\langle R|$:

$$\langle \mathcal{O} \rangle = Z^{-1} \int dR dR' \rho(R, R'; \beta) \langle R | \mathcal{O} | R \rangle \tag{8.3}$$

where $R = \{\vec{r}_1, ..., \vec{r}_2\}$ ($\vec{r}_i \equiv$ position of ith particle). The partition function then becomes:

$$Z = \int dR \rho(R, R; \beta) \tag{8.4}$$

The matrix element:

$$\rho(R, R'; \beta) = \langle R | e^{-\beta \mathcal{H}} | R' \rangle = \sum_i \phi_i^*(R) \phi_i(R') e^{-\beta E_i} \tag{8.5}$$

is called the position space *density matrix* and contains all necessary information about the system in order to calculate operator averages.

The crucial, exact property that makes density matrices computationally useful is the exponential dependence on temperature:

$$\rho(R_1, R_3; \beta_1 + \beta_2) = \langle R_1 | e^{-\beta_1 \mathcal{H}} e^{-\beta_2 \mathcal{H}} \rangle R_2 = \int dR_2 \rho(R_1, R_2; \beta_1) \rho(R_2, R_3; \beta_2) \quad (8.6)$$

By applying this property M times one obtains:

$$\rho(R_0, R_M; \beta) = \int ... \int dR_1 dR_2 ... dR_{M-1} \rho(R_0, R_1; \tau) \rho(R_1, R_2; \tau) ... \rho(R_{M-1}, R_M; \tau) \quad (8.7)$$

and the density matrix $\rho(R, R'; \tau)$ can be very well approximated for small "time" step $\tau = \beta/M$. Note that τ is the imaginary time from Feynman's path integral formalism [234] at finite temperature.

If we assume that the hamiltonian can be written as $\mathcal{H} = \mathcal{T} + \mathcal{V}$ ($\mathcal{T}$ and $\mathcal{V}$ being the kinetic and potential operators) the following exact operator identity holds [199]:

$$e^{-\tau(\mathcal{T}+\mathcal{V}) + \frac{\tau^2}{2}[\mathcal{T},\mathcal{V}]} = e^{-\tau\mathcal{T}} e^{-\tau\mathcal{V}} \quad (8.8)$$

For small enough τ that the commutator term can be neglected, the *primitive approximation* is obtained:

$$e^{-\tau(\mathcal{T}+\mathcal{V})} \simeq e^{-\tau\mathcal{T}} e^{-\tau\mathcal{V}} \quad (8.9)$$

and therefore the density matrix is the product of the density matrices for the kinetic ($\mathcal{T}$) and potential operators ($\mathcal{V}$).

More rigorously, it can be proved that [235],[199]:

$$e^{-\beta(\mathcal{T}+\mathcal{V})} = \lim_{M \to \infty} \left[e^{-\tau\mathcal{T}} e^{-\tau\mathcal{V}} \right] \quad (8.10)$$

In the primitive approximation the density matrix in position space is therefore:

$$\rho(R_0, R_2; \tau) \simeq \int dR_1 \langle R_0 | e^{-\tau\mathcal{T}} | R_1 \rangle \langle R_1 | e^{-\tau\mathcal{V}} | R_2 \rangle \quad (8.11)$$

The potential operator part is straightforward:

$$\langle R_1 | e^{-\tau\mathcal{V}} | R_2 \rangle = e^{-\tau V(R_1)} \delta(R_2 - R_1) \quad (8.12)$$

and the kinetic contribution can be calculated using the eigenfunctions of $\mathcal{T}$. Let us assume there are N distinguishable particles in a cube of side L with periodic boundary conditions. The eigenfunctions and eigenvalues of $\mathcal{T}$ are then $L^{-3N/2}e^{i\vec{K}_{\vec{n}}R}$ and $\lambda\vec{K}^2_{\vec{n}}$, with $\vec{K}_{\vec{n}} = 2\pi\vec{n}/L$ and $\vec{n}$ a $3N$-dimensional integer vector. In that case:

$$\begin{aligned}\langle R_0|e^{-\tau\mathcal{T}}|R_i\rangle &= \sum_{\vec{n}} L^{-3N} e^{-\tau\lambda\vec{K}^2_{\vec{n}} - i\vec{K}_{\vec{n}}(R_0-R_1)} && (8.13)\\ &= (4\pi\lambda\tau)^{-3N/2}\exp\left[-\frac{(R_0-R_1)^2}{4\lambda\tau}\right] && (8.14)\end{aligned}$$

Here we have approximated the sum by an integral, which is only acceptable for:

$$\lambda\tau \ll L^2 \tag{8.15}$$

Putting everything together (eqs 8.7, 8.9, 8.12 and 8.13) one obtains the discrete path-integral expression for the density matrix in the primitive approximation:

$$\rho(R_0, R_M;\beta) = \int dR_1...dR_{M-1}(4\pi\lambda\tau)^{-3NM/2}\exp\Bigg(-\sum_{m=1}^{M}\left[\frac{(R_{m-1}-R_m)^2}{4\lambda\tau}+\tau V(R_m)\right]\Bigg) \tag{8.16}$$

This high dimensional integral cannot be performed exactly without severe approximations, but its value can be approximated to any necessary accuracy by sampling the integrand. This means randomly choosing points of the $R_1 \times R_2 \times ... \times R_M$ phase space with a probability given by the exponential in equation 8.16 (see section 8.1.4).

In the continuum limit $M \to \infty$ equation 8.16 yields the Feynman-Kacs formula [234]:

$$\rho(R_0, R_F;\tau) = \int[\mathcal{D}R(t)]\exp\left[-\int\left(\frac{1}{4\lambda}\dot{R(t)}^2 + V(R(t))\right)dt\right]$$

where the boundary conditions are $R(0) = R_{m-1}$ and $R(t_{end}) = R_m$.

Defining the action S as the minus logarithm of the exact density matrix:

$$S \equiv -\ln\left[\rho(R_0, R_F)\right] \Rightarrow \rho(R_0, R_F) \equiv e^{-S}, \tag{8.17}$$

the kinetic action as the exact kinetic action:

$$K \equiv \frac{3N}{2} \ln\left[4\pi\lambda\tau\right] + \frac{(R_0 - R_F)^2}{4\lambda\tau} \tag{8.18}$$

and the interaction as the remainder:

$$U \equiv S - K \tag{8.19}$$

we can write the interaction U as:

$$e^{-U(R_0,R_F;\tau)} = \left\langle \exp\left[-\int V(R(t))dt\right]\right\rangle_{RW} \tag{8.20}$$

Here $\langle ... \rangle_{RW}$ indicates averaging over all Gaussian random walks from R_0 to R_F:

$$\begin{aligned}\left\langle \exp\left[-\int V(R(t))dt\right]\right\rangle_{RW} &\equiv \int [\mathcal{D}R(t)] \exp\left[-\int \left(\frac{1}{4\lambda}\dot{R(t)}^2 + V(R(t))\right) dt\right] \\ &\times \left(\int [\mathcal{D}R(t)] \exp\left[-\int \left(\frac{1}{4\lambda}\dot{R(t)}^2\right)\right]\right)^{-1}\end{aligned} \tag{8.21}$$

Using this definition of U for each time step, and the definition of S in equation 8.17, equation 8.7 becomes:

$$\rho(R_0, R_M; \beta) \ = \int ... \int dR_1 dR_2 ... dR_{M-1} \exp\left[-\sum_{m=1}^{M} S^m\right] \tag{8.22}$$

where

$$S^m = \frac{3N}{2}\log(4\pi\lambda\tau) + \frac{(R_{m-1} - R_m)^2}{4\lambda\tau} + U(R_{m-1}, R_m; \tau) \tag{8.23}$$

Here, in contrast with equation 8.16, we have not made the primitive approximation. Equation 8.22 is valid for any value of τ and M, whereas equation 8.16 is only accurate for small enough τ.

8.1.2 Bose symmetry

Nothing stated above took into account the indistinguishability of particles. For a bose system only eigenfunctions $\phi_i(R)$ symmetric with respect to the permutation of particles can contribute to the density matrix. Thus, we require that $\phi_i(PR) = \phi_i(R)$, where P is a permutation of particle labels: $PR = (\vec{r}_{P_1}, \vec{r}_{P_2}..., \vec{r}_{P_N})$.

The symmetrization operator:

$$\mathcal{P} = \frac{1}{N!} \sum_P \phi(PR) \tag{8.24}$$

projects out only these bose states, because for hamiltonians symmetric under particle exchange, all states are even or odd with respect to a given permutation [199].

Applying the symmetrization operator to the density matrix in the position representation (eq. 8.5) we find:

$$\rho_B(R_0, R_1; \beta) = \frac{1}{N!} \sum_P \rho(R_0, PR_1; \beta) \tag{8.25}$$

where ρ_B is the bosonic density matrix.

The number of terms in this sum is $N!$, which makes it unfeasible to calculate for large N. As in the case of paths one must sample this sum. Therefore, for a bosonic system, it is necessary to execute a random walk both through the path space *and* the permutation space.

The partition function for a bosonic system is then:

$$Z_B = \frac{1}{N!} \int dR_0 ... dR_{M-1} \exp\left(- \sum_{m=1}^{M} S^m \right) \tag{8.26}$$

with the boundary condition that $PR_m = R_0$.

8.1.3 The pair action

The primitive approximation expression for the density matrix in equation 8.16 is certainly valid and it is possible to obtain correct results with it, but at the cost of using a very small τ and an enormous number of time slices. This is particularly true for very singular potentials of the type found in Helium (Lennard-Jones) or the case of the vortex lattice (where the potential addition due to the vortex phase $\simeq 1/r^2$ for small r, see section 8.1.8).

In order to understand why, take equation 8.22 for a time step (R_{m-1}, R_m). Approximating $\int V(R(t))dt \simeq V(R_m)\tau$ equation 8.16 is recovered:

$$e^{-U(R_{m-1},R_m;\tau)} = \left\langle \exp\left[-\int V(R(t))dt\right]\right\rangle_{RW} \simeq e^{-V(R_m)\tau} \tag{8.27}$$

In doing so, though, we are washing out the details of the potential between R_{m-1} and R_m by assuming that the potential at $V(R_m)$ is representative of the slice. For very localized potentials it is necessary to use very small τ to take their structure into account.

There are, however, better ways to calculate the action. The one used for this work is the pair-product action, although there are other possibilities [199]. The idea is to determine the action exactly for two interacting atoms and to use that to get the full action. Assume that the potential is a pairwise sum of terms of the form:

$$V(R) = \sum_{i<j} v(r_i - r_j) \tag{8.28}$$

The Feynman-Kacs formula (eq. 8.20) then gives:

$$e^{-U} = \left\langle \prod_{i<j} \exp\left[-\int_0^\tau v(\vec{r}_{ij}(t))\right]\right\rangle \tag{8.29}$$

For low enough τ such that three body effects can be neglected, equation 8.29 can be approximated by a product of pair density matrices:

$$e^{-U} \simeq \prod_{i<j} \left\langle \exp\left[-\int_0^\tau v(\vec{r}_{ij}(t))\right]\right\rangle \tag{8.30}$$

The advantage of this approach is that the appropriate value of τ is set by the length scale of the density of particles (so three body effects are negligible) and not by the length scale of the variation of the potential. In practical terms, this means that PIMC simulations become feasible.

The pair density matrices are obtained through an exact matrix squaring method [199].

8.1.4 Monte Carlo algorithm

As mentioned before, the integral in equation 8.7 cannot be performed exactly without severe uncontrolled approximations. Instead, we sample the configuration space of:

$$s = [P, R_1, ..., R_M] \tag{8.31}$$

where $R_k = \{\vec{r}_{1k}, ..., \vec{r}_{Nk}\}$ are the path variables and P is the permutation of particles that closes the path: $R_{M+1} = PR_1$. Each point s of the configuration space is sampled with probability:

$$\pi(s) = \frac{\exp[-\sum_{k=1}^{M} S_k]}{Z} \tag{8.32}$$

in order to calculate the integral.

The sampling is done through a generalization of the Metropolis et al. [236] rejection algorithm. A fixed transition rule $P(s \to s')$ is used to generate a random walk through the configuration space:$\{s_0, s_1, s_2, ...\}$. This rule must be set so that it is *ergodic*: all the configuration states can be accessed with a nonzero probability. Usually, the transition probabilities are chosen so they satisfy detailed balance:

$$\pi(s)P(s \to s') = \pi(s')P(s \to s) \tag{8.33}$$

In this way, if the probability of a walk being at states s and s' is proportional to $\pi(s)$ and $\pi(s')$ respectively, the walk is in equilibrium in the sense that, on average, there are as many transitions going from s to s' as from s' to s. Under these conditions, the walk is guaranteed to sample $\pi(s)$ in the limit of many steps [199].

This is the gist of the sampling techniques used in PIMC. Further refinements are necessary for a good performance which are thoroughly explained in reference [199].

8.1.5 Calculating energies, superfluid fractions and correlation functions

The calculation of scalar operators such as densities and correlation functions is straightforward. By inserting $\mathcal{O} = \sum_i \delta(\vec{r} - \vec{r}_i)$ in equation 8.1 we get:

$$\rho(\vec{r}) = \frac{1}{M} \sum_{i,t} \langle \delta(\vec{r} - \vec{r}_{it}) \rangle \tag{8.34}$$

where we have averaged over slices, too, since they are equivalent.

The superfluid fraction ρ_s is calculated through the winding number W [199]:

$$\frac{\rho_s}{\rho} = \frac{m}{\hbar^2} \frac{\langle W^2 \rangle L^2}{3\beta N} \tag{8.35}$$

where L is the size of the system and the winding number is defined by:

$$\sum_{i=1}^{N} (\vec{r}_{P_i} - \vec{r}_i) = WL \tag{8.36}$$

which essentially counts how many times the paths of the N particles have wound around the periodic cell. For a normal system, the permutation P_i is the identity permutation and the winding number is zero. For a superfluid, particle exchange becomes important and configurations with permutations other than the identity have nonnull probabilities and contribute to the winding number. This is a measure of the response of the system to moving boundary conditions [237].

Energy calculations are based on the thermodynamic estimator:

$$E_T = -\frac{1}{Z} \frac{dZ}{d\beta} \tag{8.37}$$

which in terms of the path variables becomes (see equation 8.26):

$$E_T = \frac{1}{M} \sum_m \left\langle \frac{3N}{2\tau} - \frac{(R_m - R_{m-1}).^2}{4\lambda\tau^2} + \frac{dU^m}{d\tau} \right\rangle \tag{8.38}$$

These averages are over paths.

The actual estimator used is the virial estimator, which is an improved version of the thermodynamic estimator averaged over several slices [199].

In this work, we do not calculate energies. Since the energy is proportional to the gas density and this is typically very small for BECs, the current implementation of the pair density matrices lacks the required accuracy for such calculations.

8.1.6 Adding vortices to the PIMC

The Path Integral Monte Carlo method as explained above in the previous sections is isotropic: all directions are equivalent. What breaks the symmetry and makes the appearance of vortices in a given direction energetically favorable is the addition of the angular momentum term to the hamiltonian (see section 7.2.1):

$$\mathcal{H}' = \mathcal{H} - \Omega \mathcal{L}_z \tag{8.39}$$

If we were to add this term to the action in eq. 8.23, vortices would appear automatically. This is not feasible, since the value of the angular momentum is not known a priori and can only be known after the sampling is done. It would be possible to do a self consistent approach to this problem but it would require major changes in the path integral code used for this work[1].

Instead, we introduce the vortices by imposing the wavefunction phase to be that of a vortex lattice. This we do via the *fixed phase method*. This is the same procedure commonly used in solving the Gross-Pitaevskii equation [198].

8.1.7 The fixed phase method

The fixed phase method [238],[239] assumes that we know the phase of the wavefunction and uses it to obtain the modulus. Let $\Phi(\mathcal{R}) = |\Phi(\mathcal{R})| \exp[i\varphi(\mathcal{R})]$ be the N-particle wavefunction, where $|\Phi(\mathcal{R})|$ and $\varphi(\mathcal{R})$ are real. The vector $\mathcal{R} = (\vec{r}_1, ..., \vec{r}_i, ..., \vec{r}_N)$ denotes a point of the $3N$-dimensional space. The imaginary and

[1] Universal Path Integral code (UPI) developed by David Ceperley

real parts of the Schrödinger equation $\mathcal{H}\Phi(\mathcal{R}) = E\Phi(\mathcal{R})$ produce the coupled equations:

$$\mathcal{H}|\Phi(\mathcal{R})| = \left[\sum_{i=1}^{N} \lambda\, \vec{\nabla}_i^2 + \mathcal{V}(\mathcal{R})\right] |\Phi(\mathcal{R})| = E|\Phi(\mathcal{R})| \tag{8.40}$$

$$\sum_{i=1}^{N} \vec{\nabla}_i \left\{ |\Phi(\mathcal{R})|^2 [\vec{\nabla}_i \varphi(\mathcal{R})] \right\} = 0 \tag{8.41}$$

where the effective potential $\mathcal{V}(\mathcal{R})$ is:

$$\mathcal{V}(\mathcal{R}) = V(\mathcal{R}) + \lambda \sum_{i=1}^{N} [\vec{\nabla}_i \varphi(\mathcal{R})]^2 \tag{8.42}$$

and $V(\mathcal{R})$ is the original potential. Hence, if we know the phase and plug it into the effective potential, we can use the exact results of the PIMC to find the modulus. The equation to be solved is exactly the same as for the full wave function, but the potential has an additional term depending on the square of the phase gradient.

Note that it is assumed that constraint equation 8.40 is automatically satisfied for the correct modulus $|\Phi(\mathcal{R})|$. For this particular case we take the phase to be the sum of phases for each particle:

$$\varphi(\mathcal{R}) = \sum_{i=1}^{N} \varphi_i(\vec{r}_i) \tag{8.43}$$

and express each phase as the sum of the contributions from the vortices in an infinite lattice:

$$\vec{\nabla} \varphi_i(\vec{r}) = \vec{\nabla} \varphi(\vec{r}) = \sum_{\alpha} \frac{1}{|\vec{r} - \vec{r}_\alpha|} \hat{u}_\theta^\alpha \tag{8.44}$$

Here $\hat{u}_\theta^\alpha$ is the unit vector in the direction perpendicular to the distance vector for the contributing vortex (fig. 8.1), and α runs over all the vortices in the lattice.

8.1.8 Summing the phase

The summation in equation 8.44 must be carried out in each PIMC step, so it needs to be fast to compute. Fast numerical methods have been developed for the evaluation of energies and forces between particles interacting logarithmically in a periodic

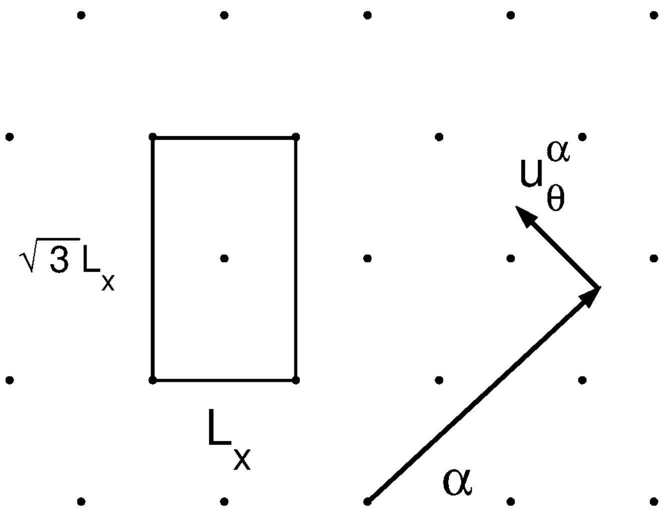


Figure 8.1: Triangular lattice disposition for vortex cores. The simulation cell is rectangular with double the area of a single vortex. This configuration was chosen in order to use periodic boundary conditions. The vector $\hat{u}_\theta^\alpha$ is the unit vector in the azimuthal direction for the vortex contributing to the phase gradient at the given point.

lattice [240],[241]. It can be shown [241] that the necessary sums:

$$\begin{aligned}
\nabla^{(x)}\varphi(\vec{r}) &= \sum_{m=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\frac{y+kL_y}{r_{mk}^2} && (8.45)\\
\nabla^{(y)}\varphi(\vec{r}) &= \sum_{m=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}-\frac{x+mL_x+k\xi}{r_{mk}^2} && (8.46)
\end{aligned}$$

where

$$r_{mk} = \sqrt{(x+mL_x+k\xi)^2+(y+kL_y)^2} \tag{8.47}$$

can be written in terms of standard trigonometric functions that converge fairly quickly:

$$\begin{aligned}
\nabla^{(x)}\varphi(\vec{r}) &= \frac{\pi}{L_x}\sum_{k=-\infty}^{\infty}\frac{\tanh(A)}{1-\cos(B)/\cosh(A)}-\frac{2\pi}{L_x}\frac{y}{L_y}\\
\nabla^{(y)}\varphi(\vec{r}) &= -\frac{\pi}{L_x}\sum_{k=-\infty}^{\infty}\frac{\sin(B)}{\cosh(A)-\cos(B)} && (8.48)\\
&A = 2\pi\frac{L_y}{L_x}\left(\frac{y}{L_y}+k\right)\\
&B = 2\pi\frac{x+\xi k}{Lx} && (8.49)
\end{aligned}$$

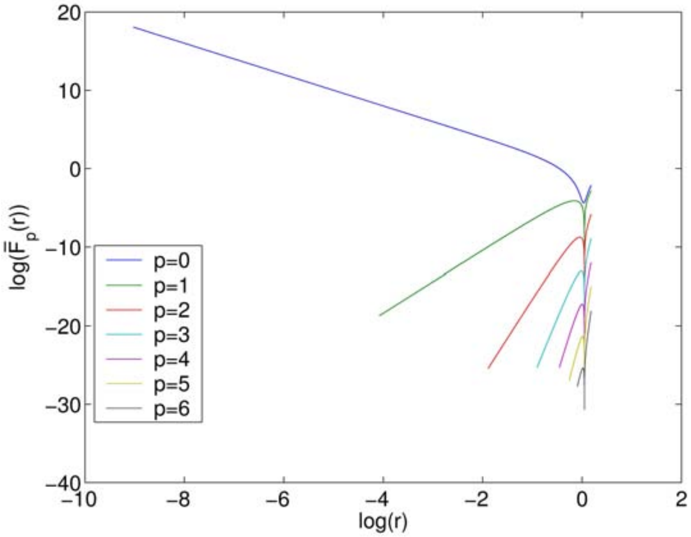


Figure 8.2: Modes of the phase gradient squared: $\bar{F}_p(r) \equiv F_p(r, 2)$. As expected, the $p = 0$ mode is much larger than the rest. The modes have been zeroed beyond a minimum threshold to avoid numerical noise.

Here the sums are meant to be carried out pairwise and for a hexagonal lattice $L_y = \sqrt{3}L_x/2$ and $\xi = L_x/2$. We have found that sums from $k = -10$ to $k = 10$ give satisfactory results.

Only the function $F(\vec{r}, L_x) \equiv |\vec{\nabla}\varphi(\vec{r})|^2$ is needed for the fixed phase method and the Gross-Pitaevskii equation in the previous chapter. Because of the 6-fold symmetry of the cell (fig. 7.2) and the vortex array (fig. 8.1) $F(\vec{r}, L_x)$ can be decomposed in harmonics:

$$F(\vec{r}, L_x) = \sum_p F_p(r) e^{i6p\theta} \tag{8.50}$$

Within a cell, it would be expected that the dominant contribution would be that of the vortex within the cell: $|\vec{\nabla}\varphi(\vec{r})|^2 \approx 1/r^2$ and therefore the modes $p \neq 0$ shouldn't have a large relative contribution. This is confirmed in figure 8.2.

An important scaling characteristic of $F(\vec{r}, L_x)$ that will be used in section 7.4 is that it can be written as:

$$F(\vec{r}, L_x) = \frac{1}{l_x^2} \bar{F}(\vec{r}/l_x) \tag{8.51}$$

for a given function $\bar{F}$, where $l_x = L_x/2$ is the distance from the vortex core to the edge.

This can be proved from the scaling properties of $F(\vec{r}, L_x)$. Since it has units of inverse length squared, a change of length scale produces:

$$F(s\vec{r}, sL_x) = \frac{1}{s^2} F(\vec{r}, L_x) \tag{8.52}$$

Taking $s = 1/l_x$ this becomes:

$$\begin{aligned} &F(\vec{r}/l_x^2) = l_x^2 F(\vec{r}, L_x) \\ &\Rightarrow F(\vec{r}, L_x) = \frac{1}{l_x^2} F(\vec{r}/l_x, 2) \equiv \frac{1}{l_x^2} \bar{F}(\vec{r}/l_x) \end{aligned} \tag{8.53}$$

8.2 Results

This section presents the results of the PIMC simulations done with a hard sphere potential of scattering length σ between particles. A bucket-like trapping potential for the whole lattice was chosen as in ref. [198] (see section 7.2.1), since this is the case in which the deviations from the Gross-Pitaevskii equation are expected. For each vortex cell this means that there is no potential within the cell, but the number of particles is constant (see section 7.2.1). Periodic boundary conditions are used for simulating the interaction with the other vortices in the lattice. The cells are of size L_z in the direction parallel to the rotation and L_x and $\sqrt{3}L_x$ (see fig. 8.1) in the perpendicular direction. In units of σ, $L_z = 10$ and $L_x = 32$, 15, 9 and 5 for the different runs. The rotation rate is determined by the area through equation 7.44.

We omitted the contribution of the centrifugal potential. These terms are very small for rapidly rotating BECs [218], and we are more interested in exploring the validity of the Gross-Pitaevskii than in accurate predictions, for the moment being. The temperature was set well below the transition temperature to obtain superfluid fractions close to 1 ($\rho_s/\rho = 0.92$ is the minimum obtained). In accordance with the arguments in section 7.2.1, we used a constant number or particles ($N = 25$) as pertains to a bucket potential where a hole develops in the center. The densities (in units of σ) ranged from $n = 1.4\ 10^{-3}$ to $n = 5.7\ 10^{-1}$, which are above the typical experimental values $n \simeq 10^{-5} - 10^{-4}$. The main reason for this limitation is technical: the interpolations used for the pair density matrix (see section 8.1.3) become inaccurate for high value of τ. At the same time, the critical value of β increases with low densities [200]:

$$\beta_c \propto n^{-2/3} \tag{8.54}$$

The combination of lower τs and higher β_c makes the number of slices $M = \beta_c/\tau$ increase very quickly. The simulations then become unfeasible, because the number of path sampling moves to reach equilibrium take more than the reasonable time expected for a single run ($\approx$ 1 day). It would be interesting to be able to reach those low densities to check some of the hypothesis presented in the last chapter and in order to do that, it would be necessary to use a better pair density matrix. Here, I present the results for higher densities.

For the lowest density ($n = 1.4\ 10^{-3}$, fig. 8.5) it is possible to make connection with the Gross-Pitaevskii equation with corrections. For these densities, the GP equation with the first correction is applicable [213] and equation 7.28 yields a modified version of equation 7.58:

$$-\nabla^2 f(\vec{r}) + \frac{1}{L_x^2}\bar{F}(\vec{r}/L_x)f(\vec{r}) + \frac{f^3(\vec{r})}{\epsilon_0^2}(1 + \frac{5}{4}a(n(\vec{R}_j)\sigma^3)^{1/2}f(\vec{r})) = kf(\vec{r}) \tag{8.55}$$

which must be solved with the boundary condition:

$$\vec{\nabla} f(\vec{r}) = 0 \quad \text{at} \quad r\cos(\theta) = L_x/2 \tag{8.56}$$

and self-consistently with the condition:

$$\mu = \frac{\lambda}{qAf_m^2}\int_A \left[|\vec{\nabla} f(\vec{r})|^2 + \frac{f^2(\vec{r})}{L_x^2}F(r/L_x) + \frac{f^4(\vec{r})}{\epsilon_0^2}\right] d^2\vec{r} \tag{8.57}$$

We look for an approximate solution using only one harmonic:

$$f(\vec{r}) \approx f_0(r) \tag{8.58}$$

$$F(\vec{r}) \approx F_0(r) \tag{8.59}$$

Under these conditions, equation 8.55 in units of ϵ_0 is:

$$-(rf_0')'/r + \frac{1}{L_x^2}F_0(r/L_x)f_0(r) + f_0^3(r)(1+5\bar{a}/4f_0(r)) = kf_0(r) \tag{8.60}$$

where $\bar{a} = a(n(\vec{R}_j)\sigma^3)^{1/2} = a(n/qf_{0m}^2)^{1/2}$ by using the normalization condition:

$$N = n^2(\vec{R}_j)\int f_0^2(r)d^2\vec{r} = n^2(\vec{R}_j)qAf_{0m}^2 \quad \Rightarrow \quad n^2(\vec{R}_j) = \frac{n}{qf_{0m}^2} \tag{8.61}$$

where n is the density of particles in units of σ: $n = N/L_xL_yL_z$.

The boundary condition is:

$$f_0'(\vec{r}) = 0 \quad \text{at} \quad r = L_x/2\epsilon_0 \equiv b \tag{8.62}$$

The value of b can be determined from the PIMC parameters as:

$$b = \sqrt{\frac{L_x^2}{4\epsilon_0^2}} = \sqrt{\frac{L_x^2 n^2(\vec{R}_j)g}{4\lambda}} = \sqrt{\frac{N}{L_z^*}}\sqrt{\frac{2\pi}{\sqrt{3}qf_{0m}^2}} \tag{8.63}$$

and must be determined self-consistently with the values of q and f_{0m} obtained from solving the equation, along with $\bar{a} = (n/qf_{0m}^2)^{1/2}$ and k:

$$k = \frac{\lambda}{qAf_{0m}^2}\int_A \left[(f_0'(r))^2 + \frac{f_0^2(r)}{L_x^2}F(r/L_x) + \frac{f_0^4(r)}{\epsilon_0^2}(1+5\bar{a}/4f_0(r))\right]d^2\vec{r} \tag{8.64}$$

Equation 8.60 is solved through a shooting finite difference method as explained in section 7.4.

The fit to the PIMC data is rather good within the approximations taken, as shown in figure 8.5. The value of k obtained from equation 8.64 is 0.83, which compares well with the value used in the finite difference solution: $k = 0.85$. The values of $\bar{a}$ and b were entirely self-consistent.

It is interesting to notice that the extra factor in the energy functional $(a\sigma^{\frac{3}{2}}f_0(r))$ has very little bearing on the vortex profile *per se* (fig. 8.6). Its main effect is on the maximum value f_{0m} (fig. 8.7) and therefore on the scale b through equation 8.63. The solutions of the unmodified Gross-Pitaevskii equation are still very close to those of its modified counterpart, as long as the right scale for b is known.

In fact, the Gross-Pitaevskii equation remains valid for much higher densities than expected if properly renormalized (by choosing an appropiate value of k) as seen in figures 8.8, 8.9 and 8.10. This explains why it yields good results even for superfluid ^{4}He, in spite of its high densities.

8.3 Conclusion

We have performed Path Integral Monte Carlo simulations for a vortex cell in a lattice with a hard sphere potential for the particles. The Path Integral Monte Carlo allowed us to access regimes where the Gross-Pitaevskii equation is, in principle, not applicable. The results obtained for densities which are within the reach of a modified version of the Gross-Pitaevskii equation are in agreement its predictions, within the approximations taken.

Surprisingly, the Gross-Pitaevskii equation is valid for much higher densities than expected if properly renormalized. This explains its validity for studying dense systems such as ^{4}He.

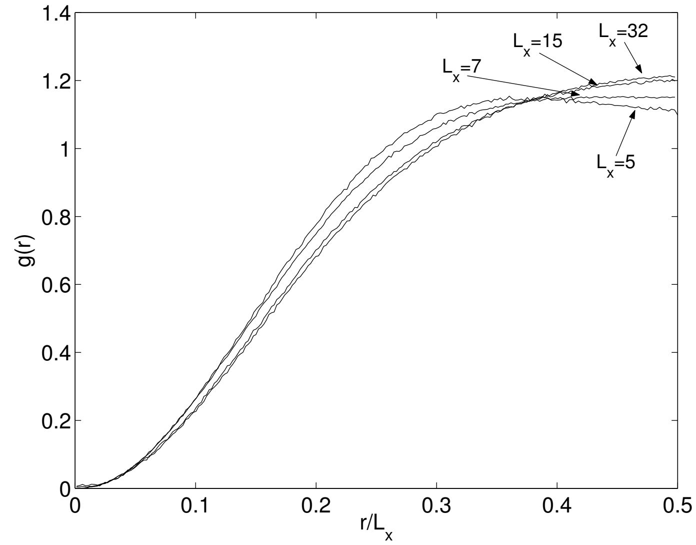


Figure 8.3: PIMC results for a single vortex cells in a lattice. Particle density profiles $g(r) \equiv f_0^2 r/nq$ as function of distance from the vortex center for different densities: $L_x = 32$ ($n = 1.41 \cdot 10^{-3}$), $L_x = 15$ ($n = 6.41 \cdot 10^{-3}$), $L_x = 7$ ($n = 2.94 \cdot 10^{-2}$), $L_x = 5$ ($n = 5.77 \cdot 10^{-2}$). For all $N = 25$. Lengths are in units of particle diameter σ.

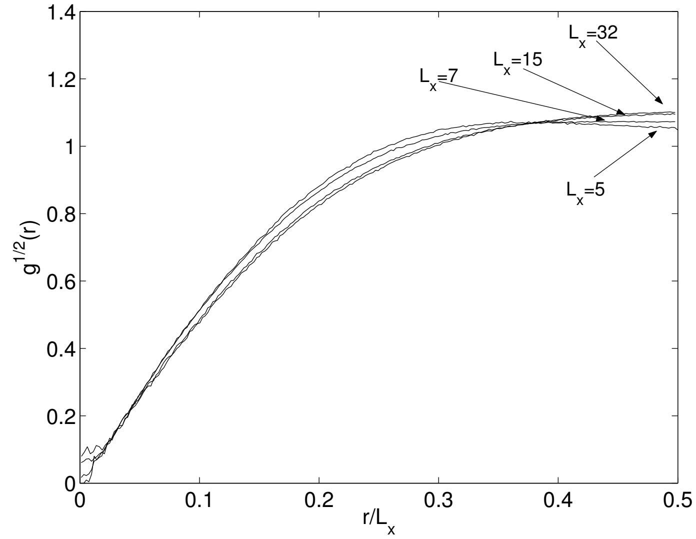


Figure 8.4: PIMC results for a single vortex cells in a lattice. Square root of density profiles $\sqrt{g(r)}$ for different cell lengths and, hence, rotation rates.

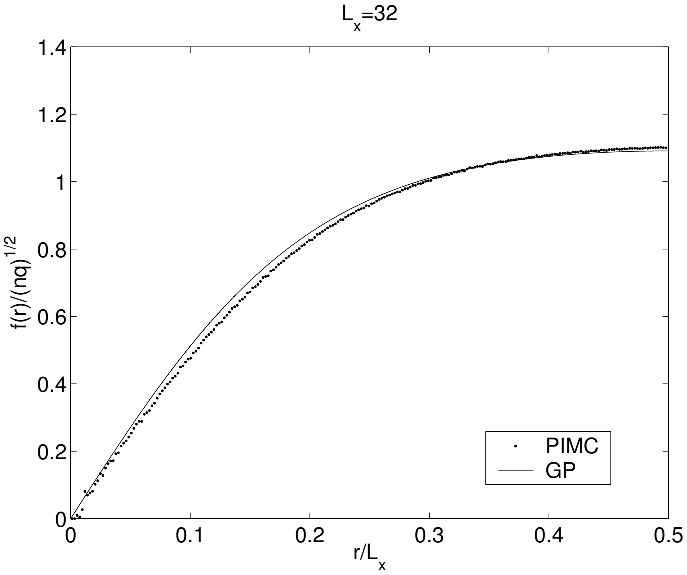


Figure 8.5: Self-consistent solution to equation 8.60 along with PIMC simulation for $L_x = 32$ ($n = 1.41 \cdot 10^{-3}$). The agreement is good, but not perfect. Using the full harmonics solution should give a more precise fit. The chemical potential used is $k = 0.85$ and equation 8.64 yields $k = 0.83$, which is in good agreement. Again, more precise results could be obtained with more harmonics. The other parameters in the self-consistent approach are: $5\bar{a}/4 = 0.31$, $b = 4.06$, $f_{0m} = 0.80$, $q = 0.84$, $f_{0m} = 0.804$, $c = 0.49986$.

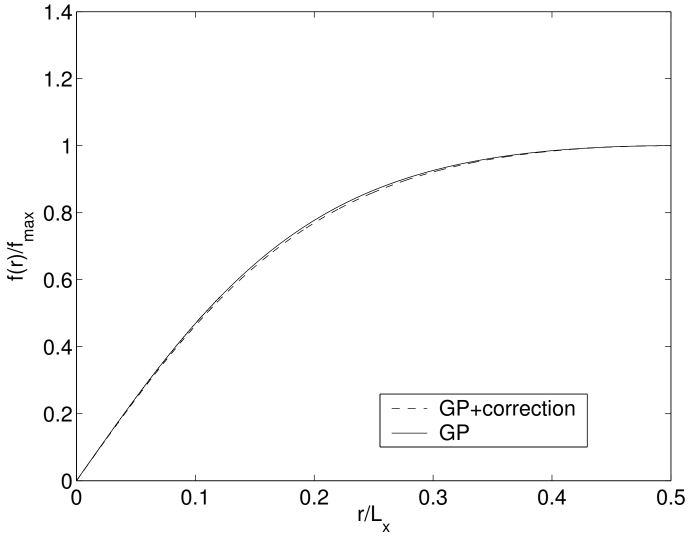


Figure 8.6: Scaled vortex profiles for the pure Gross-Pitaevskii equation and the modified version. For both $b = 4.06$ and for the modified version $5\bar{a}/4 = 0.31$ (same fit as in fig. 8.5). Notice that the profile is virtually the same. The only difference comes in the value of the maximum of $f_0(r)$ (f_{0m}) as can be seen in figure 8.7.

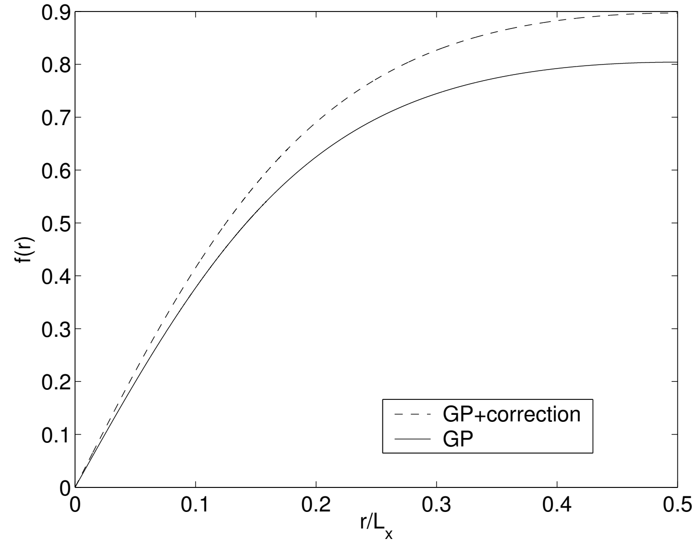


Figure 8.7: Same plot as figure 8.6, but figures have not been scaled.

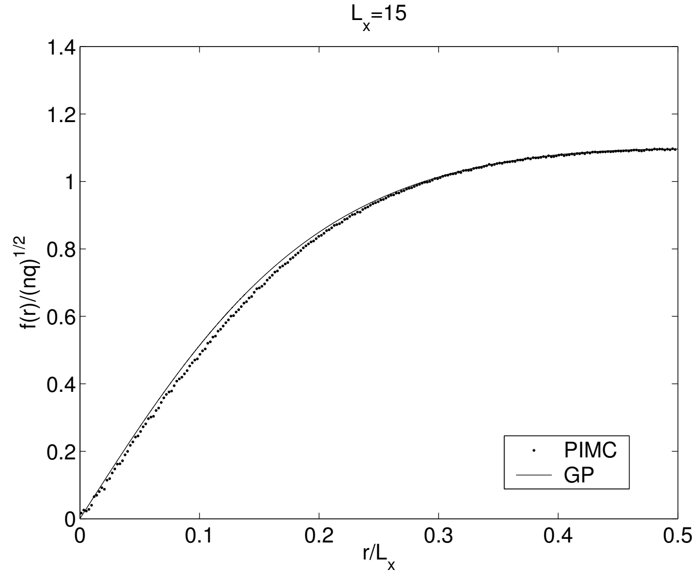


Figure 8.8: Fit to PIMC data by the unmodified GP equation for $L_x = 15$ ($n = 6.41 \cdot 10^{-3}$). Parameters are: $b = 3.71$, $q = 0.84$, $f_{0m} = 1.00$, $c = 0.68302$, $k = 1.06$. The fit is still quite accurate even though the unmodified GP equation should not be applicable.

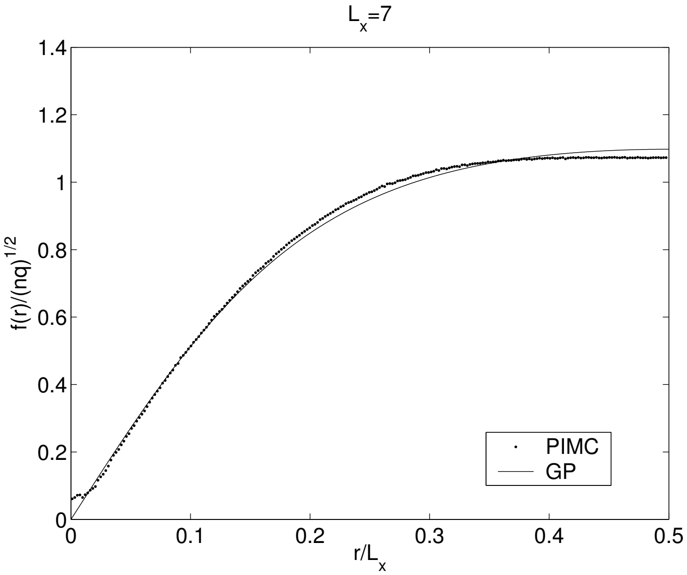


Figure 8.9: Fit to PIMC data by the unmodified GP equation for $L_x = 7$ ($n = 2.94 \cdot 10^{-2}$). Parameters are: $b = 3.44$, $q = 0.84$, $f_{0m} = 1.07$, $c = 0.78645$, $k = 1.42$. The unmodified equation still seems to hold.

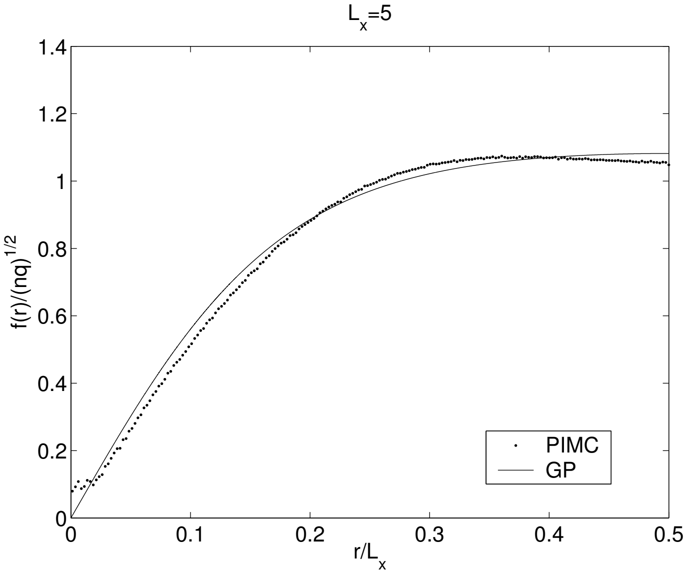


Figure 8.10: For $L_x = 0.5$ ($n = 5.77 \cdot 10^{-2}$), it is no possible to find a self consistent solution, but it is still possible to fit the profile with: $b = 2.85$, $f_{0m} = 1.55$, $q = 0.86$, $c = 1.5542$, $k = 2.48$. The values of f_{0m} and q are not consistent with the value of b through equation 8.63.

Chapter 9

Conclusions

In this dissertation I have presented work relating the statistical study of highly correlated systems in three fields of science: Ecology, microbial Ecology and Physics.

In the field of Ecology, I have proposed an explanation for one of the oldest and most widely observed patterns in ecology: the fractal Species Area Rule, which relates a given area and the number of species it harbors through a power law. I have shown that one should expect this functional form whenever the individuals for each species cluster (highly correlated positions) and the abundance distribution is similar to a lognormal distribution, but exhibiting higher rarity. Both characteristics are typically observed in ecosystems. The specific details of the clustering characteristics and the abundance distribution have little effect on the appearance of the power law SAR, which explains its widespread occurrence.

I have also shown that the exponent for the power law SAR is mainly determined by the abundance distribution, and that the most commonly observed values for this exponent are those more robust to differences in the clustering patterns and abundance distribution.

This work suggests a new and robust mechanism for the emergence of power laws that is not related to previous proposals.

My interest in Ecology has taken me to its study in a nascent subfield that has lately become much more accessible thanks to the irruption of molecular biology techniques: microbial ecology. I have formed part of a multidisciplinary team that studies the possible influence of microbes on the formation of travertine terraces in Yellowstone Hot Springs. Besides taking part in the field trips, I have used ecological methods to characterize the microbial biodiversity of our study site and found that, in spite of not having sampled more than 40% of the total biodiversity, the most abundant organisms seem to have been detected. I have also developed a new bootstrap method for extracting abundance information out of clone libraries and used it for this system. Its application singled out the most abundant bacteria and, hence, most likely to influence the formation of terraces. It also identified the abundance distributions for this microbial system, one of which turned out to be a power law, and suggested that the encrustment of microbes by the quick carbonate precipitation is not random. This, in turn, suggests a non-passive role of microorganisms in carbonate precipitation.

Convinced that many of the tools commonly used in Physics may have future applications outside of the realms of this field, I have undertaken research in a topic of current interest in Physics, dealing with highly correlated systems: rotating Bose-Einstein condensates.

Within this topic I have used finite difference techniques to solve the Gross-Pitaevskii equation to obtain the structure of a vortex in a lattice. Surprisingly, I have found that, in order to understand this structure, it is necessary to add a correction to the Gross-Pitaevskii equation which introduce a dependence on the particle scattering length. This result should provide a basis for perturbative techniques to yield a semi-analytical description of vortex structure in lattices formed in rotating BECs.

I have also used Path Integral Monte Carlo techniques to go beyond the Gross-Pitaevskii equation to study these vortices. Interestingly, the Gross-Pitaevskii equation seems to be valid for much higher densities than expected if properly renormalized. This explains its validity for studying dense systems such as ^{4}He.

References

[1] National Research Council, *Physics in a new era* (National Academic Press, Washington, 2001).

[2] R. Gallagher and T. Appenzeller, Science **284**, 79 (1999).

[3] C. R. Woese, Microbiology and molecular biology reviews **68**, 173 (2004).

[4] R. F. Service, Science **284**, 80 (1999).

[5] J. Gould, New York Times (2001).

[6] P. W. Anderson, Science **177**, 393 (1972).

[7] N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group* (Perseus Books, Reading Massachusetts, 1992).

[8] J. H. Brown and G. B. West, eds., *Scaling in Biology* (Oxford University Press, New York, 2000).

[9] P. S. Dodds, D. H. Rothman, and J. S. Weitz, J. Theor. Biol. **209**, 9 (2001).

[10] C. R. White and R. S. Seymour, PNAS USA **100**, 4046 (2003).

[11] R. B. Azevedo and A. M. Leroi, PNAS **98**, 5699 (2001).

[12] M. Rosenzweig, *Species Diversity in Space and Time* (Cambridge Univ. Press, Cambridge, 1995).

[13] H. J. Jensen, *Self-Organized criticality: emergent complex behavior in physical and biological systems* (Cambridge University Press, Cambrige, U.K., 1998).

[14] P. Bak, *How Nature works: The science of self-organized criticality* (Copernicus, New York, 1996).

[15] J. M. Carlson and J. Doyle, PNAS **99**, 2538 (2002).

[16] R. Pastor-Satorras and R. V. Sole, Physical Review E **64** (2001).

[17] J. Harte, A. Kinzig, and J. Green, Science **284**, 334 (1999).

[18] C. J. Krebs, *Ecology* (Harper & Row, New York,NY, 1985).

[19] D. T. Krohne, *General Ecology* (Brooks/Cole, Pacific Grove, CA, 2001).

[20] S. P. Hubbell, *The unified neutral theory of biodiversity and biogeography* (Princeton Univ. Press, Princeton,NJ, 2001).

[21] M. Tokeshi, Advances in ecological research **24**, 111 (1993).

[22] D. Rabinowitz, S. Cairns, and T. Dillon, in *Conservation Biology: The Science of Scarcity and Diversity* (Sinauer Associates, Sunderland,MA, 1986), pp. 182–204.

[23] S. Hubbell and R. Foster, in *Conservation Biology: The Science of Scarcity and Diversity* (Sinauer Associates, Sunderland,MA, 1986), pp. 205–231.

[24] K. J. Gaston, *Rarity* (Chapman and Hall, London, 1994).

[25] M. Soule, *Conservation Biology: The Science of Scarcity and Diversity* (Sinauer Associates, Sunderland,MA, 1986).

[26] R. May, in *Ecology and evolution of communities* (Harvard University Press, Cambridge, Massachussets, 1975), pp. 81–120.

[27] F. Preston, Ecology **29**, 254 (1948).

[28] R. H. MacArthur, PNAS USA **43**, 293 (1957).

[29] R. MacArthur, American Naturalist **94**, 25 (1960).

[30] A. McKane, D. Alonso, and R. V. Solé, Physical Review E **62**, 8466 (2000).

[31] S. Pigolotti, A. Flammini, and A. Maritan, Physical review E **70** (2004).

[32] J. Whitfield, Nature **417**, 480 (2002).

[33] A. Corbet, Proceedings of the Royal Society, London, Series A **16**, 101 (1941).

[34] R. Fisher, A. Corbet, and C. Williams, Journal of Animal Ecology **12**, 42 (1943).

[35] R. Kempton and L. Taylor, Journal of Animal Ecology **43**, 381 (1974).

[36] L. Taylor, R. Kempton, and I. Woiwood, Journal of Animal Ecology **45**, 255 (1976).

[37] C. Williams, *Patterns in the Balance of Nature and Related Problems in Quantitative Ecology* (Academic Press, London, 1964).

[38] F. Preston, Ecology **43**, 185 (1962).

[39] C. Williams, Journal of Animal Ecology **22**, 14 (1953).

[40] R. Whittaker, Science **147**, 250 (1965).

[41] R. Whittaker, Taxon **21**, 213 (1972).

[42] G. Batzli, Journal of Animal Ecology **38**, 531 (1969).

[43] R. Whittaker, *Communities and Ecosystems* (Macmillan, New York, 1970).

[44] G. Sugihara, American Naturalist **116**, 770 (1980).

[45] D. W. Gibbons, J. B. Reid, and R. A. Chapman, *The new atlas of breeding birds in Britain and Ireland* (T & A. D. Poyser, London, 1993).

[46] C. King, Ecology **45**, 716 (1964).

[47] M. S. Longuet-Higgins, Theor. Pop. Biol. **2**, 271 (1971).

[48] A. J. Kohn, Ecol. Monogr. **29**, 47 (1959).

[49] C. Goulden, in *Diversity and Stability in Ecological Systems* (U.S. Department of Commerce, Springfield, 1969), 22, pp. 96–102.

[50] E. S. J. Deevey, in *Diversity and Stability in Ecological Systems* (U.S. Department of Commerce, Springfield, 1969), 22, pp. 224–241.

[51] M. Tskukada, Trans. Conn. Acad. Arts Sci. **44**, 337 (1972).

[52] I. Motomura, Zoological Magazine, Tokyo **44**, 379 (1932).

[53] W. E. Kunin, Science **281**, 1513 (1998).

[54] F. He and K. J. Gaston, American Naturalist **156**, 553 (2000).

[55] W. E. Kunin, The American Naturalist **156**, 560 (2000).

[56] M. Cody, in *Ecology and evolution of communities*, edited by M. L. Cody and J. M. Diamond (Belknap Press of Harvard University Press, Cambridge, Mass, 1975), pp. 214–257.

[57] E. Goldstein, Evolution **29**, 750 (1975).

[58] M. L. Rosenzweig and E. A. Sandlin, Oikos **80**, 172 (1997).

[59] J. Lawrey, Bryologist **94**, 377 (1991).

[60] J. Leps, in *Spatial processes in plant communities*, edited by F. Krahulec, A. D. Q. Agnew, S. Agnew, and J. H. Willems (Academia, Prague, 1990), pp. 1–11.

[61] E. O. Wilson, American Naturalist **95**, 169 (1961).

[62] M. L. Rosenzweig, Journal of Mammalogy **73**, 715 (1992).

[63] R. O. Bierregard, T. E. Lovejoy, V. Kapos, A. Augusto dos Santos, and Hutchings R. W., Bioscience **42**, 859 (1992).

[64] J. M. Diamond, in *Theoretical ecology: principles and applications*, edited by R. M. May (Blackwell, Oxford, 1981), pp. 163–186.

[65] M. L. Rosenzweig, Science **284**, 276 (1999).

[66] M. Huston, Science **262**, 1676 (1993).

[67] F. Vuilleumier and D. Simberloff, Evolutionary Biology **12**, 235 (1980).

[68] S. J. Wright, American Naturalist **118**, 726 (1981).

[69] J. Green, J. Harte, and A. Ostling, Ecology letters **6**, 919 (2003).

[70] E. Tjorve, Journal of Biogeography **30**, 827 (2003).

[71] O. Arrhenius, Journal of ecology **9**, 95 (1921).

[72] H. A. Gleason, Ecology **3**, 158 (1922).

[73] H. A. Gleason, Ecology **6**, 66 (1925).

[74] J. Monod, Annales de l'Institut Pasteur **79**, 390 (1950).

[75] P. de Caprariis, R. Lindemann, and C. Collins, J. Intl. Assn. Math. Geol **8**, 575 (1976).

[76] H. K. Clench, Journal of Lepidopterists' Society **33**, 216 (1979).

[77] L. R. Holdridge, W. C. Grenke, W. H. Hatheway, T. Liang, and J. A. Tosi, *Forest environments in tropical life zones* (Pergamon Press, Oxford, 1971).

[78] R. I. Miller and R. G. Wieger, Ecology **70**, 16 (1989).

[79] D. A. Ratkowsky, *Handbook of nonlinear regression models* (Marcel Dekker, New York, 1990).

[80] D. A. Ratkowsky, *Nonlinear regression modelling: a unified approach* (Marcel Dekker, New York, 1983).

[81] M. Williamson, in *Analytical Biogeography*, edited by A. A. Myers and P. S. Giller (Chapman and Hall, New York, 1988), chap. 6, pp. 91–115.

[82] C. Rigby and J. Lawton, Journal of Biogeography **8**, 125 (1981).

[83] F. W. Preston, Ecology **41**, 785 (1960).

[84] D. Turcotte, *Fractals and chaos in geology and geophysics* (Cambridge University Press, Cambridge, 1992).

[85] B. J. Fox, in *Mediterranean-Type Ecosystems: the role of nutrients*, edited by F. J. Kruger, D. T. Mitchell, and J. U. M. Jarvis (Springer-Verlag, Berlin, 1983), pp. 473–489.

[86] R. H. MacArthur and E. O. Wilson, Evolution **17**, 373 (1963).

[87] H. C. Watson, *Remarks on the Geographical Distribution of British Plants* (Longman, Rees, Orme, Brown, Green and Longman, London, 1835).

[88] W. A. Leitner and M. L. Rosenzweig, Oikos **79**, 503 (1997).

[89] A. Ostling and J. Harte, Science **290**, 671a (2000).

[90] J. Harte and A. Kinzig, Oikos **80**, 417 (1997).

[91] J. Harte, S. McCarthy, K. Taylor, A. Kinzig, and M. L. Fischer, Oikos **86**, 45 (1999).

[92] A. Ostling and J. Harte, Oikos **103**, 218 (2003).

[93] A. Ostling, J. Harte, J. L. Green, and A. P. Kinzig, The American Naturalist **163**, 627 (2004).

[94] J. Lennon, W. E. Kunin, and S. Hartley, Oikos **97**, 378 (2002).

[95] J. Banavar, J. Green, J. Harte, and A. Maritan, Phys. Rev. Lett. **83**, 4212 (1999).

[96] R. D. Maddux and K. Athreya, Science **286**, 1647a (1999).

[97] R. D. Maddux, The American Naturalist **163**, 616 (2004).

[98] H. García Martín and N. Goldenfeld, Physical Review E **65**, 032901 (2002).

[99] A. L. Sizling and D. Storch, Ecology letters **7**, 60 (2004).

[100] M. Ney-Nifle and M. Mangel, Journal of Theoretical Biology **196**, 327 (1999).

[101] J. B. Plotkin, M. D. Potts, N. Leslie, N. Manokaran, J. LaFrankie, and P. S. Ashton, Journal of theoretical biology **207**, 81 (2000).

[102] M. Hoyle, Proceedings of the Royal Society of London B **271**, 1159 (2004).

[103] U. Brose, A. Ostling, K. Harrison, and N. D. Martinez, Nature **428**, 167 (2004).

[104] R. E. Ricklefs and E. Bermingham, The American Naturalist **163**, 227 (2004).

[105] L. Borda-de agua, S. P. Hubbell, and M. McAllister, The American Naturalist **159**, 138 (2002).

[106] B. D. Coleman, Mathematical Biosciences **54**, 191 (1981).

[107] M. R. Williams, Ecology **76**, 2607 (1995).

[108] R. Durrett and S. Levin, Journal of Thoretical Biology **179**, 119 (1996).

[109] R. V. Solé, D. Alonso, and A. McKane, Philosophical Transactions of the Royal Society of London, B **357**, 667 (2002).

[110] U. Bastolla, M. Lassig, S. C. Manrubia, and A. Valleriani, **212**, 11 (2001).

[111] R. Condit, S. Hubbell, and R. Foster, Journal of Tropical Ecology **12**, 231 (1996).

[112] G. Bell, Science **293**, 2413 (2001).

[113] H. Caswell, Ecological Monographs **46**, 327 (1976).

[114] G. Bell, M. J. Lechowicz, and M. J. Waterway, Journal of ecology **88**, 67 (2000).

[115] A. Kause, I. Saloniemi, J. P. Morin, E. Haukioja, S. Hanhimaki, and K. Ruohomaki, Evolution **55**, 1992 (2001).

[116] A. A. Agrawal and P. A. Van Zandt, Trends in ecology and evolution **17**, 62 (2002).

[117] I. Volkov, J. R. Banavar, S. P. Hubbell, and A. Maritan, Nature **424**, 1035 (2003).

[118] R. Condit, N. Pitman, E. G. J. Leigh, J. Chave, J. Terborgh, R. B. Foster, P. Nunez, S. Aguilar, R. Valencia, G. Villa, et al., Science **295**, 666 (2002).

[119] A. Ishikawa, *Pareto law and Pareto index in the income distribution of Japanese companies*, cond-mat/0409145 (2004).

[120] J. Harte, T. Blackburn, and A. Ostling, The American Naturalist **157**, 374 (2001).

[121] C. Cohen-Tannoudji, B. Diu, and F. Laloë, in *Quantum Mechanics* (John Wiley & Sons, New York, 1977), vol. 2, p. 1470.

[122] E. Castillo, *Extreme value theory in engineering* (Academic Press, London, 1988).

[123] E. Gumbel, *Statistics of extremes* (Columbia University Press, New York, 1958).

[124] R. Condit, P. S. Ashton, P. Baker, S. Bunyavejchewin, S. Gunatilleke, N. Gunatilleke, S. . P. Hubbell, R. B. Foster, A. Itoh, J. V. LaFrankie, et al., Science **288**, 1414 (2000).

[125] R. Fisher and L. Tippett, Proc. Cambridge Phil. Soc. **24**, 181 (1928).

[126] G. Strang, *Linear algebra and its applications* (Harcourt Brace Jovanovich, San Diego, 1988).

[127] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C* (Cambridge University Press, Cambrige, 1995).

[128] G. B. West, J. H. Brown, and B. J. Enquist, Science **276**, 122 (1997).

[129] P. Bak, C. Tang, and K. Wiesenfeld, Physical Review Letters **59**, 381 (1987).

[130] G. Bonheyo, B. Fouke, H. Garcia Martin, J. Veysey, N. Goldenfeld, and J. Frias-Lopez (2004), submitted to Applied and Environmental Microbiology.

[131] H. García Martín and N. Goldenfeld, in preparation.

[132] D. Mittlefehldt, Meteoritics **29**, 900 (1994).

[133] D. S. McKay, E. K. Gibson Jr, K. L. Thomas-Keprta, H. Vali, C. S. Romanek, S. J. Clemett, X. D. F. Chillier, C. R. Maechling, and R. N. Zare, Science **273**, 924 (1996).

[134] R. A. Kerr, Science **282**, 1398 (1998).

[135] NASA, http://www.jsc.nasa.gov/er/seh/marslife.html.

[136] W. Whitman, D. Coleman, and W. Wiebe, PNAS **95**, 6578 (1998).

[137] H. L. Ehrlich, Earth-Science Reviews **45**, 45 (1998).

[138] D. Newman and J. Banfield, Science **296**, 1071 (2002).

[139] J. F. Kasting and J. L. Siefert, Science **296**, 1066 (2002).

[140] M. Alexander, *Introduction to Soil Microbiology* (Wiley, New York, 1977).

[141] E. A. Paul and F. E. Clark, *Soil Microbiology and Biochemistry* (Academic Press, San Diego, CA, 1989).

[142] BBC, *The Secret Life of Caves*, http://www.bbc.co.uk/science/horizon/2003/lifeofcaves.shtml (2003).

[143] D. Northup and K. Lavoie, Geomicrobiology Journal **18**, 199 (2001).

[144] J. F. Luhr, *Smithsonian Earth* (DK publishing, New York, 2003).

[145] B. Fouke, J. Farmer, D. D. Marais, L. Pratt, N. Sturchio, P. Burns, and M. Discipulo, Journal of Sedimentary Research, Section A **70**, 565 (2000).

[146] R. Given and B. Wilkinson, Journal of Sedimentary Petrology **55**, 109 (1985).

[147] H. Chafetz, P. Rush, and N. Utech, Sedimentology **38**, 107 (1991).

[148] E. Burton, Chemical Geology **105**, 163 (1993).

[149] B. Fouke, G. Bonheyo, B. Sanzenbacher, and J. Frias-Lopez, Canadian Journal of Earth Sciences **40**, 1531 (2003).

[150] I. Friedman, Geochimica et cosmochimica acta **34**, 1303 (1970).

[151] J. M. Good and J. L. Pierce, *Interpreting the landscapes of Grand Teton and Yellowstone National Parks: Recent and Ongoing Geology* (Grand Teton Natural History Association, Grand Teton National Park, WY, 1998).

[152] A. Pentecost, Geological Magazine **127**, 159 (1990).

[153] W. Dreybrodt, L. Eisenlohr, B. Madry, and S. Ringer, Geochimica et Cosmochimica Acta **61**, 3897 (1997).

[154] S. Castanier, G. Le Metayer-Levrel, and J.-P. Perthuisot, Sedimentary Geology **126**, 9 (1999).

[155] F. Hammes and W. Verstraete, Re/Views in Environmental Science & Bio/Technology **1**, 3 (2002).

[156] S. Douglas and T. J. Beveridge, FEMS Microbiology Ecology **26**, 79 (1998).

[157] C. Rodriguez-Navarro, M. Rodriguez-Gallego, K. B. Chekroun, and M. T. Gonzalez-Munoz, Applied and environmental microbiology **69**, 2182 (2003).

[158] H. Felbeck and D. L. Distel, *The prokaryotes* (Springer-Verlag, New York, 1992), vol. 4, chap. Prokaryotic symbionts of marine invertebrates, pp. 3891–3906.

[159] R. I. Amann, W. Ludwig, and K.-H. Schleifer, FEMS Microbiology reviews **59**, 143 (1995).

[160] I. Head, J. Saunders, and R. Pickup, Microbial Ecology **35**, 1 (1998).

[161] M. T. Madigan, J. M. Martinko, and J. Parker, *Brock biology of microorganisms* (Prentice Hall, NJ, 1997).

[162] C. Woese, O. Kandler, and M. Wheelis, PNAS **87**, 4576 (1990).

[163] N. R. Pace, Science **276**, 734 (1997).

[164] G. J. Olsen, D. J. Lane, S. J. Giovannoni, and N. Pace, Annual Review of Microbiology **40**, 337 (1986).

[165] N. R. Pace, D. A. Stahl, D. J. Lane, and G. J. Olsen, Advances in Microbial Ecology **9**, 1 (1986).

[166] R. J. Redfield, Nature **2**, 634 (2001).

[167] Genbank, http://www.psc.edu/general/software/packages/genbank/.

[168] R. Rossello-Mora and R. Amann, FEMS Microbiology Reviews **25**, 39 (2001).

[169] C. Krebs, *Ecological methodology* (Harper and Row, New York,N.Y., 1989).

[170] M. Palmer, Ecology **71**, 1195 (1990).

[171] A. Chao and S.-M. Lee, Journal of American Statistical Association **87**, 210 (1992).

[172] S.-M. Lee and A. Chao, Biometrics **50**, 88 (1994).

[173] R. K. Colwell and J. Coddington, Philosophical Transactions of the Royal Society (Series B) **345**, 101 (1994).

[174] J. Raaijmakers, Biometrics **43**, 793 (1987).

[175] J. F. Heltshe and N. E. Forrester, Biometrics **39**, 1 (1983).

[176] E. P. Smith and G. Van Belle, Biometrics **40**, 119 (1984).

[177] J. F. Heltshe and N. Forrester, Biometrics **39**, 1073 (1983).

[178] J. J. Hellman and G. W. Fowler, Ecological applications **9**, 824 (1999).

[179] A. Baltanás, Oikos **65**, 484 (1992).

[180] J. B. Hughes, J. J. Hellman, T. H. Ricketts, and B. J. M. Bohannan, Applied and Environmental Microbiology **67**, 4399 (2001).

[181] K. A. Keating, J. F. Quinn, M. A. Ivie, and L. L. Ivie, Ecological applications **8**, 1239 (1998).

[182] K. L. J. Heck, G. Van Belle, and D. Simberloff, Ecology **56**, 1459 (1975).

[183] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap* (Chapman & Hall/CRC, Boca Raton Florida, 1993).

[184] A. C. Davison, *Bootstrap methods and their application* (Cambridge University Press, Cambridge ; New York, 1997).

[185] B. Smith and G. Van Belle, Biometrics **40**, 119 (1984).

[186] B. Efron, Annals of Statistics **7**, 1 (1979).

[187] B. Harris, Annals of Mathematical Statistics **30**, 521 (1959).

[188] M. J. Soberon and B. J. Llorente, Conservation Biology **7**, 480 (1993).

[189] F. v. Wintzingerode, U. B. Göbel, and E. Stackebrandt, FEMS Microbiology Reviews **21**, 213 (1997).

[190] G. Muyzer, E. C. de Waal, and G. A. Uitterlinden, Applied and Enviromental Microbiology **59**, 695 (1993).

[191] C. Ding and C. R. Cantor, Journal of Biochemistry and Molecular Biology **37**, 1 (2004).

[192] E. G. Zoetendal, C. T. Collier, S. Koike, R. I. Mackie, and H. R. Gaskins, Journal of Nutrition **134**, 465 (2004).

[193] J. Shao and D. Tu, *The Jackknife and Bootstrap* (Springer, New York, 1995).

[194] M. R. Chernick, *Bootstrap methods, a practitioner guide* (Wiley, New York, 1999).

[195] V. Pareto, *Le Cours d'Economie Politique* (Macmillan, London, 1897).

[196] W. Heisenberg, Z. Phys. **33**, 879 (1925).

[197] A. L. Fetter, Physical Review A **64**, 063608 (2001).

[198] U. R. Fischer and G. Baym, Physical Review Letters **90**, 140402 (2003).

[199] D. M. Ceperley, Reviews of Modern Physics **67**, 279 (1995).

[200] C. J. Pethick and H. Smith, *Bose-Einstein condensation in dilute gases* (Cambridge University Presss, Cambridge, 2002).

[201] S. N. Bose, Z. Phys. **26**, 178 (1924).

[202] A. Einstein, Sitzungberichte der Preussischen Akademie der Wissenschaften, Physikalisch-mathematische Klasse p. 261 (1924).

[203] F. London, Nature **141**, 643 (1938).

[204] M. H. Anderson, J. R. Ensher, M. R. Matthews, C. E. Wieman, and E. A. Cornell, Science **269**, 198 (1995).

[205] K. B. Davis, M.-O. Mewes, M. R. Andrews, N. J. van Druten, D. S. Durfee, D. M. Kurn, and W. Ketterle, Physical Review Letters **75**, 3969 (1995).

[206] C. C. Bradley, C. A. Sackett, J. J. Tollett, and R. G. Hulet, Physical Review Letters **75**, 1687 (1995).

[207] C. C. Bradley, C. A. Sackett, and R. G. Hulet, Physical Review Letters **78**, 985 (1997).

[208] V. Bretin, S. Stock, Y. Seurin, and J. Dalibard, **92**, 050403 (2004).

[209] T.-L. Ho, Physical Review Letters **87**, 060403 (2001).

[210] K. Huang and P. Tommasini, J. Res. Natl. Inst. Stand. Technol. **101**, 435 (1996).

[211] K. Huang and C. N. Yang, Physical Review **105**, 767 (1957).

[212] K. F. Riley, M. P. Hobson, and S. J. Bence, *Mathematical methods for physics and engineering* (Cambridge University Press, Cambridge, UK, 1998).

[213] S. Giorgini, J. Boronat, and J. Casulleras, Physical Review A **60**, 5129 (1999).

[214] N. M. Hugenholtz and D. Pines, Physical Review **116**, 489 (1959).

[215] P. Engels, I. Coddington, P. C. Haljan, and E. A. Cornell, Physical Review Letters **89**, 100403 (2002).

[216] R. P. Feynman, *Application of quantum mechanics to liquid helium* (Interscience, New York, 1955), vol. 1, chap. 2.

[217] A. A. Abrikosov, Soviet Physics JETP **5**, 1174 (1957).

[218] G. Baym, *Rapidly rotating Bose-Einstein condensates*, cond-mat/0408401 (2004).

[219] I. Coddington, P. C. Haljan, P. Engels, V. Schweikhard, S. Tung, and E. Cornell, cond-mat/0405240 (2004).

[220] J. Sinova, C. B. Hanna, and A. H. MacDonald, Physical Review Letters **89**, 030403 (2002).

[221] G. Baym, Physical Review A **69**, 043618 (2004).

[222] S. A. Gifford and G. Baym, cond-mat/0405182 (2000).

[223] K. Kasamatsu, M. Tsubota, and M. Ueda, Physical Review A **66**, 053606 (2002).

[224] G. M. Kavoulakis and G. Baym, New Journal of Physics **5**, 51.1 (2003).

[225] A. D. Jackson and G. M. Kavoulakis, Physical Review A **70**, 023601 (2004).

[226] A. D. Jackson, G. M. Kavoulakis, and E. Lundh, Physical Review A **69**, 053619 (2004).

[227] E. Lundh, Physical Review A **65**, 043604 (2002).

[228] J. B. Marion and S. T. Thornton, *Classical Dynamics* (Harcourt Brace Jovanovich College Publishers, Orlando, FL, 1988).

[229] G. Baym and E. Chandler, J. Low temp. Phys. **50**, 57 (1983).

[230] G. Baym and C. J. Pethick, Physical Review A **69**, 043619 (2004).

[231] G. I. Barenblatt, *Scaling, self-similarity, and intermediate asymptotics* (Cambridge University Press, Cambridge, 1996).

[232] R. L. Burden and J. D. Faires, *Numerical Analysis* (Brooks/Cole Publishing Company, Pacific Grove, CA, 1997).

[233] L.-Y. Chen, N. Goldenfeld, and Y. Oono, Physical Review Letters **73**, 1311 (1994).

[234] R. P. Feynman, *Statistical Mechanics* (Addison-Wesley, 1990).

[235] H. F. Trotter, Proc. Am. Math. Soc. **10**, 545 (1959).

[236] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Journal of chemical physics **21**, 1087 (1953).

[237] E. L. Pollock and D. M. Ceperley, Physical Review B **36**, 8343 (1987).

[238] G. Ortiz and D. M. Ceperley, Physical Review Letters **75**, 4642 (1995).

[239] G. Ortiz, Physical Review Letters **71**, 2777 (1995).

[240] M. Unge, Master's thesis, University of Manchester (2001).

[241] N. Gronbech-Jensen, Computer Physics Communications **119**, 115 (1999).

Vita

Héctor García Martín was born on March 23, 1976 in Bilbao, Spain (Basque Country). He graduated with honors from high school (Matricula de Honor global) and enrolled in the Physics program at the University of the Basque Country (UPV/EHU) in 1993. He specialized in condensed matter Physics and obtained his degree of "Licenciado en Ciencias Fisicas" in June 1999. Craving to travel a bit and see what the world had to offer, he won a BBK scholarship to take part in a TASSEP exchange program at the University of Texas at Austin, where he did his last year of undergrad studies the academic year 1998-1999. In August 1999, Héctor entered the graduate program in the Department of Physics at the University of Illinois at Urbana-Champaign. In 2000, he obtained the "Excellence in Teaching" award and was included in the "Incomplete List of Teachers Ranked as Excellent". That same year, he joined Nigel Goldenfeld's research group and in 2002 he obtained the "Renato Bobone Award to the Outstanding European Graduate Student in Physics".