2014

Individual-Based Modeling and Nonlinear Analysis for Complex Systems with Application to Theoretical Ecology

Abbas Ghadri Golestani
University of Windsor

Individual-Based Modeling and Nonlinear Analysis for Complex Systems with Application to Theoretical Ecology

By

Abbas Ghadri Golestani

A Dissertation
submitted to the Faculty of Graduate Studies
through the Department of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy
at the University of Windsor

Windsor, Ontario, Canada
2014

Individual-Based Modeling and Nonlinear Analysis for Complex Systems with Application to Theoretical Ecology

by

Abbas Ghadri Golestani

APPROVED BY:

_____

Dr. J. Liu, External Examiner
Hong Kong Baptist University

_____

Dr. D. Heath
Great Lakes Institute for Environmental Research

_____

Dr. B. Boufama
School of Computer Science

_____

Dr. Z. Kobti
School of Computer Science

_____

Dr. R. Gras, Advisor
School of Computer Science

December 5, 2014

Declaration of Co-Authorship / Previous Publication

I. Co-Authorship Declaration

I hereby declare that this dissertation incorporates material that is result of joint research with Dr. Robin Gras, my supervisor. This dissertation also incorporates the outcome of a joint research undertaken in collaboration with Dr. M. Cristescu and Dr. A.P Hendy under the supervision of professor Robin Gras. The collaboration is covered in Chapter 4 of the dissertation. In all cases, the key ideas, primary contributions, experimental designs, data analysis and interpretation, were performed by the author, and the contribution of co-author was primarily through the provision of required background biological information. In Chapter 3, Ms. Khater also contributed in explaining the materials.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my dissertation, and have obtained written permission from each of the co-author(s) to include the above material(s) in my dissertation.

I certify that, with the above qualification, this dissertation, and the research to which it refers, is the product of my own work.

II. Declaration of Previous Publication

This dissertation includes 13 original papers and patents that have been previously published/submitted for publication in peer reviewed journals, as follows:

Dissertation Chapter	Publication title/full citation	Publication status*
Chapter 2	R. Gras, A. Golestani, M. Hosseini, M. Khater, Y.M. Farahani, M. Mashayekhi, M. Sina, A. Sajadi, E. Salehi and R. Scott, EcoSim: an individual-based platform for studying evolution, European Conference on Artificial Life, pp 284-286, 2011.	Published
Chapter 2	M. Mashayekhi, A. Golestani, Y.M. Farahani, R. Gras, An enhanced artificial ecosystem: Investigating emergence of ecological niches, International Conference on the Simulation and Synthesis of Living Systems (ALIFE 14), pp 693-700, 2014.	Published
Chapter 4	A. Golestani, R. Gras, and M. Cristescu, Speciation with gene flow in a heterogeneous virtual world: can physical obstacles accelerate speciation?, *Proceedings of the Royal Society B: Biological Sciences*, vol. 279 no. 1740, pp 3055-3064, 2012.	Published
Chapter 4	A. Golestani, R. Gras, A New Species Abundance Distribution Model Based on Model Combination, *International Journal of Biostatistics (IJB)*, 9(1): 1–16, 2013.	Published
Chapter 4	A. Golestani, R. Gras, Using Machine Learning Techniques for Identifying Important Characteristics to Predict Changes in Species Richness in EcoSim, an Individual-Based Ecosystem Simulation", International Conference on Machine Learning and Data Analysis (ICMLDA'12), vol 1, pp 465-470, San Francisco, 2012.	Published
Chapter 4	R. Gras, A. Golestani, M. Cristescu, and A.P. Hendry, Speciation without pre-defined fitness functions, *PLOS ONE,* 2014.	Submitted
Chapter 5	A. Golestani, R. Gras, Regularity Analysis of an individual-based Ecosystem Simulation, journal of Chaos: An *Interdisciplinary journal of Nonlinear Science, CHAOS* 20, 043120. pp 1-13, 2010.	Published
Chapter 5	A. Golestani, R. Gras, Identifying Origin of Self-Similarity in EcoSim, an Individual-Based Ecosystem Simulation, Using Wavelet-based Multifractal Analysis, International Conference on Modeling, Simulation and Control (ICMSC'12), vol 2, pp 1275-1282, San Francisco, 2012.	Published
Chapter 5	Y.M. Farahani, A. Golestani, R. Gras, Complexity and Chaos Analysis of a Predator-Prey Ecosystem Simulation, COGNITIVE, ISBN: 978-1-61208-001-7, pp: 52-59, 2010.	Published
Chapter 5	A. Golestani, R. Gras, "Multifractal Phenomena in EcoSim, a Large Scale Individual-Based Ecosystem Simulation", ICAI (International Conference on	Published

	Artificial Intelligence), Las Vegas, USA, pp 991-999, 2011.	
Chapter 6	A. Golestani, R.Gras, Can We Predict the Unpredictable?, Scientific Reports 4, 6834; DOI:10.1038/srep068342014.	Published
Chapter 6	A. Golestani, R. Gras, System and Process for Predictive Chaos Analysis , Filed under US Patent Application Serial Number 61/882863.	In press
Chapter 6	A. Golestani, R. Gras, Method And Apparatus For Prediction Of Epileptic Seizures, Filed under US Patent Application Serial Number 62/042535.	In press

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my dissertation. I certify that the above material describes work completed during my registration as graduate student at the University of Windsor.

I declare that, to the best of my knowledge, my dissertation does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my dissertation, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my dissertation.

I declare that this is a true copy of my dissertation, including any final revisions, as approved by my dissertation committee and the Graduate Studies office, and that this dissertation has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

One approach to understanding the behaviour of complex systems is individual-based modeling, which provides a bottom-up approach allowing for the consideration of the traits and behaviour of individual organisms. Ecosystem models aim to characterize the major dynamics of ecosystems, in order to synthesize the understanding of such systems and to allow predictions of their behaviour. Moreover, ecosystem simulations have the potential to help scientists address theoretical questions as well as helping with ecological resource management. Because in reality biologists do not have much data regarding variations in ecosystems over long periods of time, using the results of ecological computer simulation for making reasonable predictions can help biologists to better understand the long-term behaviour of ecosystems. Different versions of ecosystem simulations have been developed to investigate several questions in ecology such as how speciation proceeds in the absence of experimenter-defined functions. I have investigated some of these questions relying on complex interactions between the many individuals involved in the system, as well as long-term evolutionary patterns and processes such as speciation and macroevolution.

Most scientists now believe that natural phenomena have to be looking as a chaotic system. In the past few years, chaos analysis techniques have gained increasing attention over a variety of applications. I have analyzed results of complex models to see whether chaotic behaviour can emerge, since any attempt to model a realistic system needs to have the capacity to generate patterns as complex as the ones that are observed in real systems. To further understand the complex behaviour of real systems, a new algorithm for long-term prediction of time series behaviour is also proposed based on chaos analysis. We evaluated the performance of our new method with respect to the prediction of the Dow-Jones industrial index time series, epileptic seizure and global temperature anomaly.

DEDICATION

To my parents for the courage

To my wife for the patience

To my friends for the laughter

ACKNOWLEDGEMENTS

I would like to gratefully thank Dr. Robin Gras my supervisor, for giving me the opportunity and continued guidance to apply and extend my knowledge in a cross disciplinary field, working at the forefront of computer science. I would like to thank my committee members Dr. Liu, Dr. Heath, Dr. Boufama and Dr. Kobti for accepting to allocate part of their valuable time to evaluate my research.

This work was made possible by the facilities of Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada.

Contents

List of Tables

List of Figures

Chapter 1

1. Introduction

The world presented by Darwin demonstrates that the existence of living creatures can basically be described in terms of a small number of fundamental processes [1]. These arguments suggest that it might be possible to create an artificial world that exhibits these simple processes on a computer. Some would argue that life is a process, which is fundamentally associated with the physical world. For example, they might argue that some of the processes associated with living organisms, such as metabolism, could not be simulated on a computer [2]. On the other hand, others would argue that life is fundamentally a process (or a set of processes) and is quite independent of its specific implementation; they would be quite happy to accept that artificial life could evolve on a computer. We believe an attempt to create artificial life is worth pursuing for a number of reasons. The approach of building an evolutionary system is very different from the traditional methods taken in theoretical biology of analyzing evolution by tracking changes in population-level measures using simple mathematical models [3]. Different approaches allow one to look at a system from different angles; one approach might suggest answers whose significance is not apparent from another approach [4]. In this way, artificial life approaches can complement the more traditional approaches of theoretical biology, and lead us to ask different sorts of questions about evolution and life [3].

A complex ecosystem is composed of organisms living in a given habitat. There are biotic components of the ecosystem such as plants or animals, while the geographical conditions are part of the abiotic components. In an ecosystem, the biotic components and the abiotic ones establish a set of relationships with each other that characterize the ecosystem itself and bring it to a balanced state [5]. One approach for understanding the behaviour of complex ecosystems is Ecological modeling [6]–[8]. This is still a growing field at the crossroad between theoretical ecology, mathematics and computer science [9]. Ecosystem models aim to characterize the major dynamics of ecosystems, in order to synthesize the understanding of such systems, and to allow predictions of their behaviour. Because natural ecosystems are very complex (in terms of number of species and of ecological interactions), ecosystem models typically simplify the systems they are representing to a limited number of components. One approach for understanding the behaviour of complex ecosystems is individual-based modeling (IBM), which provides a bottom-up approach allowing for the consideration of the traits and behaviour of individual organisms.

Individual-based models are simulations aiming to study the global consequences of local interactions of members of a population (see chapter 2). One of the main interests of IBM ecosystem simulations is that they both offer a global view of the evolution of the system, which is difficult to observe in nature and a detailed view, which cannot be considered by mathematical modeling (see section 2.1). Although, how much these models are realistic is under question. How can we measure the similarity between models and real ecosystems? Is there a measure for quantifying the complexity of real ecosystems and models? Can mathematical equations accurately simulate real systems?

Sir Isaac Newton was a pioneer in modeling of the motion of physical systems with mathematical equations. It was necessary to have calculus along the way, since basic equations of motion comprise velocities and accelerations, are derivatives of position. His major achievement was his finding that the motion of the planets and moons of the solar system emerged from a fundamental source: the gravitational force between bodies [10], [11]. Later generations of researchers expanded the method of using differential equations to explain how physical systems evolve. But the method had a limitation. While the differential equations were sufficient to characterize the behaviour, it was mostly difficult to detect what that behaviour would be [11]. When solutions could be discovered, they described very regular motion. Scientists comprehended the sciences from textbooks filled with examples of differential equations with usual topics. If the solutions stayed in a confined area of space, they settled down to either (1) a steady state, mostly because of energy loss by friction, or (2) an oscillation that was either periodic or quasiperiodic. Around 1975, after three centuries of study, many scientists around the world suddenly became aware that there is a third kind of motion, a type (3) motion, that we now call "chaos". The new motion is erratic, but not simply quasiperiodic with a large number of periods, and not necessarily due to a large number of interacting particles. This type of behaviour can emerge from very simple systems (see chapter 3) [10].

Most scientists believe that chaotic behaviour can be observed in many natural systems [10], [12]–[22]. It is therefore interesting to study natural phenomena considering them as chaotic systems [23]. In the past few years, chaos analysis techniques have gained increasing attention in medical signal and image processing. For example, analyses of encephalographic data and other biosignals are among its applications [24], [25]. It has been shown that the evaluation of chaoticity (level of chaos) is an important issue in several such applications. There are many publications that justify that, without chaoticity, biological systems might be unable to get

discriminated between different stages and thereby different modes of operation [26], [27] (for example epileptic seizures can be detected by using measure of chaoticity [28]).

The expectations of scientists and mathematicians are different. Mathematicians prove theorems while scientists seek for pragmatic models fitting and explaining their data and not for a mathematical proof for their model. The first studies indicating chaotic behaviour in computer studies of very simple models were unpleasant to both folks. The mathematicians feared that nothing was proved so nothing was learned. Scientists pointed out that models without physical quantities like mass, charge, energy, or acceleration could not be linked to physical studies. But further studies led to a change in point of view. Mathematicians realized that these studies could lead to new ideas that slowly led to new theorems. Scientists found that computer studies of much more complicated models yielded behaviours similar to those of the simplistic models, and that perhaps the simpler models captured the key phenomena [11].

Modeling (simulation) is a well-known approach for studying natural phenomena. Our research focused on the modeling of ecosystem alongside with the chaos analysis of the resulting complex models. By modeling organisms with varying characteristics (such as age, mating preferences, and role in the ecosystem), the properties of the system can emerge from their complex interactions possibly avoiding the issue of having pre-included into the model the very things that one would like to study. Ecosystem simulations, for example, can help scientists to understand theoretical questions and could have some significance in ecological resource management. Because in reality biologists do not have much data regarding variation of ecosystems over long periods of time, using the results of a logical simulation for making reasonable predictions can help biologists to better understand long-term behaviour of ecosystems.

Different versions of a large evolving ecosystem simulation (EcoSim [29]) have been developed to investigate several biological questions. We investigated questions relying on complex interactions between the multiple individuals involved in the system, as well as long-term evolutionary patterns and processes such as speciation and macroevolution. For instance, we investigated how small, randomly distributed physical obstacles influence the distribution of populations and species [30]. We also investigated forces influencing speciation by considering the formation of genetic clusters and the level of hybridization between them [31]. The results of EcoSim have been used to predict variation in the number of species in EcoSim by applying machine learning techniques. Identifying important features for species richness prediction and the relationship between them could be beneficial for future conservation studies.

In order to show that these simulations can be considered as reasonable models for simple real ecosystems, we analyzed them to see whether complex chaotic behaviour can emerge. This is because any attempt to model a realistic system need to have the capacity to generate patterns as complex as the ones that are observed in real systems. We analyzed the results of EcoSim [29], to evaluate its complexity [32]. We also examined multifractal patterns in the results of EcoSim, for example time series corresponding to the variation of the number of prey and predator individuals and individuals' positions [33]. We wanted to investigate whether the data generated by EcoSim present the same kind of multifractal properties as the ones observed in real ecosystems. We also analyzed different parameters of the simulation to detect which ones cause the multifractal behaviour since one important issue for ecologists is to understand where these structures come from [34].

Analysis and prediction of complex systems (coming from either models or real systems) is always a serious challenge for scientists. Using chaos analysis is a fine answer to address this challenge since it reveals simple and logical principles behind complex behaviour. With chaos analysis, dealing with some of the most challenging complex system analysis problems, intractable using traditional mathematical or physical approaches, seems to be realistic. Analysis of complex data and the leveraging of that analysis towards making reasonable predictions is an important goal. For this reason, a new algorithm for long-term prediction of time series' behaviour is also proposed based on measures of level of chaos [35]. The new method has been used to address different open problems like prediction of epileptic seizure and long-term prediction of financial market' trends (couple of months in advance). What follows is a summary of the important contributions made by this dissertation:

- First, two important theoretical questions in ecology that are hard to study in nature have been investigated using computer simulations. We investigated how small, randomly distributed physical obstacles influence the distribution of populations and species, the level of population connectivity (e.g., gene flow) as well as the mode and tempo of speciation in a virtual ecosystem (see section 4.1). We modified EcoSim to examine complex predator-prey dynamics and coevolution in spatially homogenous and heterogeneous worlds. Further we investigated if and how speciation proceeds in the absence of experimenter-defined functions (see section 4.2). To address this key knowledge gap, we used EcoSim to explore speciation in the absence of pre-defined fitness functions. In our model, speciation results from emergent properties arising from

interactions between individuals in spatial landscapes where abiotic parameters are initially invariant.

- Second, we proposed a new species abundance distribution model based on an ensemble of base models, combined using a genetic algorithm (see section 4.3). Species abundance distribution is a component of biodiversity and refers to how common or rare a species is relative to other species in a defined location or community. It is one of the main characteristics investigated in ecological studies.

- Third, we predicted changes in the number of species in EcoSim using several important features by applying machine learning techniques, such as using different feature selection algorithms and decision trees (see section 4.4).

- Fourth, we analyzed the output of EcoSim, such as population time series and spatial distribution of individuals (see sections 5.1 and 5.2). These analyses showed that not only the overall behaviours of patterns generated by EcoSim is deterministic but also EcoSim is capable of generating patterns as complex as patterns that have been observed in natural phenomena.

- Finally an algorithm for time series prediction has been developed leading to highly accurate long-term predictions of nonlinear time series (see chapter 6). We evaluated its performance with respect to the prediction of the long-term behaviour of the Dow-Jones Industrial Index (DJIA) time series, EEG time series for epileptic seizure prediction and prediction of global temperature anomalies.

The Outline of this dissertation is as follows:

- Chapter 2 reviews existing literature regarding evolutionary systems and the use of individual-based modeling (IBM) in ecology, with a particular focus on ALife evolutionary simulations

- Chapter 3 reviews the basic concepts in chaos theory as well as useful methods for analyzing chaotic systems.

- Chapter 4 presents the results obtained by studying the importance of different parameters on speciation in EcoSim. This chapter presents a new species abundance

distribution model along with presenting the results of machine learning techniques applied to the outputs of EcoSim for species richness prediction

- Chapter 5 presents the results of nonlinear analysis on various outputs of EcoSim.
- Chapter 6 presents a new method for time series predictions with three different applications.

Chapter 2

2. Review of Ecosystem Modeling

Modeling has turned into a vital apparatus in the investigation of ecological systems. Powerful computational resources and graphical software packages have overcome a great part of the drudgery of creating models with a programming language and opened new perspectives of model development. Models give a chance to investigate ideas regarding ecological systems that it may not be conceivable to field-test for financial or logistical reasons. The procedure of forming an ecological model is very beneficial for organizing one's thinking, shedding light on concealed assumptions, and recognizing information needs [36]. Ecologists employ models for different purposes, including explaining existing data, formulating predictions, and guiding research [37].

Ecological models can guide research in various ways. Sensitivity analysis of a model can uncover which procedures and coefficients have the most impact on observed results, and along these lines, proposes how to prioritize sampling efforts [38], [39]. Most importantly, models make an interpretation of ecological hypotheses into predictions that could be assessed in light of existing or new information. The type of models and details will rely on the system examined, the questions asked, and the data available. Models can rapidly become complex and clear problem definition is crucial to keeping the model focused [36]. Once the general type of ecological model has been chosen, the ecologist must determine the appropriate level of abstraction for the model.

Ecologists have an interest for deeper understanding of concepts such as: the evolutionary process, the emergence of species, the emergence of learning capacities, the usage of energetic resources of individuals in different stress conditions, and the effect of climatic variations or catastrophic events in evolution. All of these studies have significance in ecological resource management, epidemiology, or in studying the impact of human behaviour on ecosystems. Ecosystem simulations can help scientists understand theoretical questions.

Many models have been designed to investigate biological hypotheses. However, a common feature of most of these models is the reliance on a pre-defined fitness function, which makes the system function as an optimization process with bounded convergence properties. Different ecological models based on a pre-defined fitness function will be discussed in this chapter.

2.1. Mathematical modeling and IBM approaches with pre-defined fitness function

Artificial evolving systems with pre-defined fitness functions, or fitness landscapes, have been well studied. The fitness function evaluates how good a potential solution is relative to other potential solutions. The fitness function is used in a process of selection to choose which potential solutions will continue on to the next generation, and which will die out. In the 1960s, John Holland introduced Genetic Algorithm (GA) [40] as a tool to model the adaptation of organisms to their environment and to develop ways in which complex evolutionary processes can be investigated in computer systems [41]. Since then, many empirical and theoretical studies have been undertaken to determine the behaviour of such artificially evolving systems [42].

To model evolutionary systems, fitness is assigned to individuals based on some genetic or phenotypic properties associated with the individual. In general, the fitness function is an *a priori* feature of the model leading to what Packard called extrinsic adaptation [43] and Channon et al. called artificial selection [44]. The dynamic of such a system (exploring the set of all possible genomic populations) is well understood, leading to convergence of the population towards the peaks of the fitness function [42]. Systems such as Genetic Algorithms with fitness sharing (also called niching) allow the modeling of competition for resources and can generate population distributions not centered on the peaks [45], [46]. The behaviours of such approaches are also well known showing convergence, with possible cycles, towards the peaks of the fitness function modified by the fitness sharing process. Problems in which the fitness function varies through time have also been studied, showing that the population distribution converges towards multiple successive points each one linked to the peaks of the new version of the fitness function [47]. In addition, the effect of linkage between loci (also called 'dependency') has been extensively studied (see the Chen review [48]), which leads to a new type of genetic algorithm called Estimation Distribution Algorithm [49] that also considers the hierarchy of linkages [50]. More complex systems based on artificial selection were designed and discussed in [44], [51].

All of these systems are optimization processes, meaning that the fate of the system is directly determined by its pre-defined fitness function with the convergence behaviour described above. Packard was the first to design a simple model not governed by extrinsic adaptation (i.e not using a pre-defined fitness function), which demonstrated that the evolutionary dynamic could be an emerging property of an intrinsic model [43], [52]. Unfortunately, their organisms and systems were too simple for new species to emerge (for a discussion about emergence see section 3). Moreover, the more complex systems, such as Geb [51] and Polyworld [53], rely heavily on

learning very complex neural networks to model behaviour, and the associated computational requirement constrains the total population to a few hundred individuals for a few hundred generations. These models are thus inefficient in dealing with processes, such as speciation, that can span large ecological and long temporal scales.

To the best of our knowledge, all pervious ecological modeling studies rely on one form or another of an *a priori* fitness function. Obviously, all purely mathematical models are also based on pre-defined fitness functions, and so here we focus our discussion on individual-based models (IBMs), providing a few representative examples. In these studies, the pre-defined fitness function is generally defined as a fitness landscape, which is a classical representation also used to study the properties of Genetic Algorithms.

- Gavrilets proposed a simple model in which L loci are each assigned a binary fitness value *fit* or *unfit*, later extended to a continuous range of fitness [54]. As these fitness values are initially set and do not evolve during the simulation, the fitness landscape is predefined.

- Gavrilets [55] used a bidimensional IBM approach with two types of cells with different resources. The fitness of an individual is modeled by two Gaussian functions with pre-defined parameters corresponding to the two types of cells, which generate a fixed multimodal fitness landscape.

- Dieckmann [56], Kirkpatrick [57] and Bolnick [58] used a Gaussian fitness function coupled with a Gaussian genomic competition function similar to the fitness sharing processing analyzed in Genetic Algorithm with niching. Both functions used fixed pre-defined parameters generating a fixed landscape.

- Drossel [59] and Doebeli [60] associated an IBM with Lotka-Volterra competition equations that predefined phenotypic fitness. Doebeli [61] designed a bidimensional IBM in which the fitness of individuals is governed by a Lotka-Volterra model with a fixed fitness landscape, composed of a succession of peaks, defined by a linear gradient of resources and associated with fitness sharing based on genomic similarity.

- Higashi [62] proposed a model where an *a priori* fitness function is computed on L additive loci.

- Takimoto [63] proposed a model with an *a priori* fitness function computed on a deterministic combination of the alleles of three loci.

- Gravilets (21, 22) and Thibert-Plante [64]–[67] defined IBMS based on pre-defined multimodal Gaussian distribution of resources associated with a normalized competition function. Even though the resulting dynamic of such probabilistic complex systems can lead to non-stationary and non-converged population distributions, their overall behaviours are pre-determined and can be studied as in Débarre [68].

- The approaches based on individual-based evolutionary game models (IBEG models; see the review of Allen [69]) integrate complex competition models but are still based on a pre-defined fitness function because of the pre-defined pay-off function.

Relying on pre-defined fitness functions, all of these methods correspond to one form or another of a genetic algorithm and they perform an optimization process with predictive convergence properties. We suggest that, for this reason, previous studies did not allow the emergence of intrinsic adaptations in the sense of Packard. In the following, two well-known ecological models with pre-defined fitness function are explained more into details.

2.1.1. Tierra

Tierra [70] is a complex simulation designed by Thomas S. Ray in 1990 where computer programs contest for central processing unit (CPU) time and access to the memory. The computer programs in Tierra are recognized to be evolvable and can mutate, self-replicate and recombine. Tierra has the ability to examine the basic procedures of evolutionary and ecological dynamics. A significant variation between Tierra and other models of evolutionary computation, such as genetic algorithms, is that it is said not to have fitness function built into the model. In these kind of models, there is the notion of a function being "optimized": there is simply survival and death. If there is no explicit fitness function embedded into the model, this may allow for more "open-ended" evolution, where the dynamics between evolutionary and ecological procedures can vary during time. However, Russell K. Standish has measured the informational complexity of Tierran 'organisms', and has not observed complexity growth in Tierran evolution [71]. Moreover, the mechanism used to determine which individual will die and which individual will perform more instructions is biased by some external rewards.

2.1.2. Avida

Christoph Adami, Charles Ofria, and C. Titus Brown developed the artificial life model, Avida [72] at the California Institute of Technology in 1993. Avida is an efficient model to explore biological questions using evolving computer programs (digital organisms) [72], which was extended from the Tierra system. Avida allocates each digital individual its own preserved area of memory, and executes it with a another virtual CPU. Normally, other digital individuals cannot access this memory space, neither for reading nor for writing, and cannot run code that is not in their own memory space. In Avida, the virtual CPUs of various individuals can work at various speeds, such that one individual runs, for example, twice as many instructions in the same time interval as another individual. The speed of virtual CPU is specified by different elements, but particularly, by the tasks that the organism performs: logical computations that the organisms can carry out to reap extra CPU speed as a bonus. In Avida, scientists can describe the existing tasks and place the consequences for individuals upon successful computation. When individuals are given with extra CPU cycles, their replication rate grows. Adami and Ofria, in collaboration with others, have used Avida to conduct research in digital evolution [73], [74]. For example: the 2003 paper, "The Evolutionary Origin of Complex Features" describes the evolution of a mathematical equals operation from simpler bitwise operations [73]. The individuals being specifically rewarded for executing pre-defined instructions, this system, as Tierra, is also an optimization process.

2.2. IBMs without Pre-defined fitness function

Models based on use of individuals as a basic unit, have been used in ecology since 40 years ago, but only since the excellent review of Huston et al. (1988) [75] emerged a decade ago, individual-based modeling has been considered as a useful approach for ecological modeling.

2.2.1. Echo

Echo which is an agent based model, was created to catch the essential characteristics of ecological systems. All of the elements and interactions in Echo are abstract, and it is not yet known whether Echo can be used to simulate real world phenomena efficiently. Echo expands the classical genetic algorithm in several significant directions: (A) fitness in endogenous, (B) agents have both a genome and a local state that evolves during time, (C) genomes are profoundly structured. Echo is a “genetic ecosystem model in which evolving agents are simulated in a resource-limited environment” [76]. Each agent in Echo, replicates itself with possible mutations when they obtain enough resources to copy its genome. The agents can gain resources during

interaction with other agents (combat, trade or mating) or from the environment. This system for endogenous reproduction is much closer to the way fitness is faced in natural settings than fitness functions in genetic algorithms. It has been shown that Echo exhibits the same relative species abundance pattern as natural ecological systems [77]. Echo was intended to be a general model of an intrinsic adaptive system rather than modeling and answering specific questions in evolutionary biology. Due to the high abstraction level of the Echo model, the degree of fidelity to real systems is uncertain.

2.2.2. Polyworld

Polyworld is another ecosystem model developed by Larry Yaeger [78] to study evolution. Polyworld, which is a computational ecology, was designed to examine issues in artificial life. Simulated individuals reproduce, fight and hunt and eat each other, eat the food that grows inside the world, and try to find successful tactics for survival. An individual's entire behaviour is controlled by its neural network "brain". Each brain's structure is specified from its genetic code, in terms of number, size, and combination of neural clusters and the types of connections between those clusters. Synaptic efficacy is adjusted via Hebbian learning, so, in fact, the individuals have the capability to learn during their lifetimes. The individuals perceive their world through their vision, prepared by a computer graphic rendering of the world from each individual's standpoint. The Individual's physiologies are also encoded genetically, thus all components of behaviour (including brain and body), evolve over several generations. Polyworld demonstrates a visual environment in which a population of individuals search for food, mate and have offspring. The population is typically in the hundreds, as each organism is complex and consumes considerable computer resources. In this platform, each individual makes its behaviour based on a neural network, which is coded in each individual's genome. The genome specifies the organism' size, speed, color, mutation rate and a number of other factors and is randomly mutated at a set probability, which are also changed in descendant organisms. Polyworld addresses a common behavioural ecology/evolutionary biology issue—how agents distribute themselves given limited, patchy resources [79]. Lack of semantics in the genomic structure (nodes) in Polyworld, makes it difficult to reason and link together different aspects of the model. Another criticism of PolyWorld, in the context of perpetual evolutionary emergence, is that learning during the life of the individuals appears to be overwhelmingly responsible for the results. This integrated learning process adds to the computational complexity of the model. Furthermore, the high complexity of the neural networks agents limits their number making it difficult to study large ecosystem phenomena's.

2.2.3. Framsticks

Framsticks presented by Komosinski et al in 1999 [46] is a 3D life simulation platform addressing both research and education. The platform consists of modules that facilitate the design of various experiments in optimization, coevolution, open-ended evolution and ecosystem modeling. Agents have both mechanical structure (bodies) consisting of connected sticks and control system (brain) using artificial neural network. The neural network brain collects data from sensors and sends signals to the joints, which control motion activities. The world is enriched with complex topology and a water level along with energy balls consumed by agents. Although some locomotion behaviours have evolved, the high complexity of the model did not present any different results than those obtained from much simpler evolutionary systems. This model is more concerned with the study of emerging motor behaviour rather than modeling a multiple level interacting ecosystem.

2.3. Other Predator-prey ecological simulations

Some of the above mentioned systems like Polyworld and Echo model predators. Other predator-prey models have also been presented focusing more on the ecological predator prey dynamics and interactions [80]. Smith (1991) [81] uses Volterra [82] model, which exhibits constant population dynamics, both in terms of oscillations in global populations as well as dynamic patchiness. The model integrated 2D spatial representation to study migration under different predation strategies. He showed that detailed movement patters in predator and prey can affect their interaction. Smith only models simple predator prey behaviour with simple genomic representation as only migration parameters are able to mutate. In [83] digital predator-prey organisms were used to study the evolution of trophic structure represented by the food web. Bell showed how different energy flow levels among organisms affect species richness and diversity. In another study [84] Lotka-Volterra equations were integrated in an IBM to examine how evolution of prey is used by predators affects community stability and whether complexity of food web increases stability of the predator prey system. The results demonstrated that number of existing species decreases with the increasing complexity.

A predator-prey simulation based in a spatial collection of individual finite state machine animate agents was first presented in [85]. This model can locate hundreds of thousands of individuals evolving in a two-dimensional featureless spatial plain. Every animate carries a small set of rules that direct its microscopic behaviour and at each time-step of the simulation, each animate executes one of these rules, causing it to: move; eat; or breed. In one study the effect of introducing camouflage behaviour as an available option for predators was investigated([86]). It

was shown that individuals who adopt this behaviour are relatively successful in obtaining prey and thus prolonging their lives against threat of dying of hunger [80]. This in turn led to higher numbers of successful older predators, which caused a crash in the population of prey. At each time step, every individual needs to change its state based on the locations and state of its neighbours. It is this process of finding the nearest neighbours that dramatically increases the time required to perform a useful run of the model. This expensive computational cost limits its number of individual and making it difficult to study large ecosystem phenomena's.

In another study a time-delayed gestation period was introduced into the predator-prey selection and adaptation mechanisms ([87]). The temporal behaviour of individual animates was affected by the gestation period parameter and hence the macroscopic behaviours of the species was also affected.

2.4. EcoSim, an Individual-based predator-prey Model without Pre-defined Fitness Function

Since, in this dissertation, EcoSim has been used to investigate several different biological questions, we give in this section a detailed description of EcoSim using the updated 7-points Overview-Design concepts-Details (ODD) standard protocol [88] for describing individual-based models.

2.4.1. Purpose

EcoSim is an individual-based predator-prey ecosystem simulation, which was designed to simulate agents' behaviour in a dynamic, evolving ecosystem [29], [89]. The main purpose of EcoSim is to study biological and ecological theories by constructing a complex adaptive system, which leads to a generic virtual ecosystem with behaviours similar to those found in nature. EcoSim uses, for the first time, a fuzzy cognitive map (FCM) to model each agent behaviour (see section 2.4.4.1). The FCM of each agent, being coded in its genome, allows the evolution of agents' behaviour throughout the epochs of the simulation.

In EcoSim, all the factors determining the reproductive success of an individual are free of pre-defined fitness functions. The overall fitness of an individual, measured as its reproductive success and that of its offspring, depends only on the interaction between its phenotype (behavioural type) and the environment. These interactions result from the usage of the behavioural models of the individuals under various environmental circumstances. At each time step, the individuals in EcoSim consume some energy. This consumption is determined by a cost

function that takes into account the complexity of the behavioural model of the individual (the number of edges it contains) and the action it performs. The more complex the model is, the faster the movements performed by the individual (such as escape and exploration) are, and the more the energy is consumed. This cost function is pre-defined. Nevertheless, a cost function is not a fitness function since it does not determine the success of a particular behavioural model. A cost function is a 'fix penalty', which is assigned to behavioural models and actions independently of the environment in order to avoid an obvious continuous increase in the behavioural model complexity and to model energy depletion with time. The success of a behavioural model relies on the tradeoff between the decisions it makes, knowing the current environment and the cost of the actions that are performed throughout the life of the individual. However, this tradeoff is not arbitrated by a predefined extrinsic function but results from the consequence of the actions undertaken.

As a consequence, decisions made by individuals with distinct behavioural models do not rely on any external evaluation (pre-defined fitness function) in the interest of the action. Instead, decisions rely on the knowledge 'learned' from the environment in the behavioural model by the evolutionary process, tuning behaviours to a particular state of the local world, and on the individual perception of the local environment. The model determining the reproductive success of an individual is thus intrinsic to the simulation in the sense that no external information is involved for determining fitness.

2.4.2. Entities, state variables, and scales

Individuals: There are two types of individuals: predators and prey. Each individual possesses several life-history characteristics (see Table 2-1) such as age, minimum age for breeding, speed, vision distance, level of energy, and amount of energy transmitted to the offspring. Energy is provided to the individuals by the resources (food) they find in their environment. Prey consume primary resources, which are dynamic in quantity and location, whereas predators hunt for prey. Each individual performs one unique action during a given time step, based on its perception of the environment. Each agent possesses its own FCM coded in its genome and its behaviours are determined by the interaction between the FCM and the environment (see section 2.4.4.1). Energy is provided by the primary or secondary resources found in their environment. For example, prey individuals gain 250 units of energy by eating one unit of grass and predators gain 500 units of energy by eating one prey. At each time step, each agent spends energy depending on its action (e.g. breeding, eating, running) and on the complexity of its behavioural model (number of existing edges in its FCM). On average, a movement action, such as escape and exploration,

requires 50 units of energy whereas a reproduction action uses 110 units of energy and the choice of no action results in a small expenditure of 18 units of energy.

Table 2-1. Several physical and life history characteristics of individuals from 10 independent EcoSim runs.

Characteristic	Predator	Prey
Maximum age	42 time steps (+/- 6)	46 time steps (+/-18)
Minimum age of reproduction	8 time steps	6 time steps
Maximum speed	11 cells / time step	6 cells / time step
Vision distance	25 cells maximum	20 cells maximum
Level of energy at initialization	1000 units	650 units
Average speed	1.4 cells / time step (+/- 0.3)	1.2 cells / time step (+/- 0.2)
Average level of energy	415 units (+/- 82)	350 units (+/- 57)
Maximum level of energy	1000 units	650 units
Average number of reproduction action during life	1.14 (+/- 0.11)	1.49 (+/- 0.17)
Average length of life	16 time steps (+/- 5)	12 time steps (+/- 3)

Cells and virtual world: The smallest units of the environment are cells. Each cell represents a large space, which may contain an unlimited number of individuals and/or some amount of food. The virtual world consists of torus-like discrete 1000 × 1000 matrix of cells.

Time step: Each time step involves the time needed for each agent to perceive its environment, make a decision, perform its action, as well as the time required to update the species membership, including speciation events and record relevant parameters (e.g. the quantity of available food). In terms of computational time, the speed of a simulation per generation is proportional to the number of individuals. An execution of the simulation with an average of 250 000 individuals simultaneously present in the world produced approximately 15 000 time steps in 35 days.

Population and Species: On average, in each time step, there are about 250,000 individuals, members of one or more species. A species is a set of individuals with a similar genome relative to a threshold.

2.4.3. Process overview and scheduling

The possible actions for the prey agents are: exploring the environment to gain information regarding food, predators, and sexual partners, evasion (escape from predator), search for food (if there is not enough grass available in its habitat cell, prey can move to another cell to find grass), socialization (moving to the closest prey in the vicinity), exploration, resting (to save energy), eating and breeding. Predators also perceive the environment to gather information used to choose an action from amongst: hunting (to catch a prey), search for food, socialization, exploration, resting, eating and breeding. After each action, the individuals' energy is adjusted and their age in incremented by one. There are also two environmental processes: after all individuals perform their actions, the amount of grass and meat are adjusted.

At each time step, the value of the state variables of individuals and cells are updated. The overview and scheduling of every time step is as follows (algorithm):

1. For prey individuals:
    1.1. Perception of the environment
    1.2. Computation of the next action
    1.3. Performing actions and updating the energy level
2. Updating the list of prey (it's done once for all prey individuals)
3. Updating prey species (it's done once for all prey individuals)
4. For predator individuals:
    4.1. Perception of the environment
    4.2. Computation of the next action
    4.3. Performing their action and update of the energy level
5. Updating the list of predator individuals (it's done once for all predator individuals)

6. Updating predator species (it's done once for all predator individuals)

7. For each cell in the world:

    7.1. Updating the grass level

    7.2. Updating the meat level

8. Updating of the age of the individuals

The complexity of the simulation algorithm is mostly linear with respect to the number of individuals. If we consider that there are N_1 prey and N_2 predators and we exclude the sorting parts, which have a complexity of $O(N_1 log N_1)$ and $O(N_2 log N_2)$ but are negligible in the overall computational time as they are only performed once per time step, then the complexity of part 1 and part 2 of the above algorithm, including the clustering algorithm used for speciation, will be $O(N_1)$ and $O(N_2)$ respectively (Aspinall and Gras, 2010). The virtual world of the simulation has 1000×1000 cells, therefore the complexity of part 3 will be $O(k = 1000 \times 1000)$. The complexity of part 4 will be $O(N_1+N_2)$. As a result, the overall complexity of the algorithm is $O(2N_1+ 2N_2+ k)$, which is $O(N = 2N_1 +2N_2)$.

2.4.4. Design concepts

2.4.4.1. Basic principles

To observe the evolution of individual behaviour and ultimately ecosystems over thousands of generations, several conditions need to be satisfied: (i) every individual should possess genomic information; (ii) this genetic material should affect the individual behaviour and consequently its fitness; (iii) the inheritance of the genetic material has to be done with the possibility of modification; (iv) a sufficiently high number of individuals should coexist at any time step and their behavioural model should allow for complex interactions and organizations to emerge; (v) a model for species identification, based on a measure of genomic similarity, has to be defined; and (vi) a large number of time steps need to be performed. These complex conditions pose computational challenges and require the use of models that combine the compactness and ease of computation with a high potential for complex representation.

In EcoSim, a Fuzzy Cognitive Map (FCM) [90] is the base for describing and computing the agent behaviours. Each agent possesses an FCM to compute its next action. The FCM is integrally coded in their genomes and therefore heritable and subject to evolution. FCMs are

weighted graphs representing the causal relationship between concepts, allowing the observation of evolutionary patterns and inference of underlying processes (Figure 2-1) (see section 2.4.4.2 and 2.4.4.6). When a new offspring is created, it is given a genome, which is a combination of the genomes of its parents with some possible mutations.

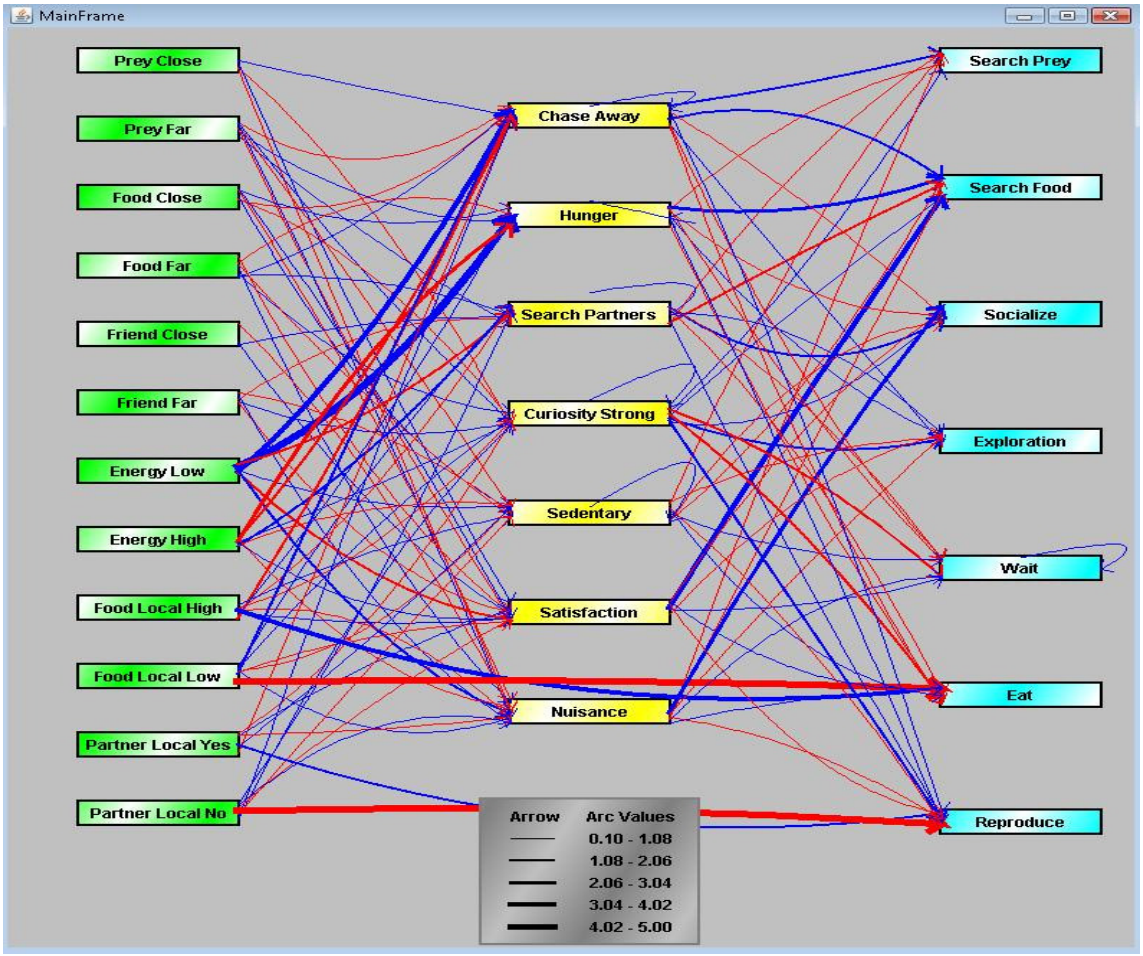


Figure 2-1. A sample of a predator's FCM including concepts and edges. The width of each edge shows the influence value of that edge. Color of an edge shows inhibitory (red) or excitatory (blue) effects.

Formally an FCM is a graph, which contains a set of nodes C, each node C_i being a concept, and a set of edges I, each edge I_{ij} representing the influence of the concept C_i on the concept C_j. A positive weight associated with the edge I_{ij} corresponds to an excitation of the concept C_j from the concept C_i, whereas a negative weight is related to an inhibition (a zero value indicates that there is no influence of C_i on C_j). The influence of the concepts in the FCM can be represented in an n×n matrix, L, in which L_{ij} is the influence of the concept C_i on the concept C_j. If $L_{ij} = 0$, there is

no edge between C_i and C_j. In EcoSim, each individual genome code for its proper FCM, with one gene coding for one weight L_{ij}.

2.4.4.2. Emergence

In each FCM, three kinds of concepts are defined: sensitive (such as distance to foe or food, amount of energy, etc.), internal (fear, hunger, curiosity, satisfaction, etc.), and motor (evasion, socialization, exploration, breeding, etc.). The activation level of a sensitive concept is computed by performing a fuzzification of the information the individual perceives in the environment. For an internal or motor concept, C, the activation level is computed by applying the defuzzification function on the weighted sum of the current activation level of all the concepts having an edge directed toward C. Finally, the action of an individual is selected based on the maximum value of motor concepts' activation level. Activation levels of the motor concepts are used to determine the next action of the individual. For example, Figure 2-2 represents two sensitive concepts (foeClose and foeFar), one internal (fear), and one motor (evasion). There are also three influence edges: closeness to a foe excites fear, distance to a foe inhibits fear, and fear causes evasion. Activations of the concepts foeClose and foeFar are computed by fuzzification of the real value of the distance to the foe, and the defuzzification of the activation of evasion tells us about the speed of the evasion (see section 2.4.4.6).

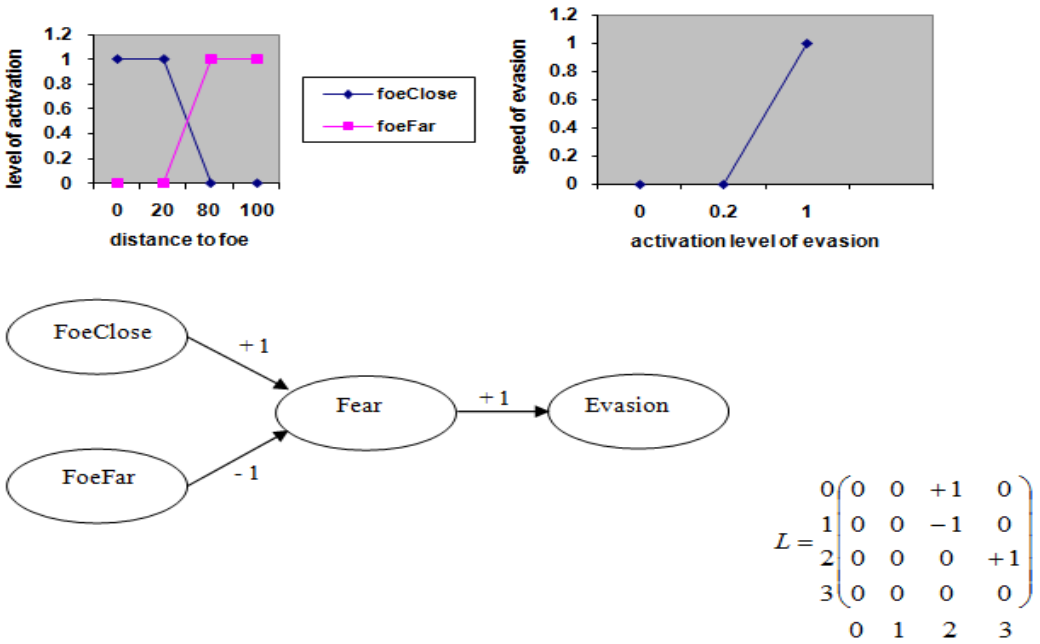


Figure 2-2. An FCM for detection of foe (predator) and decision to evade, with its corresponding matrix (0 for 'Foe close', 1 for 'Foe far', 2 for 'Fear' and 3 for 'Evasion') and the fuzzification and defuzzification functions [91]**.**

The behavioural model of individuals coded in FCM can react to the changes in the environment for example, it has been shown that the contemporary evolution of prey behaviour owing to predator removal is also accompanied by prey genetic change [92]. At the initiation of the simulation, prey and predators are scattered randomly all around the virtual world. Through the epochs of the simulation, the distribution of the individuals in the world is changed drastically based on many different factors: prey escaping from predators, individuals socializing and forming groups, individuals migrating gradually to find sources of food, species emerging, etc. The size of the world is large enough to accommodate population structures and the emergence of migrations. For example, an individual moving at its maximum speed could barely cross half of the world during its life span. Moreover, previous studies demonstrate that the usage of behavioural models lead to a non-random distribution of individuals and species in which individuals form populations that contain agents with similar genomes [30], [33]. Figure 2-3 shows an example of a snapshot of the virtual world after thousands of time steps with emerging grouping patterns.

It has been shown that the data generated by EcoSim present the same kind of multifractal properties as those observed in real ecosystems [93]. Individuals' distribution forming spiral waves is one property of prey-predator models (Figure 2-3). Prey near the wave break have the capacity to escape from the predators sideways. A subpopulation of prey then finds itself in a region relatively free from predators. In this predator-free zone, prey starts expanding extensively, forming a circularly expanding region. The whole pressure process and spiral formation will be applied to this subpopulation of prey and predators, leading to the formation of a second scale [34]. This process repeats many times, which is a common property of self-similar processes [94]. Because there are consecutive interactions between prey and predators over time, the same pattern repeats itself over and over. The result of this pattern repetition is the emergency of self-similarity in the spatial distribution of individuals. In addition, migration phenomena can be observed, since the relocation of individuals leads to the redistribution in the population [95].

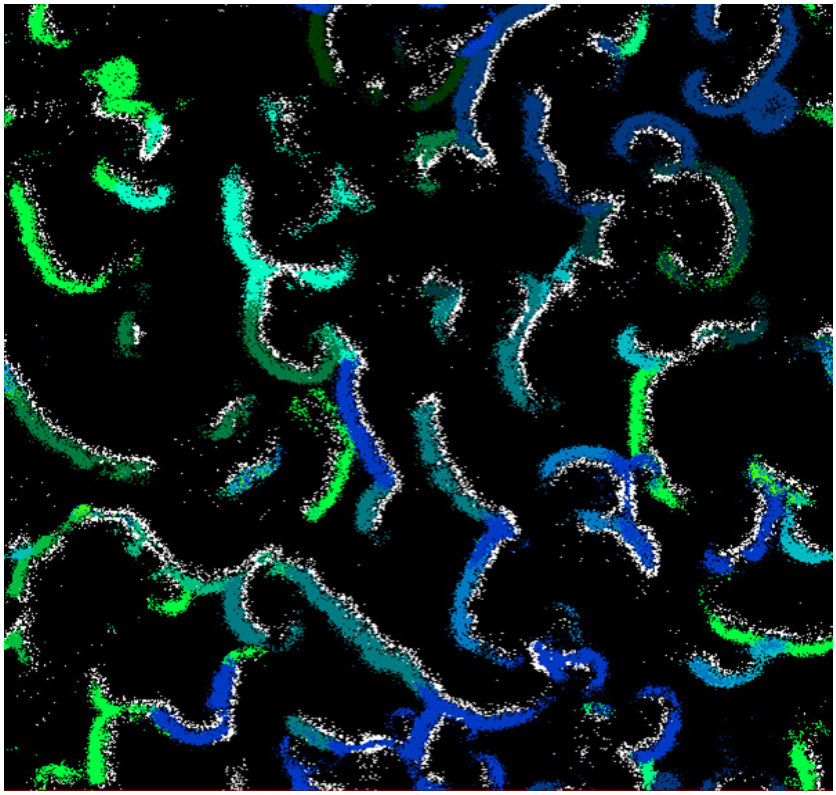


Figure 2-3. A snapshot of the virtual world in one specific time step, white color represents predator species and the other colors show different prey species.

2.4.4.3. Adaptation

The genome maximum length is fixed (390 sites), where each site is a real number and corresponds to an edge between two concepts of the FCM and code for the weight associated to this edge. However, as many edges have an initial value of zero, only 114 edges for prey and 107 edges for predators exist at initialization (see section 2.4.4.1). One more gene is used to code for the amount of energy, which is transmitted from the parents to their child at birth. The value of a site, which is a real number, corresponds to the intensity of the influence between the two concepts. The genome of an individual is transmitted to its offspring after being combined with the genome of the other parent and following the possible addition of some mutations. To model linkage, the weights of edges are transmitted by blocks from parents to the offspring. For each concept, its entire incident edges' values are transmitted together from the same randomly chosen parent. The behavioural model of each individual is therefore unique. Step after step, as more individuals are created, changes in the FCM occur due to the formation of new edges (with probability of 0.001), removal of existing edges (with probability of 0.0005) and changes in the weights associate to existing edges (with probability of 0.005). These low probabilities, compare

to crossover probability, reflects the fact that change in genome should be relatively slow to avoid random evolution. Therefore, new genes may emerge from among the 265 initial edges of zero value.

2.4.4.4. Fitness

We calculated the fitness for each species as the average fitness of its component individuals. In order to realistically represent the capacity of an individual to survive and produce offspring that can also survive, fitness was calculated as the sum of age at death of the focal individual with the death age of its children (a post-processing computation). Since the sum involves all direct offspring, it is representative of the fertility and survivability of the individual.

2.4.4.5. Prediction

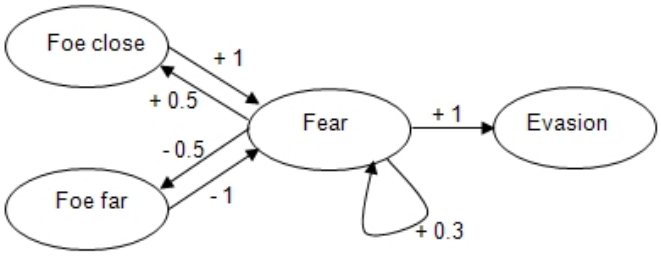
So far, there is no learning mechanism for individuals during their life and they cannot predict the consequences of their decision. The only available information for every individual to make decisions is the information coming from their perceptions at that particular time step and the value of the activation level of the internal and motor concepts at the previous time steps. The activation levels of the concepts of an individual are never reset during its life. As the previous time step activation level of a concept is involved in the computation of its next activation level, this means that all previous states of an individual during its life participate in the computation of its current state. Therefore, an individual has a basic memory of its own past that will influence its future states.

2.4.4.6. Sensing

Every individual in EcoSim is able to sense its local environment inside its range of vision. For instance, each prey can sense its five closest foes, cells with food units, mates within its range of vision, the number of grass units in its cell and the number of possible mates in its cell. Moreover, each individual is capable of recognizing its current level of energy.

It should be noted that the FCM process explained in section 2.4.4.2, enables, for example, distinguishing between perception and sensation: sensation is the real value coming from the environment, and perception is sensation modified by an individual's internal states. For example, it is possible to add three edges to the map presented in Figure 2-2: one auto excitatory edge from the concept of fear to itself, one excitatory edge from fear to foeClose, and one inhibitory edge from fear to foeFar (Figure 2-4). A given real distance to the foe seems higher or lower to the individual depending on the activation level of fear. Also, the fact that the individual is frightened

at time t influences the level of fear of the individual at time t + 1. This kind of mechanism makes possible the modeling of the degree of stress for an individual. It also enables the individual to memorize information from previous time steps: fear maintains fear. It is therefore possible to build very complex dynamic systems involving feedback and memory using an FCM, which is needed to model complex behaviours and abilities to learn from evolution.



$$L = \begin{pmatrix} 0 & 0 & +1 & 0 \\ 0 & 0 & -1 & 0 \\ +0.5 & -0.5 & +0.3 & +1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Figure 2-4. An FCM for detection of foe (predator) - difference between perception and sensation [91]**. This map shows different kind of interactions between three kinds of concepts: perception concept (Foe close and Foe far), internal concept (Fear) and motor concept (Evasion).**

2.4.4.7. Interaction

The only action that requires a coordinate decision of two individuals is reproduction. For reproduction to be successful, the two parents need to be in the same cell, to have sufficient energy, to choose the reproduction action and to be sufficiently genetically similar. The individuals cannot determine their genetic similarity with their potential partner. However, if they try to mate and the potential partner is too dissimilar (the difference between the two genomes is greater than a specified threshold (half of the speciation threshold)), then the reproduction fails.

Predator's hunting introduces another type of interaction in the simulation. For a predator to succeed in the hunting action, its distance to the closest prey is required to be less than one cell. When a predator's hunting action succeeds, a new meat unit is added to the corresponding cell, and the energy level of the predator is also increased by one unit of meat energy.

Furthermore, there is a competition for prey and predators for food. For example, if in a given cell there is only one food unit and two agents have chosen the action of eating, the younger will act first, and so it will be the only one that can eat (in this cell) at this time step. This is a way to simulate the fact that older species members help younger species members to survive.

2.4.4.8. Stochasticity

To produce variability in the ecosystem simulation, several processes involve stochasticity. For instance, at initialization, the number of grass units is randomly determined for each cell. Moreover, the maximum age of an individual is determined randomly at birth from a uniform distribution centered at a value associated with the type of agent (see section 2.4.5). Stochasticity is also included in several kinds of actions of the individuals such as evasion and socialization. If there is no predator or partner respectively in the vision range of the individual, the direction of the movement would be random. Furthermore, the direction of the exploration action is always random.

However, to understand the extent of randomness in EcoSim, Golestani et al. (2010) examined whether chaotic behaviour exists in signals (time series) generated by the simulation. They concluded that the EcoSim is capable of generating non-random and chaotic pattern (time series) [32]. For a more detailed description of these studies see chapters 3 and 5.

2.4.4.9. Collectives

In EcoSim, the notion of species is implemented in a way that species emerge from the evolving population of agents. EcoSim implements a species concept directly related to the genotypic cluster definition [96] in which a species is a set of individuals sharing a high level of genomic similarity. In addition, in EcoSim, each species is associated with the average of the genetic characteristics of its members, called the 'species genome' or the 'species center'. The speciation method involves a 2-means clustering algorithm [97] in which an initial species is split into two new species, each of them containing the agents that are mutually the most similar (see section 4.2). Over time, a species will progressively contain individuals that are increasingly genetically dissimilar up to an arbitrary threshold where the species splits. After splitting, the two sister species are sufficiently similar that hybridization events can occur. Therefore, two individuals can interbreed if their genomic distance is smaller than an arbitrary threshold (half of the speciation threshold) even if they are designated as members of two sister species by our clustering algorithm. The information about species membership is only a label. It is not used for any purpose during the simulation but only for post-processing analysis of the results. Several studies have been conducted to analyze the concept of species in EcoSim. Devaurs & Gras (2010) [98] compared the species abundance patterns emerging from EcoSim with those observed in natural ecosystems using Fisher's logseries [99]. Species abundance is a key component of macroecological theories and Fisher's logseries is one of the most widely known classic models

of species abundance distribution. The results of this study proved that at any level in sample size, EcoSim generates coherent results in terms of relative species abundance, when compared with classical ecological results [100]. In another study, Golestani et al. (2012) [30] investigated how small, randomly distributed physical obstacles influence the distribution of populations and species, showing that there is a direct and continuous increase in the speed of evolution (e.g. the rate of speciation) with the increasing number of obstacles in the world. We also investigated one of the most difficult questions in biology that refers to the astonishing fact that species are an inevitable byproduct of evolution and it has been shown that natural selection is the leading factor of speciation [31]. For a more detailed description see section 4.2.

2.4.4.10. Observation

EcoSim produces a large amount of data in each time step, including number of individuals, new and extinct species, geographical and internal characteristics of every individual, and status of the cells of the virtual world. Information regarding each individual includes position, level of energy, choice of action, specie, parents, FCM, etc. There is also the possibility to store all of the values of every variable in the current state of the simulation in a separate file, making possible the restoration of the simulation from that state onwards. All of the data is stored in a compact special format, to facilitate the storage and future analysis.

2.4.5. Initialization and input data

A parameter file is used to assign the values for each state variable at initialization of the simulation. These parameters are as follows: width and height of the world, initial numbers of individuals, threshold of genetic distance for prey/predator speciation, maximum age, energy, speed, vision range, and initial values of FCM for prey/predator. Any of these parameters can be changed for specific experiments and scenarios. An example of a list of the most common user-specified parameters is presented in Table 2-2. For other initial parameters see Table 2-3 to Table 2-8.

Different values of initial parameters can lead to an extinction of either the prey or the predators or both of them. The current values lead to stable runs for the simulation. Some parameters like number of individuals are less sensitive than other. However, the whole system is quite stable and many different combinations of values still tested have led to stable runs. Moreover, as far as the runs are stables, all the general patterns behaviour described in section 2.4 emerged and have been observed systematically.

Table 2-2. Values for user-specified parameters in EcoSim.

User Specified Parameter	Used Value
Number of Prey	12000
Number of Predators	500
Grass Quantity	5790000
Maximum Age Prey	46
Maximum Age Predator	42
Prey Maximum Speed	6
Predator Maximum Speed	11
Prey Energy	650
Predator Energy	1000
Distance for Prey Vision	20
Distance for Predator Vision	25
Reproduction Age for Prey	6
Reproduction Age for Predator	8

2.4.6. Submodels

As mentioned earlier, each individual performs one unique action during a time step based on its perception of the environment. EcoSim iterates continuously, and each time step consists of the computation of the activation level of the concepts, the choice and application of an action for every individual. A time step also includes the update of the world: emergence and extinction of species and growth and diffusion of grass, or decay of meat.

At initialization time there is no meat in the world and the number of grass units is randomly determined for each cell. For each cell, there is a probability, probaGrass, that the initial number

of units is strictly greater than 0. In this case, the initial number is generated uniformly between 1 and maxGrass. Each unit provides a fixed amount of energy to the agent that eats it. The preys can only eat the grass, and the predators have two modes of predation: hunting and scavenging. When a predator's hunting action succeeds, a new meat unit is added in the corresponding cell and the predator is considered consuming another one. When a predator's eating action succeeds (which can be viewed as a scavenging action), one unit of meat is removed in the corresponding cell. The amount of energy is energyGrass for one grass unit when eaten by a prey and is energyMeat for one meat unit eaten by a predator. The number of grass units grows at each time step, and when a prey dies in a cell, the number of meat units in this cell increases by 2. The number of grass units in a cell decreases by 1 when a prey eats, and the number of meat units decreases by 1 when a predator eats. The number of meat units in a cell also decreases at each time step, even if no meat has been eaten in this cell.

1. Evasion (for prey only). The evasion direction is the direction opposite to the direction of the barycenter of the 5 closets foes within the vision range of the prey, with respect to the current position of the prey. If no predator is within the vision range of the prey, the direction is chosen randomly. Then the new position of the prey is computed using the speed of the prey and the direction. The current activation level of fear is divided by 2.

2. Hunting (for Predator only). The predator selects the closest cell (including its current cell) that contains at least one prey and moves towards that cell. If it reaches the corresponding cell based on its speed, the predator kills the prey, eating one unit of food and having another unit of food added to the cell. When there are several preys in the destination cell, one of them is chosen randomly. If the speed of the predator is not enough to reach the prey, it moves at its speed toward this prey. If there is no prey in the current cell and in the vicinity or it does not have enough energy to reach to a prey, hunting action is failed.

3. Search for food. The direction toward the closest food (grass or meat) within the vision range is computed. If the speed of the agent is high enough to reach the food, the agent is placed on the cell containing this food. Otherwise, the agent moves at its speed toward this food.

4. Socialization. The direction toward the closest possible mate within the vision range is computed. If the speed of the agent is high enough to reach the mate, the agent is placed on the cell containing this mate, and the current activation level of sexualNeeds is divided by 3. Otherwise, the agent moves at its speed toward this mate. If no possible mate is within the vision range of the agent, the direction is chosen randomly.

5. Exploration. The direction is computed randomly. The agent moves at its speed in this direction. The activation level of curiosity is divided by 1.5.

6. Resting. Nothing happens.

7. Eating. If the current number of grass (of meat) units is greater than 1, then this number is decreased by 1 and the prey's (predator's) energy level is increased by energyGrass (energyMeat). Its activation level for hunger is divided by 4. Otherwise nothing happens.

8. Breeding. The following algorithm is applied to the agent A:

if A.energyLevel > 0.125 × maxEnergyPrey then

for all A of the same type in the same cell

if A.energyLevel > 0.125 × maxEnergyPrey and D(A,A) < T and

A′ has not acted at this time step yet and

A's choice of action is also breeding

then

interbreeding(A,A)

A.sexualNeeds ← 0

A.sexualNeeds ← 0

If A′ satisfies all the criteria, the loop is canceled

If none of the A′ agents satisfies all the criteria, the breeding action of A fails.

For every action requiring that the agent move, its speed is computed by the formula

Speed = Ca × maxSpeedPrey => for the preys

Speed = Ca × maxSpeedPredator => for the predators

with Ca the current activation level of the motor concept associated with this action.

The process of generating a new offspring (interbreeding function) consists of following steps. First, the value of birthEnergyPrey is transmitted with possible mutations from one randomly chosen parent to the offspring. Second, the edges' values are transmitted with possible mutations, and the initial energy of the offspring is computed. To model the crossover mechanism, the edges are transmitted by block from one parent to the offspring. For each concept, its incident edges' values are transmitted together from the same randomly chosen parent. Third, the maximum age of the offspring is computed. Finally, the energy level of the two parents is updated.

Table 2-3. The initial parameters of the EcoSim at the first time step of the simulation. There are 42 parameters for each run of EcoSim. The value of these parameters has been obtained empirically and by biologists' expert opinion to preserve the equilibrium in the ecosystem.

Parameter	Initial Value	Comments
Width	1000	width of the world
Height	1000	height of the world
ProbaGrass	0.187	initial probability of grass per cell
ProbaGrowGrass	0.0028	probability of diffusion of grass
ValueGrass	250	energy value for a consumed grass
ValuePrey	500	energy value for a consumed prey
MaxGrass	8	maximum number of grass in a cell
SpeedGrowGrass	0.5	speed of growing grass
MaxMeat	8	maximum number of meat in a cell
NbResources	2	number of food resources in the world
ProbaMut	0.005	probability of mutation to a nonzero gene
ProbaMutLow	0.001	probability of mutation to a zero gene
MinArc	0.075	threshold for an arc to be counted as nonzero
InitNbPrey	12000	initial number of prey
InitNbPredator	2000	initial number of predator
DistanceSpeciesPrey	1.5	threshold of genetic distance for prey species
DistanceSpeciesPred	1.3	threshold of genetic distance for predator species
AgeMaxPrey	46	maximum age for prey
AgeMaxPred	42	maximum age for predator
AgeReprodPrey	6	minimum reproduction age for prey
AgeReprodPred	8	Minimum reproduction age for predator
ClusterPrey	10	number of prey per clusters at initialization
ClusterPredator	20	number of predators per clusters at initialization
RadiusCluster	5	radius in number of cell of each initial cluster
EnergyPrey	650	maximum energy of prey
EnergyPredator	1000	maximum energy of predator
SpeedPrey	6	maximum speed of prey
SpeedPredator	11	maximum speed of predator
VisionPrey	20	maximum vision of prey
VisionPredator	25	maximum vision of predator
StateBirthPrey	30	initial parental energy investment for prey
StateBirthPred	40	initial parental energy investment for predator
nbSensPrey	12	number of sensitive concepts in prey
nbConceptsPrey	7	number of internal concepts in prey
nbMotorPrey	7	number of motor concepts in prey
nbSensPredator	12	number of sensitive concepts in predator
nbConceptsPredator	7	number of internal concepts in predator
nbMotorPredator	7	number of motor concepts in predator

Restore	1	0-no restore, 1-restore
MaxSave	500	0-no save, #-save every # states
MinSave	0	0-no save, #-save every # states
WorldSave	0	0-no save, 1-save world

Table 2-4. Initial FCM values for Prey (See Table 2-5). Every prey individual has a FCM which represent its behaviour. At first time step, all prey individuals have an initial FCM. During time and during each generation with operators like crossover and mutation, the FCM of individuals change.

	FR	HG	SP	CU	SD	ST	NU	ES	SF	SC	XP	WT	ET	RP
PC	4	0	0	0.1	0	-1	1	0	0	0	0	0	0	0
PF	-4	0	0	0	0	0.5	-0.5	0	0	0	0	0	0	0
OC	0	0.5	0	-0.1	0.1	0.5	-0.5	0	0	0	0	0	0	0
OF	0	0	-0.4	0.2	-0.2	-0.7	0.7	0	0	0	0	0	0	0
FC	0	0	0.5	-0.1	0.1	0.5	-0.5	0	0	0	0	0	0	0
FF	0	0	-0.4	0.2	-0.2	-0.5	0.5	0	0	0	0	0	0	0
EL	0.4	4	-1.5	0	0	-2.2	2.2	0	0	0	0	0	0	0
EH	0	-1	1.5	0.2	-0.2	1.5	-1.5	0	0	0	0	0	0	0
OH	0	-0.2	0	-0.3	0.3	1.1	-1.1	0	0	0	0	0	2.6	0
OL	0	0.2	0	1	-1	-1.1	1.1	0	0	0	0	0	-4	0
PY	0	0	0	-0.4	0.4	0.5	-0.5	0	0	0	0	0	0	1.5
PN	0	0	0.5	0.3	-0.3	-0.8	0.8	0	0	0	0	0	0	-4
FR	0.5	0	0	0	0	0	0	1.5	-0.8	-1	0.3	-1	-1	-1
HG	0	0.3	0	0	0	0	0	-0.8	2.1	-0.7	0.7	-0.5	4	-1.8
SP	0	0	0.2	0	0	0	0	-0.2	0	1.5	0.5	-0.3	-0.4	3
CU	0	0	0	0.1	0	0	0	-0.1	0.5	0.3	1.5	-0.2	-0.3	-0.2
SD	0	0	0	0	0.1	0	0	0	-0.5	-0.3	-1.2	0.2	0.3	0.2
ST	0	0	0	0	0	0	0	-0.1	-0.8	-0.2	-2	1.5	0.8	0.7
NU	0	0	0	0	0	0	0	0.4	1	0.2	2	-1.2	-0.7	-0.7
ES	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SF	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SC	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XP	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WT	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0
ET	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2-5. Prey/predator FCM abbreviation table. The abbreviation used to present concepts of FCM in EcoSim. These abbreviations have been used in other tables to show values of these concepts.

NodeName	Abbreviation	NodeName	Abbreviation
Fear	FR	PredClose	PC
Hunger	HG	PredFar	PF
SearchPartner	SP	FoodClose	OC
CuriosityStrong	CU	FoodFar	OF
Sedentary	SD	FriendClose	FC
Satisfaction	ST	FriendFar	FF
Nuisance	NU	EnergyLow	EL
Escape	ES	EnergyHigh	EH
SearchFood	SF	FoodLocalHigh	OH
Socialize	SC	FoodLocalLow	OL
Exploration	XP	PartnerLocalYes	PY

Wait	WT	PartnerLocalNo	PN
Eat	ET	PreyClose	YC
Reproduce	RP	PreyFar	YF
ChaseAway	CA		
SearchPrey	SY		

Table 2-6. Parameters of prey defuzzification function (see Figure 2-5). The function that has been used for fuzzifications uses three parameters which shape the fuzzification curve.

NodeName	Activation	Fuzzy Parameter1	Fuzzy Parameter2	Fuzzy Parameter3
PredClose	0	1	3.5	3.5
PredFar	0	2	3.5	3.5
FoodClose	0	1	6	6
FoodFar	0	2	6	6
FriendClose	0	1	5	5
FriendFar	0	2	5	5
EnergyLow	0	1	4	4
EnergyHigh	0	2	4	4
FoodLocalHigh	0	2	4	4
FoodLocalLow	0	1	4	4
PartnerLocalYes	0	2	1000	20
PartnerLocalLow	0	1	1000	20
Fear	0	0	1	3.5
Hunger	0	0	1	3
SearchPartner	0	0	1	3
Curiosity	0	0	1	2.5
Sedentary	0	0	1	2.5
Satisfaction	0	0	1	3
Nuisance	0	0	1	3
Escape	0	0	1	3.5
SearchFood	0	0	2	3
Socialize	0	0	4	3
Exploration	0	0	6	2.5
Wait	0	0	7	3
Eat	0	0	8	3.5
Reproduce	0	0	10	3.5

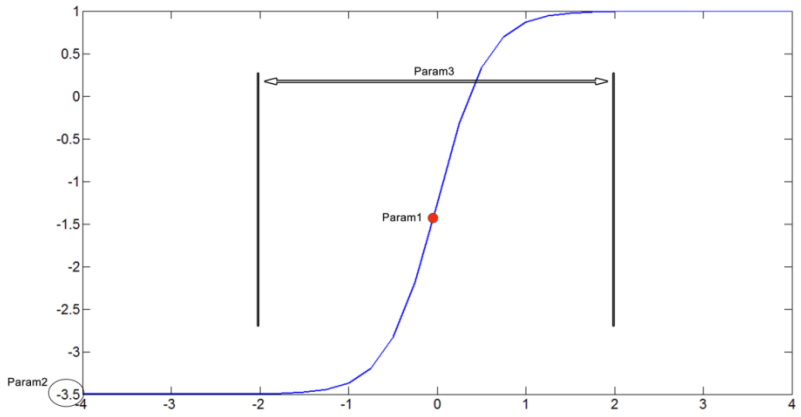


Figure 2-5. The three parameters that specify the shape of the curve. The first parameter specifies the center of curve in the horizontal axis, the second parameter specifies the lower band of curve in the vertical axis and the third parameter specifies the width of curve.

Table 2-7. Initial FCM for Predator (See Table 2-5). Every predator individual has a FCM which represent its behaviour. At first time step, all predator individuals have an initial FCM. During time and during each generation with operators like crossover and mutation, the FCM of individuals change

	CA	HG	SP	CU	SD	ST	NU	SY	SF	SC	XP	WT	ET	RP
YC	0.7	0	0	-0.1	0	0.5	-0.5	0	0	0	0	0	0	0
YF	-0.5	0.7	0.1	0.4	-0.4	-0.5	0.5	0	0	0	0	0	0	0
OC	-0.5	0.7	0	-0.1	0.1	0.5	-0.5	0	0	0	0	0	0	0
OF	0.8	-0.2	0.1	0.2	-0.2	-0.6	0.6	0	0	0	0	0	0	0
FC	0	0	0.7	0	0	0.4	-0.4	0	0	0	0	0	0	0
FF	0	0	-0.5	0.3	-0.3	-0.4	0.4	0	0	0	0	0	0	0
EL	3.5	5	-1.2	0	0.2	-1.5	1.5	0	0	0	0	0	0	0
EH	-2	-3	1.4	0.3	-0.3	1	-1	0	0	0	0	0	0	0
OH	-1.5	0.3	-0.2	-0.3	0.3	1	-1	0	0	0	0	0	4	0
OL	1.7	0	0.2	1	-1	-1	1	0	0	0	0	0	-5	0
PY	-0.3	0	0	-0.4	0.4	0.8	-0.8	0	0	0	0	0	0	2
PN	0.3	0	0.5	0.3	-0.3	-0.8	0.8	0	0	0	0	0	0	-5
CA	0.2	0	0	0	0	0	0	1.5	-0.2	-0.4	0.3	-0.4	0	-0.4
HG	0	0.3	0	0	0	0	0	4	2.5	-1.2	0.3	-0.4	3.5	-0.8
SP	0	0	0.2	0	0	0	0	-0.8	-0.8	1.5	0.3	-0.5	-0.6	3
CU	0	0	0	0.1	0	0	0	0.3	0.3	0.3	1.5	-0.4	-0.3	-0.2
SD	0	0	0	0	0.1	0	0	-0.3	-0.3	-0.3	-1.5	0.4	0.3	0.2
ST	0	0	0	0	0	0	0	-0.8	-0.8	-0.2	-1.8	1	0.8	0.8
NU	0	0	0	0	0	0	0	1	0.8	0.2	2	-1	-0.6	-0.8
SY	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SF	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SC	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XP	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WT	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0
ET	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RP	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2-8. Parameters of predator defuzzification function (see Figure 2-5). The function that has been used for fuzzifications uses three parameters which shape the fuzzification curve.

NodeName	Activation	Fuzzy Parameter1	Fuzzy Parameter2	Fuzzy Parameter3
PreyClose	0	1	4	4
PreyFar	0	2	4	4
FoodClose	0	1	5	5
FoodFar	0	2	5	5
FriendClose	0	1	5	5
FriendFar	0	2	5	5
EnergyLow	0	1	4.5	4.5
EnergyHigh	0	2	4.5	4.5
FoodLocalHigh	0	2	1000	20
FoodLocalLow	0	1	1000	20
PartnerLocalYes	0	2	1000	20
PartnerLocalNo	0	1	1000	20
ChaseAway	0	0	1	3
Hunger	0	0	1	3.5
SearchPartner	0	0	1	3
Curiosity	0	0	1	2.5
Sedementary	0	0	1	2.5
Satisfaction	0	0	1	3
Nuisance	0	0	1	3
SearchPrey	0	0	1	3
SearchFood	0	0	3	3.5
Socialize	0	0	5	3
Exploration	0	0	7	2.5
Wait	0	0	8	3
Eat	0	0	9	3.5
Reproduce	0	0	11	3.5

2.5. Randomized version of EcoSim

2.5.1. The randomized version of EcoSim

In order to have 'null hypothesis' model for comparison with our complex model, we also develop a version of our simulation in which the behavioural models are unplugged and as a result of that there is no effects of natural selection. Instead, we apply a random walk process to our system. Without a behavioural model, the spatial distribution of individuals is random. In this version of the simulation, we used a random walk model with no behavioural model for individuals.

A random walk, sometimes denoted RW, is a mathematical formalization of a trajectory that consists of taking successive random moves [101].

This model is derived from the "unified neutral theory of biodiversity" by ecologist Stephen Hubbell [102]. Hubbell's theory treats individuals in the population as essentially identical in their per capita probabilities of giving birth, dying, migration, and speciation. This implies a random behaviour at the individual level.

In the randomized version of the simulation, the behavioural model responsible for different actions of each individual is removed and the actions of the individuals are narrowed down to movement and reproduction:

- Movement of the individuals in the virtual world is random; however, the distribution of movements and the size of the world are kept the same as in the EcoSim.
- Predator-Prey population dynamics are determined by the Lotka-Volterra competition model [82], [103], [104]. This model controls the number of births and deaths of individuals at each time step. The following formulas have been used to compute the variation in number of both of prey and predators:

$$\frac{dn_1}{dt} = r_1.\left(1-\frac{n_1}{k_1}\right).n_1 - a_1.n_1.n_2$$
$$\frac{dn_2}{dt} = -r_2.n_2 + a_2.n_1.n_2 \qquad (2\text{-}1)$$

Where n_2 is the number of predators, n_1 is the number of prey, dn_1/dt and dn_2/dt represent the variation of the two populations with time, t represents the time; and r_1, a_1, r_2, a_2 and k_1 are parameters representing the interaction between the prey and predators. The individuals that die are randomly selected. Reproduction action is also random, and unlike EcoSim there is no need for genetic similarity of the parents. The parents and the offspring's initial location are also randomly chosen.

Chapter 3

3. Nonlinear and Chaos Analysis

The important point in science is the assumption that experiments are predictable and repeatable [105]. Thus it surprised most scientists when simple deterministic systems were found that were neither predictable nor repeatable. Instead, they exhibited chaos, in which the tiniest change in the initial conditions produces a very different outcome, even when the governing equations are known exactly [105].

Over the past 30 years, in the field of mathematics and modern physics, a new scientific method and very interesting theory called "Chaos Theory" has appeared [106], [107]. Chaos theory studies complex dynamical systems such as the atmosphere, animal populations, flow, heart palpitations, and geological processes [108]. The key idea of chaos theory is that in any irregularities, order lies. This means that regularity should not be sought just at one scale. Phenomenon which at the local scale seems to be completely random and unpredictable, perhaps on a larger scale, is highly stationary and predictable [105].

There is a similarity between chaos theory and the science of statistics. Statistic is also seeking for regularity in irregularity. Outcome of a coin toss every time is random and uncertain, because it is considered in a local domain. But the expected consequences of this phenomenon, when the event is repeated, is steady and predictable [109]. Existence of such a regularity allows the gambling industry to survive, otherwise no investor would be willing to invest in such an industry. In fact, gambling for someone who gambles is random (because he lies in the local scale) but for the owner of the casino, the gambling phenomenon is predictable and reliable (because he lies in the larger (global) scale and because the gambling phenomenon has order at that scale). Many historical events in the scale of 20 years may seem completely random and stochastic, whereas it is possible that in the scale of 200 years, 2,000 years, or 20,000 years, there is a specific period or an order that emerge. Scientific approaches build on chaos theory can change the scale in which the events are considered in a way that their structural order can be discovered [106].

In the area of mathematics, complex systems has been assumed as a deterministic systems with chaos [105], [110], [111], although complex systems can be defined in many other ways. There are different definitions depending on the field of science. Complex systems in mechanics is defined: "A complex system is a damped, driven system (for example, a harmonic oscillator)

whose total energy exceeds the threshold for it to perform according to classical mechanics but does not reach the threshold for the system to exhibit properties according to chaos theory" [112]. Complex systems have many degrees of freedom, many elements that are partially but not completely independent. The study of complex systems is concerned with both the structure and the dynamics of systems and their interaction with their environment [113]. In other words, in order to have a complex system, two or more components are needed, which are joined in such a way that it is difficult to separate them [105].

Another important term in the field of complexity is Emergence, which is:

1. How behaviour at a larger scale of the system arises from the detailed structure, behaviour and relationships on a finer scale.

2. What a system does by virtue of its relationship to its environment that it would not do by itself.

3. Act or process of becoming an emergent system.

Emergence refers to all the properties that we assign to a system that are really properties of the relationship between a system and its environment [114]. For these kind of systems, statistical techniques, agent based models, individual-based models and evolutionary models are useful [115], [116]. Different studies have been shown that agent-based modeling (ABM) can reveal the emergence phenomenon [116], [117]. An ABM is a model for simulating the actions and interactions of agents with a view to assessing their effects on the system as a whole. It combines elements of complex systems, emergence, multi-agent systems, and evolutionary programming [118].

In this dissertation, like much research in the area of mathematics, complex systems have been assumed to be deterministic systems with chaos. We tried to approach complex systems with this specific perspective. The key is to find what is the main characteristic of chaotic systems and how we can characterize and quantify key features in chaotic systems. In order to show how a chaotic behaviour emerges, we present a simple chaotic system as following.

3.1. Simple chaotic system

The logistic equation is a common model of population growth where the rate of reproduction is proportional to both the existing population size and the amount of available resources [105]. A population size in year $n+1$ is proportional to the size of the population in year n [105]. It is

perceived that a given population has a limit: the maximum size of population beyond which the population cannot sustain itself. This limit is called the "carrying capacity" of the environment for this population. Observation has suggested that a population close to the carrying capacity will emerge to have a large drop in size the following year, because of limited amount of resources.

The logistic model, commonly used in population modeling, suggests that the size of a population in year $n+1$ should be not only proportional to the size in year n, but also proportional to how close the population is to the carrying capacity. Consider the size of a population as a fraction of the carrying capacity: The sized of population will be a number between 0 and 1, 1 representing the maximum size, and 0 representing extinction. Thus, the logistic model is:

$$f(x) = a\ x(1-x) \qquad (3\text{-}1)$$

Assume that the constant a has the value 1.7 (initial population size is 0.1). The population begins to oscillate and then tend to stabilize in the following years. After the 17th year, the population has stabilized at 0.411. With the function $f(x) = 2.8\ x(1-x)$ and an initial population size of 0.2, stable population is 0.634 after about 15 years. For a specific value of a, it doesn't matter what is the size of population, the population will finally stabilize at the same size.

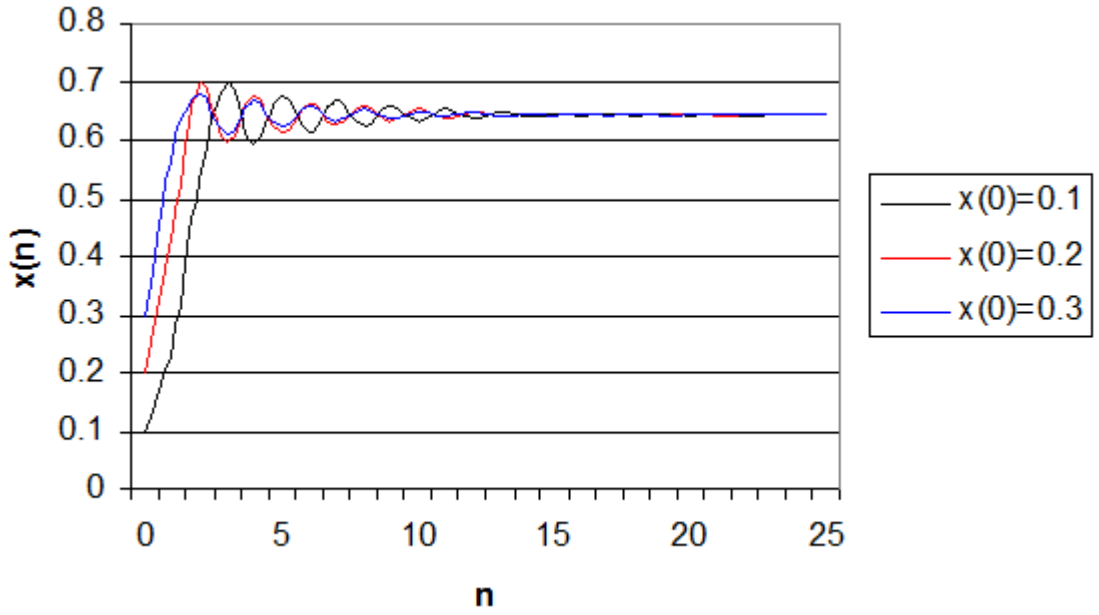


Figure 3-1. Logistic Map with *a* = 2.8. The population will eventually stabilize.

Now consider the logistic population model $f(x) = 3.2\ x(1-x)$, with an initial population size of 0.15. After couple of iterations, the size of population alternates between two values: 0.799 and 0.513. In fact, the initial population size is once again irrelevant. These population sizes form what is called an attracting cycle of period 2. This "period 2" behaviour occurs with all values of a between (approximately) $a = 3$ and $a = 3.4$.

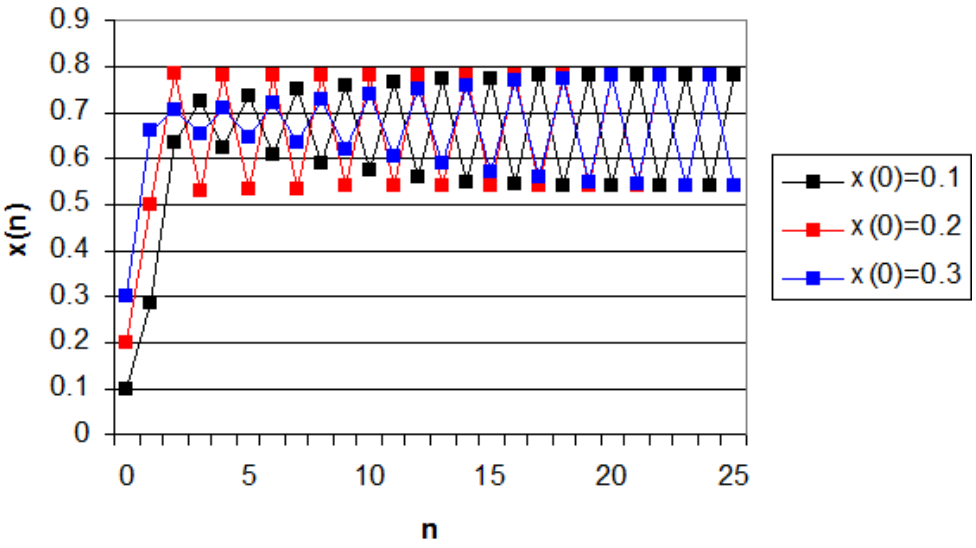


Figure 3-2. Logistic Map with *a* = 3.2. The population oscillate between two points.

Now we consider the function $f(x) = 3.5\ x(1-x)$, starting with any initial x value. The orbit (the population growth curve during time) is attracted to a cycle of period 4, specifically, 0.826, 0.501, 0.874, and 0.382.

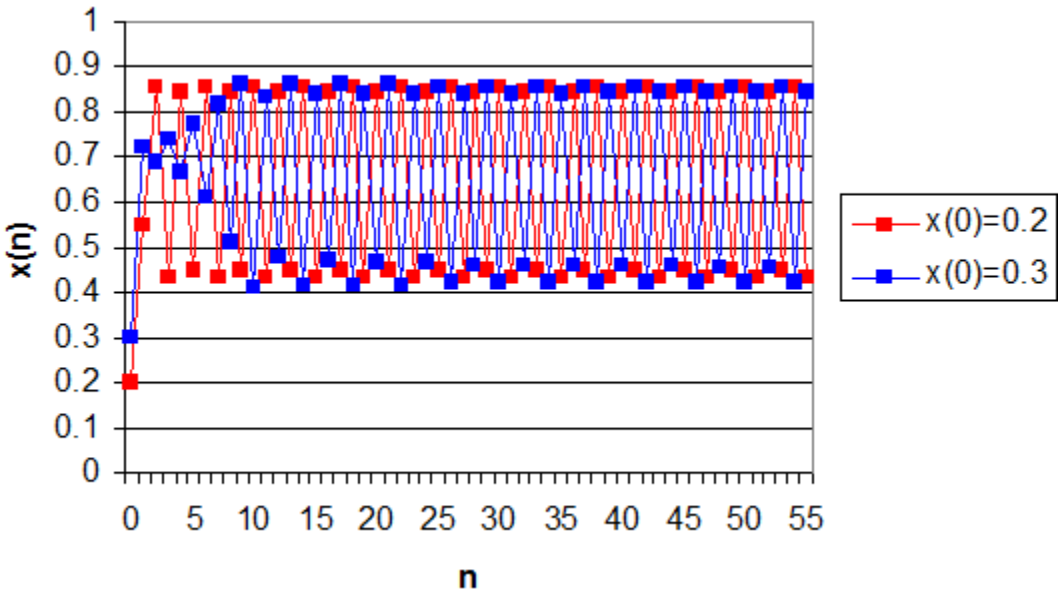


Figure 3-3. Logistic Map with *a* = 3.5. The population oscillate between four points.

Not all values of *a* lead to stable or periodic behaviour. For instance, *a* = 3.8 leads to completely different behaviour of the population. For a population with *a* = 3.8, there is absolutely no regularity to the population sizes in subsequent years. All possible population sizes between approximately 0.25 and 0.92 will be taken on by this population, if we wait long enough. Moreover, the size of the population in any given year now seems to be very highly dependent on the initial population size, a situation totally unlike the regular, periodic behaviour that we noticed for the other values of *a* that we have considered. As an example, consider one population with *a* = 3.8 and initial population size 0.2, and a second population with *a* = 3.8, but initial size 0.3 (see Figure 3-4) [105].

Although the populations start off fairly close in size, after about ten years they are quite different. And by 12 or 13 years, they are so different that there is no relationship whatsoever. Rather than having stable or periodic behaviour, as we observed in the logistic model with values of *a* equal to 1.7, 2.1, 3.2 and 3.5, the behaviour observed with *a* = 3.8 is what is called chaotic. In the above example, the two populations differed by only a small amount, an amount that could easily be due to an error in measurement, and then progressed in different ways.

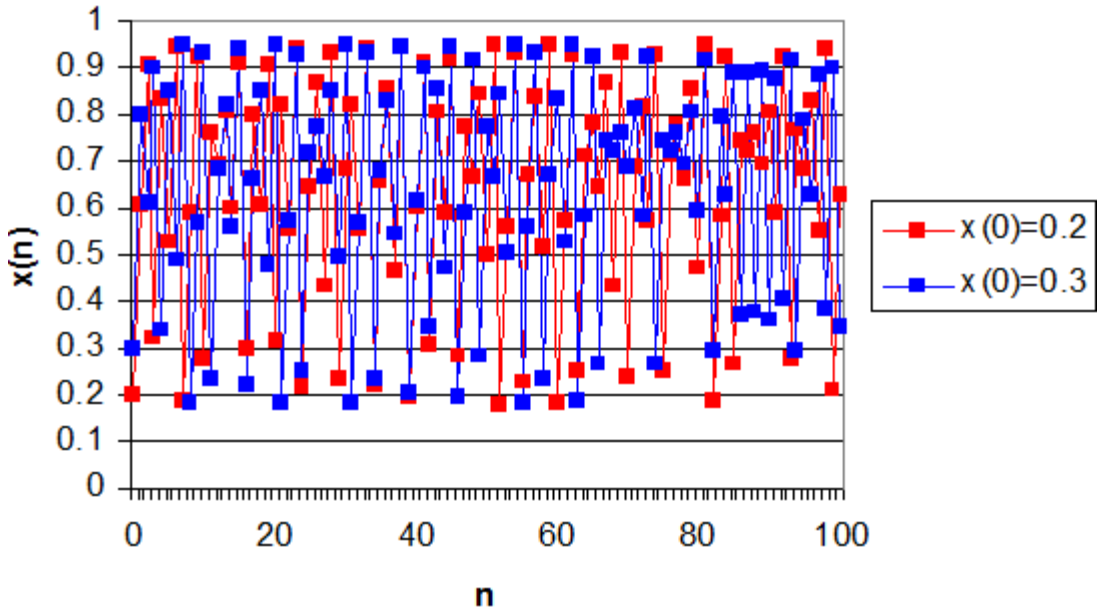


Figure 3-4. Logistic Map with *a* = 3.8. The behaviour of population is non-periodic, bounded and deterministic (chaotic).

When observing the long-term behaviour of dynamical systems, the states of the system approach particular possible solutions. In other words, the phase space of the system evolves to a comparatively small region, which is indicated by the attractor. Geometrically, simple attractors may be fixed points. Another form would be the limit cycle in which the solution oscillates between a sequence of values periodically. These simple attractors have in common that they have an integer dimension in the phase space. The structure of so-called strange attractors reflects the behaviour of chaotic systems – they cannot be described with a closed geometrical form and therefore, since they have a non-integer dimension, are fractals (see below) [119]. Well-known examples of strange attractors as a representation for the limiting values of nonlinear equation systems are the Henon attractor, the Rossler attractor and the Lorenz attractor (Figure 3-7) [105].

In order to analyze chaotic systems, there are two main approaches. First, we can analyze the strange attractor of chaotic systems in terms of fractal analysis. Second, we can analyze the time series behaviour to see how sensitive they are to initial conditions to extract the level of chaos.

3.2. Self-Similarity

Self-similarity strongly suggests that a process, or a structure, and a part of it appear to be the same when compared. A self-similar process is infinite and it is not, in general, something that can be directly visually detected. When an object does not display perfect self-similarity, it is said to have an approximate self-similarity. For instance, a coastline is a self-similar object, but it does

not have perfect self-similarity[120]. It is not only natural fractals that show approximate self-similarity; the Mandelbrot set is another example. Identical patterns are not visible straight away, but when magnified, smaller versions of the same patterns appear at all levels of magnification (see Figure 3-5) [121]. Geometrical similarity is a trait of the space-time metric, whereas physical similarity is a property of the matter fields. The classic forms of geometry do not have this trait; a circle if on a large enough scale will look like a straight line. This is why people believed that the world was a flat cookie, the earth just looks that way to humans [113], [120]. One well-known example of self-similarity and scale invariance is fractals, patterns that form of smaller objects that look the same when magnified. Many natural forms, such as coastlines, fault and joint systems, folds, topographic features, turbulent water flows, drainage patterns, trees, leaves, bacteria cultures, blood vessels, roots, lungs and even universe, look alike on many scales [113], [120]. It appears as if the underlying forces that produce the network of rivers, creeks, streams and rivulets are the same at all scales, which results in the smaller parts and the larger parts looking alike, and these looking like the whole [120], [122].

3.2.1. Self-organization

Self-organization is a system where some kind of global order arises out of the local interactions between the elements of an initially disordered system [120], [122]. Self-organization presents useful models for many complex characteristics of the natural world, which are characterized by fractal geometries, self-similarity of structures, and power law distributions. Openness to the environment and coherent behaviour are necessary conditions for self-organization [123].

Because of a common conceptual framework, self-organizing processes are characterized by self-similarity and fractal geometries, in which similar patterns are repeated with different sizes or time scales without changing their essential meaning. Similar geometric patterns are repeated at different sizes and are expressive of a fundamental unity of the system such as braiding patterns ranging from streambeds to root systems and the human lung [123].

Systems as diverse as metabolic networks or the world wide web are best described as networks with complex topology. A common property of many large networks is that the vertex connectivity follows a scale-free power-law distribution (see section 3.2.2). This feature is a consequence of two generic mechanisms shared by many networks: networks expand continuously by the addition of new vertices, and new vertices attach preferentially to already well connected sites. A model based on these two mechanisms reproduces the observed stationary

scale-free distributions, indicating that the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems [124].

3.2.2. Power laws

Power law is one of the common signatures of a nonlinear dynamical system. With power laws it is possible to express self-similarity of the large and small, i.e., to unite different sizes and lengths. In fractals, for example, there are many more small structures than large ones [119]. Their respective numbers are represented by a power law distribution. A common power law for all sizes demonstrates the internal self-consistency of the fractal and its unity across all boundaries. The power law distributions result from a commonality of laws and processes at all scales [123]. The scaling relationship of power laws applies widely and brings into focus one important feature of the systems considered [125]. When using the power laws, one must notice that statistical data for a phenomenon that obeys one of the power laws (exponential growth) is biased towards the lower part of the range, whereas that for a phenomenon with saturation (logarithmic growth) tends to be biased towards the upper part of the range [105].

The natural world is full of power law distributions between the large and small: Earthquakes, words of the English language, and coastlines of continents. For example power laws define the distribution of catastrophic events in Self-Organized Critical (SOC) systems. If a SOC system shows a power law distribution, it could be a sign that the system is at the edge of chaos, i.e., going from a stable state to a chaotic state. A power law distribution is also a litmus test for self-organization, self-similarity and fractal geometries [105], [123].

The power laws and fractal dimensions are just two sides of a coin and they have a tight relationship joining them the together [119]. The relationships can be clarified with a mathematical discussion. The general equation for power law is shown in (3-2). It is a mathematical pattern in which the frequency of an occurrence of a given size is inversely proportional to some power n of its size:

$$y(x) = x^{-n} \tag{3-2}$$

Note that

$$y(\lambda x) = (\lambda x)^{-n} = \lambda^{-n} x^{-n} = \lambda^{-n} y(x) \tag{3-3}$$

It turns out that the power law can be expressed in “linear form” using logarithms:

$$\log(y(x)) = -n\log(x) \quad (3\text{-}4)$$

where the coefficient n represents the fractal dimension [2]. The mathematical relationship connecting self-similarity to power laws and to fractal dimension is the scaling equation. For an self-similar observable $A(x)$, which is a function of a variable x, a scaling relationship holds:

$$A(\lambda x) = \lambda^s A(x) \quad (3\text{-}5)$$

where λ is a constant factor and s is the scaling exponent, which is independent of x. Looking at (3-3), it is clear that the power law obeys the scaling relationship.

The data emerging from the combination of self-similarity and self-organization cannot be described by either Normal or exponential distribution. The reason is, that emergence of order in complex systems is fundamentally based on correlations between different levels of scale. The organization of phenomena that belong at each level in the hierarchy rules out a preferred scale or dimension. The relationships in this type of systems are best described by power laws and fractal dimension [122].

3.2.3. Fractal Dimension

Fractals are characterized by three concepts: Self-similarity, response of measure to scale, and the recursive subdivision of space. Fractal dimension can be measured by many different types of methods. A common property of all these methods is that they all rely heavily on the power law plotted to logarithmic scale, which is the property relating fractal dimension to power laws [105], [126]. One definition of fractal dimension D is the following equation:

$$D = \log N / \log(1/R) \quad (3\text{-}6)$$

where N is the number of segments created, when dividing an object, and R is the length of each of segments. This equation relates to power laws as follows:

$$\log(N) = D.\log(1/R) = \log(R^{-D}) \quad (3\text{-}7)$$

so that

$$N = R^{-D} \quad (3\text{-}8)$$

It is simple to obtain a formula for the dimension of any object provided. The procedure is just to determine in how many parts it gets divided up into (N) when we reduce its linear size, or scale it down ($1/R$).

By applying the equation to line, square and cubicle, we get the following results; For a line divided in 4 parts, N is 4 and R is 1/4, so dimension $D = \log(4)/\log(4) = 1$. For a square divided in four parts N is 4, R is 1/2, and dimension $D = log(4)/log(2) = 2 \cdot \log(2)/\log(2) = 2$. And for a cubicle divided in 8 parts, N is 8, R is 1/2 and dimension $D = log(8)/log(2) = 3 \cdot log(2)/log(2) = 3$.

The following series of pictures (Figure 3-5) represents iteration of the Koch curve. By applying equation (3-7) to the Koch curve, it is evident, that the dimension is not an integer, but instead between 1 and 2. In fact, the dimension is always 1.26185, regardless of the iteration level. Hence,

$$D = \log N/\log(1/R) = 1.26185, \qquad (3\text{-}9)$$

Which can also be written as

$$N = (1/R)^{1.26185} \qquad (3\text{-}10)$$

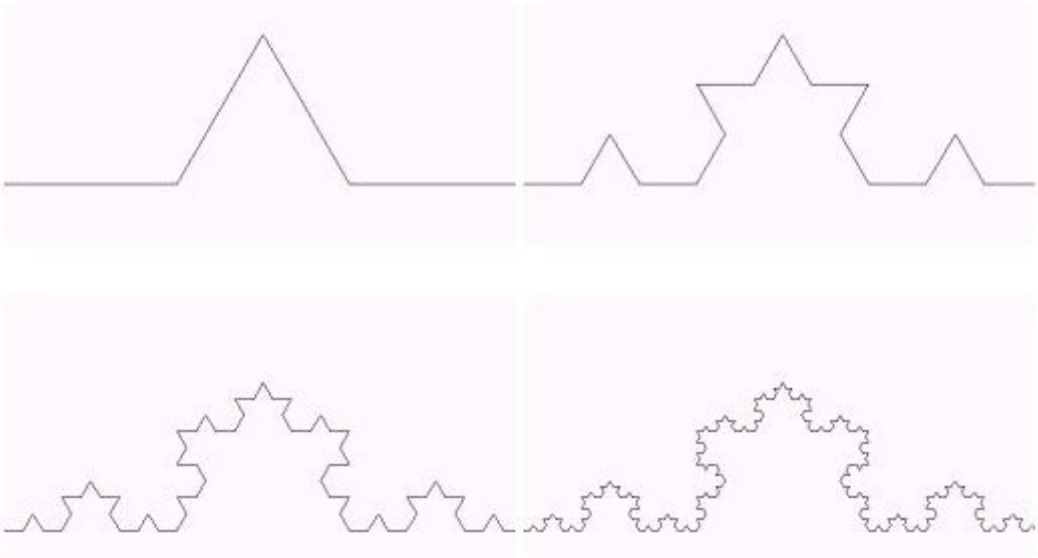


Figure 3-5. The Koch curve illustrates self-similarity. As the image is enlarged, the same pattern re-appears.

The formulas above indicate that N and R are related through a power law. In general, a power law is a nonlinear relationship, which can be written in the form $N = a(1/R)^D$, where D is

normally a non-integer constant and a is a numerical constant, which in the case of the Koch curve is 1.26.

Another way of defining the fractal dimension is box counting. In box counting the fractal is put on a grid, which is made of identical squares having size of side *h*. Then the amount of non-empty squares, *k*, is counted. The magnification of the method equals to 1*/h* and the fractal dimension is defined by equation [113]:

$$D = \log(k) / \log(1/h) \qquad (3\text{-}11)$$

Some studies have shown that when the state-space dimension of a system can be express by a fractal dimension function, deterministic properties of the system can be determined by low values of fractal dimension and stochastic properties of a system can be determined by high values of fractal dimension [127]. Among different versions of fractal dimension that are used for regularity analysis (i.e. detection of deterministic behaviour versus stochastic behaviour), the Higuchi fractal dimension is a precise and applicable mechanism to estimate self-similarity that also gives a stable value for the fractal dimension [128]–[130]. In addition, this method is considered to be the standard for representing the irregularity of time series [131]. The Higuchi fractal dimension can be calculated directly in the time domain and is therefore simple and fast. It has also been proved to be an accurate estimator of fractal dimension for samples as short as 150-500 data points.

In the past few years, fractal analysis techniques have gained increasing attention in medical signal and image processing. For example, analyses of encephalographic data (EEG) and other bio-signals are among its applications [24], [132]. The same method may also be used in other biomedical applications [131], [132]. The correlation dimension [133] has also been used associated with the Gaussian kernel Algorithm (GKA) method [133] for analysis with results similar to those of Higuchi fractal dimension.

3.2.3.1. Higuchi Fractal Dimension method

In Higuchi's method [128], which is used for fractal dimension calculation, a new time series, x_m^k, needs to be constructed from the input time series, *x(1), x(2), …, x(N)* as follows:

$$x_m^k = \left\{ x(m), x(m+k), x(m+2k), \ldots, x\left(m+\left\lfloor \frac{N-m}{k} \right\rfloor .k\right) \right\} \quad , \quad m=1,2,\ldots,k \qquad (3\text{-}12)$$

Where both *m* and *k* are integers and $\lfloor \ \rfloor$ is Gauss' notation. *m* is the initial time, and *k* is the interval time. For a time interval *k*, we get *k* sets of new time series (for example if k=3, there are 3 time series: x_1^3, x_2^3, x_3^3). The length of the curve $L_m(k)$ is defined as follows:

$$L_m(k) = \frac{\sum_{i=1} \left| x(m+ik) - x(m+(i-1)k) \right| \ (N-1)}{\left\lfloor \frac{N-m}{k} \right\rfloor k} \qquad (3\text{-}13)$$

N is the number of samples and $\frac{N-1}{\left\lfloor \frac{N-m}{k} \right\rfloor k}$ is the normalization factor. The length of the curve for the time interval *k, L(k),* is defined as the average value over k sets of $L_m(k)$. If $L(k) \propto k^{-D}$, then the curve is a fractal with dimension *D*.

Deterministic time series are identified by low values of Higuchi fractal dimension and stochastic properties of a time series are identified by high values of Higuchi fractal dimension. In the current implementation, the range is between 0 and 1, with the values close to 0 for deterministic time series and the values close to 1 for random time series.

3.2.3.2. Correlation Dimension

Correlation dimension is an extension of the usual notion of dimension to objects with a fractional dimension [134]. In dimensions one, two, three or more it is easily established, and intuitively obvious, that a measure of volume V(ε) (e.g. length, area, volume and hyper-volume) varies as $V(\varepsilon) \propto \varepsilon^d$

where ε is a length scale (e.g. the length of a cube's side or the radius of a sphere) and d is the dimension of the object. For a general fractal, it is natural to assume a relation like $V(\varepsilon) \propto \varepsilon^d$ holds true, with its dimension given by:

$$d \propto \frac{\log V(\varepsilon)}{\log \varepsilon} \qquad (3\text{-}14)$$

And therefore:

$$d \approx \lim_{\varepsilon \to 0} \frac{\log V(\varepsilon)}{\log \varepsilon} \qquad (3\text{-}15)$$

Let $\{z_t\}$ be an embedding of a time series in R^d. We therefore define the correlation function, $C_N(\varepsilon)$, by

$$C_N(\varepsilon) = \binom{N}{2}^{-1} \sum_{0 \le i < j \le N} I\left(\left\| z_i - z_j \right\| < \varepsilon\right) \qquad (3\text{-}16)$$

Here I(X) is the indicator function, which, has a value of 1 if condition X is satisfied and 0 otherwise, and ‖ . ‖ is the usual distance function in R^d. N is the total number of points, The correlation dimension d is then defined as the slope of the line log(CN(ε)) versus log(ε) at small scales where $\varepsilon \to 0$.

$$C_N(\varepsilon) \propto \varepsilon^d,$$
$$d = \lim_{\varepsilon \to 0} \lim_{N \to \infty} \frac{\log C_N(\varepsilon)}{\log \varepsilon} \qquad (3\text{-}17)$$

Although this definition of correlation dimension is valid, for reliable estimation of correlation dimension a Gaussian Kernel algorithm (GKA) method [135] should be used.

Gaussian Kernel Algorithm takes an entirely different approach to modeling the noise in a signal. We return to the original definition of correlation dimension, i.e. Eqs. (5) and (6). Since the observations x_n are contaminated by noise, one cannot know z_n precisely [44]. Therefore, computation of $I(|z_i - z_j| < e)$ in Eq. (5) is actually somewhat fuzzy. Rather than adopting contemporary density estimation to improve the estimate of the distribution, one can model this uncertainty by replacing the hard indicator function I(.) with a continuous one. The choice (implied by its title) of the Gaussian Kernel algorithm is the Gaussian basis function

$\exp\frac{-\|z_i - z_j\|^2}{4\varepsilon}$. Details of this algorithm are described by [136] and an efficient implementation of this technique is presented by [135].

Correlation dimension estimates from the GKA suggest the complexity of the system attractor (the number of active degrees of freedom) [134]. The correlation dimension gives an estimate of system complexity. The GKA separates the data into purely deterministic and stochastic components [133].

3.2.4. Multifractal Analysis

A multifractal process is a generalization of a fractal process (monofractal) in which one exponent (fractal dimension) is not enough to explain its dynamics; instead, a continuous spectrum of exponents (singularity spectrum) is needed. Multifractal analysis uses the mathematical basis of multifractal theory to explore datasets, often in conjunction with other methods of fractal analysis. The technique requires to illustrate how scaling varies over the dataset. The techniques of multifractal analysis have been applied in a variety of practical situations such as predicting earthquakes and interpreting medical images [137]–[139].

In a multifractal system *S*, the behaviour around any point is described by a local power law:

$$S(x+a) - S(x) \approx a^{h(x)} \qquad (3\text{-}18)$$

The exponent *h(x)* is called the singularity exponent, as it demonstrates the local degree of singularity around the point *x*. The ensemble formed by all the points that share the same singularity exponent is called the singularity manifold of exponent *h*, and is a fractal set of fractal dimension D(h). The curve D(h) versus h is called the singularity spectrum and fully describes the statistical distribution of the variable *S*.

3.2.4.1. The Continuous Wavelet Transform (CWT) and wavelet-based multifractal analysis

Multifractal analysis using the wavelet transform is a powerful tool for detecting self-similarity [140]. The wavelet transform is a convolution product of the data sequence (a function f(x), where x is usually a time or space variable) with the scaled and translated version of the mother wavelet, ψ(x) [140]. The scaling and translation are performed by two parameters; the scale

parameter s stretches (or compresses) the mother wavelet to the required resolution, while the translation parameter b shifts the analyzing wavelet to the desired location:

$$(Wf)(s,b) = \frac{1}{s}\int_{-\infty}^{+\infty} f(x).\psi^*(\frac{x-b}{s})dx \qquad (3\text{-}19)$$

where s, b are real, s > 0 for the continuous version (CWT) and ψ* is the complex conjugate of ψ. The wavelet transform acts as a microscope: it reveals more and more details while going towards smaller scales, i.e. towards smaller s values. The mother wavelet (ψ(x)) is generally chosen to be well localized in space (or time) and frequency [141].

Usually, ψ(x) is only required to be of zero mean, but for the particular purpose of multifractal analysis ψ(x) is also required to be orthogonal to some low order polynomials, up to the degree n:

$$\int_{-\infty}^{+\infty} x^m \psi(x)dx = 0, \qquad \forall m, \quad 0 \le m < n \qquad (3\text{-}20)$$

Thus, while filtering out the trends, the wavelet transform can reveal the local characteristics of a signal, and more precisely its singularities. (The Hölder exponent can be understood as a global indicator of the local differentiability of a function.) By preserving both scale and location (time, space) information, the CWT is an excellent tool for mapping the changing properties of non-stationary signals.

It can be shown [141] that the wavelet transform can reveal the local characteristics of f at a point x_o. More precisely, we have the following power law relation:

$$W^{(N)} f(s, x_0) \sim |s|^{h(x_0)} \qquad (3\text{-}21)$$

where h is the Hölder exponent (or singularity strength). The symbol "(N)", which appears in the above formula, shows that the wavelet used (ψ(x)) is orthogonal to polynomials up to degree n (including n). The scaling parameter (the so-called Hurst exponent) is estimated when analyzing time series by using "monofractal" techniques. It is a global measure of self-similarity in a time series, while the singularity strength h can be considered a local version (i.e. it describes "local similarities") of the Hurst exponent. In the case of monofractal signals, which are characterized by the same singularity strength everywhere (h(x) = ct), the Hurst exponent equals h. Depending on the value of h, the input series could be long-range correlated (h > 0.5), uncorrelated (h = 0.5) or anti-correlated (h < 0.5).

To characterize the singular behaviour of functions, it is sufficient to consider the values and position of the Wavelet Transform Modulus Maxima (WTMM) [142]. The wavelet modulus

maxima is a point (s_0, x_0) on the scale-position plane, (s,x), where $|Wf(s_0, x_0)|$ is locally maximum for x in the neighborhood of x_0. These maxima are located along curves in the plane (s,x). The WTMM representation has been used for defining the partition function based multifractal formalism [143], [144].

Let $\{u_n(s)\}$, where n is an integer, be the position of all local maxima at a fixed scale s. By summing up the q's power of all these WTMM, we obtain the partition function Z:

$$Z(q,s) = \sum_n \left|Wf(u_n, s)\right|^q \qquad (3\text{-}22)$$

By varying q in Eq. (4), it is possible to characterize selectively the fluctuations of a time series: positive q's accentuate the "strong" inhomogeneities of the signal, while negative q's accentuate the "smoothest" ones. We have employed a slightly different formula to compute the partition function Z by using the "supremum method", which prevents divergences from appearing in the calculation of Z(q,a), for $q < 0$ [144].

Often scaling behaviour is observed for Z(q,s) and the spectrum $\tau(q)$, which describes how Z scales with s can be defined:

$$Z(q,s) \sim s^{\tau(q)} \qquad (3\text{-}23)$$

If the $\tau(q)$ exponents define a straight line, the analyzed signal is a monofractal; otherwise the fractal properties of the signal are inhomogeneous, i.e. they change with location, and the time series is a multifractal. By using the Legendre transformation we can obtain the multifractal spectrum D(h) from $\tau(q)$.

3.3. Chaoticity Analysis

Nowadays, measures based on the deterministic chaos are effective tools for characterizing the time series behaviour [145],[28]. Deterministic chaos shows the sensitivity of deterministic behaviour to slight change of initial condition [25], [146]. It has been shown that the evaluation of chaoticity (level of chaos) is an important issue in several applications. There are lots of publications that justify without chaoticity, biological systems might be unable to get discriminated between different stages and thereby different modes of operation [26] (for example epileptic seizures can be detected by using measure of chaoticity [28]). As some researches pointed out, methods based on the largest local Lyapunov exponent can detect the changes of the chaoticity in the biological time series [27]. The chaoticity of the process has to

characterize the predictability of future values of the time series. In the stationary case, the chaoticity quantity can be directly distinguished by the largest Lyapunov exponent (LLE) [147]. Other than LLE method, a few studies have been focused on chaoticity measure to investigate the level of chaos in different time series. In many cases, the duration of a biological time series does not allow for the generated signal to be treated as stationary [98], [148]. Therefore, the application of the standard method of nonlinear system theory is often questionable. The data that come from real world application always have significant amount of noise, which make it major challenge for analysis. Among methods that used for chaoticity measure, it is unclear how presence of noise influence the performance of those methods.

3.3.1. P&H Method

This method is a new and efficient method for detecting the random signals from deterministic signal, which is based on the Poincaré section and the Higuchi fractal dimension [149]. Moreover, this method can be used to detect chaotic behaviour in a signal. This method recently has been applied to some biomedical data [150]. P&H method is composed of several steps that can be summarized as follows.

The first step is to intersect the time series trajectory with the Poincaré section. This intersection induces a set of points that indicate dynamic flow. This intersection leads to a series (P) specified by these points. Applying the Higuchi fractal dimension to the P series yields to a vector $L(k)$. This vector is a basis for decision making about the time series properties. Figure 3-6 gives an example of a Poincaré section for a 3-D flow (Γ).

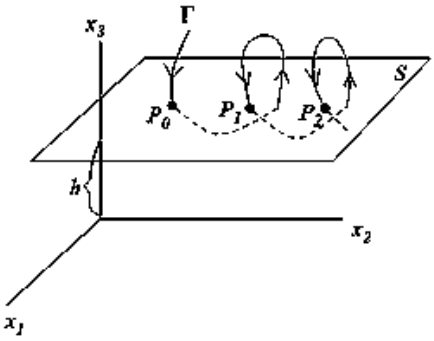


Figure 3-6. Intersection between the flow (Γ) and the Poincaré section (S) generating the set of points P = (P0, P1, P2).

P_0, P_1, P_2, and so on come from intersections of the flow with the Poincaré section (see Figure 3-7). This intersection leads to a set of points P = (P_0, P_1, P_2, ...) that indicate the dynamic flow behaviour [151]–[153]. Many of the time series features, such as periodicity and quasi-periodicity of the original curve (Γ), could still be extracted from P.

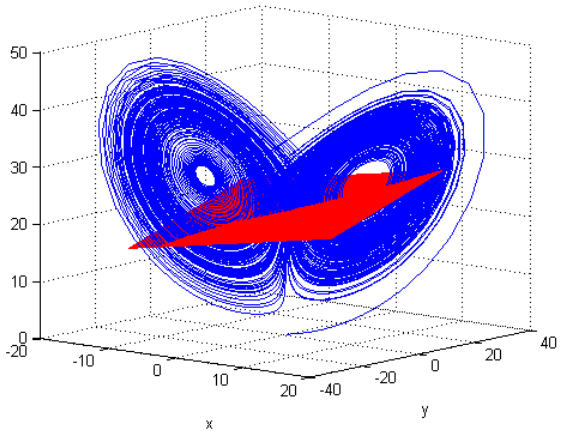


Figure 3-7. Intersection of Lorenz attractor and Poincaré section. The Poincaré map is product of this stage.

In order to use P&H method on one-dimensional signals, these time series must be embedded according to the embedding theorem [154]. Then Higuchi fractal dimension is computed based on formula (3-3) and (3-4) applied on time series P. The resulting $L_m(k)$ contain information about the properties of Γ. A figure with $Log(L_m(k))$ as the vertical axis and k as the horizontal axis could allow defining criteria for detecting stochastic signals from deterministic signals.

$L_m(k)$ is obtained by summing approximately (N/k) terms. If the normalization factor is ignored, then as k increases, the number of summed terms decreases. If the x_m^k time series were random, then the positive subtraction of consequent terms yields a value that is also random. In the value of $L_m(k)$, (N/k) random terms are summed, and in the value of $L_m(k+1)$, N/k+1 random terms are summed. As k increases, the number of summands decreases. Therefore, the value of $L_m(k)$ decreases. This property allows to propose a criterion to distinguish between random time series and deterministic time series [149]. For random time series, a decreasing pattern (Figure

3-8b) is observed, and for chaotic time series, a non-decreasing pattern (such as in Figure 3-8a) is observed.

For chaotic time series, it has been shown that, because of the stretching and folding property [105], there is at least one k value for which the value of $L_m(k+1)$ is greater than $L_m(k)$ and consequently the $L_m(k)$ vector indicates a zigzag pattern [149]. According to this property, $L(k)$ is greater for any odd values of k than for the previous even values of k.

As it is apparent in Figure 3-8, the P&H method generates a decreasing monotonic pattern for all stochastic signals. For chaotic time series, because of the stretching and folding property, the $L(k)$ vector indicates a zigzag pattern. There are different chaotic signals based on route to chaos (period-doubling or quasi-periodic or intermittent) [105]. This method has been tested to examples from all types of route to chaotic signals [149]. The P&H method output can be easily mapped to negative and positive numbers for stochastic and chaotic time series respectively. While other criterion [155] does not have the ability of identifying discrete maps such as Henon map [156], P&H method shows satisfying results [149].

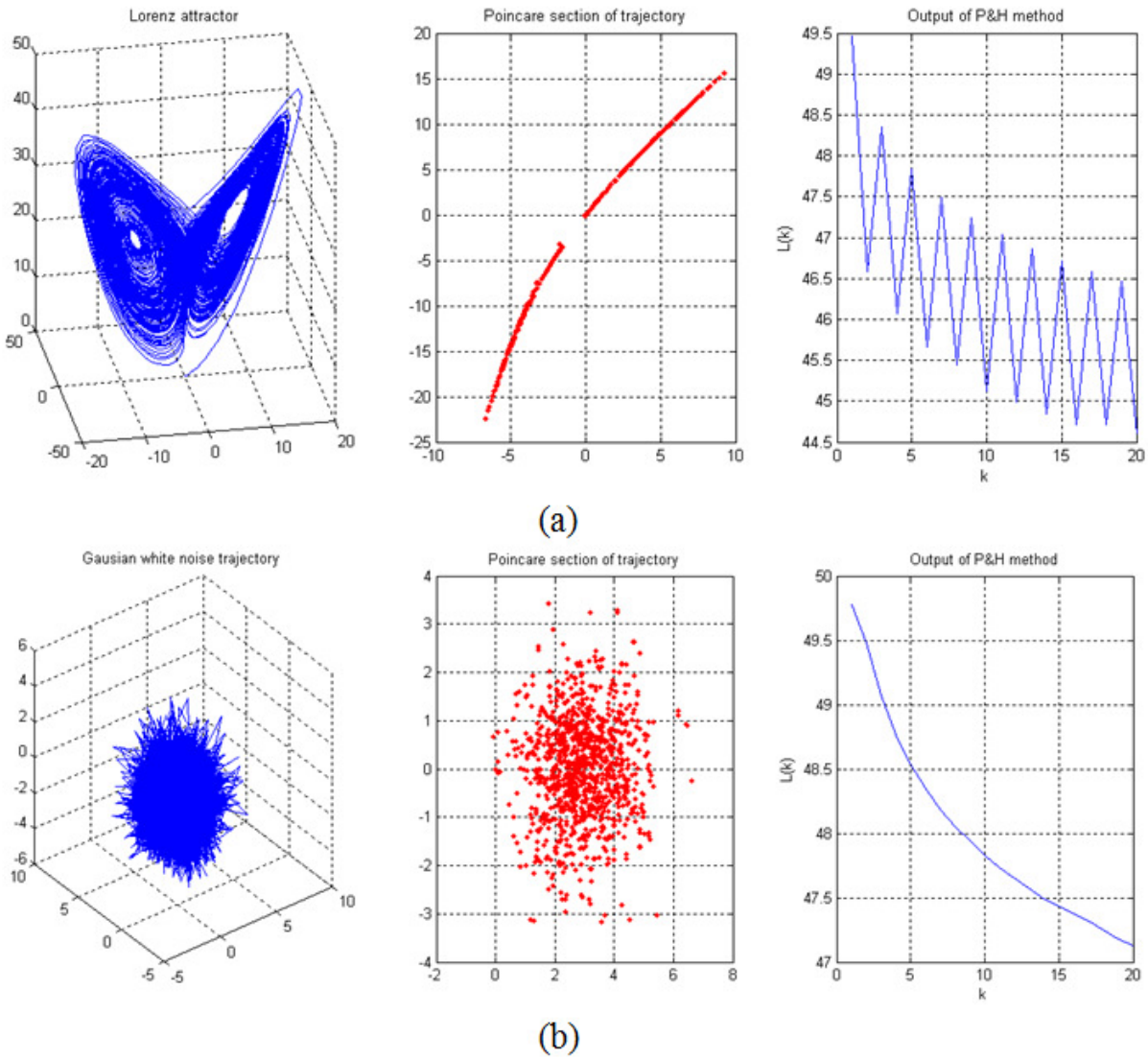


Figure 3-8. Applying of P&H method over Lorenz time series and random time series.

3.3.2. Lyapunov Exponent

Most experts would agree that chaos is the aperiodic, long-term behaviour of a bounded, deterministic system that demonstrates sensitive dependence on initial conditions. For that purpose, quantifying the sensitivity to initial condition [105] could be a good strategy to detect a chaotic behaviour.

Lyapunov exponents quantify the exponential divergence of initially close state-space trajectories and estimate the amount of chaos in a system [157]. A bounded dynamical system with a positive largest Lyapunov exponent is chaotic [105].

Suppose a one-dimensional function $X_{n+1} = f(X_n)$. Imagine two nearby initial points at X_0 and $X_0 + \Delta X_0$, respectively. After one application of the function $f(X_n)$, the points are separated by

$$\Delta X_1 = f(X_0 + \Delta X_0) - f(X_0) \cong \Delta X_0 f'(X_0) \tag{3-24}$$

Where $f' = df/dX$. Now the local Lyapunov exponent λ at X_0 is defined such that $e^{\lambda} = |\Delta X_1 / \Delta X_0|$, or

$$\lambda = \ln|\Delta X_1 / \Delta X_0| = \ln|f'(X_0)| \tag{3-25}$$

To obtain the largest Lyapunov exponent, an average of the above equation is taken over many iterations.

$$\lambda = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \ln|f'(X_n)| \tag{3-26}$$

The largest Lyapunov exponent determines the average exponential rate of separation of two nearby initial conditions, or the average stretching of the space. A positive value shows chaos since it shows high sensitivity to initial conditions [105].

The different methods that have been proposed for computing Lyapunov exponents from time series can be divided in two classes: Jacobian-based methods and direct methods. Direct methods directly estimate the divergent motion of the reconstructed states without fitting a model to the data [158]. The implementation of Lyapunov method, which has been used, was proposed by Sato et al. [159] and Kurths and Herzel [160]. The average exponential growth of the distance of neighboring orbits is studied on a logarithmic scale, via the prediction error:

$$p(k) = \frac{1}{Nt_s} \sum_{n=1}^{N} \log_2 \left(\frac{\|X(n+k) - X(nn+k)\|}{\|X(n) - X(nn)\|} \right) \tag{3-27}$$

Where *X(nn)* is the nearest neighbor of *X(n)* (*nn* is a neighbor indices of *n*, which could be *n-1*, *n+1*,...). The dependence of the prediction error *p(k)* on the number of time steps k may be divided into three phases [54]. Phase I is the transient where the neighboring orbit converges to the direction corresponding to the largest Lyapunov exponent. During phase II the distance grows exponentially until it exceeds the range of validity of the linear approximation of the flow. Then

phase III begins where the distance increases slower than exponentially until it decreases again because to the folding in the state space. In phase II, a linear segment with slope λ_1 appears in *p(k)* vs. *k* diagram. This allows an estimation of the largest Lyapunov exponent λ_1 [161]. Figure 3-9 gives an example to determine the largest Lyapunov exponent λ_1 of data by this method [158].

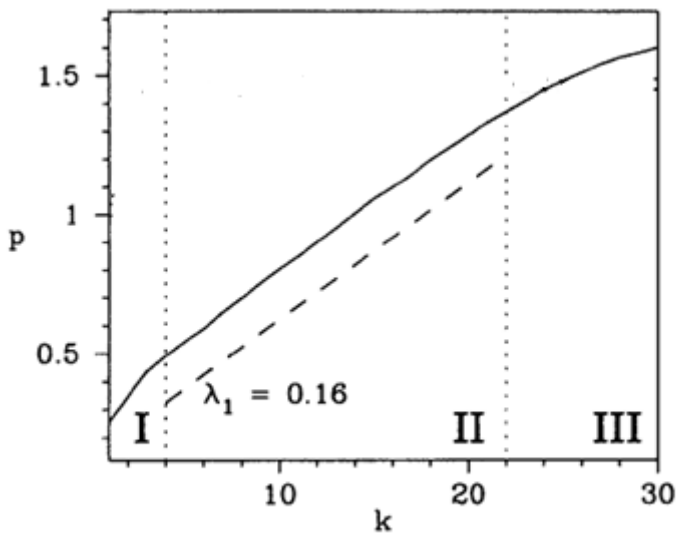


Figure 3-9. Prediction error p for experimental data vs. the number of time steps k. the slope of the solid line in the intermediate range of k gives the largest Lyapunov exponent $\lambda_1 = 0.16$.

With some modification, there are several implementations of Lyapunov exponent when time series are non-stationary [162].

3.3.3. Surrogate data test Method

A simple chaotic system may produce an output time series that passes randomness testing. In contrast, a completely random process with non-uniform power spectrum (correlated noise) may be appear to be a chaotic system. To solve the problem of separating such systems from each other, time series are tested by means of surrogate data. It is rather simple to generate a time series with similar probability distribution as the original data, it suffices to make random sampling in the original time series. Randomizing samples preserve the probability distribution, while it does not preserve the power spectrum and autocorrelation function of the time series. Even if the original data is not uncorrelated and white, surrogate data will have these characteristics. To constitute surrogate data, which have the same power spectrum of X_n, a

surrogate data time series Y_n, with the same Fourier coefficients but with random phase, is generated:

$$X_n \cong \frac{a_0}{2} + \sum_{m=1}^{N/2} (a_m Cos\frac{2\pi mn}{N} + b_m Sin\frac{2\pi mn}{N})$$

$$a_m = \frac{2}{N}\sum_{n=1}^{N} X_n \; Cos\frac{2\pi mn}{N}$$

$$\mathrm{b}_m = \frac{2}{N}\sum_{n=1}^{N} X_n \; Sin\frac{2\pi mn}{N} \qquad (3\text{-}28)$$

$$\mathrm{S}_m = a_m{}^2 + b_m{}^2$$

$$\Rightarrow Y_n = \frac{a_0}{2} + \sum_{m=1}^{N/2} \sqrt{S_m} \; Sin2\pi(\frac{mn}{N} + r_m)$$

Where r_m is a random number between zero and one ($0 \le r_m \le 1$). In this method, the power spectrum does not change, but the probability distribution becomes a Gaussian distribution [163]. A simple solution to generate the same probability distribution is to sort the original and the random phase data samples separately, from smallest to largest. Then the smallest value of surrogate data is assigned to the smallest value of original data and so on, until the end of samples. Finally, a random time series with the identical distribution function of original time series and with only slightly altered power spectrum [155], [158] is obtained.

3.4. Prediction methods

Prediction of future values in a complex time series is a major concern for scientists [164], [165] with applications to various fields of science [165]–[168]. There are many natural phenomena, such as variation in population size, orbits of astronomical objects, and Earth's seismic waves, that require a prediction algorithm for answering important questions. Prediction is an ongoing and pressing problem in the forecasting of economic time series [169]. In the medical sciences there are also many applications for which an efficient prediction algorithm could save lives. A large number of time series gained from the human body can be used as an origin of the decision making process to treat or prevent grave diseases such as epilepsy or Alzheimer's [170]–[173]. Time series analysis of Earth's seismic waves can be used for earthquake prediction [174]. Also prediction of twenty-first century global temperature rise is a valuable information for policy makers and planners [175]. Another application of time series prediction is population projection. Population projections may be used to predict species extinction before they reach a crisis point [4], [176], [177].

It has been shown that data generated by such natural phenomena often follow chaotic behaviour [23]. They are well known to be strongly dependent on initial conditions; small changes in initial conditions can possibly lead to immense changes in subsequent time steps, and are particularly difficult to predict. Since the exact conditions for many natural phenomena are not known and the properties of a chaotic time series are very complex, it is difficult to model these systems.

Most of the existing methods for complex time series prediction are based on modeling the time series to predict future values, although there are other types of methods such as agent-based simulation that models the system generating the time series [178]. Model-based approaches can mainly be classified into two main domains: linear models like ARIMA (AutoRegressive Integrated Moving Average) [164] and nonlinear models like MLP (Multi-Layer Perceptron) [179] and GARCH (Generalized AutoRegressive Conditional Heteroskedasticity) [180]. However, other studies concluded that there was no clear evidence in favor of nonlinear models over linear models in terms of forecast performance [181]. Still, there is no robust procedure that can estimate an accurate model for chaotic time series. For all of these methods, the prediction error increase dramatically with the number of time points predicted [178], [181], [182]. This is why most of the existing methods focus on very short-term prediction to reach a reasonable level of accuracy. None of the existing methods show acceptable accuracy for long-term prediction [183]. For example, for financial time series prediction, most methods can predict only one step ahead, which is not very helpful for acting against a financial recession beforehand [178], [183], [184].

Time series forecasting has fundamental importance on various numbers of problem domains including prediction of earthquakes, financial market prediction, and prediction of epileptic seizures [164], [165]. But a few active research works is going on in long-term time series forecasting [169], [184].

3.4.1. Existing methods

The number of papers that concern time series forecasting has been fairly stable over time. The classic approach is to build an explanatory model from first principles and measure initial data. There are many methods for time series prediction from mathematical models to individuals-based simulation neural networks. The main existing methods are briefly presented in the following subsections.

3.4.1.1. Exponential smoothing

Three decades ago, exponential smoothing methods were considered for extrapolating various types of time series. These methods originated in [185]–[187]. Exponential smoothing was first suggested by Robert Goodell Brown in 1956 [186]. The formulation below, which is the one commonly used, is attributed to Brown and is known as "Brown's simple exponential smoothing". The simplest form of exponential smoothing is given by the formulae:

$$s_t = \alpha x_{t-1} + (1-\alpha) s_{t-1} \qquad (3\text{-}29)$$

where α is the *smoothing factor*, and $0 < \alpha < 1$. In other words, the smoothed statistic s_t is a simple weighted average of the previous observation x_{t-1} and the previous smoothed statistic s_{t-1}. Simple exponential smoothing can be used easily, and it generates a smoothed statistic as soon as two observations are available. By direct substitution of the defining equation for simple exponential smoothing back into itself we find that

$$\begin{aligned} s_t &= \alpha x_{t-1} + (1-\alpha) s_{t-1} \\ &= \alpha x_{t-1} + \alpha(1-\alpha) x_{t-2} + (1-\alpha)^2 s_{t-2} \\ &= \alpha[x_{t-1} + (1-\alpha) x_{t-2} + (1-\alpha)^2 x_{t-3} + (1-\alpha)^3 x_{t-4} + \ldots] + (1-\alpha)^t s_0 \end{aligned} \qquad (3\text{-}30)$$

In other words, as time passes the smoothed statistic s_t becomes the weighted average of a greater and greater number of the past observations x_{t-n}, and the weights assigned to previous observations are in general proportional to the terms of the geometric progression $\{1, (1-\alpha), (1-\alpha)^2, (1-\alpha)^3, \ldots\}$.

3.4.1.2. ARMA Model

The autoregressive moving average (ARMA) models provide a parsimonious description of a stationary stochastic process in terms of two polynomials, one for the auto-regression and the second for the moving average [166]. The ARMA model was explained in the 1951 by Peter Whittle, and it was popularized in the 1971 by George E. P. Box and Gwilym Jenkins [166].

Given a time series of data X_t, the ARMA model is a method for predicting future values in this series. The model consists of two parts, an autoregressive (AR) part and a moving average (MA) part. The model is usually then referred to as the ARMA(p,q) model where p is the order of the autoregressive part and q is the order of the moving average part (as defined below).

The notation AR(p) refers to the autoregressive model of order p. The AR(p) model is written

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \qquad (3\text{-}31)$$

where $\varphi_1,\ldots,\varphi_p$ are parameters, c is a constant, and the random variable ε_t is white noise.

The notation MA(q) refers to the moving average model of order q:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \qquad (3\text{-}32)$$

where the θ_1, ..., θ_q are the parameters of the model, μ is the expectation of X_t (often assumed to equal 0), and the ε_t , ε_{t-1} ,... are again, white noise error terms.

3.4.1.3. ARCH/GARCH Models

A key feature of financial time series is that large (small) absolute returns tend to be followed by large (small) absolute returns, where there are periods, which display high (low) fluctuation. The class of autoregressive conditional heteroscedastic (ARCH) models, introduced by Engle (1982) [188], describe the dynamic changes in conditional variance as a deterministic (typically quadratic) function of past returns. Because the variance is known at time t-1, one-step-ahead forecasts are readily available. Next, multi-step-ahead forecasts can be computed recursively. A more parsimonious model than ARCH is the so-called generalized ARCH (GARCH) model [189] where additional dependencies are permitted on lags of the conditional variance. A GARCH model has an ARMA-type representation, so that the models share many properties.

The GARCH (p, q) model [180] (where p is the order of the GARCH terms σ^2 and q is the order of the ARCH terms ε^2) is given by

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \ldots + \alpha_q \varepsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \ldots + \beta_p \sigma_{t-p}^2 = \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 \qquad (3\text{-}33)$$

α and β are parameters and have to be estimated.

3.4.1.4. Regime-switching models

Vector autoregression (VAR) is an econometric model used to catch the linear interdependencies among multiple time series [190], [191]. VAR models are used for multivariate time series [192].

The structure is that each variable is a linear function of past lags of itself and past lags of the other variables (unlike ARMA which is linear function of past lags of itself).

As an example, suppose that we measure three different time series variables, denoted by $x_{t,1}$, $x_{t,2}$, and $x_{t,3}$.

The vector autoregressive model of order 1, denoted as VAR(1), is as follows:

$$x_{t,1} = \alpha_1 + \varphi_{11} x_{t-1,1} + \varphi_{12} x_{t-1,2} + \varphi_{13} x_{t-1,3} + w_{t,1}$$

$$x_{t,2} = \alpha_2 + \varphi_{21} x_{t-1,1} + \varphi_{22} x_{t-1,2} + \varphi_{23} x_{t-1,3} + w_{t,2} \qquad (3\text{-}34)$$

$$x_{t,3} = \alpha_3 + \varphi_{31} x_{t-1,1} + \varphi_{32} x_{t-1,2} + \varphi_{33} x_{t-1,3} + w_{t,3}$$

the α, φ and w are parameters that has to be computed.

3.4.1.5. Summary

Although linearity is a useful and powerful tool in many areas, it became increasingly clear that linear models are insufficient in many applications. For example, sustained animal population size cycles, weather cycles, energy flow, and amplitude–frequency relations were found not to be suitable for linear models. Increasingly several useful nonlinear time series models were proposed in this period [184]. De Gooijer and Kumar (1992) [193] provided an overview of the developments in this area to the beginning of the 1990s. These authors argued that the evidence for the superior forecasting performance of nonlinear models is patchy.

There is no clear evidence in favour of nonlinear over linear models in terms of forecast performance. The poor forecasting performance of nonlinear models calls for substantive further research in this area[181]. The problem may simply be that nonlinear models are not mimicing reality any better than simpler linear approximations

Chapter 4

4. Modeling applications

In this chapter, we propose several different approaches to investigating questions in theoretical biology. Section 4.1 is related to the study of the effect of partial geographical barriers on speciation rate. In section 4.2, we investigate whether speciation can occur in an artificial system without experimenter-defined functions. We then consider what the main driving forces of speciation are. To answer these questions. several variants of EcoSim has been developed (see section 2.4) and used to explore speciation in the absence of a pre-defined fitness function. In section 4.3, we investigate whether an ensemble method can attain higher accuracy levels for the estimation of species abundance distribution. Finally, section 4.4 develops a methodology to predict the changes in species richness of an ecosystem from its general characteristics.

4.1. Effect of geographical barrier on speciation

The relative contribution of geography and ecology to speciation remains one of the most controversial topics in evolutionary biology. Models of speciation that involve geographically unrestricted gene flow (sympatric speciation) or limited gene flow (parapatric speciation) are often considered unrealistic. The major theoretical problems with models that assume gene flow stem from the antagonism between selection and recombination and from the problem of coexistence [194]. It is generally assumed that while selection acts to maximize the fitness optimum of populations, generating genetic and phenotypic divergence, and recombination continuously shuffles the co-adapted gene complexes and brings populations together. Moreover, sister species that are not sufficiently ecologically divergent are believed to experience competitive exclusion that leads to rapid extinction of emerging lineages [194]. The wide range of theoretical conditions that diminish these major conflicts [195], [196] are considered by many critics to be biologically unrealistic, maintaining the long-lasting debate over the likelihood of speciation with gene flow.

While placing the level of gene flow at the center of speciation debates has been extremely successful in shaping research programs and directions [194], the simple dichotomy of sympatry and allopatry along with the static spatial (biogeographic) context also has the potential to hinder progress in the field. It has been suggested that conclusions reached about the relative importance of various mechanisms of speciation can be drastically different if investigators use explicit geographical-pattern (biogeographic) concepts versus more demographic (population genetic) criteria that imply a strict condition of original panmixia outside the geographic context (e.g.,

sympatric, allopatric) [197]. Cleary, many species have strong schooling or homing behaviours or strong ecological preferences that result in a non-random distribution of genetic diversity at the onset of speciation. Moreover, habitat heterogeneity can often enhance the local structuring of genetic variation. Such strategies make the distinction between sympatric and micro-allopatric speciation scenarios hard to disentangle without very good knowledge of the early stages of speciation. Moreover, the intense debate over the geography of speciation has often left biologically relevant scenarios, such as the intermediate parapatric conditions, out of the research context.

There is no doubt that the complexity of natural systems poses a great challenge when one tries to assign speciation cases to discrete categories. Most species exhibit dynamic changes in distribution that involve population expansions and contractions, fragmentations and secondary contacts across evolutionary relevant time scales [197]–[199]. Moreover, macro-geographic barriers are also ephemeral on a larger geological scale. At a fine local scale, the effect of micro-geographic barriers depends largely on how important the structure of the habitat is for dispersal rates. It has recently been proposed that the complex context of speciation can be better understood outside the framework of a classical geographical definition by focusing on the important evolutionary forces such as gene flow, selection and genetic drift [197]. This is the hypothesis that we are investigating in this chapter using our modelling approach.

4.1.1. Experiment Design

In this study, we adapted EcoSim, to allow fine tuning of the gene flow's level between populations by adding various numbers of obstacles in the world. In order to measure the effect of the raggedness of the environment on population fragmentation and the speciation process, we included small physical obstacles uniformly in the world that obstruct the movement (dispersal) of agents. The presence of obstacle cells in the world is expected to impede the movement of our agents, change their spatial distribution, and in turn influence dispersal and ultimately the gene flow between populations. Three important changes have been made in the simulations that involve small, random obstacles. First, because of obstacles, the vision system of the agents has been modified. Obstacle cells are considered impenetrable and opaque and therefore affect not only the movement of species but also the capacity of the agents to locate food resources, potential partners for reproduction or potential danger. The perception concepts for food and friends were modified to stop the information coming from the other side of the obstacles. The prey perception of foes concept was also modified.

The second main modification concerns the action of movement performed by the agents. Obviously there is a big limitation in the agent's movement because they cannot pass through obstacles. As a consequence, a few movement actions were modified. As the agent cannot perceive food or potential mating partners through the obstacle, when the actions of movement towards food or potential reproduction partners are performed, it means that there is no obstacle between the agent and its destination. Therefore, these actions have been kept unchanged. The only action concerning both prey and predators that was changed was the action of exploration. The destination of the movement is still chosen randomly but a path toward it, circling the eventual obstacles, needs to be found. When different paths to reach the destination point exist, the shortest path algorithm is applied.

Prey individuals often need to escape predation. However, during the escape action, prey agents try to avoid collision with the obstacle cells. To compute the escape direction two criteria are considered. First, the barycenter of the five closest foes is computed. Second, the closest obstacle position is found. Then, two vectors (V1 for predator and V2 for obstacle in Figure 4-1) pointing at the opposite direction from these two positions are computed with a length proportional to the desire of the corresponding action. The final destination position is then computed by the addition of these two vectors. Finally, the same process used for exploration action, including the computation of the shortest path toward the desired final position, is applied to avoid the other possible obstacles (see Figure 4-1).

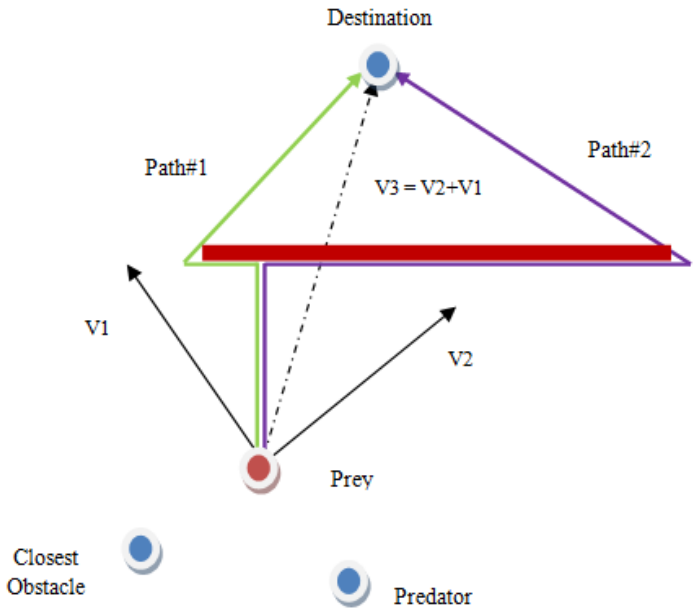


Figure 4-1. Computation of final direction of the escape route for prey. The prey agent takes into account the position of the closest obstacle as well as the position of the predators and the shortest path (path#1) is used to avoid another obstacles (red line).

The last modification is related to the model of food diffusion. Normally during the evolution of our ecosystem simulation, the grass present in a cell could diffuse in adjacent cells that do not contain grass. This process generates a dynamic distribution of food in the world that can form non uniform and non static patterns. To take into account the presence of obstacles, the diffusion mechanism has therefore been modified to prevent diffusion towards cells that contain obstacles.

The reduction of gene flow should be proportional to the raggedness of the world. We control the level of gene flow by changing the percentage of obstacle cells. We investigated how this impediment in the movement of organisms, without any complete extrinsic barrier separating two subpopulations of an initial species, affects the speciation frequency and the number of coexisting species. We compared three different situations. We considered a neutral configuration with no obstacles (the "Density of Obstacles (0%)" experiment) (see section 4.1.1). We also considered two virtual worlds with various numbers of obstacles: 1% and 10%. For example, in the experiment "density of obstacle (10%)", ten percent of cells in the world are obstacles. For each experiment we conducted ten independent runs using the same parameters and averaged the results. To ensure that our results are not dependent on special parameter values, several speciation threshold for prey (0.65, 1.3, 2.6) and predator (0.75, 1.5, 3) were used. As the results for the three speciation thresholds were very similar, we averaged them. All the results represent the average of 30 experiments (10 runs x 3 speciation thresholds). To avoid any bias due to a variation in the number of free cell available after the addition of obstacles, we maintained the number of free cells constant by increasing the world size accordingly. We analyzed the variation of the number of species for both prey and predator during the simulation process for several numbers of obstacles. We also observed other properties of the whole system, such as individual behaviours, spatial distribution of species, and the assembly and dynamics of ecological communities.

4.1.2. Results and Discussions

4.1.2.1. Global patterns

We investigated the global behaviour of our system in different situations, by varying the number of obstacles. We measured and monitored several representative characteristics of the system, such as the number of species, individual behaviour and spatial distribution of the individuals that

can give insight on the evolutionary processes that shape the biodiversity of our virtual world. An overview of the distribution of species reveals that individuals show a strong clustering distribution with circular or spiral shapes. Individuals forming spiral waves is a common property of predator-prey models (see Figure 4-2).

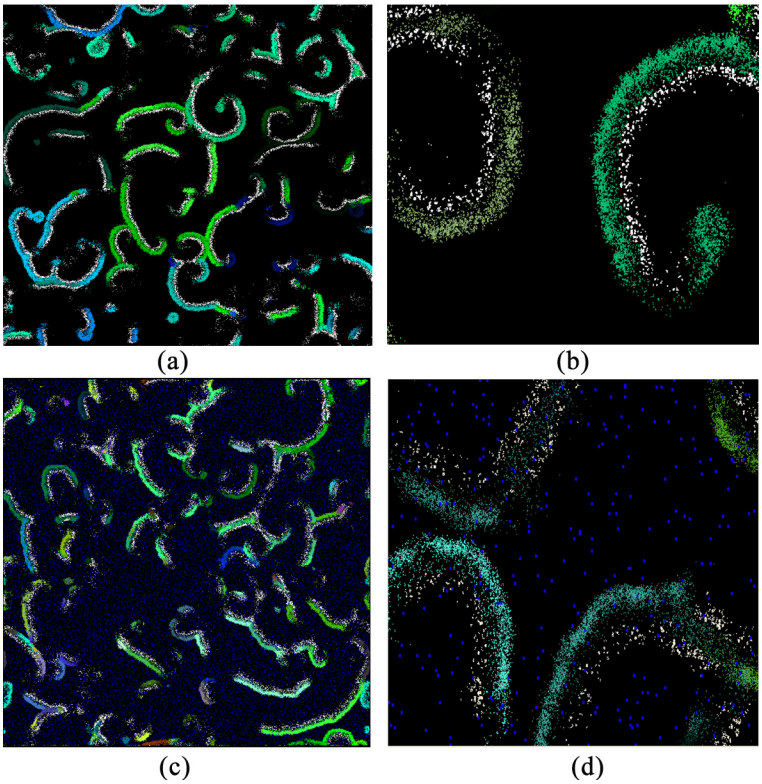


Figure 4-2. An overview of the distribution of species and populations in the world with density of obstacles (10%) and the density of obstacles (0%) experiment. (a) View of the whole world in the density of obstacles (0%) experiment. (b) Magnified part of the world in density of obstacles (0%) experiment. (c) View of the entire world with obstacles. (d) Magnified part of the world with obstacles. The blue squares are obstacle cells and dots are individuals. Different colored dots represent different prey species and white dots represent predator species.

The prey near the wave break have the capacity to escape from the predators sideways. A subpopulation of prey then finds itself in a region relatively free from predators. In this predator-free zone, prey individuals start dispersing rapidly forming a circular expanding region. The predation pressure creates successive interactions between prey and predators over time. The same pattern repeats over and over again, leading to the formation of spirals. Strong and robust spiral waves have been commonly observed in complex and dynamic biological systems [200]. Self-organized spiral patterns have been seen not only within chemical reactions [201] but also among host-parasitoid or predator-prey systems [95], [200], [202], [203] even when the world is

uniform in terms of environment's raggedness [95], [202]. It has been observed that the size and number of spirals in all the experiments are almost the same (Table 4-1). Therefore, the existence of such patterns is unlikely to explain the differences in speciation rates between different experiments.

Table 4-1. Average and standard deviation of the number and size of spirals in 30 independent runs of every configuration.

	Number of Spirals		Size of Spirals in cells	
	Mean	STD	Mean	STD
Density of Obstacle (0%)	41	5	186	67
Density of Obstacle (1%)	38	6	188	58
Density of Obstacle (10%)	43	8	181	73

4.1.2.2. Species richness and relative species abundance

The most important quantities of the system that we monitored through time were species richness (the total number of species), as well as species abundance (the number of individuals per species) in the world during the entire simulation process and across multiple simulations. Our results indicate that the number of species for both prey and predators increases directly with the number of obstacles (Figure 4-3).

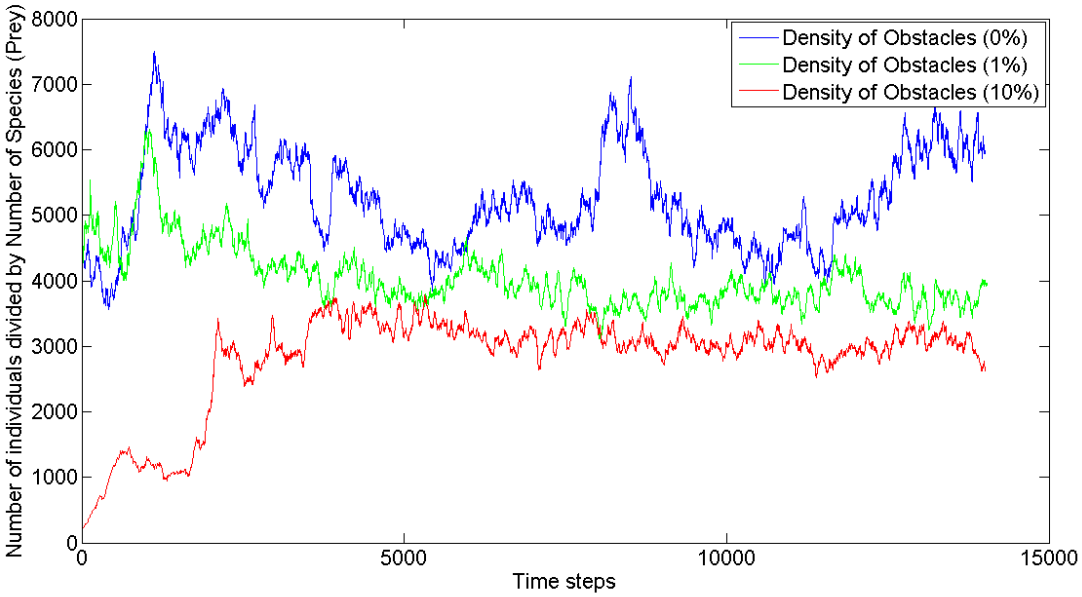


Figure 4-3. Comparison between numbers of prey species in the whole world during 16,000 time steps. Every curve represents an average value obtained from 30 independent runs with three different speciation thresholds.

Our results reveal that the total number of prey and predator species in the world is higher in the two configurations with obstacles compared with the no obstacles configuration. Moreover, the number of species in the configuration with 10% obstacles is much higher than the number of species for another configuration with obstacles. The computed average and standard deviation of the number of species during the whole process for prey population (Table 4-2) reveal clear differences between the three experiments. This suggests that the speciation rate is directly dependent to the restriction of movement and therefore gene flow between populations even though the relationship is not linear. As the total numbers of individuals in the three configurations are almost the same, it follows that the number of individuals per species decreases when obstacles are added in the world.

Table 4-2. Average and standard deviation of the number of species in the 30 independent runs for every configuration

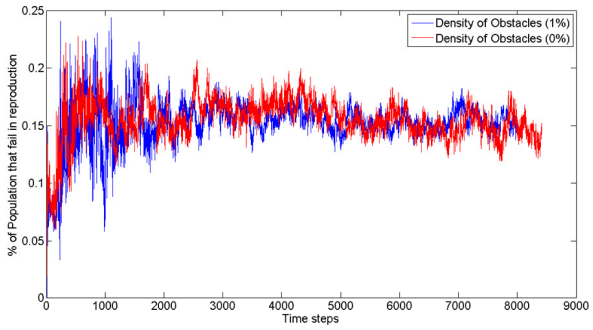
Version	Nr Prey Species Mean	Nr Prey Species STD	Nr Prey Species Median
Density of Obstacle (0%)	27	9	25

Density of Obstacle (1%)	68	16	61
Density of Obstacle (10%)	94	19	82

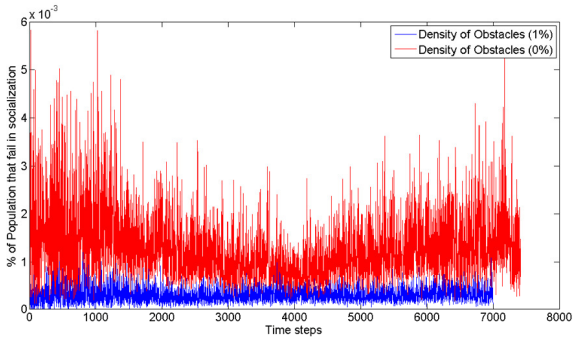
Our results clearly show that the addition of obstacles in the world is associated with an increase in the number of species. Population genetic theory predicts that natural selection and genetic drift cause populations to diverge from each other while migration resulting in gene flow acts in an opposite direction creating genetic homogeneity. We suggest that obstacles lead to an impediment in dispersal, more geographic isolation, less migration and gene flow. This overall lower level of population connectivity leads to rapid differentiation. Eventually, populations will contain individuals with genome dissimilarities higher than the speciation threshold, leading to speciation.

4.1.2.3. Variation in individual behaviours

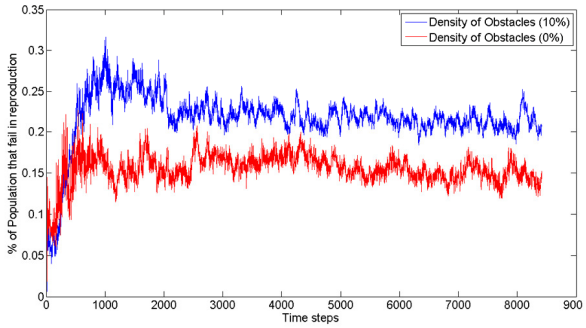
The results on species richness confirm that the restriction in the movement of species due to scattered physical obstacles is strongly correlated with the frequency of speciation events. In order to verify that the increased speciation rate cannot be explained by other factors such as the change in the behaviour of the agents, we monitored the actions performed by prey individuals in the three configurations. Most of the actions chosen by the individuals (e.g., feeding, predator avoidance, prey chasing) were the same in the three configurations although a slight difference was identified on reproduction and socialization (Figure 4-4). The reproduction action fails either because the agent cannot find a partner for reproduction, or has insufficient energy, or the two partners are genetically too different. The action of socialization fails because the agent cannot reach the place where the chosen partners are located. Our results suggest that the number of failed socialization events is much higher and much more variable in the density of obstacles (0%) configuration than in all obstacle experiments (see Figure 4-4b,d). This is likely a direct consequence of the fact that the species have much smaller spatial distribution in obstacle configurations, making the socialization action easier to perform (the genetic similarity between agents is not considered for this action). However, the number of failed reproduction actions is significantly higher when the raggedness of the world increases (Figure 4-4c). This is likely due to the higher genetic distance between individuals often found in close proximity (data not shown). This result enforces the hypothesis that the presence of obstacle cells in the world increases the genetic distance and therefore the speciation rate even in situations where the heterospecific individuals are more spatially compact.



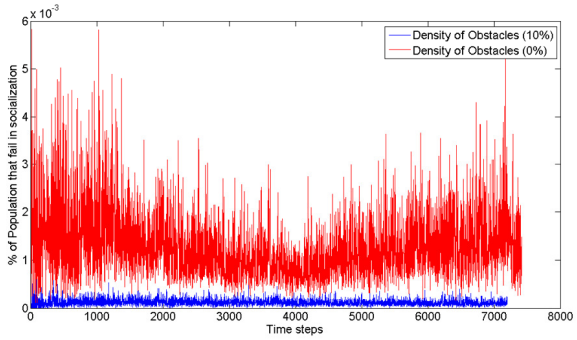
(a) Reproduction fail (density of obstacles 1%)



(b) Socialize fail (density of obstacles 1%)



(c) Reproduction fail (density of obstacles 10%)



(d) Socialize fail (density of obstacles 10%)

Figure 4-4. Percentage of prey individuals that fail in reproduction action (a,c) and socialize action (b,d) between the various density of obstacles (1%, 10%) configuration and the density of obstacles (0%) configurations. The red curves represent the density of obstacles (0%) experiment and the blue curves represent the experiments with various densities of obstacles (1%, 10%). Every curve is an average value obtained from 30 independent runs with three different speciation thresholds.

4.1.2.4. Spatial distribution of populations and species

To evaluate the spatial distribution of the species, we used a measure based on an average distance of all the members of a species to its physical center. This measure is expressed in number of cells and gives an accurate evaluation of the distribution area of a particular species in the world. The average and standard deviation of individuals' average distances around the center of each species taken from ten independent runs shows that the species have a more compact distribution in the obstacle versions (Table 4-3).

Table 4-3. The average and standard deviation of individuals' average distances around the spatial center of the species in the 30 independent runs corresponding to the 3 speciation thresholds for the 3 configurations.

	Time Steps

		3000	7000	10000	13000
Density of Obstacle (0%)	Mean of Spatial average distance	173.11	184.609	154.80	118.69
	STD of Spatial average distance	73.29	104.33	82.32	94.84
Density of Obstacle 1%	Mean of Spatial average distance	135.80	107.68	93.89	79.15
	STD of Spatial average distance	63.19	54.37	51.11	52.30
Density of Obstacle 10%	Mean of Spatial average distance	97.57	81.11	64.74	62.03
	STD of Spatial average distance	49.62	43.19	46.38	41.37

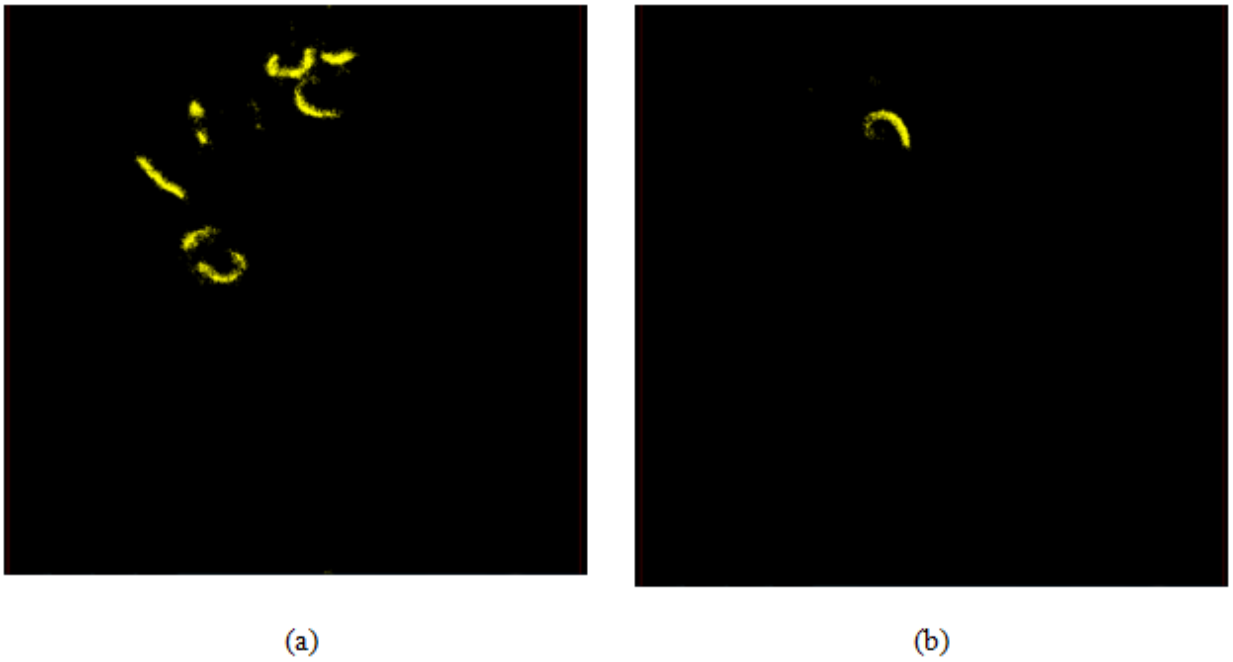
Knowing that in a torus world of size 1000x1000 cells the largest possible distance between two points is about 700 cells, the average values observed for the density of obstacles (0%) configuration, which can be more than 180 cells, are quite large. This suggests that, in the density of obstacles (0%) configuration, many species have a widespread spatial distribution covering a large part of the world. In contrast, in the world with obstacles, species show a much more restricted geographic distribution, which means that the species' spatial distribution decreases proportionally with the increase in number of obstacles. These results are also confirmed by the strong negative correlation between number of obstacles and the maximum observed spatial distribution of a species (Table 4-4). The high value of the standard deviation for all configurations can be easily explained by the high variability of the number of individuals by species.

Table 4-4. The median of maximum distances between individuals around the center of species in the 30 independent runs corresponding to the 3 speciation thresholds for the 3 configurations.

	Density of Obstacle (1%)	Density of Obstacle (1%)	Density of Obstacle (10%)

Maximum Spatial Distribution	310.45	186.13	129.58

Our communities of species shows a log-normal distribution pattern commonly found in nature [98]. This property leads to an important diversity in terms of number of individuals per species, which in turn explains the observed high variance in spatial distribution. For example, the graphical overview of the spatial distributions of individuals belonging to the same species (Figure 4-5) illustrates how obstacles strongly affect the spatial distribution of the individuals by reducing the total number of subdivided populations that constitute a species. For most of the species, several spatially separated populations are observed in the density of obstacles (0%) configuration whereas only one compact population is observed in the obstacle configurations.



(a) (b)

Figure 4-5. Spatial distribution of individuals that belong to one species (a) in a world with density of obstacles (0%) and (b) in a world with density of obstacles (10%).

4.1.2.5. FCM Evolution

The composition of species in our virtual world depends on the fine balance between speciation and extinction. The high species richness in the obstacle worlds could be due to an accelerated speciation rate, a decelerated extinction rate or a combination of both. In order to investigate the factors driving the biodiversity of the three virtual worlds we analyzed the level of genetic divergences between the initial genome and the genome of all individuals at every time steps. To evaluate the speed of evolution in our simulation, we compared the average distance [29] between

all existing prey or predator genomes at any time step with the two initial genomes of prey and predators.

This average distance computed for a total of 5000 time steps (Figure 4-6), indicates that the overall genetic divergence of the community of prey and predator species is greater in obstacle trials than in the density of obstacles (0%) experiment. The more obstacles in the world, the steeper the slope of the curve was. This suggests that evolution accelerates with the number of obstacles.

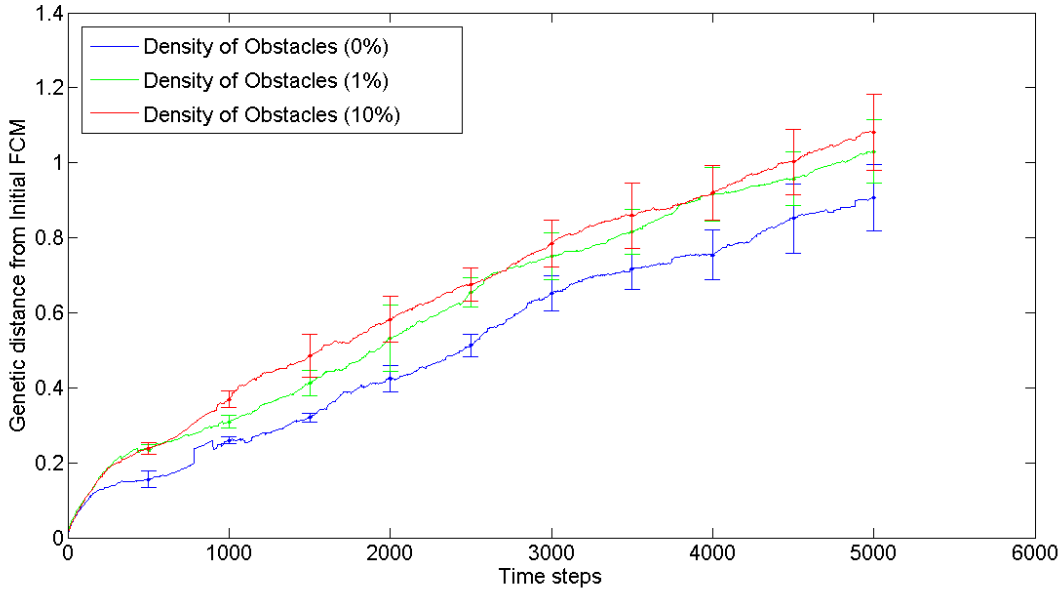


Figure 4-6. Average genetic distance between the community genomes (all individuals of prey or predators) at time zero and time x for the three configurations. Every curve is an average value obtained from 30 independent runs with three different speciation thresholds.

Given that this result shows only global information (at the community level) about speciation patterns, changes at the intraspecific level or between closely related species are also very informative. We measured the speed of divergence between two sister species after a speciation event. In EcoSim, a species is associated with a genome, which corresponds to the average genome of all its individuals allowing us to compute a distance between the 'genome' (called center) of two species. We considered 20 independent speciation events for each of the three configurations. We then computed the distance between the centers of the two new species after the speciation event occurred across 250 time steps (Figure 4-7). It can be noticed that, as expected, sister species diverge quite quickly after speciation. After speciation, hybridization events are quite rare because in each of the newly emerged species the individuals are highly

similar, but dissimilar from the ones of the sister species. As a result, gene flow between the two species is likely very low and leads to fast divergence. More noteworthy, the speed of divergence is much higher when there are obstacles in the world. Once again, it is clear that this phenomenon is continuous, as the speed of divergence increases proportionally with the number of obstacles.

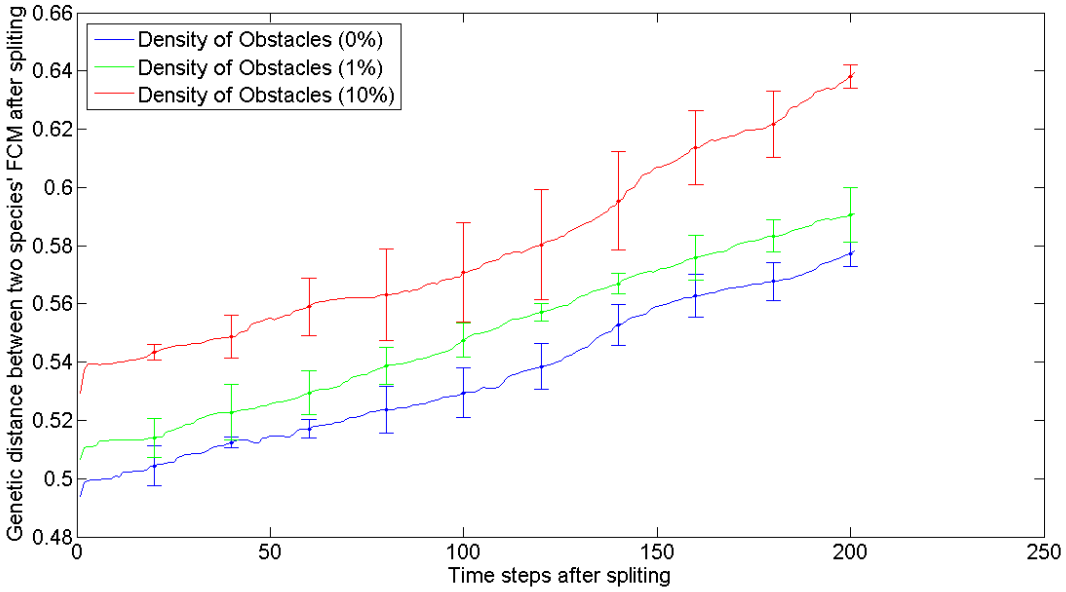


Figure 4-7. Average genetic divergence between the FCMs of sister species after their splitting for the three configurations. Each curve is an average of 600 couples of sister species (30 runs x 20 couples of sister species).

To understand if the reduction in gene flow is enough to explain the speed of divergence, we also considered the effect of obstacles on spatial distribution of sister species. We computed the average geographical distance between the physical centers of the emerging species after speciation events occurred and across 250 time steps for 20 speciation events (Figure 4-8). We observed that the physical distance between species is smaller in the world with obstacles than without. This result can be correlated to the fact that the number of individuals per species is smaller and the spatial distribution of individuals is more compact for obstacle than density of obstacles (0%) configurations. It is interesting to notice that even if there is high variation in these spatial distances, there are no visible trends to an increase of distance between species after speciation in the three configurations.

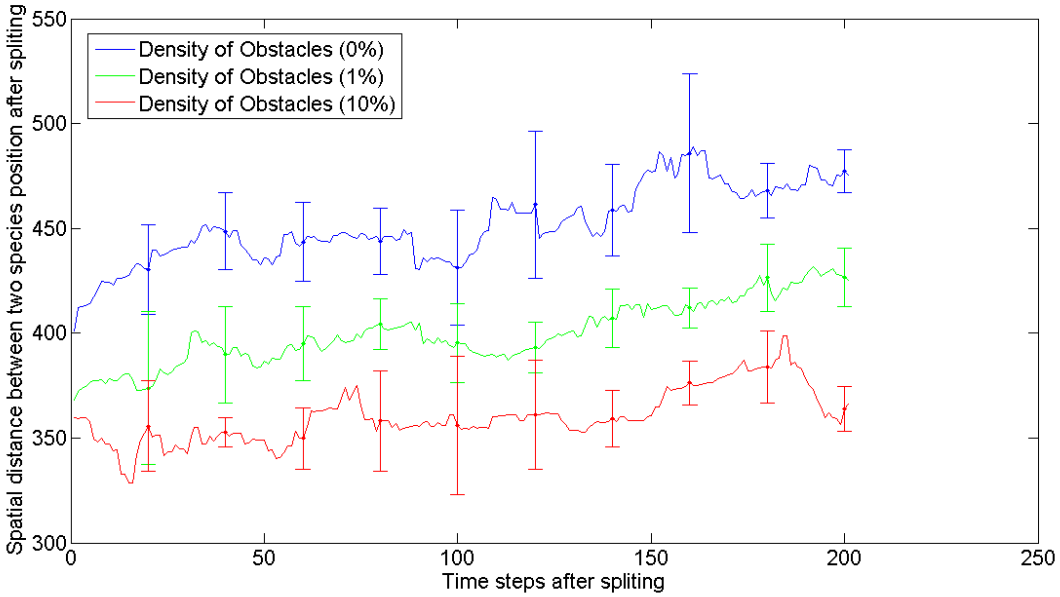


Figure 4-8. Average spatial distance between the spatial center of 2 sister species after their splitting for the three configurations. Each curve is an average of 600 couples of sister species (30 runs x 20 couples of sister species).

We observed a direct and continuous increase in the speed of evolution (e.g. the rate of speciation) with the increasing number of obstacles in the world [30].

4.1.2.6. Conclusion

Using a modified EcoSim individual-based platform to implement various degrees of physical obstacles that restrict the movement of individuals and likely reduce gene flow, we compared three different configurations with different densities of obstacles. It is clear that the speciation rate and species diversity is directly proportional with the roughness of the physical environment. Our study also reveals that species are more spatially compact in the configurations with obstacles than in the world of experiment with density of obstacles (0%). Moreover, the reduction in spatial distribution as the number of obstacles increases results in low levels of gene flow between sister species. Therefore, the rapid genomic divergence between species should be directly linked to the reduction of movement due to obstacles that result in low gene flow and rapid divergence between subdivided populations. We investigated several factors that could be involved in the increase of speciation rate, such as the individual's behaviours, the spatial distribution of the species or the overall speed of evolution (increase in genetic divergence between sister species). We show that the faster divergence between populations and accelerated speciation cannot be explained by an increase of spatial separation during the initial stage of speciation or different behaviours of the individuals. We suggest that this is likely due to the

significantly lower population sizes in obstacle configurations. This reduced size, results in more pronounced genetic drift and rapid differentiation between population that experience relatively low levels of gene flow. Similarly slowing down dispersal of individuals can have the same increasing effect on speciation since there will be less gene flow between individuals [92].

It is well accepted that the effect of micro-geographic barriers (e.g., the raggedness of the environment) to maintain population cohesion and the genetic homogeneity of a species depends heavily on the intrinsic properties of the species (e.g., dispersal ability, intra- and inter- specific interactions) (see section 4.1.1). We suggest that the complex context of speciation can be better understood outside the framework of a classical geographical definition of speciation (e.g. sympatric, allopatric) by focusing on the complex interactions at the community level.

Speciation and extinctions are very important processes that influence the species composition of an ecosystem at a particular time and the long-term dynamics of ecological communities. Our approach allows testing the unified neutral theory of biodiversity and biogeography proposed by [35], which suggests that the persistence of ecologically equivalent species in sympatry across relevant time scales might not depend strictly of complex niche differences. Our results suggest that factors affecting demographic stochasticity (e.g., factors shaping the density of individuals in a local area, the extent of the distribution of species in space) can influence speciation and extinction rates and ultimately the distribution of relative species abundance. Our approach has demonstrated its utility to model several important biological problems, and it seems possible to modify it to represent many new ones. However, the FCM model has some limitations because it cannot evolve new sensory inputs or new actions. The complexity of the model also grows with the square of the number of such concepts, limiting this application to relatively simple behavioural models. Since our simulation takes into account spatial information and individual behaviour while allowing the creation of new species and, more importantly, the growth rate of species is not fixed, it is difficult to determine if varying the growth rates of species or predator pressure would lead to different distributional patterns without testing it. A more in depth analysis of the effect of reproduction rates and predator pressure on the spiral formation is needed.

4.2. Exploring the nature of species in a virtual ecosystem

Darwin's "mystery of mysteries," the origin of species, is difficult to study in nature because – in most cases – the process is rare, protracted, and unreplicated [194]. Mechanisms of speciation – and the forces influencing them – are therefore frequently studied in theoretical models [54],

[204], [205]. These models can be grouped into several broad classes – a summary of which will set the stage for how ours differs. (1) A single starting population is subject to a pre-defined intra-specific competition function on a pre-defined resource distribution that would favour a single phenotype in the absence of competition: i.e., "adaptive or competitive speciation" [56], [206]. (2) Geographically isolated populations, with or without gene flow, are subject to different selective environments, which are typically specified a priori as favouring or disfavouring particular phenotypes or genotypes: i.e., "ecological speciation" [61], [206], [207]. (3) Geographically isolated populations are subject to a single pre-defined selective pressure (or no selection at all); in response to which they can evolve different and incompatible mutations: e.g., "mutation order speciation" [208]. (4) Different groups are subject to similar pre-defined natural selection but experience different patterns of sexual selection, which can be pre-defined or can evolve owing to pre-defined fitness consequences (see section 2.2) [209], [210].

Previous speciation models thus take a diversity of forms and are implemented in a diversity of ways; yet a feature common to all of them, which we have emphasized above, is reliance on pre-defined fitness functions. This reliance on investigator-specified functions suggests the possibility that outcomes are heavily dependent on the specific functions used (see the discussion regarding pre-defined functions in section 2.1). Thus, although existing models have taught us much about speciation, they have left open the question of how speciation proceeds in the absence of experimenter-defined functions. To address this key knowledge gap, we here used individual-based simulations to explore speciation in the absence of pre-defined fitness functions. In our model, speciation must instead proceed owing to emergent properties of interactions between individuals in spatial landscapes where abiotic parameters are initially invariant.

4.2.1. Experiment Design

To investigate forces influencing speciation in the virtual world, we considered the formation of genetic clusters and the level of hybridization between them. Four main forces could lead to clusters with limited hybridization: (1) enforced reproductive isolation due to a rule that allows only genetically similar individuals to mate, (2) spatial isolation due to low dispersal ability, (3) natural selection as a result of behavioural divergence that causes hybrids to have low fitness (inappropriate combinations of behaviours) and (4) genetic drift where the persistence of the new mutations is governed by chance and become clustered owing to dispersal limitation. To analyze these potential contributors to speciation, we conducted five experiments in EcoSim. With the exception of the **Selection, Enforced Reproductive Isolation, and Low Dispersal** experiment conducted as a control, all other experiments have one or more of the described features

deactivated (Table 4-5). In total, we conducted 50 independent runs, 10 for each experiment, with an overall computational time of 65,000 hours and about 175 TB (Terabytes) of data. The first experiment, the classical version of EcoSim (**Selection, Enforced Reproductive Isolation, and Low Dispersal**), maintains the four features implemented in Gras et al. (2009) [29] and defined above (see Table 4-5).

Table 4-5. Overview of the five experiments and their respective features.

Experiment	Enforced reproductive isolation	Spatial isolation	Natural selection
1. Selection, Enforced Reproductive Isolation, and Low Dispersal	Yes	Yes	Yes
2. Selection and Low Dispersal	No	Yes	Yes
3. Selection and High Dispersal	No	No	Yes
4. No Selection and High Dispersal	No	No	No
5. No Selection and Low Dispersal	No	Yes	No

The enforced reproductive isolation has been removed from all other experiments, which means that the genetic similarity of two individuals is not considered when organisms attempt to mate and that no extrinsic force prevents two divergent individuals from mating if they encounter each other. While the first two experiments involve a complex and evolvable behavioural model that allows agents to make decisions that directly influence their survival and reproductive success, for the last two experiments, the behavioural model is not used, such that individuals make random decisions. Consequently, their 'genomes' contain information that is not utilized during the simulation, so that natural selection is not possible.

In the second experiment (**Selection and Low Dispersal**) only the enforced reproductive isolation is removed and all other settings are maintained as in the **Selection, Enforced Reproductive Isolation, and Low Dispersal** experiment. The relatively low dispersal ability of agents allows for strong geographic clustering of individuals and can potentially enhance local adaptation.

The third experiment (**Selection and High Dispersal**) implements an extreme dispersal ability that facilitates high levels of gene flow. In addition to removing the enforced reproductive isolation, we increased the dispersal rate of the individuals, while conserving the behavioural model of individuals. In this simulation, newborn individuals are placed in randomly chosen cells in the world instead of in the cell of its parents. As the behavioural model is used in the context of very limited geographic isolation between populations, it is possible to evaluate the effect of natural selection on the genetic clustering of individuals and ultimately on the speciation process.

In the **No Selection and High Dispersal** experiment, we used a randomized version of EcoSim (see section 2.5). As individuals do not use their behavioural model, their actions are random. Therefore, in this version, all the evolutionary forces except genetic drift are considered to be deactivated and all other parameters have been kept as close as possible to those of EcoSim. Since there is no behavioural model that governs the agent decision and they have extreme dispersal ability, the evolutionary process will be only driven by genetic drift.

Finally, in the **No Selection and Low Dispersal** experiment, we forced the creation of groups (herds) of individuals based on the random walk model. We aimed to obtain groups of individuals as similar as possible to the ones observed in the **Selection, Enforced Reproductive Isolation, and Low Dispersal** experiment. To enforce the grouping, we placed the new-born individuals in one of the parent's positions. However, the movement of the individuals was random since individuals do not use their behavioural model (see section 2.5). We conducted 10 simulations for each of the above five experiments.

We used the Spatiotemporal Complexity (STC) measure to obtain a quantitative comparison of the level of grouping between individuals. This measure has been developed for the analysis of individual patches and is typically used to measure the dispersion or "clumpiness" of different patches of individuals [211]. The values close to 'one' correspond to a random uniform distribution of individuals in the world and values close to 'zero' correspond to a unique group of individuals. For each experiment we conducted ten independent runs using the same physical characteristics (see Table 4-6).

Table 4-6. Several physical and life history characteristics of individuals averaged over 10 independent runs for every experiment. Exp1 stands for Selection, Enforced Reproductive Isolation, and Low Dispersal, Exp2 for Selection and Low Dispersal, Exp3 for Selection and High Dispersal, Exp4 for No Selection and High Dispersal and Exp5 for No Selection and Low Dispersal. In the experiments without natural selection, because there is no behavioural model, some characteristics do not exist.

Characteristic	**Predator**					**Prey**				
	Exp1	Exp2	Exp3	Exp4	Exp5	Exp1	Exp2	Exp3	Exp4	Exp5
Maximum age (time steps)	39 (± 5)	38 (± 6)	44 (± 7)	20 (± 2)	22 (± 1)	44 (± 15)	42 (± 17)	42 (± 17)	22 (± 2)	25 (± 3)
Minimum age of reproduction (time steps)	8	8	8	1	1	6	6	6	1	1
Maximum speed (cells / time step)	11	11	11	n/a	n/a	6	6	6	n/a	n/a
Vision distance (cells maximum)	25	25	25	n/a	n/a	20	20	20	n/a	n/a
Level of energy at initialization of the system (units)	1000	1000	1000	n/a	n/a	650	650	650	n/a	n/a
Average speed (cells / time step)	4	4.2	4.1	...	...	3.1	3.2	3.1	...	...
Average level of energy (units)	445	448	432	n/a	n/a	278	271	268	n/a	n/a
Maximum level of energy (units)	1000	1000	1000	n/a	n/a	650	650	650	n/a	n/a
Average number of reproduction action during life	1.14	1.21	1.18	...	...	1.49	1.37	1.41	...	...
Average length of life (time steps)	9	10	10	4	5	13	13	12	4	5
Number of individuals (in thousands)	31 (± 7)	36 (± 9)	33 (± 11)	38 (± 4)	37 (± 5)	307 (± 24)	293 (± 28)	285 (± 17)	261 (± 8)	266 (± 9)
Level of patchiness	0.26 ±	0.25 ±	0.84 ±	0.92 ±	0.38 ±	0.24 ±	0.26 ±	0.82 ±	0.9 ±	0.41 ± (0.09)

	(0.05)	(0.04)	(0.11)	(0.08)	(0.08)	(0.07)	(0.06)	(0.12)	(0.05)	

4.2.2. Measure for cluster quality

In order to explore the causality of species formation, we investigated the conditions that lead to the emergence of strong phenotypic/genotypic clusters. We investigated whether our species concept implemented in EcoSim is consistent with the genotypic cluster definition. To achieve this we analyzed the degree of compactness and isolation of the generated clusters of genomes, called species-clusters. Then, we compared the species-clusters obtained at selected time steps (12000, 14000, 16000, 18000 and 20000) with the K-means-clusters and random-clusters. The implemented speciation mechanism in EcoSim can be viewed as an online hierarchical clustering process, with each species being a cluster of genomes (see section 2.4.4.9). Since clustering is a difficult and time-consuming task, it is impossible to apply it to the whole population of individuals at each time step. For example, at some time steps, EcoSim supports more than 500,000 individuals. We have therefore chosen a heuristic hierarchical approach in which clustering is done greedily along the evolutionary process. Therefore, in a given time step, only a small subset of individuals is effectively clustered by our species splitting mechanism. For ease of comparison, the k-means-clusters were obtained by directly applying a k-means clustering algorithm [212], with k being the number of clusters at that particular time step, to the whole population of genomes using the same number of clusters. As a reference, for a lower bound of cluster quality, we also conducted random clustering using the same number of clusters k and randomly assigning every individual to one of the clusters.

To evaluate the separation between clusters and their compactness we calculated the distance between every species' genetic center, which represents the average of all individual genomes of a particular species and their farthest individuals. Furthermore, we calculated the mean and standard deviation of the distances between every two species' genetic center at four different time steps. In addition, we used the Davies-Bouldin index [213], a commonly used method for measuring the quality of clustering algorithms. We also used this index to compare the quality of different clustering results for the five experiments.

In order to analyze the level of reproductive isolation of species obtained in the five experiments, we measured the percentage of mating events that generate hybrid offspring, that is, offspring for which each parent is a member of a different species, across all mating events. These analyses were performed on all replicates of the five experiments separately for the first 10,000 time steps

and for the time steps between 10,000 and 20,000. This approach was used given that the behaviour of individuals takes time to stabilize.

4.2.3. Results and Discussions

To explore the causality of species formation, we first investigated the conditions that led to the emergence of strong genetic clusters. EcoSim includes such clusters, called species-clusters, by implementing a heuristic divisive hierarchical clustering process for all individuals in the entire virtual world in a given time step. We then evaluated the emergent clusters based on their compactness and separation from other clusters and compared these results to those obtained using a K-means-clustering algorithm and randomized clusters. A good way to assess the organization of genotype groups that emerged is the number of individuals per cluster: if genotype groups exist, then the simulations should generate and maintain clusters with many individuals. Other measures of compactness and separation are detailed in section 4.2.2 and include genomic distance between and within clusters and the Davies-Bouldin index (which is a combination of these two previous measures). Our key results are as follows: all experiments involving natural selection (evolving behaviour model) led to compact and distinct clusters; experiments involving geographic isolation without selection generated less compact and more overlapping clusters; experiments with genetic drift alone did not generate clusters (see Figure 4-10). All the comparisons, except for the rate of hybrid production and fitness of the hybrids, were performed on the average and standard deviations of ten runs taken at time steps 12000, 14000, 16000, 18000 and 20000.

In the experiments with natural selection, the number of individuals per species was much higher than in the experiments without natural selection from time step 14000 (one-way ANOVA, $P = 0.0001$; Tukey post hoc test, $P < 0.05$; Figure 4-9). Moreover, this metric in the **Selection and High Dispersal** and the **Selection and Low Dispersal** experiments converge toward those obtained for the **Selection, Enforced Reproductive Isolation and Low Dispersal** experiment, indicating that the three experiments involving natural selection exhibit the same long-term result. By contrast, the two experiments without natural selection generate only clusters that contain two or three individuals, showing that no organization of genotype groups emerged.

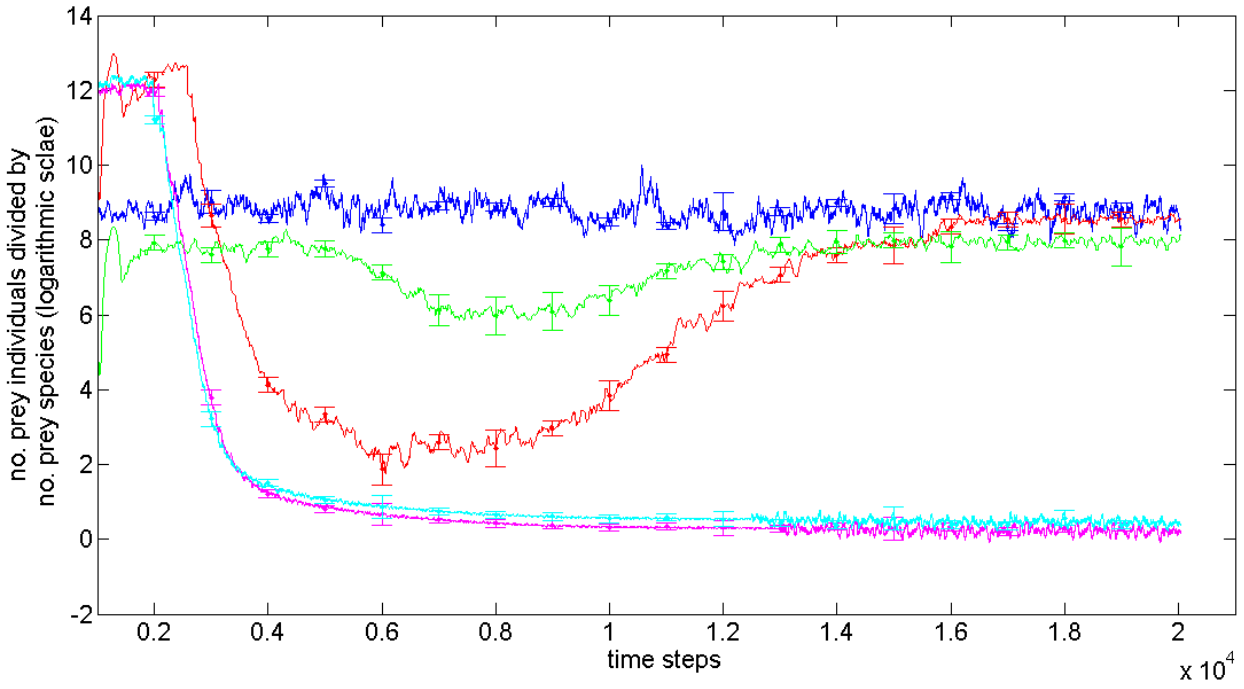


Figure 4-9. The number of individuals to number of species ratios (logarithmic scale) in the different simulation experiments (blue line, Selection, Enforced Reproductive Isolation and Low Dispersal experiment; red line, Selection and Low Dispersal experiment; green line, Selection and High Dispersal experiment; clay line, Selection and Low Dispersal experiment; magenta line, No Selection and High Dispersal experiment).

Our other metrics support the above assertion: experiments with natural selection led to clusters that were more discrete, in terms of both compactness and separation (genomic distance and the Davies-Bouldin index) than was the case in random clustering (Figure 4-10). Further, we found no difference in these properties between the **Selection, Enforced Reproductive Isolation and Low Dispersal** experiment that involves an pre-determined extrinsic mating rule based on genetic distance and the **Selection and Low Dispersal** experiment where agents make free reproductive decisions (one-way ANOVA, $P = 0.6$) (Figure 4-10). This important result indicates the emergence of genetic clusters in the absence of extrinsic barriers to gene flow but in the presence of natural selection.

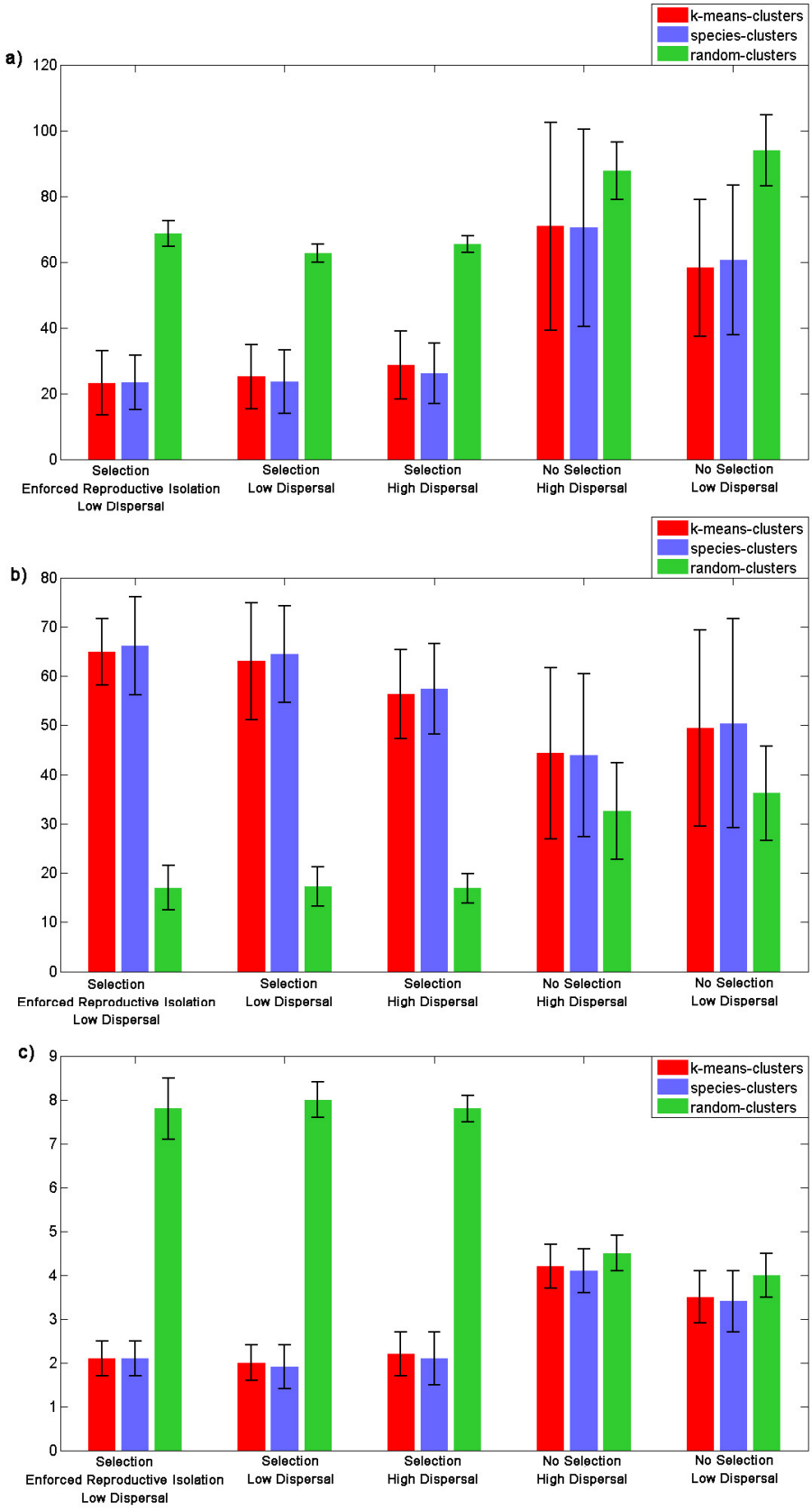


Figure 4-10. Average and standard deviation (error bars) of the distance of the farthest individual from its cluster's genetic center (a), the distance between the genetic centers of all pairwise clusters (b) and Davies-Bouldin index (c) for the five experiments. For (a) and (c) the lower the value the more compact the cluster and the more it is separated from other clusters. For each experiment, the values are given for a global k-means clustering algorithm (blue), the species-clusters generated by the simulation (red) and randomized clusters (green).

If the genetic clusters uncovered in our simulations represent species under the biological species concept, then reproductive barriers between them should be evident. We tested this possibility by quantifying and averaging the rate of hybrid production (Figure 4-11a) and the fitness of hybrids (Figure 4-11b) measured at every 100 time steps. These metrics demonstrated that all simulations that involve selection led to reduced mating success between clusters and reduced hybrid fitness. Beyond time step 10000, results for the **Selection and High Dispersal** and the **Selection and Low Dispersal** experiments converged toward those obtained in the **Selection, Enforced Reproductive Isolation and Low Dispersal** experiment. Similar reproductive barriers were not evident in the simulations without selection (one-way ANOVA, $P = 0.001$; Tukey post hoc test, $P < 0.05$ for all pairs of selection/no selection experiments after time step 10000 for rate of hybrid production and before and after time step 10000 for fitness of hybrids). These results confirm that the genetic clusters under selection correspond to local fitness (see section 2.4.4.4) maxima in genotypic space, whereas genotypes outside of the clusters have lower fitness. A few dozen time steps after their formation, new clusters are fully reproductively isolated with no additional hybridization events. These large compact groups of locally high fitness phenotypes, reproductively isolated from the others, can be reasonably considered as separate species.

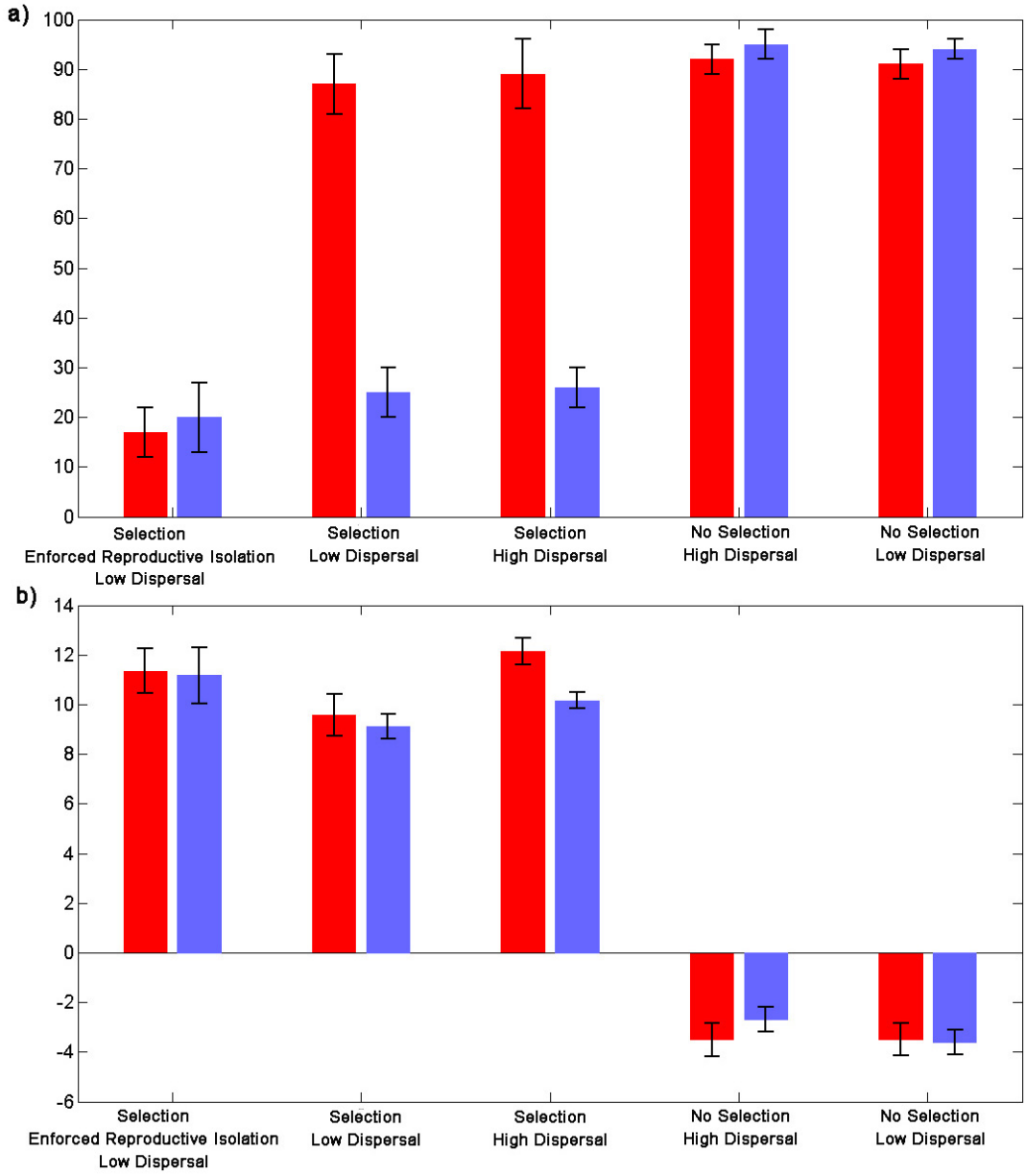


Figure 4-11. (a) Average and standard deviation (error bars) of the rate of hybrid production before (red) and after (blue) 10000 time steps. (b) Average and standard deviation of the percentage of decrease in fitness of the hybrid individuals compared to non-hybrid individuals before (blue) and after (red) 10000 time steps. We averaged the fitness value of hybrid and non-hybrid individuals at every 100 time steps.

These results confirmed the role of natural selection in speciation by showing its importance even in the absence of pre-defined fitness functions [31].

4.2.4. Conclusion

Hundreds of mathematical models have been developed to study the role of selection in speciation [54], [204], [205], and the general view to have emerged is that selection causes speciation only under a specific subset of conditions. These previous models used pre-defined fitness functions that left open the question of whether or not results are particular to those functions. Our model did not include such functions and instead allowed selection to emerge as a result of complex behavioural interactions. Under these conditions, speciation occurred in many different configurations, thus providing further support for the role of selection in driving speciation [214], [215].

In our model, speciation occurred due to biotic interactions, both within and between species. These biotic interactions drove the evolution of a diversity of behavioural types and these different types formed discrete clusters. Mating between these clusters rapidly decreased and hybrids between them had low fitness. Although abiotic conditions can certainly drive speciation, our results support assertions that biotic interactions could be particularly important drivers of the selection that causes the formation of new species [214]–[216]. In addition, although speciation can be driven by morphological or physiological divergence, our results support arguments that speciation might proceed particularly rapidly as a result of behavioural divergence [217], [218]. The next challenge will be to determine how such interactions shape speciation in natural systems.

4.3. A New Species Abundance Distribution Model

Species Abundance Distributions (SADs) follow one of ecology's universal laws – every community shows a hollow curve or hyperbolic shape on a histogram with many rare species and just a few common species. The species abundance distribution is one of the important measures of biodiversity and one of the most significant concepts in ecology communities. Using this concept, the biologists can infer a lot of information from their collected data [219]. There are several definitions of SAD which has been proposed by different authors. McGill gave the following explanation about the SAD [219]:

"A species abundance distribution is a description of the abundance (number of individuals observed) for each different species encountered within a community. As such, it is one of the most basic descriptions of an ecological community. When plotted as a histogram of number (or percent) of species on the y-axis vs. abundance on an arithmetic x-axis, the classic hyperbolic,

'lazy J-curve' or 'hollow curve' is produced." This J-curve represents the power law relationship, which has been shown are emergent quantitative features of biodiversity [119]. In this study, we proposed a new species abundance distribution model which can fit the SAD in ecological communities more accurately.

Many models have been proposed to predict (estimate) SADs. McGill provided a helpful survey of different models [219]. Unfortunately, most of the models that have been proposed contain some weakness mentioned by McGill. We have considered them carefully for proposing a new method.

1- For most SAD models, no comparison with other existing models has been proposed. In other words, for the existing models there is no comparison of how their predictions fit data in comparison to other models.

2- According to McGill, another weakness is that different inconsistent methods have been used to measure goodness of fit. Unfortunately the different methods used for evaluation, all emphasize different facets of fit. For example, "By-class Good Fit" fits data to the logged-bin and emphasizes fitting rare species. Therefore, the "By-class Good Fit" method and lognormal family methods work on similar features. Thus any claim of an exceptional fit must be robust by being superior on multiple measures.

3- Even when consistent methods are used, most of the new models will fit some datasets well and other datasets poorly. In other words, for most of the models that have been proposed, the authors have evaluated their method on specific datasets well designed for their methods. So it could be helpful if we can test our method over several datasets.

In practice, one might come across a case where no single model can achieve an acceptable level of accuracy. In such cases, it would be better to combine the results of different models to improve the overall accuracy. Every model operates well on different aspects of the dataset. For example the lognormal family methods emphasize fitting rare species more than other methods. As a result, assuming appropriate conditions and combining multiple models may improve prediction performance when compared with any single model [220], [221]. In this study, we proposed a new method called FPLP, based on the combination of existing models: Fisher's logseries, Power-law, Logistic-j and Poisson-lognormal [222] (see section 4.3.1). The main idea came from a combination of techniques that are used in different fields. According to our knowledge, it is first time that a model based on the combination of other models using genetic

algorithm has been proposed for the species abundance distribution problem. In this study we evaluate several SAD models, including the FPLP model, with three different goodness-of-fit measures and applied to eight different datasets. By using the FPLP method, we investigated how the combination of different models' behaviour is important in characterizing different aspects of SAD.

4.3.1. SAD Models

SAD typically represents the way *N* individuals are partitioned into *S* species [223]. An example of how SADs can be represented graphically is given in Figure 4-12.

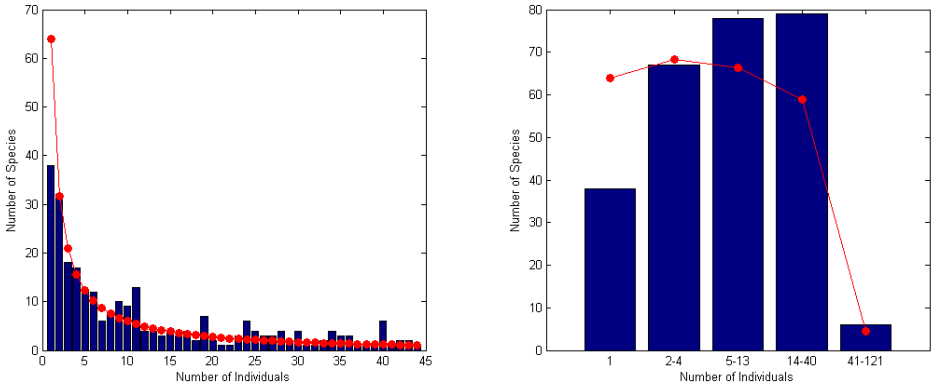


Figure 4-12. There are different representations for Species Abundance Distribution. (Left) The histogram is the observed SAD and red curve is the predicted SAD, (Right) A histogram with abundance on a log-scale.

Besides Figure 4-12, there are different ways to plot SADs. The complete set of ways to plot SADs have been presented in [219]. The origin of the SAD model points to 1932 in which the first model for prediction of SAD was proposed. Since that time, many models have been proposed. We give here several of them among others that are representatives of different modelling families (see Table 4-7).

Table 4-7. Different families of SADs.

Family	**SAD**
Statistical	Fisher's logseries [99] Lognormal - Preston [102]

Spatial distribution of individuals	Power Law [224] Fractal distribution [225] Multifracta [226]
Population dynamics (metacommunity models)	Logistic-J [227] Neutral model [228]
Metacommunity models	ZSM [229] Poisson Lognormal [230]
Niche partitioning	Broken stick [231] Sugihara [232]

As mentioned above, they correspond to various methods based on different concepts, which lead to variable results depending on the dataset. The main idea of this study is to use a combination of models belonging to different families of methods for SAD modeling in order to have more flexibility in the final model. In this section we introduce the four basic models that we used in our model.

4.3.1.1. Fisher's Logseries

In the 1940s, researchers proposed different statistical models to describe patterns of species abundance [99], which still stimulate a great deal of interest today [233]. Given a sample of a community, Fisher has defined a series expressing the species abundance distribution of this sample. Let N and S be respectively the numbers of individuals and of species in the sample. If n_i is the number of species that contain i individuals in the sample, then:

$$\forall i \in N,\ n_i = \alpha x^i / i \qquad (4\text{-}1)$$

The series is thus represented by

$$SAD_{Fisher} = n_1, n_2, ..., n_k = \alpha x,\ \alpha\frac{x^2}{2},\ \alpha\frac{x^3}{3}, ..., \alpha\frac{x^k}{k} \qquad (4\text{-}2)$$

Where α and x, the two parameters of the model, satisfy the equations:

$$S = \alpha \ln(1 + \frac{N}{\alpha}) \qquad \text{and} \qquad x = \frac{N}{N + \alpha} \qquad (4\text{-}3)$$

Therefore, if N and S are known, α and x can easily be calculated. The first parameter, α, is constant for all samples from a given community (it is a characteristic of the community and not of the sample). α is correlated with the total number of species in the considered community and is called the "index of diversity" of the community [98].

4.3.1.2. Logistic-J

The logistic-J distribution arises from a dynamic, individual-based model of species [227]. The resulting Probability Density Function (PDF) can be written as following:

$$f(x) = c(\frac{1}{x} - \delta) \qquad \varepsilon \le x \le \Delta$$
$$0 \qquad \text{Otherwise} \qquad (4\text{-}4)$$

where the abundance x runs from ε to a maximum Δ. The constants $\delta = 1/\Delta$ and ε are parameters of the distribution and c is a constant of integration that gives a value of 1 to the area under the curve of the PDF. The constant c is a function of Ɛ and δ. The parameters $\mathcal{E}$ and Δ are called the inner and outer limits of the distribution, respectively (see Figure 4-13).

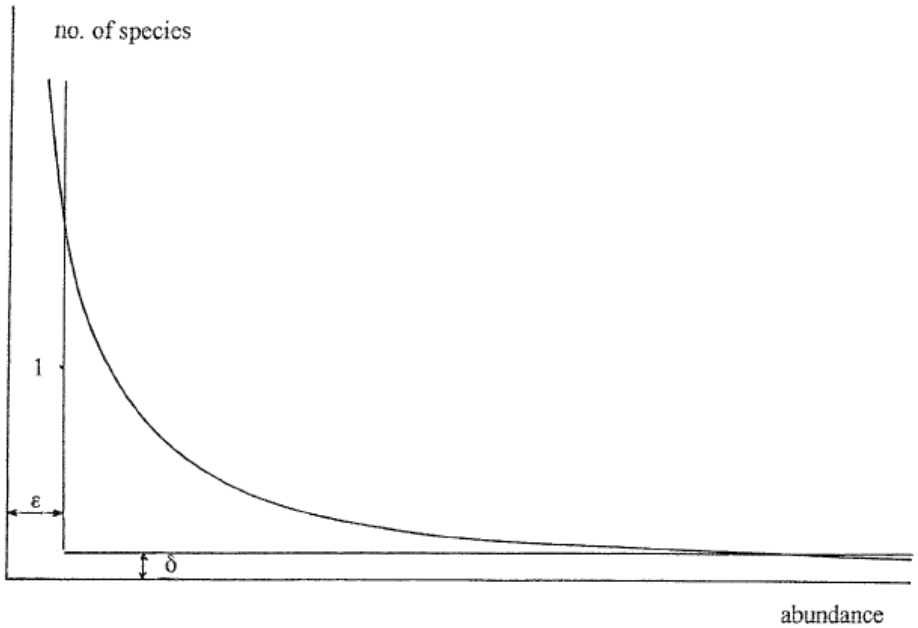


Figure 4-13. The probability distribution function of the general logistic-J distribution.

The distribution function F of the logistic-J probability density function f is obtained by multiplying it by R, the number of species in a sample or in a community (F(x) = Rf(x)) leading to the following prediction by logistic-J:

$$SAD_{Logistic-J} = F_1, F_2, ..., F_k \qquad (4\text{-}5)$$

4.3.1.3. Power law

One of the best-known patterns in ecology is the power-law form of the species-area relationship. Such a general pattern is important not only for fundamental aspects of ecological theory but also for ecological applications such as the design of reserves and the estimation of species extinction [226]. We consider here a SAD, which decays with the power-law from the minimum number of individuals [224], $x=1$, to the maximum value, $x=X$ as

$$f(x) = (S+1)\alpha x^{-(1+\alpha)}$$
$$SAD_{Power-law} = f_1, f_2, ...f_k \qquad (4\text{-}6)$$

The relation between the total number of species S and the maximum number of individuals X is obtained as following:

$$X^{\alpha} = S + 1 \qquad (4\text{-}7)$$

4.3.1.4. Poisson Lognormal

This model mixes the lognormal with the Poisson distribution. One possible way to generalize the univariate Poisson distribution is to use a variable that follows a univariate lognormal distribution. If the abundances, λ, are lognormally distributed (which mean that $\log x$ is normally distributed) with mean M and variance V, then the compound Poisson Lognormal distribution is the probability function [230]:

$$P_r = \frac{(2\pi V)^{-1/2}}{r!}\int_0^{\infty} x^{r-1}\ e^{-x}\ e^{-(\log x - M)^2/2V}\ dy \quad , \quad r=1,2,... \qquad (4\text{-}8)$$
$$\mathrm{SAD}_{\text{Poisson-Lognormal}} = P_1, P_2, ...P_k$$

Where r specifies the number of individuals. The distribution can be fitted to observed data by estimating the parameters, M and V, by the method of maximum likelihood.

4.3.2. Goodness-of-fit

The goodness-of-fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. In this section, several criteria that we used to compare the observed abundance distribution and the calculated abundance distribution have been introduced.

4.3.2.1. Squared prediction error (SPE)

Squared prediction error (SPE) is a frequently-used measure of the differences between values predicted by a model and the actual observed values [234]. SPE is a good measure of precision. The formula used is as follows:

$$SPE = \sqrt{\sum [g(x_i) - \hat{g}(x_i)]^2} \qquad (4\text{-}9)$$

$\hat{g}(x_i)$ and $g(x_i)$ are observed values and predicted (or calculated) values respectively.

4.3.2.2. Acceptable fit

Modeled distribution provides an acceptable fit to an observed species abundance distribution if the absolute difference between the observed and the calculated values of n_1 is less than 15% of the observed value, which means that $\left| n_1^{obs} - n_1^{calc} \right| \leq 0.15 n_1^{obs}$ [98]. n_1 is historically meaningful; it has always been an important statistic of a sample, and the SAD models had to give good approximations of n_1 to be validated (Fisher et al. 1943). Finally, from a practical standpoint, ignoring the role of n_1 can lead to unacceptable results of a goodness-of-fit test. For example, the distributions depicted in Fig. 1 (left) do not present an acceptable fit, since the error on n_1 is 68% ($n_1^{obs} = 38$, $n_1^{calc} = 64$).

4.3.2.3. Basic Good fit

Based on the notion of acceptable fit, we can define what we call a "basic good fit". We say that the model distribution provides a basic good fit to an observed species abundance distribution, if it presents an acceptable fit, and if a basic X^2 test (chi-square test), applied on both distributions, gives a X^2 that is not significant (for example 5% significance) [98].

The X^2 test is then performed, calculating the observed and expected counts and computing the chi-square test statistic.

$$X^2 = \sum_{i=1}^{N} (O_i - E_i)^2 / E_i \qquad (4\text{-}10)$$

Where O_i are the observed counts and E_i are the expected counts. The statistic has an approximate chi-square distribution.

To make a direct comparison between test results possible, all the data in the result section includes a chi square test with corresponding scores adjusted to a degrees of freedom equal to *S-1* (*S* is number of species or number of classes).

4.3.2.4. By-class Good Fit

Because of the problem of statistical invalidity of the X^2 test when applied on too small values, another solution consists in grouping the terms of a usual species abundance distribution into classes, in order to produce a grouped species abundance distribution [98].

Analyzing some geometrically varying data is more convenient if they are transferred on a logarithmic scale [235]. The naive approach would be to use the base 2 for the logarithm, but this presents the disadvantage to violate the independence of data points. Traditionally, a base 3 logarithm is used to transform the abundance data. This is done by grouping data into "× 3 classes" (C_k) $k \geq 0$: class C_k has its center at 3^k, and its edges at $3^k/2$ and $3^{k+1}/2$. When used with integer values, it gives the following classes: class C_0 contains only 1; class C_1 contains 2, 3 and 4; class C_2 contains integers from 5 to 13; class C_3 contains integers from 14 to 40; class C_4 contains integers from 41 to 121; etc. After log-scaling, we can use formula 9 and 10 to measure difference between observed and predicted values.

4.3.3. The FPLP model

The FPLP model is based on a combination of other models and follows the stacking approach, which uses a combination of models to generate performing predictors [220]. By combining models, we expect a more accurate prediction at the expense of an increased complexity of the final model.

Suppose we have a sample dataset *Z* and a number of different models with a good performance on *Z*. We can pick a single model as the solution, running into the risk of making a bad choice for the problem. Instead of picking just one model, a safer option would be to use them all and "average" their outputs. The new model might not be better than the single best model but will diminish or eliminate the risk of picking an inadequate single model [220]. This approach will

lead to a more robust predictive method. The proposed new method is based on the combination of four different models:

Fisher's logseries,

Power-Law,

Logistic-J, and

Poisson Lognormal distribution

The main reason of selecting these four models is that they represent different families of methods for SAD modeling and we tried to pick one method from each family. Each family has a specific approach for modeling the species abundance distribution and we wanted to include most of these approaches to have enough flexibility to generate all possible predictions for modeling. We expect to enrich our global model if we choose our base models from different families. Moreover, as a FPLP model uses a learned weighted combination of base models, it provides information on the relative importance of each base model on each sub-range of species abundance distribution. The FPLP method can be view as a post-processing process that combines several fits, using a pre-computed weighted combination, to generate a new fit.

Another important point is that we combine the basic models in three equal sub-ranges of the whole range of values. We found that the SAD pattern is so complex that it could not be modeled by a single formula. In addition, partial combination gives us more flexibility to use all aspects of combination of models. We noticed that every single base model that we used obtained good prediction levels in specific sub-ranges of SAD. Therefore, we chose to build our model using combinations of the basic models, one for each sub-range we considered. We have chosen three sub-ranges because of a trade-off between two extreme situations: having too many sub-ranges leads to the over-fitting problem; more sub-ranges enhance the flexibility capability of new model and therefore improve the quality of the match. A division in three sub-ranges seems to be a good initial compromise but more study is needed to be done to see what the impact of the number of sub-ranges and their positions is. Moreover, it has been shown in [236] that the community abundance distribution might have at least three modes and it could be another justification of having three sub-ranges in our combinatorial model. Since there is no limitation on sum of weights (see below) and we have three sub-ranges, it is possible to have multimodal patterns and even, varying the number of sub-ranges, the exact number of modes can be specified. Consequently, to predict a SAD for a specific community, we divide it into three sub-ranges and

combine the four basic models in each sub-range independently leading to twelve weights to be learned. The process for making and evaluating the FPLP method is: 1) For each dataset, knowing the population and the number of species, the parameters of each base model (Fishers,...) is computed. In the training part we just computed the weights of the combination of functions for the FPLP model. 2) Then we used the computed weights for the rest of datasets to predict the SAD by FPLP model. 3) We compared the accuracy of predicted SADs by base models with the FPLP model.

We used the genetic algorithm [237] to estimate the weights; however other optimization method could also easily be applied. Genetic algorithms are combinatorial optimization methods belonging to the class of stochastic search methods [238]. Whereas most stochastic search methods operate on a single solution of the problem at a time, genetic algorithms operate on a population of solutions. They can be viewed as a kind of hill-climbing optimization approach [239] applied simultaneously to a population of solutions. In our problem, a solution is a combination of the values of the weights associated to each base model that minimize the error between SAD's real values and predicted SAD values. For the learning process, as we try to minimize the error, we used the SPE method to evaluate the performance of each weight combination. More precisely, in our case 12 weights are used: one for each three sub-ranges for each four base models (see Table 4-8).

Table 4-8. Interpretation of weights in three sub-range combinations for four base models

	W1 (Fisher)	**W2 (Logistic-J)**	**W3 (Power-law)**	**W4 (Poisson-Log)**
Sub-range 1	w_{11}	w_{12}	w_{13}	w_{14}
Sub-range 2	w_{21}	w_{22}	w_{23}	w_{24}
Sub-range 3	w_{31}	w_{32}	w_{33}	w_{34}

The weights can be obtained by fitting the FPLP model to a dataset. In this study, we measured the weights with using two different datasets and we evaluated the performance of this method to show how robust this method is no matter what type of dataset is used for estimating the weights for the combination. As the weights only represent the relative importance of each of the 4 base

models for each of the three sub-ranges, the weights learned on a particular dataset are still valid in different conditions and can therefore be used on different datasets.

Problem of over fitting, which is traditional pitfall of learning methods, is bypassed by setting some stop criterion. For example, during the genetic algorithm process, the learning process was stopped in generation number 15. In other words, we do not allow the process to go through over and over until it perfectly matches on training dataset with a risk of loss of generality.

In order to show how the FPLP model works based on the combination of other models, we showed the prediction of every model as below in which the indices 1, 2, 3, 4 have been chosen for Fisher's Logseries, Logistic-J, power-law and Poisson-lognormal respectively.

$$\begin{aligned} SAD_{Fisher} &= n_{11}, n_{12}, ..., n_{1k} \\ SAD_{Logistic-J} &= n_{21}, n_{22}, ..., n_{2k} \\ SAD_{Power-Law} &= n_{31}, n_{32}, ..., n_{3k} \\ SAD_{Poission-Lognormal} &= n_{41}, n_{42}, ..., n_{4k} \end{aligned} \quad (4\text{-}11)$$

Based on the characteristics of every dataset, such as number of species and number of individuals, the species abundance distribution can predicted by the base models (see formula 11). Based on the combination of models for the FPLP model, we have:

$$\begin{aligned} SAD_{FPLP} = W_1.SAD_{Fisher} + W_2 SAD_{Logistic-J} + W_3 SAD_{Power-Law} + W_4 SAD_{Poission-Lognormal} = \\ (w_{11}.SAD^{part1}_{Fisher} + w_{12}.SAD^{part1}_{Logistic-J} + w_{13}.SAD^{part1}_{Power-Law} + w_{14}.SAD^{part1}_{Poission-Lognormal}) + \\ (w_{21}.SAD^{part2}_{Fisher} + w_{22}.SAD^{part2}_{Logistic-J} + w_{23}.SAD^{part2}_{Power-Law} + w_{24}.SAD^{part2}_{Poission-Lognormal}) + \\ (w_{31}.SAD^{part3}_{Fisher} + w_{32}.SAD^{part3}_{Logistic-J} + w_{33}.SAD^{part3}_{Power-Law} + w_{34}.SAD^{part3}_{Poission-Lognormal}) = \\ w_{11}.n_{11} + w_{12}.n_{21} + w_{13}.n_{31} + w_{14}.n_{41}, ..., w_{21}.n_{1\frac{2k}{3}} + w_{22}.n_{2\frac{2k}{3}} + w_{23}.n_{3\frac{2k}{3}} + w_{24}.n_{4\frac{2k}{3}} \\ , ..., w_{31}.n_{1k} + w_{32}.n_{2k} + w_{33}.n_{3k} + w_{34}.n_{4k} \end{aligned} \quad (4\text{-}12)$$

The information on the weights for each sub-ranges and for each model of the combination is given in Table 4-8.

As we have access to the value of the learned weights, we have some information about what model is important (higher weight) for a particular sub-range. It is very important in terms of interpretation of the results. It gives the possibility to discover some specific properties of particular sub-ranges. For example, seeing that the weight of the Fisher's logseries model is very

high for the second sub-range can tell us that the set of species that have an average number of individuals have a distribution, which closely follows the Fisher's logseries distribution. It gives also the possibility to compare, for each sub-range, the relative predictive capacity of each model used and therefore have a better understanding of their relative importance. As a consequence, our approach could be very helpful for a more precise analysis of the properties of the observed distributions and contribute to build a better ecological theory to explain the distribution patterns observed in a given community. In the results presented in the next section we always use three combinations of the four basic models for the FPLP model.

4.3.4. Results and Discussions

In this section we made a comparison between our new model and other models. We compared them according to the several goodness-of-fit to see how general different methods are in modeling of SAD in different cases. The species abundance distributions are different in terms of environment and other ecological factors. This means that different datasets have different characteristic. In order to see how general SAD models are in modeling different datasets, we used a dataset for training the FPLP method (computation of weights) and then use the same weights to get the SAD for other datasets.

From the results (see Table 4-10), it appears that the Fisher's model make better prediction than Logistic-J, Power-law and classic Poisson lognormal, therefore we only use Fisher's logseries and two recent proposed methods: ZSM [229] and advanced Poisson lognormal [240] for comparison purposes. An interesting thing that can also be seen in Figure 4-14 is that the combination of a Fisher's logseries model with other base models leads to a more accurate global model. In order to give a first visual comparison, the outputs for the four selected base models and FPLP model over Mudamali dataset are shown in Figure 4-14.

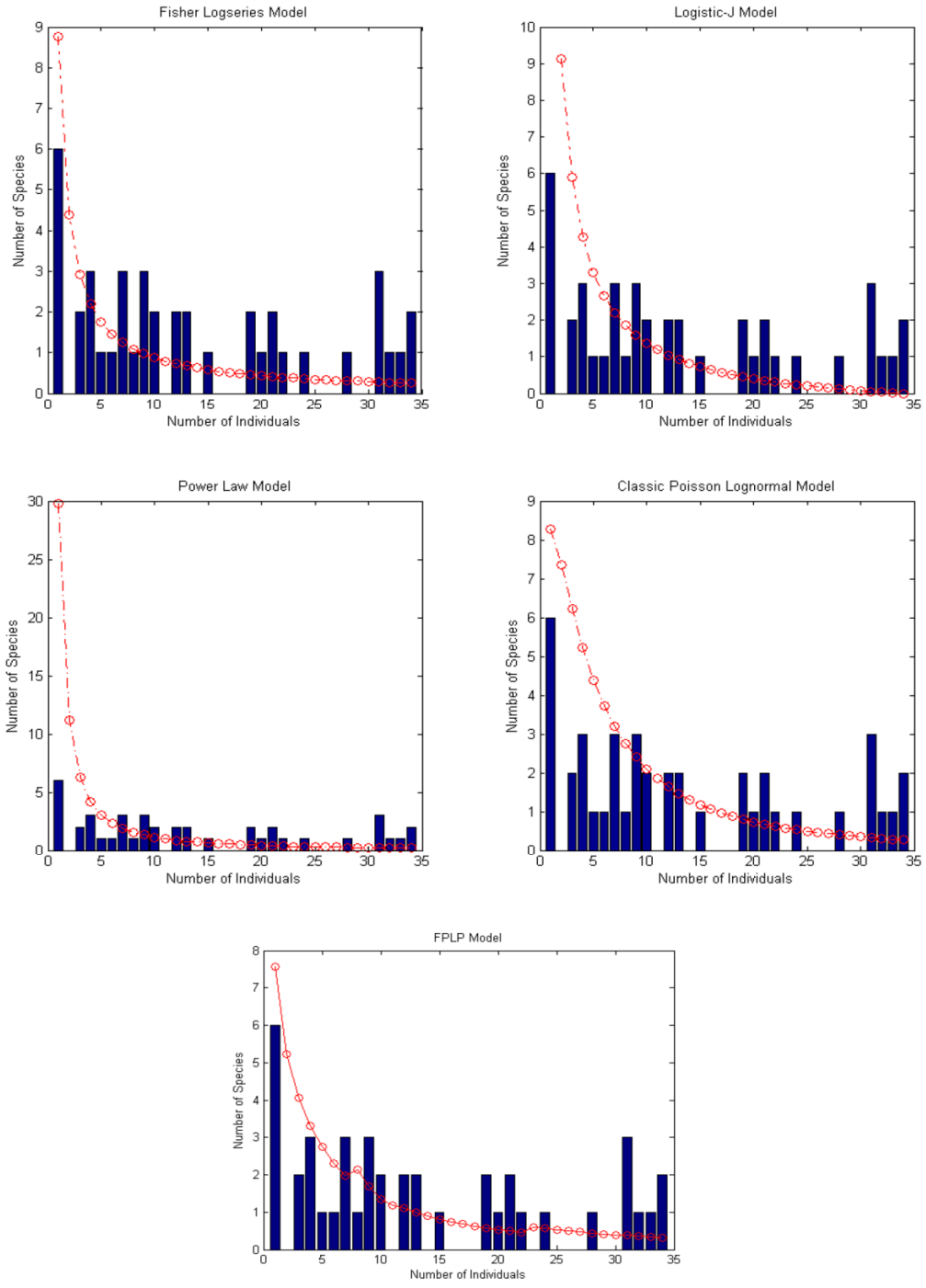


Figure 4-14. Prediction of Fisher's logseries model, Logistic-J model, Power-law model, Classical Poisson Lognormal model and FPLP model on Mudamali dataset (dataset from real ecosystem).

In the following we compare more in depth our model with Fisher's model, extended version of Poisson Lognormal model (PN) that was proposed in [240] and model based on Neutral theory

(ZSM), which was proposed in [229]. We test these models over eight different datasets (see Table 4-9). With these comparisons we can make a better judgment about the relative efficiency of the models (this is an important feature according to [219].

It is worth mentioning that the diversity measures can be affected by the sampling process. For example rare species are less likely to be observed in small samples than in large samples. So the sample size could be critical in estimation of species richness [236]. For this reason we consider different datasets with different size of sample.

Table 4-9. Characteristics of different real datasets from nature (S: Number of Species, N: Number of Individuals).

Dataset	Description	N	S
Sherman [241]	trees of the Sherman 6 ha forest plot, Panama	22000	230
Dirks [242]	sample of Lepidoptera, Maine, USA	55539	349
Fushan[241]	trees of the Fushan 25 ha forest plot, Taiwan	114511	110
HKK [243]	trees of the Huai Kha Khaeng 50 ha forest plot, Thailand	78444	287
Bell[243]	bird community of lowland rainforest in New Guinea	27112	165
Thiollay [244]	Birds in French Guiana, 1986	8507	315
Mudumalai1988 [245]	Trees, Mudumalai 50 ha plot, 1988	25551	70
Malaysian Butterflies [99]	Malaysian butterflies	9029	620

The eight datasets have been used for the evaluation process (see Table 4-9). For each dataset there is the number α, which reflects diversity of that community sample. In order to ensure a

complete evaluation process, we trained the FPLP model on one dataset with low α value, computing the combination of weights, and then we compare the performance of all models on all datasets. We also repeat this experiment using a dataset with high α value for computing the combination of weights and then we compare the performance of all models on all datasets.

4.3.4.1. Learning with a low α value dataset

In this experiment, the FPLP model has been trained over the Fushan dataset, as it has been explained in section 4.3.3, to compute the weights for combining the basic models in three sub-ranges of the whole range of values. We divided the Fushan dataset in three equal size sub-ranges and then learned the weights (w1, w2, w3, w4) to combine the four basic models in each sub-range independently. We give here the 12 weights learned:

Sub-range 1:	Sub-range 2:	Sub-range 3:
$W_{logseries} = 0.464$	$W_{logseries} = 0.963$	$W_{logseries} = 0.805$
$W_{Logistic\text{-}j} = 0.18$	$W_{Logistic\text{-}j} = 0.439$	$W_{Logistic\text{-}j} = 0.398$
$W_{power\text{-}law} = 0$	$W_{power\text{-}law} = 0.185$	$W_{power\text{-}law} = 0.116$
$W_{PN} = 0.282$	$W_{PN} = 0.02$	$W_{PN} = 0.356$

From the value of the weights, it can be deduced that the Fisher's logseries model has a much better predictive capacity than the other models. It seems also that the power-law is not a good predictor for these data and that the third sub-ranges is more complex to describe as it needs a more homogeneous combination of the four models to reach a high predictive level. Values in Table 4-10 indicate the prediction's error of the different methods for every dataset and every model. The bold numbers are used in Table 4-10 when FPLP model outperforms all other methods. For all these results, lower the value is better the fit is.

Table 4-10. Different errors of the four selected models over eight various datasets. FPLP models is trained over the Fushan dataset

Dataset	Model	Accepted Fit(%)	SPE	Basic Good Fit	By-class (SPE)	By-class (Chi-square)
Sherman 1996	Fisher's	26.12	10.85	19.66	11.27	4.28

α = 35.3709	logseries					
	PN	6.08	20.54	127.14	47.60	71.64
	ZSM	47.51	22.57	67.99	45.96	67.30
	FPLP Model	14.21	**9.44**	23.44	14.53	6.97
Driks α = 49.7198	Fisher's logseries	30.72	21.32	25	32.99	16.40
	PN	5.46	21.66	60.83	35.82	18.08
	ZSM	80.24	55.35	177.42	107.04	179.25
	FPLP Model	8.73	**19.07**	**21.36**	**23.97**	**7.82**
Fushan α = 12.0045	Fisher's logseries	50.04	5.5	4.17	7.03	6.69
	PN	57.89	13.85	108.62	31.74	125.6
	ZSM	135.66	14.35	57.75	28.44	141.86
	FPLP Model	**15.44**	**5.16**	6.41	11.43	9.28
HKK α = 37.5398	Fisher's logseries	63.13	20.29	28.73	24.39	19.06
	PN	43.29	23.63	84.31	45.87	44.86
	ZSM	77.64	36.73	113.97	69.69	115.84
	FPLP Model	**17.28**	**12.98**	**18.55**	**9.62**	**2.67**
Bell α = 23.3823	Fisher's logseries	66.87	12.02	15.06	10.75	7.65
	PN	35.33	17.52	87.74	35.05	55.19
	ZSM	12.65	11.64	39.25	29.82	40.13

	FPLP Model	16.71	**7.63**	**13.81**	**7.95**	**3.11**
Thiollay α = 64.4038	Fisher's logseries	68.21	29.41	50.63	34.86	24.98
	PN	4.81	17.96	55.19	24.17	8.46
	ZSM	68.88	50.2	190.17	115.23	189.62
	FPLP Model	5.67	**16.27**	**36.14**	24.94	8.49
Mudumalai1988 α = 8.7754	Fisher's logseries	46.20	7.85	6.41	10.88	11.1
	PN	33.97	10.35	20.25	14.12	33.15
	ZSM	363.8	23.84	31.7	30.9	110.4
	FPLP Model	**6.72**	**7.07**	**3.75**	**8.11**	**7.21**
Malaysian butterflies α = 150.92	Fisher's logseries	25.79	36.04	39.27	50.61	19.2
	PN	39.66	55.15	68.84	60.18	27.82
	ZSM	94.84	152.32	493.29	263.9	493.15
	FPLP Model	**24.11**	40.67	44.8	**50.19**	**18.41**

According to Table 4-10, the FPLP combination model produces more accurate results for most of datasets, even in dataset with high value of α. When the FPLP method is not the best model, the accuracy of FPLP method is still reasonable and close to the best one, which shows the robustness of this method. In this evaluation process, there are 5 goodness-of-fit methods and 8 datasets. Therefore, 40 different comparison tests have been performed. FPLP method outperforms, Fisher's Logseries on 32 of these comparisons, PN on 35 and ZSM on 39. The average percentage of improvement for FPLP method is summarized in Table 4-11. For example, we computed the average percentage of improvement of FPLP on PN model for each dataset, and

then we average the results over all datasets. The percentage values on different measures are not in same scale.

Table 4-11. The average percentage of improvement of FPLP method compared to other methods in various measures in the case of using "Fushan" dataset as a training data set.

Criterion	FPLP vs Fisher's Logseries	FPLP vs PN	FPLP vs ZSM
Accepted Fit(%)	352%	112%	1109%
SPE	28%	75%	182%
Basic Good Fit	16%	456%	573%
By-class (SPE)	30%	157%	334%
By-class (Chi-square)	131%	745%	2005%

From statistical point of view, we can investigate the result of Table 4-10 to see how significant the difference between the FPLP model and the other models is. For this reason, we used the t-test [246]. The t-test assesses whether the means of two groups are statistically different from each other. The results of applying t-test to the data of Table 4-10 are presented in Table 4-12.

Table 4-12. The p-value for the distance between error rates of different models for each measure in the case of using "Fushan" dataset as a training data set.

Criterion	FPLP vs Fisher's Logseries	FPLP vs PN	FPLP vs ZSM
Accepted Fit(%)	0.0001	0.001	0.00001
SPE	0.04	0.03	0.0001
Basic Good Fit	0.06	0.0001	0.00001
By-class (SPE)	0.04	0.0001	0.00001
By-class (Chi-square)	0.0001	0.00001	0.00001

Except when we consider the FPLP model with Fisher's Logseries model for the "Basic Good Fit", in all other cases the p-value is less than 0.05, which means that the differences between the FPLP model with the other single models are statistically significant.

4.3.4.2. Learning with a high α value dataset

In this experiment, the FPLP model has been trained over "Malaysian butterflies" dataset, which has high value of α. We divided the Malaysian butterflies dataset in the same three sub-ranges and then estimated the weights (w1, w2, w3, w4) to combined the four basic models in each sub-range independently. We give here the 12 weights learned:

Sub-range 1:	Sub-range 2:	Sub-range 3:
$W_{logseries} = 0.464$	$W_{logseries} = 0.788$	$W_{logseries} = 0.805$
$W_{Logistic\text{-}j} = 0.788$	$W_{Logistic\text{-}j} = 0.429$	$W_{Logistic\text{-}j} = 0.996$
$W_{power\text{-}law} = 0$	$W_{power\text{-}law} = 0.558$	$W_{power\text{-}law} = 0.116$
$W_{PN} = 0.337$	$W_{PN} = 0.02$	$W_{PN} = 0.325$

The values obtained for these data are quite different from the previous ones. It seems that for these data the logistic-j is a much better predictor than for the previous dataset. The power-law is also quite important for the prediction of the middle sub-range distribution. The values in Table 4-13 indicate the prediction's error of the different methods for every dataset and every measurement method. The bold numbers are used in Table 4-13 when FPLP model outperforms all other methods.

Table 4-13. Different errors of the four selected models over eight various datasets. Models trained over Malaysian butterflies dataset.

Dataset	Model	Accepted Fit(%)	SPE	Basic Good Fit	By-class (SPE)	By-class (Chi-square)
Sherman 1996	Fisher's logseries	26.12	10.85	19.66	11.27	4.28

α = 35.3709	PN	27.91	23.66	116.97	46.46	66.64
	ZSM	47.51	22.57	67.99	45.96	67.3
	FPLP Model	**18.9**	**10.75**	22.44	15.03	7.63
Driks α = 49.7198	Fisher's logseries	30.72	21.32	25	32.99	16.4
	PN	43.64	28.52	62.38	40.98	27.56
	ZSM	80.24	55.35	177.42	107.04	179.25
	FPLP Model	**2.76**	**16.99**	**21.72**	**15.88**	**3.37**
Fushan α = 12.0045	Fisher's logseries	50.04	5.55	4.17	7.03	6.69
	PN	115.05	16.84	107.74	32.16	120.2
	ZSM	135.66	14.35	57.75	28.44	141.86
	FPLP Model	**24.38**	6.2	5.59	9.3	8.58
HKK α = 37.5398	Fisher's logseries	63.13	20.29	28.73	24.39	19.06
	PN	95.16	33.32	97.84	51.69	66.48
	ZSM	77.64	36.73	113.97	69.69	115.84
	FPLP Model	**25.39**	**13.5**	**26.57**	**11.45**	**3.84**
Bell α = 23.3823	Fisher's logseries	66.87	12.02	15.06	10.75	7.65
	PN	84.33	22.33	98.14	36.82	65.48
	ZSM	12.65	11.64	39.25	29.82	40.13
	FPLP Model	24.37	**8.81**	19.57	11.95	**6.89**

Thiollay α = 64.4038	Fisher's logseries	68.21	29.41	50.63	34.86	24.98
	PN	29.65	22.79	56.13	34.17	18.74
	ZSM	68.88	50.2	190.17	115.23	189.62
	FPLP Model	**11.06**	**15.42**	**35.53**	**19.94**	**5.65**
Mudumalai1988 α = 8.7754	Fisher's logseries	46.2	7.85	6.41	10.88	11.1
	PN	82.47	12.37	31.33	16.67	48.16
	ZSM	363.8	23.84	31.7	30.9	110.4
	FPLP Model	**14.3**	**7.32**	**3.91**	**7.87**	**8.49**
Malaysian butterflies α = 150.92	Fisher's logseries	25.79	36.04	39.27	50.61	19.2
	PN	17.82	35.36	53.13	53.91	20.67
	ZSM	94.84	152.32	493.29	263.9	493.15
	FPLP Model	20.7	**35.2**	**38.93**	**39.97**	**11.87**

According to Table 4-13, the FPLP combination method produces more accurate results in most of datasets even in dataset with low value of α. In this evaluation process, there are also 5 goodness-of-fit methods and 8 datasets. Therefore, 40 different comparison tests have been performed. The FPLP method outperforms Fisher's logseries on 31 of these comparisons, PN on 39 and outperforms ZSM on 39. The average percentage of improvement for FPLP method is summarized in Table 4-14.

Table 4-14. The average percentage of improvement of FPLP compared to each measure in the case of using "Malaysian butterflies" dataset as a training data set.

Criterion	**FPLP vs Fisher's**	**FPLP vs PN**	**FPLP vs ZSM**

	logseries		
Accepted Fit(%)	280%	381%	861%
SPE	25%	97%	181%
Basic Good Fit	9%	487%	573%
By-class (SPE)	37%	173%	371%
By-class (Chi-square)	144%	754%	2429%

We also applied the t-test to compare how significance the difference is between the FPLP model and the other models (see Table 4-15).

Table 4-15. The p-value for the distance between error rates of different models for each measure in the case of using "Malaysian butterflies" dataset as a training data set.

Criterion	FPLP vs Fisher's Logseries	FPLP vs PN	FPLP vs ZSM
Accepted Fit(%)	0.0001	0.0001	0.00001
SPE	0.03	0.02	0.001
Basic Good Fit	0.08	0.0001	0.0001
By-class (SPE)	0.02	0.001	0.0001
By-class (Chi-square)	0.001	0.0001	0.00001

Like in our previous experiment, except when we consider the FPLP model with Fisher's Logseries model for the "Basic Good Fit", in all other cases the differences between the FPLP model and the other single models are statistically significant.

As it can be seen in Table 4-11 and Table 4-14, Fisher's logseries generally outperforms the recent methods PN and ZSM. It can be due to the fact that these methods have been developed for very

specifics cases and they are not robust enough for general cases. The results show clearly an important improvement of our new model compared with the three others. The improvement in average quality of prediction is very large compared with the two recent methods. The average improvement of our model is always positive for all measures and against all other tested models. What is also very important to notice is that our approach seems to be quite robust and works well even to make a prediction on distributions that are very different from the ones used to learn the parameters of our model.

4.3.5. Conclusion

In this study, the new species abundance distribution model (FPLP model) has been proposed. The FPLP model is based on the combination of several other base models. In response to the criterion defined in the McGill's survey, we have performed a large experimental comparison protocol with our model and the best existing and promising models. We also used eight different datasets with various characteristics corresponding to very different species abundance distributions. We also used 5 different criterions for evaluating the quality of fit of the models. We have shown that our model outperforms the Fisher's logseries model, which itself outperforms the two recent models PN and ZSM for all criteria used. The improvements obtained are impressive and statistically significant.

These results show that the approach based on the combination of learned models is very promising and leads to robust and accurate predictors. One important point for these kinds of method is the choice of the base models. The main factor for the efficiency of the resulting global model is the diversity of prediction of the base models. Another important component of our approach is the decomposition of the range of the distribution in three sub-ranges. It seems that this concept is very important because different sub-ranges of the distribution have different characteristics, which can hardly be represented by one unique model.

To be able to conceive ecological theory from an observed SAD, it is very important to clearly understand what distribution this SAD follows. Because we use a weighted combination of models, we know the relative importance of each basic model in each sub-range. In other words, we have a global model, which is a combination of four other basic models and, because we have the weights associated to each of them, we can deduce how close to each model, for every sub-ranges, the observed SAD is. For this reason the combination distribution significantly outperforms other approaches based on a single model as a descriptor of abundances in communities. Obviously the weights are computed for a given community (due to the training

phase) and therefore they are a good instrument to discover specific properties of a given community. However, we have also shown that the predictor we build on a specific community is still a good predictor for a large range of other different communities, outperforming every single model approaches for this task. From our experiment we have observed that the predictive capacities of the four basic models we used vary a lot depending on the value of the α parameter. For low α values, the Fisher's logseries seems to be a much better descriptor than the other models. But for high α values, the logistic-j model seems to be more important. We have also observed high variations of the relative importance of these models depending on the sub-ranges considered.

4.4. Identifying Important Characteristics to Predict Changes in Species Richness in EcoSim

Species richness is one of the important measures used by ecologists. Species richness is a critical variable for biodiversity management that has been used for decision making and prioritization of conservation efforts [247], [248]. Ecological theory assumes that species richness is determined in part by environmental gradients and resources [249]. Defining a set of environmental variables, which are recognized to entail direct or indirect responses from presence/absence of species and linking them by an ecologically-relevant statistical model enable the acquisition of significant information aimed at conservation planning [249]–[251]. Several studies have also demonstrated strong relationships between total species richness and measures of temperature, precipitation and net primary productivity [252], [253]. Developing a standardized method of predicting species richness is vital for international conservation efforts [247], [248]. Few tools are available to provide decision makers with relevant data on biodiversity patterns, ecosystem processes, and underlying forces at spatial scales from local to global [254]. Considering working with real data, it is highly expensive and time-consuming to measure species richness over extensive areas, especially for nonvascular plants and invertebrates and in tropical or marine ecosystems [255].

By using computer simulations, it would be possible to examine factors that could affect the performance of models that predict species occurrence based on environmental variables [256]. Simulation modeling explicitly incorporates the processes believed to be affecting the geographical ranges of species and generates a number of quantitative predictions that can be compared to empirical patterns. The simulation approach offers new insights into the origin and maintenance of species richness patterns, and may provide a common framework for investigating the effects of contemporary climate, evolutionary history and geometric constraints

on global biodiversity gradients [257]. But most of the simulations failed to provide a conceptual bridge between macroecology and biogeography. The problem is that those simulations contain a lot of simplifications [257]. They are not as complex as real ecosystems [23], [32], therefore in most cases the results that come from those simulations are not valid for making any conclusion for real systems.

In this study, we tried to predict the changes in the number of species and identify the most important features that can be used for such a prediction [34]. We used EcoSim [29], our multi-food chain evolving ecosystem simulation. In this study, we tried to predict the variation in the number of species in EcoSim by applying machine learning techniques. In other words, we trained a decision tree which is one of the well-known machine learning techniques from sample data to learn a model for predicting (classifying) an increase or a decrease in the number of species in the next 100 time steps (see section 4.4.1).

In this research, we tried to predict the changes in the number of species using several important features by applying machine learning techniques such as different feature selection algorithms and decision trees. To the best of our knowledge, this is the first time that a complex agent-based simulation has been used to examine the effects of different features on prediction of changes in species richness by extracting meaningful rules from environmental and genetic parameters.

For extracting rules and finding a relationship between environmental variables and species richness, different approaches using nonparametric coefficients, especially decision trees, have been demonstrated to outperform linear models since both linear and nonlinear relationships between biotic and abiotic components were well identified [258]. Therefore we used this machine learning algorithm to select potential features for the sake of species richness prediction. Our objective in this study, was to conduct a robust test of the effectiveness of our framework for identifying important features in prediction of changes in the number of species and introducing a restricted set of features that could help biologists to focus on a specific variables (since there are lots of features that can be studied by biologists). Using these simulations as a shortcut can save time and resources for biologists.

4.4.1. Development of a predictive model

In this study, the goal is the prediction of changes in species richness for next100 time steps using a set of features from EcoSim, which produces a large amount of data about the individuals and the species in each time step. We conducted three runs of the simulation with the same parameters. The prepared training dataset comes from two independent runs that contain 20,000

samples (10000 time steps for each unique run) related to about 38 species on average. Each sample is labeled ‘smaller’ or ‘bigger’ if the number of species in the world respectively has decreased or has increased (or without change) 100 time steps later. The test set contains about 10,000 samples. Both the training and the test datasets contain almost an equal number of 'smaller' labels and 'bigger' labels. The most important part for prediction is the selection of the most significant features. In each time step, every individual has a certain number of attributes (features). We started our learning process with an initial set of 49 features. These features are averaged over all individuals and are: 12 sensitive concepts’ average activation level, 7 internal concepts’ average activation level, 7 motor concepts’ average activation level, 11 actions frequency, the total amount of food in the world, the total population size, the ratio of individuals in a species to the whole population size, the number of dead individuals in the world, the genetic diversity of the whole population, the average age of individuals, the average energy and speed of individuals, the average genetic distance of all the genomes of the individuals from initial genome, the average amount of energy transmit from a parent to a child (parental investment) and the current number of species. The genetic diversity of a species measures how much diversity exists in the gene pool of the individuals of a species. The entropy measure, which we use in this project, is commonly used as an index of diversity in ecology and increasingly used in genetics [259].

We used decision tree as a predictive model, applying the C4.5 algorithm implemented in [260]. Decision trees are effective techniques for discovering the linear and non-linear structures in data and are simpler to interpret than artificial neural networks since they provide a set of binary decision rules. Even if the decision tree technique is not the best machine learning techniques in terms of accuracy of the obtained model, the possibility to understand the obtained model and to discover the effect of the variables on the prediction is what have guided our choice for this approach.

Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree [261], [262]. Learned trees can also be re-represented as sets of if-then rules to improve human readability. These learning methods are among the most popular of inductive inference algorithms and have been successfully applied to a broad range of tasks from learning to diagnose medical cases to learning to assess credit risk of loan applicants [263].

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance [261]. Each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute in the given example. This process is then repeated for the subtree rooted at the new node.

Figure 3.1 illustrates a typical learned decision tree. This decision tree classifies Saturday mornings according to whether they are suitable for playing tennis. For example, the instance:

(Outlook = Sunny, Humidity= High, Wind = Strong)

would be sorted down the leftmost branch of this decision tree and would therefore be classified as a negative instance (i.e., the tree predicts no for playing tennis (*PlayTennis : no*)).

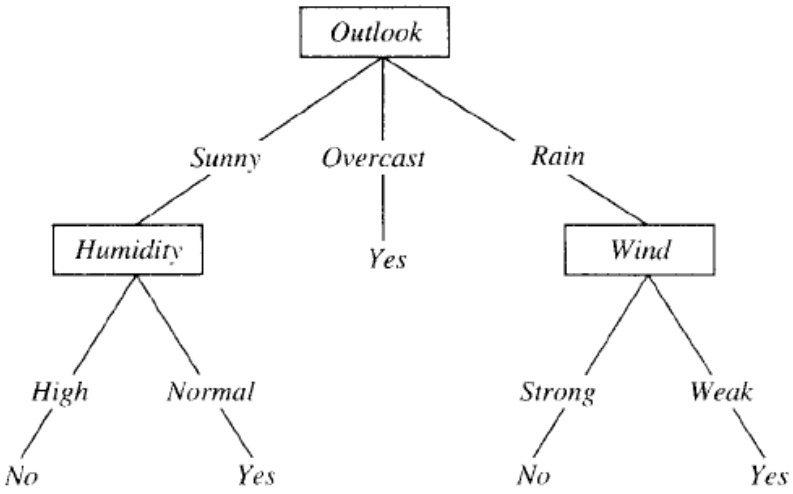


Figure 4-15. A decision tree for the concept *PlayTennis.* An example is classified by sorting it through the tree to the appropriate leaf node, then returning the classification associated with this leaf (in this case *Yes* or *No*). This tree classifies Saturday mornings according to whether or not they are suitable for playing tennis.

In general, decision trees represent a disjunction of conjunctions of constraints on the attribute values of instances. Each path from the tree root to a leaf corresponds to a conjunction of attribute tests, and the tree itself to a disjunction of these conjunctions.

The high number of features leads to very complex models, which are extremely hard to interpret and prone to over-fitting (the obtained tree has 342 rules). Therefore, we tried to reduce the

number of features by selecting the ones that have higher impact on prediction. We used different feature selection algorithms such as Linear-Forward-Selection and Greedy-Stepwise search on WEKA (V3.6.4). These algorithms rank the features by the level of importance in the prediction and eliminate all features that do not achieve any score. Both feature selection algorithms show the highest scores for only five features: Current number of species, amount of food, parental investment, genetic evolution and genetic diversity. These features have been used for learning the prediction model. Using only this subset of features, the prediction accuracy decreases by 5% on training set and increases by 9% on validation set. With these five features, the obtained tree has 35 rules, which are still hard to interpret because they are very specialized using different values of these five features. For example, there is a branch in the tree for every short range of values for a feature. In order to get a smaller tree for extracting meaningful rules with reasonable accuracy, we chose to use decision tree with the confidence factor 0.25 for pruning and 100 minimum instances per leaf [260]. This ensured that the final model neither fitted too specific to the training data set, nor was so general that it renders its predictions meaningless. With this reduction in size, the obtained tree has 10 rules (Figure 4-16). The accuracy decreased by 7% on training set and increased by 3% accuracy on validation set.

For comparing the quality of classification, four measures of accuracy, true positive (TP) rate, true negative (TN) rate, global accuracy, and ROC area have been used. The global accuracy shows the percentage of correctly classified samples. The true positive (negative) rate presents the percentage of true classified positive (negative) samples. Finally, ROC area reveals sensitivity by measuring the fraction of true positives out of the positives versus the fraction of false positives out of the negatives.

For the training and test set, using 10-fold cross-validation, the final tree model has a total accuracy of 82%, the two classes being predicted with almost the same high accuracy. The accuracy of the prediction on training data sets with 10-fold cross-validation is given shown in Table 4-16.

Table 4-16. Results of prediction of species richness for next 100 time steps by decision tree on training set.

Class	TP Rate	FP Rate	Precision	ROC Area
Smaller	0.834	0.184	0.794	0.89
Bigger	0.816	0.166	0.853	0.89

Total	0.824	0.174	0.826	0.89

For the test set, we picked a completely separate run of simulation. In this case the total accuracy is about 80%, which means that, using selected features, prediction of changes in species richness time series is possible with high level of accuracy even on data generated by an independent process (Table 4-17). This means that the rules we have discovered all quite general and could bring some interesting insight on the speciation process.

Table 4-17. Results of prediction of species richness for next 100 time steps by decision tree on test set.

Class	TP Rate	FP Rate	Precision	ROC Area
Smaller	0.777	0.169	0.798	0.872
Bigger	0.831	0.223	0.812	0.872
Total	0.806	0.198	0.805	0.872

4.4.2. Extracting the Rules from Decision Tree

The decision tree effectively modeled much of the variations in species richness, as this method was able to both select a relevant set of predictor variables and to make accurate predictions. The splitting rules used in the partitioning algorithm split the data at values that were ecologically meaningful, describing the relationship between species richness and environmental parameters. This demonstrates the utility of trees as a powerful exploratory modeling tool for building and analyzing prediction models in ecology.

Looking at the selected features and the tree obtained for prediction (Figure 4-16), we can conclude that genetic features and world productivity have an important role on variation of species richness. We can also observe that the tree is well balanced in terms of rule support and in terms of accuracy. It means that all of the rules are important and correspond to a situation characteristic of one of the two possible states we try to predict. One of the rules is about a very high amount of food availability and the number of species that is not low (Rule #3). This rule associates the high level of food to a decrease in the number of species. According to several studies [264], [265], this rule makes sense because when there is a high amount of food in the environment, there are few individuals that consume it. Low number of individuals could be a

sign for a low number of species. According to [264], richness of animal populations is determined by the abundance, distribution and diversity of food resources.

If the number of species is low and also the amount of available food is low (Rule #1), it means that the environment is particularly difficult, the fact that it leads to a decrease in the number of species is quite intuitive. However, this rule is the one with the lowest accuracy, which mean that the phenomenon is not as simple as that. This should explain the multiple rules that exist (#4 to #10) that are in the 'Middle Range' for the amount of food available. If the amount of food is high (Rule #2), it means that it is easy for the individuals to survive and reproduce and, with an increase in population size and as the number of species is currently low, we can expect an increase in the number of species. Using machine learning algorithms like the one that we used allows discovering how adjusting amount of food can be used to control the system. This mechanism could be a direction for future conservation researches.

These two cases correspond to extreme situations for the availability of food, but there are intermediate situations. These cases are trickier for prediction and need the use of other features. Our model discovers the interest of the variable describing parental investment (the average amount of energy transmit from a parent to a child). When parental investment is low and the number of species is also low, the variable describing the distance evolution become involved. Distance evolution reflects the genetic evolution of individuals from beginning. If distance evolution is high (Rule #5), which represent situation in which the evolution is fast, the possibility of an increase in number of species arises and we could expect an increase in the number of species. This rule is one of the most important one, with the highest support and a very good accuracy.

Conversely, if the distance evolution (average genomic distance between the current population and the initial genome at the time step #1) is low (Rule #4), a decrease in the number of species will happen, which make sense. This phenomenon has been found by other studies [30], [266]. They emphasize that mating can contribute to the origin of reproductive isolation by increasing genetic variance, which facilitates splitting of a population into two non-interbreeding parts. According to [266], distance evolution has a direct relationship with the speciation process. If the current number of species is high, other features are needed to make the prediction. One such feature is the amount of genetic diversity (Entropy) that we estimate with the Shannon entropy [92]. When the genetic diversity is high (Rule #7), there are many individuals that cannot mate

together anymore and speciation happens, so we can expect an increase in number of species. Conversely, when the genetic diversity is low (Rule #6), the number of species decreases.

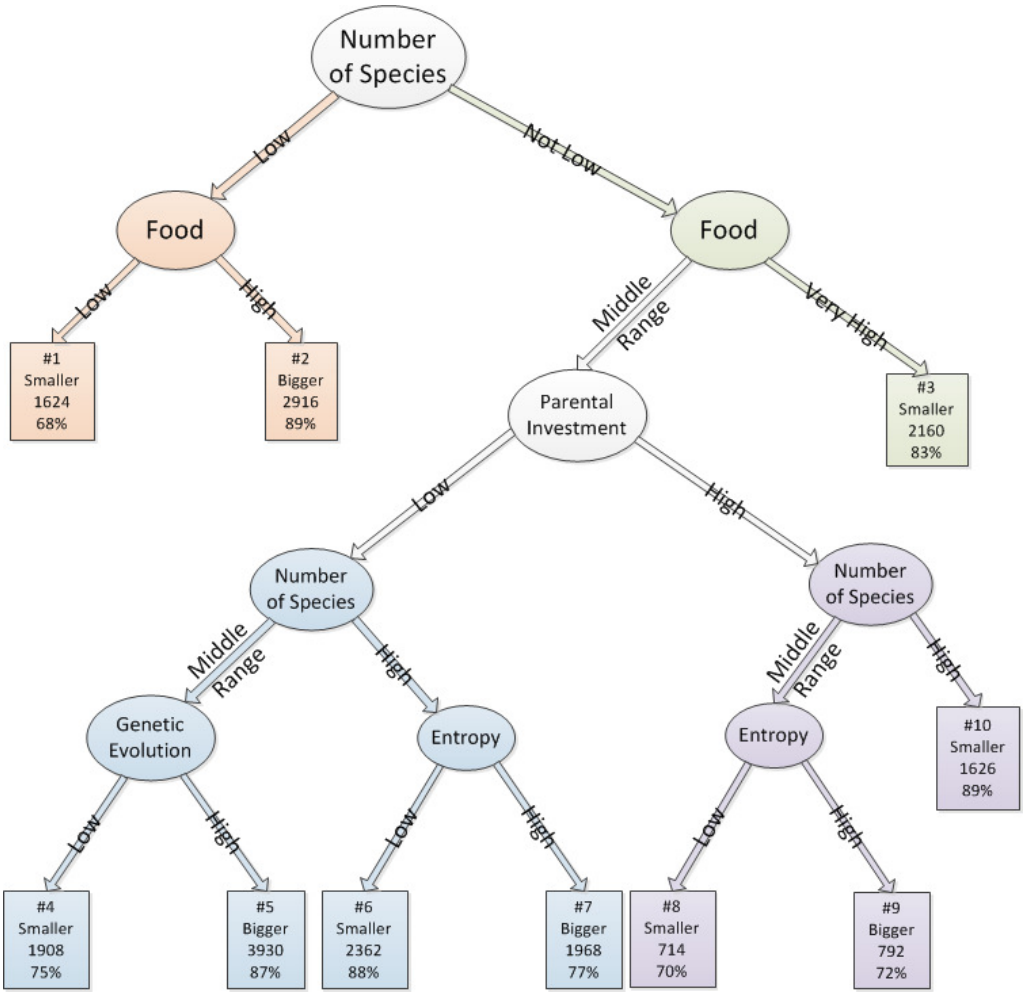


Figure 4-16. The decision tree corresponding to the partitioned feature space for prediction of changes in species richness. Number of samples covered by each rule and the accuracy are also given.

This process also was found by [266], which shows speciation through an increase in genetic variance between populations can occur by evolution over time. This phenomenon has also already been observed in EcoSim [30].

When the parental investment is high and the average number of species is in a middle range, the next important feature again is genetic diversity. High value of genetic diversity (Rule #9) could stand for more possibility of speciation in the next time steps for the same reasons that have been explained above and for low genetic diversity (Rule #8), number of species decreases as well. The parental investment feature itself stands for the amount of energy that is transferred from parents to the new-born individuals. This feature is also subject to mutation during evolutionary

process. High value of parental investment and high number of species (Rule #10, which has the highest accuracy and a good support) means that for such situation (there is also not much food available) having a high parental investment in energy to their child leads to a high probable decrease in the number of species. Other studies also emphasize the effect of balance of energy on species richness [267]. Environmental energy availability can explain much of the spatial variation in species richness [268]–[270].

By identifying the most influential variables (and the relative value for each feature that leads to specific rule), this study provides an important first step towards the development of future predictions of species richness for predator-prey ecosystems that can incorporate higher resolution data.

4.4.3. Conclusion

In this study, a machine learning techniques was applied to data generated by EcoSim, an individual-based ecosystem simulation, to predict variations in species richness. Our objective was to conduct a robust test of the effectiveness of our framework for identifying important features for species richness prediction. We initially used all possible features available to predict species richness. Then we used feature selection algorithms such as Greedy-Stepwise and Linear-Forward-Selection to detect the five most important features that guarantee maximum possible prediction accuracy. By interpreting the obtained decision tree we have been able to extract meaningful rules to enrich our knowledge about the kind of features involved and how their combination can be used to predict species richness variation.

According to the results, a specific range of amount of food available in relation to the current number of species could be critical for ecosystems. So for future records and real data, finding such a relationship could help biologists in conservation efforts. Genetic features have important roles in species richness prediction, which seems reasonable as the whole concept of species rely on the notion of similar genetic characteristics. These results confirmed that our implementation of species in EcoSim has the capacity to reflect concepts and behaviours observed in population genetics that affect the species richness of an ecosystem.

Chapter 5

5. Nonlinear and Chaos Analysis of EcoSim

Nonlinear signal processing is an important research area for many applications. Specifications and identifications of nonlinear signals can help us to detect nonlinear behaviour of dynamical systems. The discrimination of stochastic and deterministic behaviours of nonlinear time series is a basic topic in nonlinear dynamic fields [156]. This specification has attracted researchers for a long time [155], [156].

It has been shown that the evaluation of chaoticity (level of chaos) is an important issue in several applications. Level of chaos means how sensitive a system is (predictability of future values) to slight change of initial condition There are many publications that justify without chaoticity, biological systems might be unable to get discriminated between different stages and thereby different modes of operation [26], [27] (for example epileptic seizures can be detected by using measure of chaoticity [28]). As some researches pointed out, methods based on the largest local Lyapunov exponent can detect the changes of the chaoticity in the biological time series [27]. In the stationary case, the chaoticity quantity can be directly distinguished by the largest Lyapunov exponent (LLE) [147].

We used different criteria to determine whether EcoSim can generate chaotic patterns given that it is a mixture of stochastic and deterministic elements. Another purpose of this analysis is to validate whether EcoSim can generate patterns as complex as patterns that has been observed in real ecosystems. To determine if our simulation system is able to generate chaotic emerging behaviours, we would like to evaluate its level of predictability.

In section 5.1, we investigate whether EcoSim is capable of generating chaotic patterns similar to those observed in nature. In section 5.2, we investigate whether the spatial distribution of individuals in EcoSim follows the same patterns that have been observed in spatial distributions of individuals in real ecosystems along with looking at what the driving forces generating these patterns are.

5.1. Chaos analysis of EcoSim

It has been shown that EcoSim is mixture of stochastic and determinist elements to mimic real ecosystems more accurately (see section 2.4). It is important to investigate the overall behaviour of EcoSim to see if the same chaotic patterns that have been observed in real systems can emerge from EcoSim. In this study, applying nonlinear analysis on the ecosystem simulation's output,

such as population dynamics, we would like to see if EcoSim is a realistic model, capable of generating chaotic patterns [32]. We used Higuchi fractal dimension and Gaussian kernel Algorithm (GKA) method to judge whether the behaviour of population time series in EcoSim [29] is stochastic or deterministic. We also tried to evaluate the chaotic behaviour of population time series in EcoSim. We used largest Lyapunov exponent and P&H method (see section 3.3), which is based on the Poincaré section and the Higuchi fractal dimension for this purpose [149]. Applying four different independent methods will give a strong assessing of behavioural properties of population time series in EcoSim.

An important point is that a simple chaotic system can produce a time series that passes most tests for randomness. Conversely, a purely random system with a non-uniform power spectrum can masquerade for chaos. Thus, we validated our conclusion about whether the behaviour of population time series in ecosystem simulation is deterministic and chaotic or not, by applying the tests to surrogate data designed to mimic the statistical properties of original data using Higuchi fractal dimension, correlation dimension using GKA method, Lyapunov exponent and P&H methods as statistics. A brief overview of our method is as follows:

- The Higuchi fractal dimension, correlation dimension, Lyapunov exponent and P&H method value are extracted from the original population time series of EcoSim and its surrogates.

- For Higuchi fractal dimension and correlation dimension, random and deterministic distributions are constituted respectively as H0 and H1 hypotheses by means of the histograms of surrogate and original data (population time series). For Lyapunov exponent and P&H method, non-chaotic and chaotic distributions are constituted respectively as H0 and H1 hypotheses by means of the histograms of surrogate and original data.

- A desired confidence level is selected, regarding to the value of standard deviation of the distribution. Typically, the interval out of $\mu \pm 4\sigma$ is considered as proper to be the confidence interval.

Finally, the criterion value of original data is compared with the confidence interval of the distribution. If it lies on the confidence interval, the corresponding time series is considered to have deterministic behaviour when the Higuchi's fractal dimension and correlation dimension are used and have chaotic behaviour when Lyapunov exponent and P&H method are used.

5.1.1. Chaos analysis result

This part is an examination of simulation's population data and it will be shown whether these time series have a stochastic behaviour or followed a specific order. For better understanding, the methods have been applied to the simulation's population data, random time series and Lorenz time series, which is one of the well-known chaotic time series. The Higuchi fractal dimension and correlation dimension are used to show the deterministic behaviour of the population time series and largest Lyapunov exponent and P&H method are used to show the chaotic behaviour of population time series in EcoSim.

In surrogate data test method, several time series are generated randomly from original data, but with the same characteristics of original data, such as power spectrum and probability distribution. Then an appropriate statistic (Higuchi fractal dimension, correlation dimension, largest Lyapunov exponent and P&H method) is extracted from both original and surrogate data. Then we should decide if the difference between the values of original data and surrogate data are statistically meaningful or not. To this purpose, one can generate several surrogate time series with different random phases. Then mean and standard deviation for surrogate data sets are computed. Finally, the value of original data is compared to the mean, to see if it is significantly far from the mean considering the value of standard deviation. The results of hypothesis testing have been obtained by using 24 surrogate data sets applying the Theiler algorithm II (see section 3.3.3).

5.1.1.1. Higuchi Fractal Dimension

To determine if the Higuchi fractal dimension of a typical time series presents a stochastic or deterministic behaviour, a hypothesis testing procedure is adopted. Several surrogate data are generated from the population time series, but with different random phases. These surrogate signals have a completely different structure in comparison to the population time series, but similar histogram and power spectrum. The larger the number of generated surrogates is, the more accurate randomness distributions for hypothesis testing are.

The hypothesis under H_0 models randomness and the hypothesis H_1 implies determinism. These two hypotheses are specified in more details:

H0: time series has stochastic behaviour. In the other word, the time series is completely irregular without any special structure.

H1: time series corresponds to a deterministic system. In other words, the extracted statistic represents a more regular structure in compare to the pure random surrogates.

Surrogate data methods are used to construct a probability distribution function under H_0. In Surrogate data methods, several time series are generated randomly from the population time series, but with the same characteristics of population time series, such as power spectrum and probability distribution (see section 3.3.3). Then a Higuchi fractal dimension statistic is extracted from both original and surrogate data. Then, it should be decided if the difference between the values of original data and surrogate data are statistically meaningful or not. To this purpose, one can generate several surrogate time series with different random phases. Then mean and standard deviation for surrogate data sets are computed.

Finally, a confidence level for acceptance or rejection of the hypothesis H_0 can be determined. Typically, the intervals out of $\mu \pm 4\sigma$ are proper to be defined as a confidence interval. If the value of the statistic lays on the confidence level interval, the H_0 hypothesis is rejected, otherwise, H_0 is accepted.

To show the clear distinction between the results obtained with stochastic and deterministic time series, we apply this method to random time series, Lorenz time series and EcoSim's population data. Figure 5-1 shows the results of hypothesis testing on several well-known systems. The red sign indicates the value of the statistic extracted from population time series and the blue one indicates the value of the statistic extracted from surrogate data. As it can be seen in Figure 5-1(left), the hypothesis testing result is in the central range of the distribution for random time series. Therefore, hypothesis H0 is accepted, which means that the time series is random. On the other hand, for Lorenz time series Figure 5-1(right), the red signal indicates the value of the statistic extracted from original data, which is completely separable from its imposter random distribution. It should be noticed that, due to the change of abscise scale compared to the Figure 5-1(left), the surrogate distribution look like a line, but it still corresponds to the distribution Higuchi fractal dimension value for 24 surrogate time series. Therefore, hypothesis H0 is rejected, which means that the population time series is deterministic.

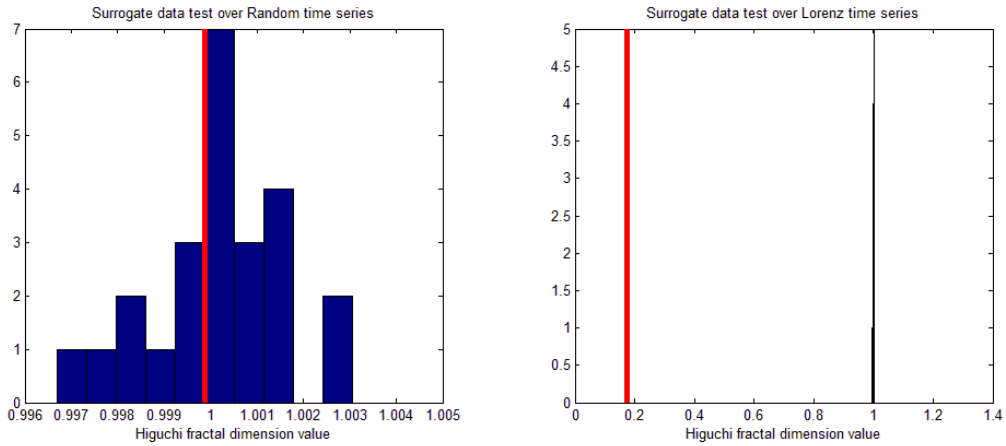
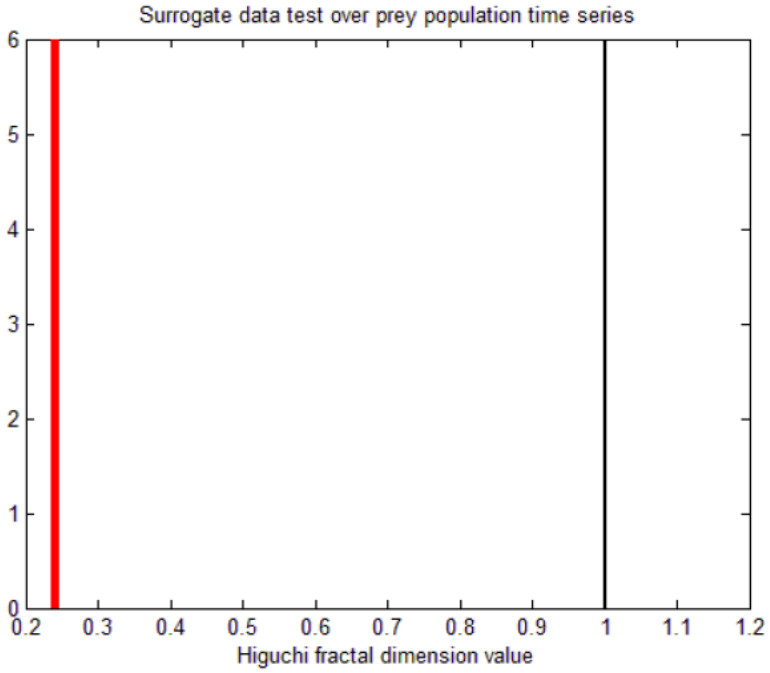
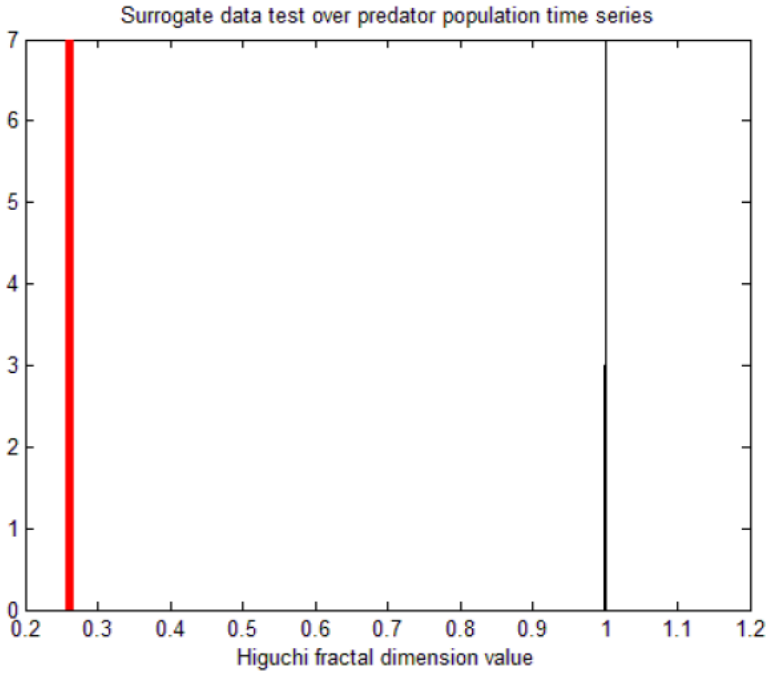


Figure 5-1. The results of hypothesis testing by using 24 surrogate data sets for random time series (left) and Lorenz time series (right) using Higuchi fractal dimension.

Having observed the result of this test for well known time series, the comparison with the results obtained with the simulation time series leads to a clear interpretation. According to Figure 5-2, the red signals indicate the value of the statistic extracted from original data both for prey and predators time series. These values are completely separable from the surrogate distribution. It means the red signals are in the confidence interval, so the null hypothesis are rejected (random behaviour rejected) and the deterministic behaviour of simulation's populations data can be concluded. This criterion therefore, shows that the simulation's population data have a deterministic behaviour.



(a) Prey



(b) Predator

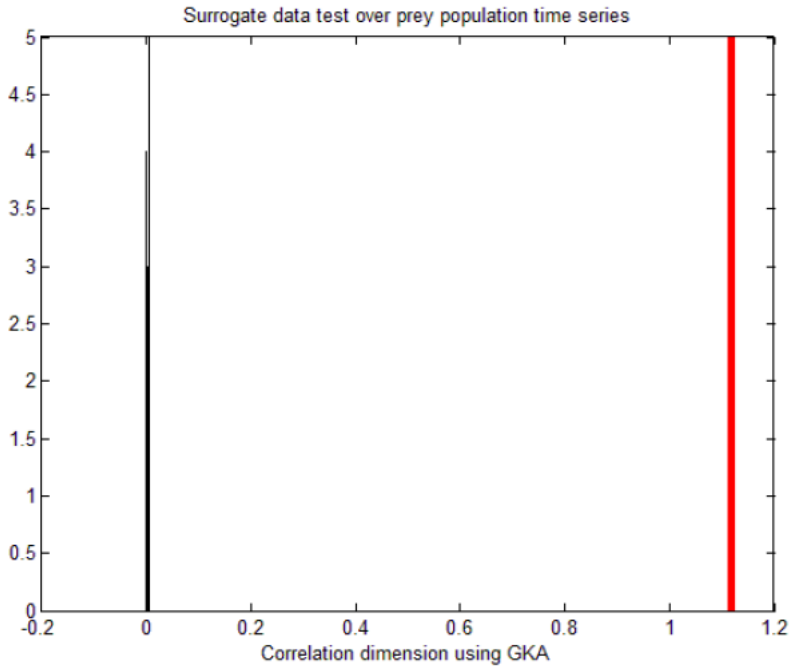
Figure 5-2. The results of hypothesis testing by using 24 surrogate data sets over simulation's population time series, (a) prey, (b) predator using Higuchi fractal dimension.

5.1.1.2. Correlation Dimension using GKA method

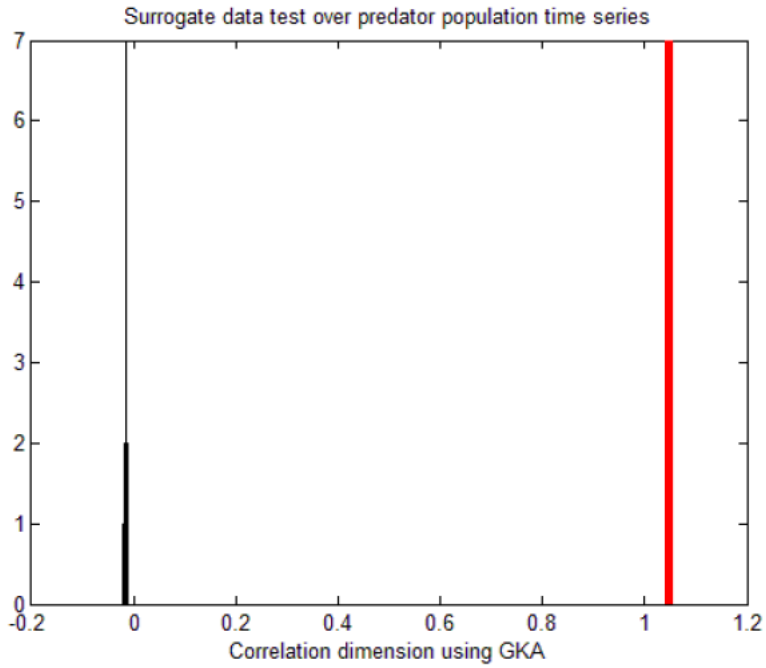
To determine if the correlation dimension (using GKA) of a typical time series presents a deterministic or stochastic behaviour, the same hypothesis testing procedure is adopted. As in previous section, the hypothesis under H_0 models stochastic and the hypothesis H_1 implies on deterministic.

The correlation dimension statistic is also extracted from both population time series and surrogate data. It should be decided if the difference between the values of original data (population time series) and surrogate data are statistically meaningful or not. To this purpose, the same process than in previous section is applied to the random time series, Lorenz time series and simulation's population data. The same results as in the previous section are observed for the random and Lorentz time series (data not shown).

In Figure 5-3, the red signals indicate the value of the statistic extracted from the original data for both prey and predators time series for GKA embedding dimension m=5. These values are completely separable from the surrogate distribution. It means that the red signals are in the confidence interval, so the null hypothesis is rejected (stochastic behaviour rejected). This criterion therefore, also confirms that the simulation's population data have a deterministic behaviour.



(a) Prey



(b) Predator

Figure 5-3. The results of hypothesis testing by using 24 surrogate data sets over simulation's population time series, (a) prey, (b) predator series using correlation dimension.

5.1.1.3. Lyapunov exponent

To determine if the largest Lyapunov exponent of a typical time series presents a non-chaotic or chaotic behaviour (see section 3.3), the same hypothesis testing procedure is adopted. The hypothesis under H_0 models non-chaotic and the hypothesis H_1 implies on chaotic.

The Lyapunov exponent statistic is also extracted from both population time series and its surrogate data. It should be decided if the difference between the values of original data (population time series) and surrogate data are statistically meaningful or not. To this purpose, the same process than in previous section is applied to the random time series, Lorenz time series and simulation's population data. Figure 5-4 shows the results obtained, as reference, for the random and Lorenz time series.

As it can be seen in Figure 5-4(left), the hypothesis testing result is falling into imposter distribution (see section 3.3.3) for random time series. Therefore, hypothesis H0 is accepted and it means the time series is non-chaotic. On the other hand, for Lorenz time series Figure 5-4(right), the red sign indicates the value of the statistic extracted from original data, which is completely separable from its imposter distribution. Therefore, hypothesis H0 is rejected and it means the time series is chaotic.

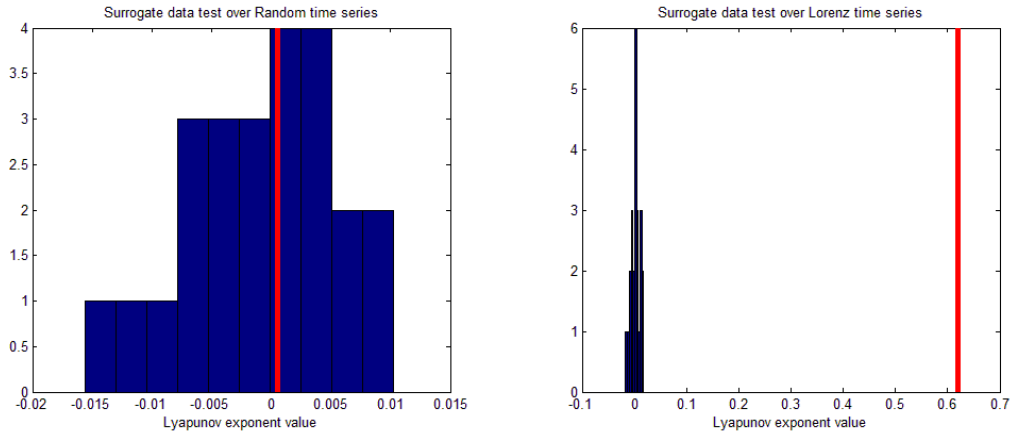
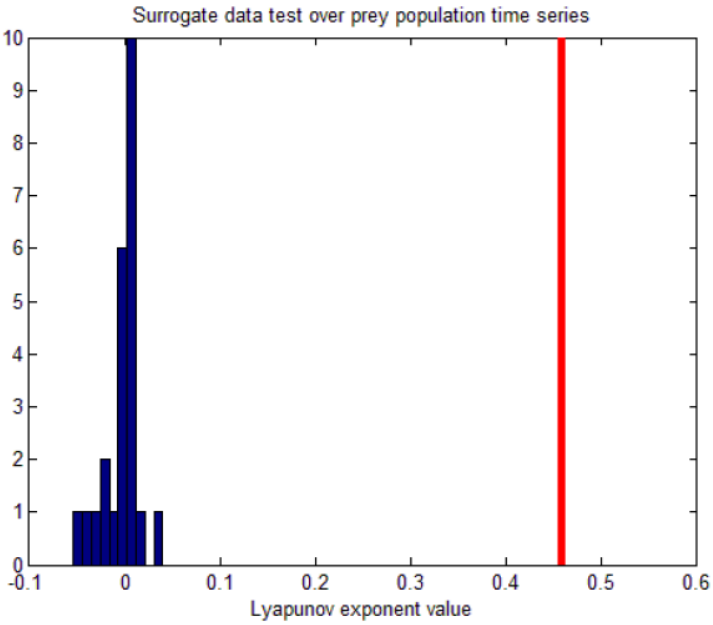


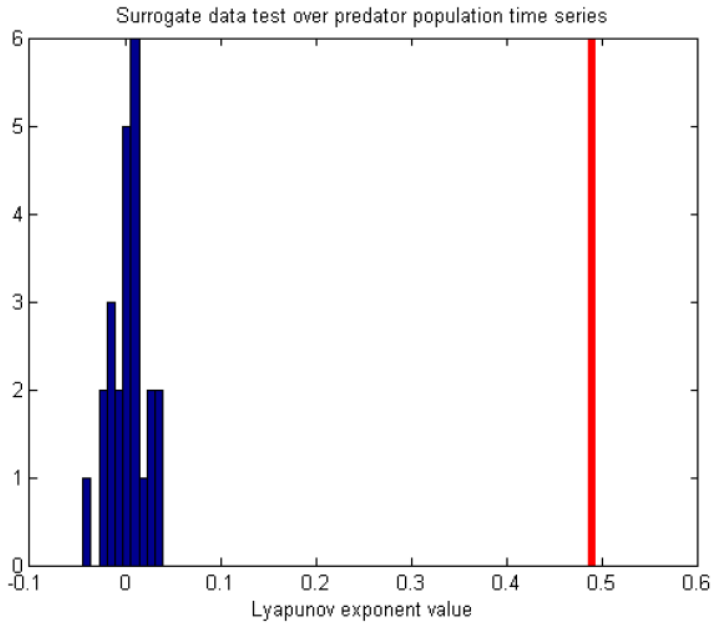
Figure 5-4. The results of hypothesis testing by using 24 surrogate data sets for random time series (left) and Lorenz time series (right) using largest Lyapunov exponent.

In Figure 5-5, the red signals indicate the value of the statistic extracted from the original data for both prey and predators time series. These values are completely separable from the surrogate distribution. It means that the red signals are in the confidence interval, so the null hypothesis is

rejected (non-chaotic behaviour rejected) this criterion therefore, shows clearly that the simulation's population data have a chaotic behaviour.



(a) Prey



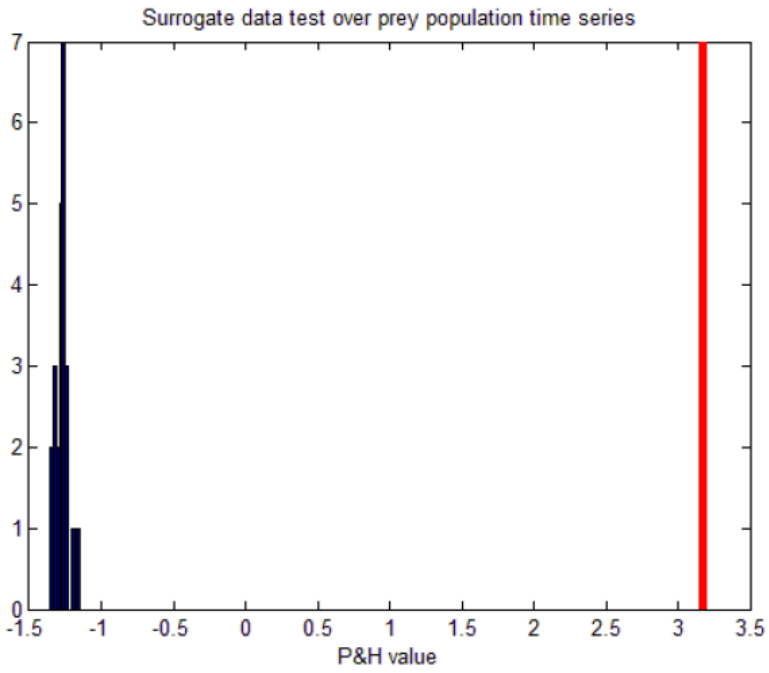
(b) Predator

Figure 5-5. The results of hypothesis testing by using 24 surrogate data sets over simulation's population time series, (a) prey, (b) predator series using largest Lyapunov exponent.

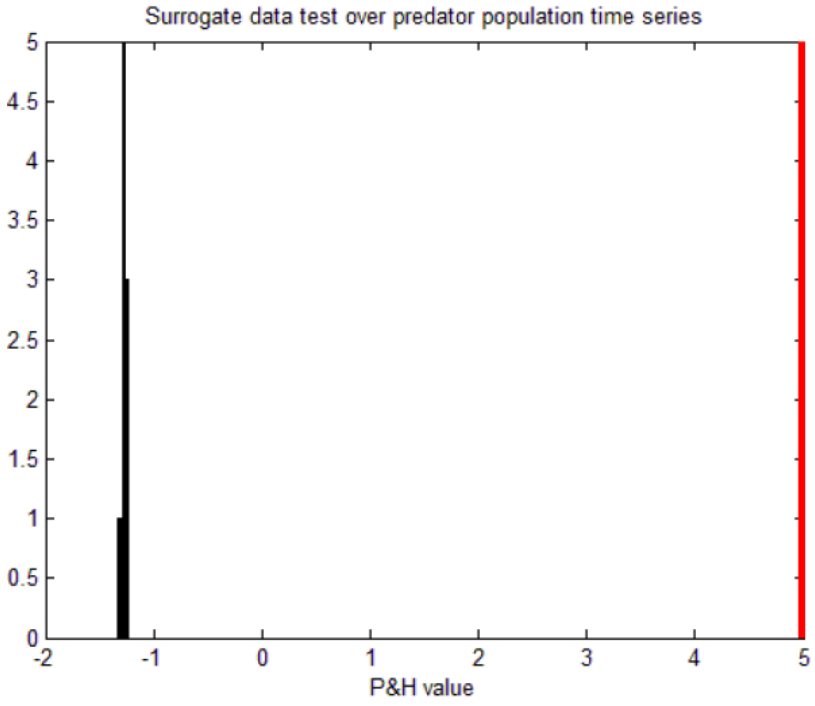
5.1.1.4. P&H method

To determine if the P&H method value of a typical time series presents a non-chaotic or chaotic behaviour, the hypothesis testing procedure is adopted. The hypothesis under H_0 models non-chaotic and the hypothesis H_1 implies on chaotic. The same results as in previous section are observed for the random and Lorenz time series.

In Figure 5-6, the red signals indicate the value of the statistic extracted from original data, both for prey and predators time series. This value is completely separable from the surrogate distribution. It means the red signals are in the confidence interval, so the null hypothesis is rejected (non-chaotic behaviour rejected). This last criterion confirms that the simulation's population data has a chaotic behaviour.



(a) Prey



(b) Predator

Figure 5-6. The results of hypothesis testing by using 24 surrogate data sets over simulation's population time series, (a) prey, (b) predator using P&H method.

5.1.2. Conclusion

The purpose of this study is the examination of the stochastic and deterministic behaviour of signals that are produced by EcoSim. To understand how close our simulation is to the random or chaotic processes, we examined whether a chaotic behaviour exists in its signals. To enforce our result, we use four different methods: Higuchi fractal dimension, correlation dimension, largest Lyapunov exponent, P&H method. For each of them, in order to obtain a statistically significant evaluation, we applied the surrogate test method on 24 samplings of the considered data.

According to the results obtained after applying these different methods, all of them providing clear predictions, we can conclude that behaviour of population time series in EcoSim is deterministic. This has been shown by Higuchi method and correlation dimension. Also among various cases of deterministic behaviour, we showed that the behaviour of population time series produced by EcoSim is chaotic. This has been shown by Lyapunov exponent and P&H methods.

5.2. Identifying Multifractal Phenomena in EcoSim

In this section, we investigated whether the spatial distribution of individuals generated by EcoSim present the same kind of multifractal properties as those observed in real ecosystems. We also analyzed different parameters of the simulation to detect which ones cause the multifractal behaviour given that one important task for ecologists is to understand where these structures originate. A wavelet-based method has been used for this analysis. Multifractal analysis of EcoSim's results demonstrates self-similarity characteristics in the spatial distribution of individuals as it has been observed in real ecosystems [34]. One important issue for ecologists is to understand where these structures come from. We analyzed different parameters of the simulation to detect, which ones cause the multifractal behaviour.

Recently, researchers have begun to recognize ecosystem data as a highly nonlinear system [271]. Analysis of time series with high complexity, such as time series resulting from the interaction between individuals' behaviours in ecosystems, requires a nonlinear dynamic approach [272][273]. Dynamic studies of nonlinear systems describe the specification of biological processes [150]. In most natural phenomena chaotic and self-similarity properties co-exist [274], [275]. Since the seminal work of Mandelbrot [276], many patterns and processes have proven to be efficiently described by fractals in many fields of the natural sciences. Fractal geometry and the resulting scaling properties have also been suggested as a way to characterize space-time heterogeneity in ecology [277].

Fractals identify the presence of patterns at multiple scales. Part of the fractals' appeal is that a single statistic can be used to describe potentially complex patterns in natural environments. The use of fractal geometry can be viewed as a tool to be used by landscape ecologists to aid in answering questions relating to scale [278]. Studies have shown that natural phenomena present self-similar property over time [279] (see section 3.2).

A multifractal system is a generalization of a fractal system in which a single exponent (the fractal dimension) is not enough to describe its dynamics; instead, a continuous spectrum of exponents is needed. Self-similarity is a typical property of fractals (see section 3.2.4). Scale invariance is an exact form of self-similarity where at any magnification, there is a smaller piece of the object that is similar to the whole [272]. Applications of multifractals to ecology still remain anecdotic, limited to forest ecology [280], [281], population dynamics [282], the characterization of species-area relationship, species diversity, and species abundance distribution [9][283], [20], and the characterization of nutrient, phyto- and zooplankton patchiness [284], [285]. Multifractal analysis techniques allow exploring features of signal distribution that are not considered very often [279]. In this study, a wavelet-based method has been used for multifractality analysis. The wavelet transform takes advantage of multifractal self-similarities, in order to compute the distribution of their singularities. This singularity spectrum is used to analyze multifractal properties [286].

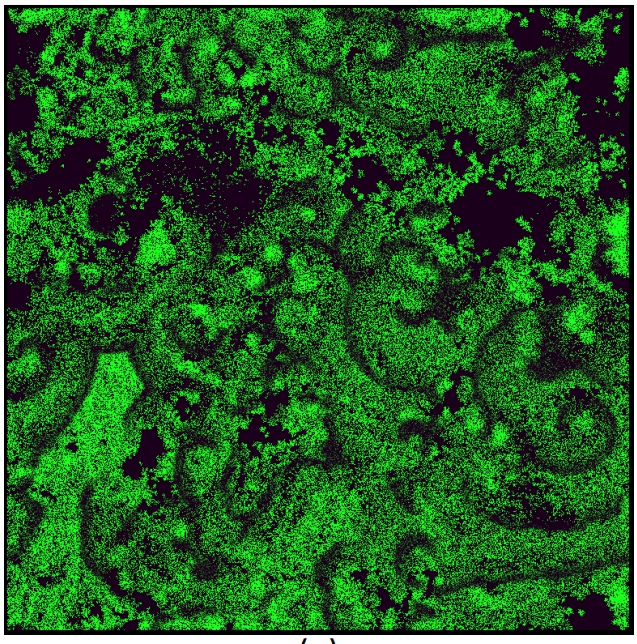
One of the issues ecologists have to deal with is not only to observe multifractal spectrum for the spatial distribution, but also to explain, from a phenomenological point of view where these structures come from. Because many environmental parameters display self-similarity, the observed biotic patterns could reflect the distribution of some abiotic factors presenting a template upon which individual operate [279], [26] (see section 3.2.4). For this reason, we analyzed different parameters of EcoSim such as the pattern of food, the predators' pressure and the raggedness of the environment to detect the factors, which can explain multifractal behaviour in spatial distribution of individuals.

Because this simulation is a logical description of how a simple ecosystem performs, this analysis can help biologists to better understanding of long-term behaviour of ecosystem. We analyzed the spatial distribution of individuals in various simulation experiments: one that used no specific pattern of food in world (EcoSim), experiment that has no predator (EcoSimNoPredator), experiments that used a specific pattern of food (EcoSimCircle, EcoSimStar) and experiments with obstacle cells in the world (EcoSimObstacle1%, EcoSimObstacle10%).

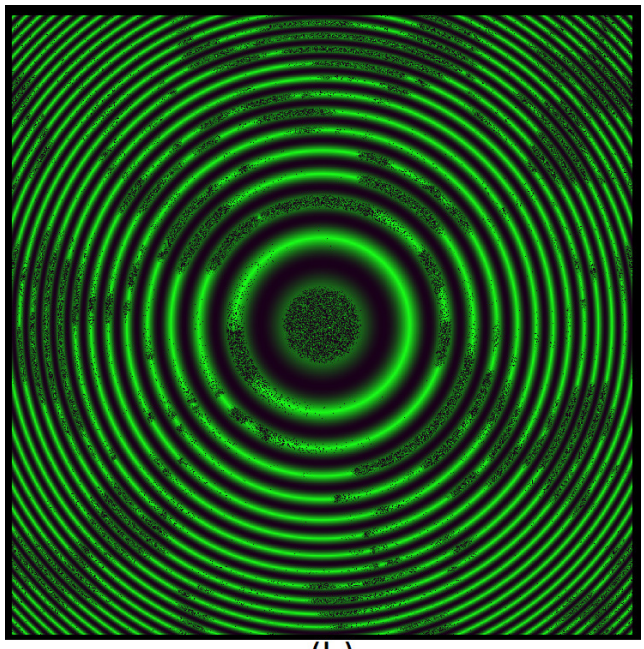
5.2.1. Experiment Design

5.2.1.1. Different food pattern

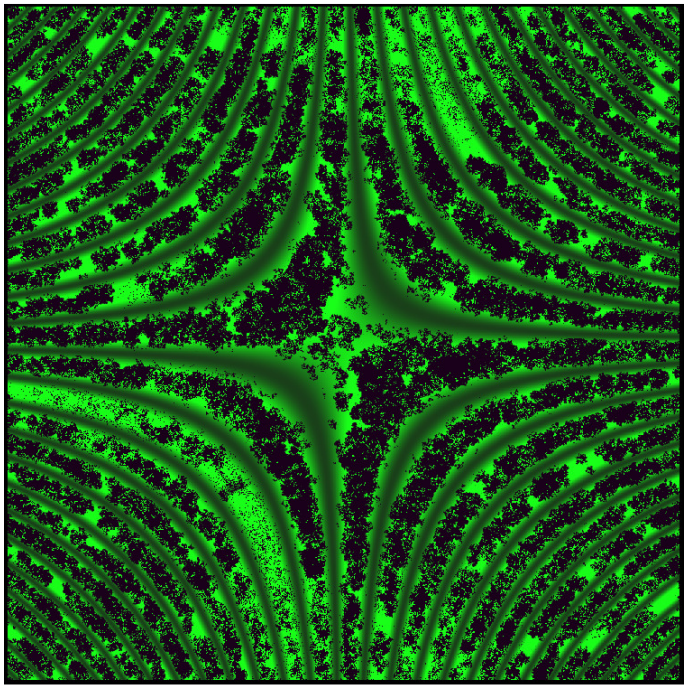
In EcoSim Each cell can contain grass. There is a limit in the amount of grass available in each cell. This allows a competition for resource between individuals to occur. At the initialization time, the number of grass units is uniformly randomly determined for each cell. The number of grass units grows at each time step. The number of grass units in a cell decreases by one when a prey eats. If the prey eats all the grass in one cell the grass cannot grow anymore unless there is still grass in an adjacent cell. This later concept models the mechanism of diffusion of resources through the world changing and renewing the interest of regions of the world (Figure 5-7a). We defined two other versions of the simulation based on specific pattern of food distribution. In the first version the food is distributed in concentric circles, we call it EcoSimCircle (see Figure 5-7b). The second one, having the star distribution is called EcoSimStar (Figure 5-7c).



(a)



(b)



(c)

Figure 5-7. Distribution of food (grass) after 10000 time steps in (a) EcoSim (b) EcoSimCircle (c) EcoSimStar.

5.2.1.2. The Raggedness of Environment

We use also another version of EcoSim simulation to measure effect of the environment's raggedness on population fragmentation and speciation processes [30]. As discussed before, small

physical obstacles are included that obstruct the movement (dispersal) of individuals. Each obstacle covers completely one cell and they also impede the vision of the individuals. The presence of obstacle cells in the world is also expected to disrupt the movement of the agents, change their spatial distribution, and in turn influence dispersal and ultimately the gene flow between populations. Two virtual worlds with various numbers of obstacles are considered: 1% and 10%. For example, in experiment "EcoSimObstacle 10%", ten percent of cells in world are obstacles. Each execution of the simulation for this analysis produced approximately 16,000 time steps in 23 days. The computed average and standard deviation for the number of prey individuals are 190,000 and 25,000 respectively (for predator 30,000 and 8,000) and the average and standard deviation for the number of prey species are 49 and 10 (for predator 58 and 9).

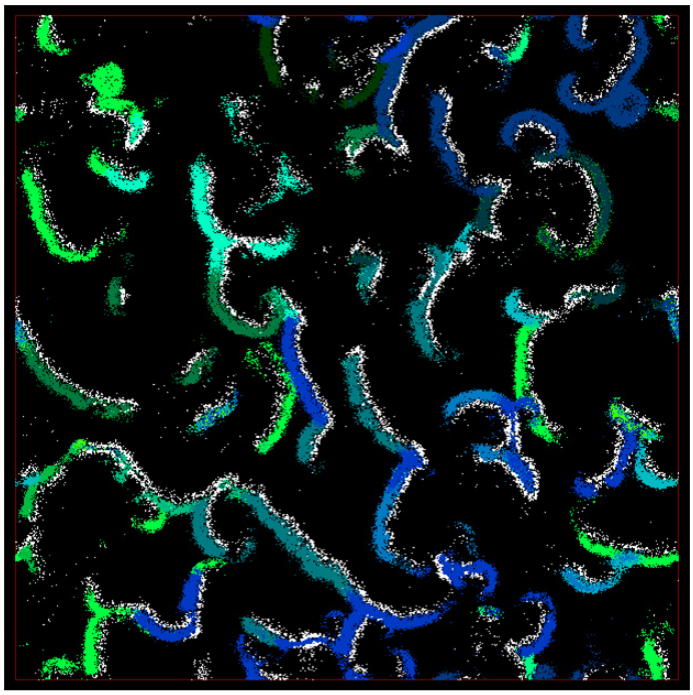
5.2.2. Multifractal Analysis using Wavelets-based method

It is essential to measure the correlation between the positions of the main biotic factors to gain new insights into the origin of distributions in biological systems. For that reason the effects of two environmental parameters and the effect of predators' pressure on prey's spatial distribution have been examined. The snapshots considered in the analysis correspond to a typical spatial distribution of the individuals, and the same results have been obtained at different time steps. For each experiment, we conducted five independent runs using the same parameters and averaged the results.

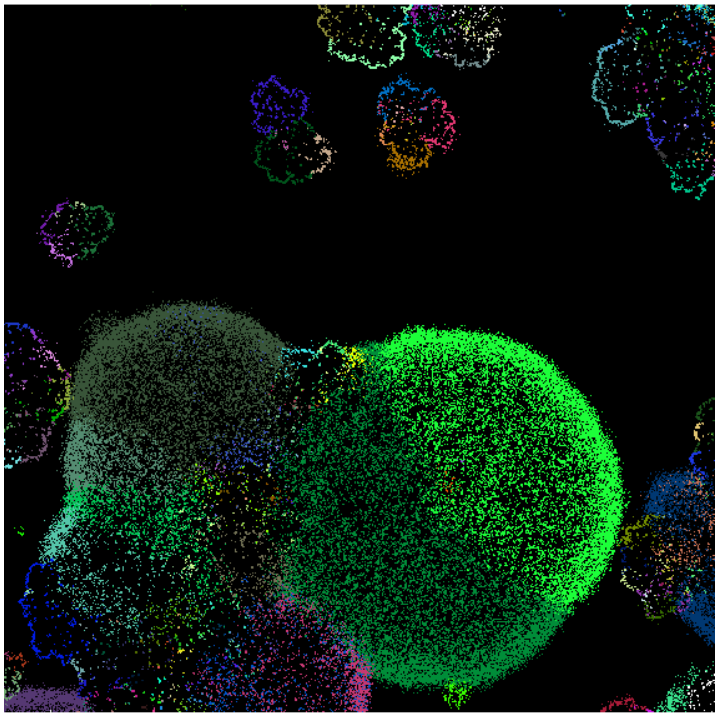
5.2.2.1. Predator pressure

This section is an analysis of the spatial distribution of prey individuals generated by two simulations: EcoSim and EcoSimNoPredator (EcoSim with no predator in the world) in order to investigate the effect of predators' pressure. Multifractal spectra have been calculated for the spatial distribution of prey individuals in both experiments. In both experiments there is an initial uniform random distribution of food. The evolution of the individuals and their interactions then shape the spatial distribution of individuals.

The spatial distribution of individuals for EcoSim and EcoSimNoPredator simulation are shown in Figure 5-8 (different color for different species). Contrary to the emerging herd patterns observed in the EcoSim simulation (Figure 5-8a), the spatial distribution of individuals in the other simulation forms simpler patterns, which prey expand in all direction in absence of predators (Figure 5-8b).



(a)



(b)

Figure 5-8. Spatial distribution of individuals in (a) EcoSim (b) EcoSimNoPredator

Figure 5-9 shows the CWT representation of the prey individuals' spatial distribution in EcoSim. From an intuitive point of view, the wavelet transform shows a "resemblance index" between the signal and the wavelet. If a signal is similar to itself at different scales, then the "resemblance

index" or wavelet coefficients also will be similar at different scales. In the coefficients plot (Figure 5-9), which shows scale on the vertical axes, this self-similarity generates a characteristic pattern. This is a good demonstration of how well the wavelet transform can reveal the fractal pattern of the behavioural activity at different times and scales.

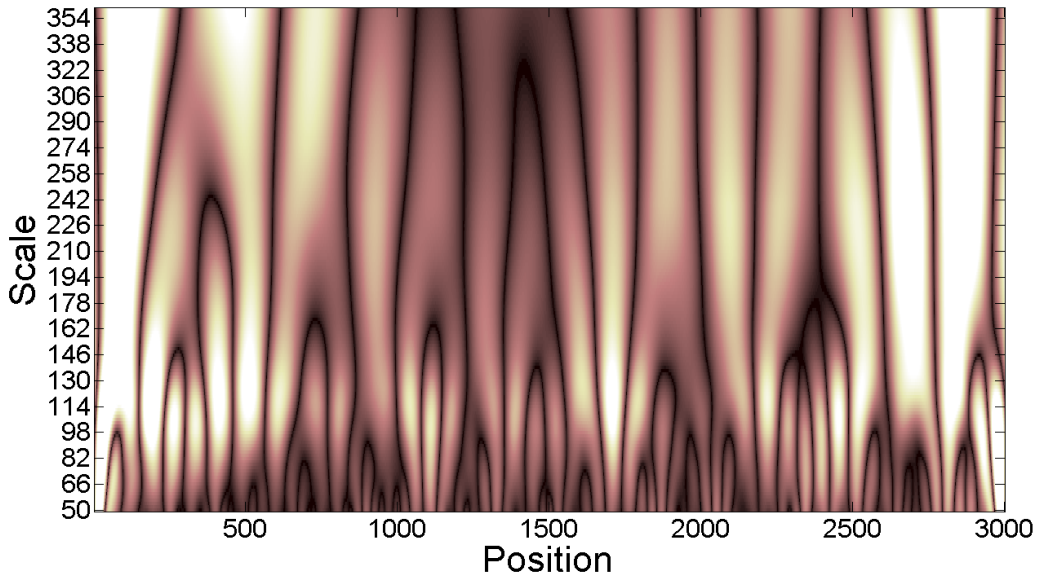


Figure 5-9. CWT coefficients plot of the spatial distribution of prey individuals in EcoSim. Scale and position are on the vertical and horizontal axis, respectively.

Figure 5-10a displays the "tau spectrum, τ(q)", obtained by using the WTMM method, applied to the spatial distribution of prey individuals in the EcoSim experiment (see section 3.2.4). The spectrum is curved, which indicates the multifractal nature of the spatial distribution. We computed the spectrum D(h), represented in Figure 5-10b, which clearly confirms the non-uniqueness of the Hölder exponent h, and thus the multifractality of the process (see section 3.2.4).

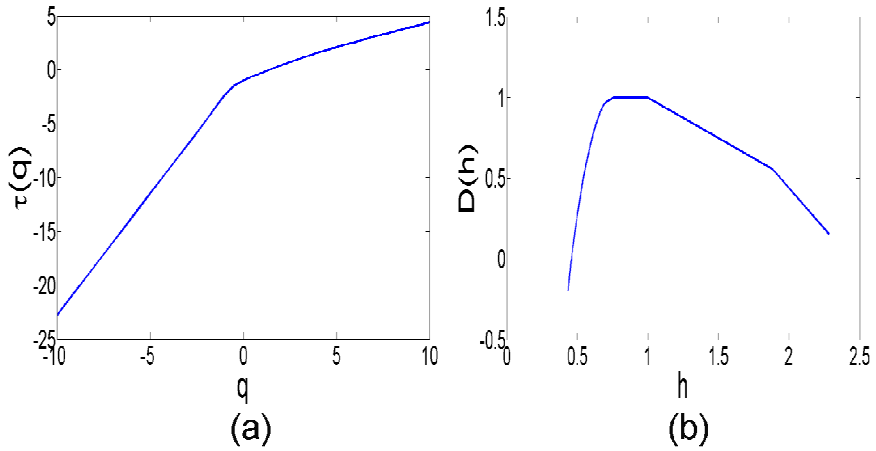


Figure 5-10. (a) "Tau spectrum" of the spatial distribution of prey individuals in EcoSim (b) Multifractal spectrum of the spatial distribution of prey individuals in EcoSim. Because of different values in the spectrum, one can assume a multifractal process. Every curve represents an average value obtained from five independent runs.

These results shown that the interaction between individuals over the time and the uniform distribution of food in the world make a complex spatial distribution of prey individuals with multifractal characteristics. As the food is initially uniformly distributed, it cannot be the leading factor that generates the fractal property. Since this is a prey-predator model, the behaviours of prey and predator have to evolve simultaneously to give them the abilities needed to survive, so the affect of predator is important in this matter. Therefore the multifractal analysis was also applied to the spatial distribution of predators. The results show that the spatial distribution of predators has the same multifractal characteristics as the spatial distribution of prey. These results confirm previous results real data, such as the population dynamics of soil microorganisms [34], the swimming behaviour of the calanoid copepod Temora longicornis, the displacements of male Daphniopsis australis and the microphytobenthos biomass distribution [279], that have multifractal properties.

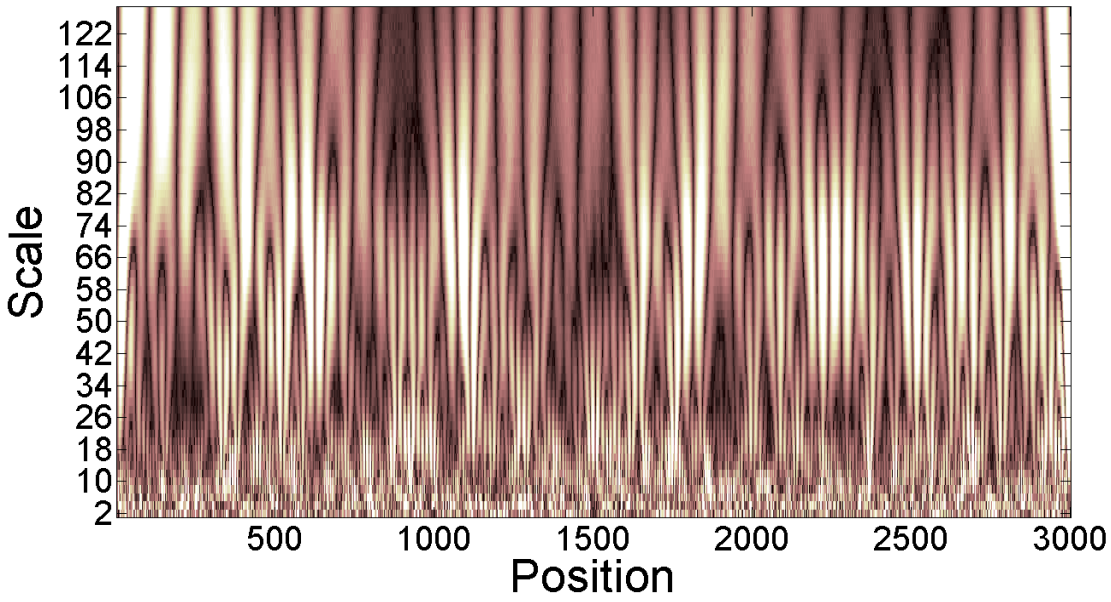


Figure 5-11. CWT coefficients plot of the spatial distribution of prey individuals in EcoSimNoPredator. Scale and position are on the vertical and horizontal axis, respectively.

The wavelet analysis has been also applied to the spatial distribution of prey individuals in EcoSimNoPredator simulation. The EcoSimNoPredator simulation's parameters are identical, with the same initial parameters and scales and population dynamic in the EcoSim. The only difference is absence of predators in the world. In the coefficients plot (Figure 5-11), there is no pattern like the patterns in Figure 5-9. Therefore, at least from a visual point of view, it seems that there is no self-similar pattern.

Figure 5-12a displays the "tau spectrum, $\tau(q)$", obtained by using the WTMM method, for the prey individuals' spatial distribution. The spectrum is not curved, confirming that there is no multifractal property in these patterns. We obtain the spectrum D(h), represented in Figure 5-12b, which clearly does not confirm the non-uniqueness of the Hölder exponent h. The figure shows just a straight line, which stands for one value thus the multifractality of the spatial distribution can be rejected.

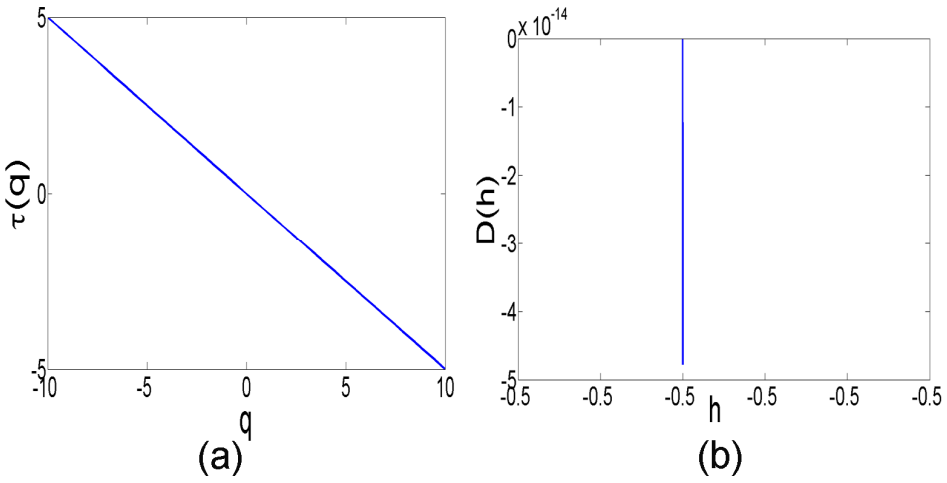


Figure 5-12. "Tau spectrum" of the spatial distribution of prey individuals in EcoSimNoPredator (b) Multifractal spectrum of the spatial distribution of prey individuals in EcoSimNoPredator. By analyzing the spectrum one can assume a multifractal process. Every curve represents an average value obtained from five independent runs.

This outcome showed that the predators' pressure can lead to a multifractal behaviour when there is no limit on mobility of individuals. With equal ease of movement in all directions, predators will be able to push prey in different scales. Individuals distribution forming spiral waves is one property of prey-predator models (like in Figure 5-8a). As we discussed before, the prey near the wave break have the capacity to escape from the predators sideways. A subpopulation of prey then finds itself in a region relatively free from predators. In this predator-free zone, prey start expanding intensively and form a circular expanding region. The whole pressure process and spiral formation will be applied to this subpopulation of prey and predators again leading to the formation of a second scale. This process repeats over and over and this is a common property of self-similar processes [35]. Because there are consecutive interactions between prey and predators during time, the same pattern repeats over and over and then self-similarity emerges in spatial distribution of individuals.

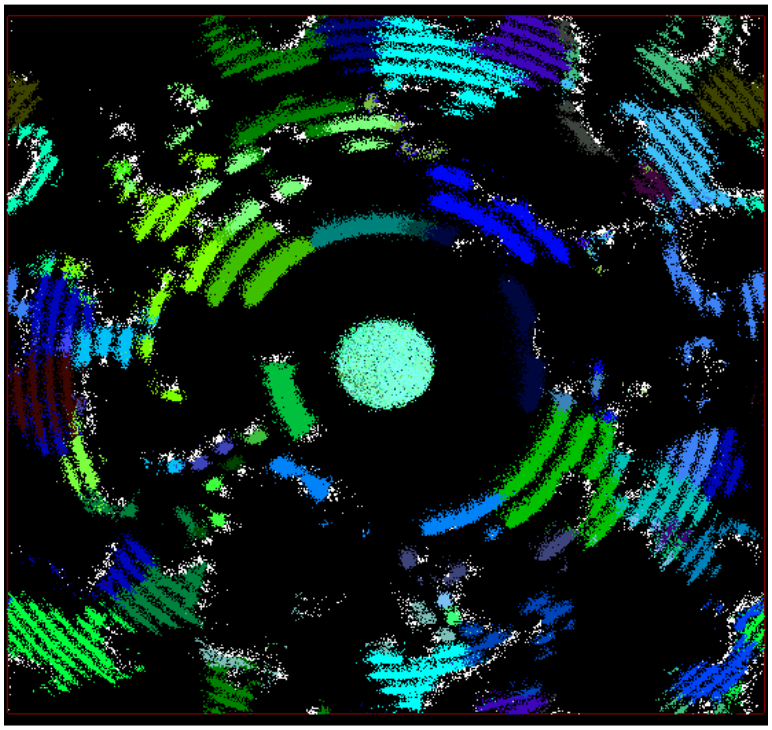
Indeed, prey distribution and food distribution are very important for predators because food availability changes depending on the fractal dimension. Non-multifractal behaviour indicates distribution of particles gathered in small numbers of patches, while multifractal behaviour indicates rough, fragmented, and space-filling distributions. When a predator has no remote detection ability (which is our case because predators don't have long range vision), prey distributions with multifractal behaviour could be efficient for predators, because available food quantity become proportional to the searched volume as multifractal behaviour increases [279].

5.2.2.2. Various Food Pattern

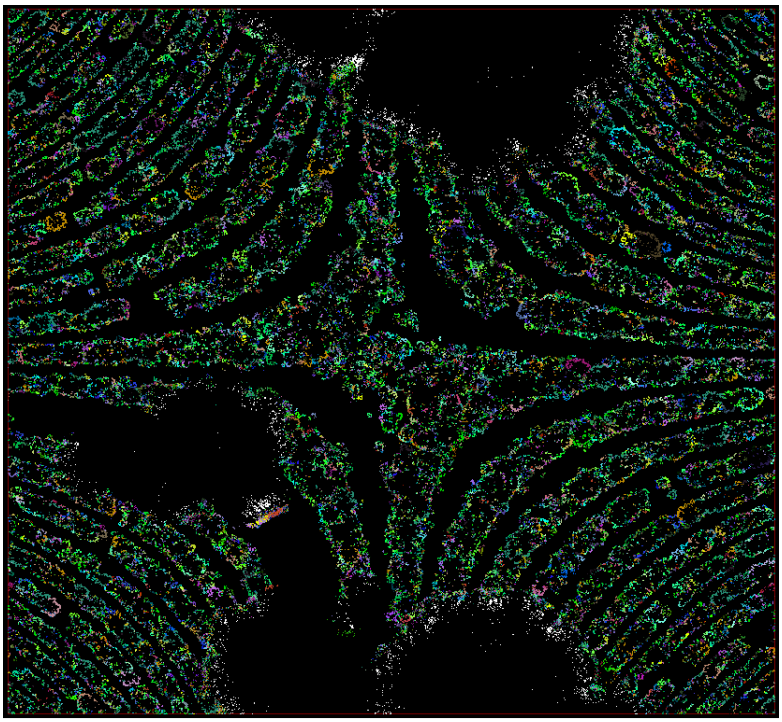
This section is an analysis of the simulation's spatial distribution of prey individuals generated by two simulations: EcoSimCircle (EcoSim with circle pattern of food) and EcoSimStar (EcoSim with star pattern of food) in order to investigate the effect of food pattern. Multifractal spectra have been calculated for the spatial distribution of prey individuals in all experiments. In EcoSimCircle and EcoSimStar, the spatial distribution of food is kept fixed during the whole simulation.

The spatial distribution of individuals for EcoSimCircle and EcoSimStar simulation are shown in Fig. 9. Contrary to the emerging herd patterns observed in the EcoSim simulation (Figure 5-8a), the spatial distribution of individuals in the these two simulations followed the circle and star food distribution respectively (Figure 5-13a,b). For space consideration, we do not present the graphs of the multifractal analysis as they are almost identical to the ones already presented.

The wavelet analysis has been applied to the spatial distribution of prey individuals in EcoSimCircle simulation. The EcoSimCircle simulation's parameters are kept the same, with the same initial parameters and scales and population dynamic in the EcoSim. The only difference is the fixed distribution of food in the world.



(a)



(b)

Figure 5-13. Spatial distribution of individuals in (a) EcoSimCircle (b) EcoSimStar

In the coefficients plot, there is no self-similar patterns like the patterns in Figure 5-9. Multifractal spectrum have been calculated for the spatial distribution of prey individuals. The "tau spectrum, τ(q)" and the spectrum D(h) (like Figure 5-12), clearly demonstrate that there is no multifractal behaviour in the spatial distribution of prey (results not shown). The multifractal analysis was also applied to the spatial distribution of predators in this experiment. The result shows there is no multifractal characteristic in spatial distribution of predators. The same wavelet-based analysis has been applied to EcoSimStar and the same results have been obtained: no multifractal pattern for grass and for spatial distribution of prey and predators. When the distribution of food in the world becomes fixed, the multifractal phenomenon vanished. Therefore, as long as there is specific fixed pattern of food in the world it seems that the complex multifractal phenomenon doesn't show up for spatial distribution of individuals. The dynamic distribution of food is needed for complex patterns to emerge as it strongly affects the spatial distribution of the prey that need this food to survive.

5.2.2.3. Various Levels of Environment's Raggedness

We are also interested in studying whether various levels of raggedness in the world, as it also has impact on the movement of the individuals, can affect the fractal properties observed. We use two

new simulation experiments with various numbers of obstacles: 1 and 10 per cent, EcoSimObstacle(1%) and EcoSimObstacle(10%). The raggedness of the world increases when the number of obstacle cells raises. Multifractal spectrum have been calculated for the spatial distribution of individuals in all these experiments and compared them with results of EcoSim.

In EcoSim, there is no obstacle cells in the world and the results of this simulation has been shown in previous section, which shows existence of multifractal behaviour in individuals' spatial distribution. We measured the CWT representation of the individuals' spatial distribution for EcoSimObstacle(1%) and EcoSimObstacle(10%). The coefficients plot (like Figure 5-9), the "tau spectrum, $\tau(q)$" and the spectrum D(h) (like Figure 5-10), clearly demonstrate the multifractality of the process.

In these two experiments, with different level of raggedness, a multifractal behaviour also emerged. We can conclude that this parameter doesn't play a major role in multifactal behaviour of spatial distribution. Regardless of the level of raggedness, individuals finally find their way to adopt and form the nested spiral pattern.

5.2.3. Conclusion

The purpose of this study is to analyze the multifractal behaviours of individuals' spatial distribution that are produced by the ecosystem simulation (EcoSim). Understanding of the origin of individuals patchiness is an important issue. It is stressed here that the knowledge of the multifractal distributions of relevant parameters such as food concentration, spatial distribution of prey and predators and density of obstacles could be the first step to infer their phenomenological links. We applied our analysis to different kinds of simulations: the ecosystem simulation with fixed specific pattern of food in the world, the world without predators' pressure and the world with several amounts of obstacles and then we compared the results with the ones obtained with the simulation without constraints.

We used a wavelet-based method for this analysis. It showed that the behaviour of the individuals without any constraints, or restricted by a limited amount of obstacles with predators' pressure can lead to the multifractal phenomena as the ones observed in real ecosystems. It is also another important confirmation of the capacity of EcoSim to model complex and realistic large scale systems. On the contrary, we have shown that when the food distribution is fixed, which strongly reduces the possibility of movement of the prey, the multifractal pattern disappears. It seems that it is the complex interaction between the predation pressure, the eating behaviour of the prey and the diffusion of food that conducts to the apparition of the multifractal phenomenon.

Chapter 6

6. Long-term prediction of complex time series

Long-term prediction of complex nonlinear time series and their application to time-ordered data is a major concern for researchers in a variety of scientific fields including physics, medicine, computer science, engineering and economics [170], [172], [174], [176]. However, the chaotic behavior of nonlinear time series makes their prediction extremely challenging. Few research projects have focused on the long-term prediction of nonlinear time series, and no satisfactory results have been achieved in this domain [178], [181], [182]. Our goal in this chapter is to investigate if such predictions are realistic and for which specific applications.

In this study, we proposed a new method, GenericPred, for time series prediction with applications to financial time series, medical diagnosis and global temperature prediction. To test our new method, we evaluated its performance with respect to the prediction of the long-term behavior of the Dow-Jones Industrial Index (DJIA) time series, EEG time series for epileptic seizure prediction and global temperature anomaly. Our new method does not rely on a complex and specialized model of time series, and so it is therefore highly general. Therefore, it has many additional possible applications such as earthquake prediction, heart attack prediction, and species extinction prediction.

6.1. Methods

Several researchers emphasise the potential of market predictions to improve important financial decisions [288], from helping businesses make sounder investment decisions to helping governments make more efficient fiscal and monetary policy decisions [169]. These time series are amongst the most complex time series because of the number of parameters involved. Our results are compared with respect to long-term predictions with ARIMA, GARCH, and VAR [289], which are the most widely used and most efficient methods for making long-term time series predictions. We also compared our results for short-term predictions with those obtained by two existing methods: the Learning Financial Agent Based Simulator (L-FABS) [178] and the MLP model [179].

For the first period, we considered the DJIA (Dow Jones Industrial Average) time series between 1993 and 2001, when markets were stable with no major changes and no financial crisis. In the second period considered, the US stock market peaked in October 2007, but by March 2009, the Dow Jones average had reached its minimum, which reflects the most serious effects of a

financial crisis. In the third period (August 2004-August 2012), the recession was in the middle of the considered range.

Another important application of time series predictions is in medical science. Approximately 1% of the world population suffers from epilepsy [290]. Epileptic seizures are the result of unusual and irregular neuronal activity in the brain [291]. Many recent methods have been proposed for predicting epileptic seizure [292]–[294] but none of them as shown their ability to perform accurate predictions more than 10 minutes in advance on a large number of patients. To evaluate the performance of our new method for predicting epileptic seizures, we examined the Electroencephalography (EEG) time series of patients with epilepsy. EEG datasets of 21 patients were chosen from the Epilepsy Center of the University Hospital of Freiburg [295]. In eleven patients, the epileptic focus was located in neocortical brain structures, in eight patients in the hippocampus, and in two patients in both. The EEG data were acquired using a Neurofile NT digital video EEG system with 128 channels, 256 Hz sampling rate, and a 16 bit analogue-to-digital converter. For each of the patients, there were datasets called "ictal" and "interictal", the former contains files with epileptic seizures and at least 50 min pre-ictal data. The latter contains approximately 24 hours of EEG-records without seizure activity [296]. The EEG signal is represented as a time series vector, X={x1, x2,..., xN} comprised of single voltage readings at various time intervals and expressed as a series of individual data points (single voltage readings by an electrode) where N is the total number of data points and the subscript indicates the time instant [297].

Predicting the monthly records of global temperature anomalies is currently one of the most pressing and controversial environmental concerns [298]. As a third experiment, we used the global temperature anomaly data from 1880 to 1983 to train for the prediction of global temperatures during 1983-2013. Global temperature anomaly data come from the Global Historical Climatology Network-Monthly (GHCN-M) data set and International Comprehensive Ocean-Atmosphere Data Set (ICOADS), which have data from 1880 to the present. These two datasets are blended into a single product to produce the combined global land and ocean temperature anomalies.

Our new method for complex time series prediction is based on the concepts of chaos theory and an optimisation process. The general idea is to extract a unique characteristic from an existing time series that somehow represents the behaviour of the time series and to subsequently generate successive new values that continue the time series, each value minimising the difference

between the characteristic of the new time series and the initial one. The details of the GenericPred method for long-term time series prediction are as follows. We consider a time series S_N:

$$S_N = \{x_1, x_2, ..., x_N\} \quad (6\text{-}1)$$

A nonlinear measure *V()* is computed on S_N. The fractal dimension [120] or the Lyapunov exponent [299] are examples of such nonlinear measures that return a single value for a time series. A possible mapping may be required, forming a new time series $S_N^m = \{y_L, y_{L+1}, ... y_N\}$, for different applications as follows:

$$y_i = V(S_{i-L+1,\,i}) \quad , \quad \mathrm{L} \le \mathrm{i} \le \mathrm{N} \quad \text{where} \quad S_{i-L+1,\,i} = \{y_{i-L+1}, y_{i-L+2}, ..., y_i\} \quad (6\text{-}2)$$

otherwise, $S_N^m = S_N$,

where 0 < L < N is the size of a sliding window used to compute the local level of chaos measured by V(). Therefore, when the mapping is applied, the new considered time series S_N^m corresponds to the variation in time of the local non-linear measure in the initial time series S_N.

We consider $V(S_N^m)$ as a reference value that will be used for predicting the next *k* values of the time series:

$$y_{N+i} \quad , \quad 1 \le \mathrm{i} \le k \quad (6\text{-}3)$$

The parameter σ of a normal distribution $N(y_i, \sigma^2)$ is estimated by computing the variation between every two consecutive values (y_i to y_{i+1}) of the time series S_N^m (Uniform distribution also has been used and the results were the same). This distribution represents the probability distribution $P(y_i \mid y_{i-1})$ (see Figure 6-1). Several data sets have been considered to determine that a normal distribution is a good approximation of the real distribution. However, the same method has been applied using other distributions without significant degradation is the prediction.

For predicting y_{N+i}, a set $Pos(y_{N+i})$ of N_{rand} random values are generated following the distribution $N(y_{N+i-1}, \sigma^2)$ (Figure 6-1):

$$Pos(y_{N+i}) = \{ y_{N+i}^j \quad , \quad 1 \le j \le N_{rand} \} \tag{6-4}$$

N_{rand} is a parameter that can impact the quality of the prediction because having more values will increase the chance of finding an optimal value. However, no significant improvement was observed for the data considered when N_{rand} was greater than 10. For this reason, we chose 10 as the value of N_{rand} for each experiment. y_{N+i} is then computed by selecting the y_{N+i}^j that makes the new nonlinear measure the closest to $V(S_N^m)$:

$$j_{\min} = \arg\min_j \left(\left| V(S_{N+i-1}^m + y_{N+i}^j) - V(S_N^m) \right| \right)$$
$$y_{N+i} = y_{N+i}^{j_{\min}} \quad , \quad \text{where} \quad S_{N+i-1}^m + y_{N+i}^j = \{ y_1, y_2, ..., y_{N+i-1}, y_{N+i}^j \} \tag{6-5}$$

The value y_{N+i}^j is chosen to make $V(S_{N+i-1}^m + y_{N+i}^j)$ as close as possible to $V(S_N^m)$.

The important point is that the reference value is always $V(S_N^m)$, which is the calculated nonlinear measure from the original time series. Therefore, the GenericPred method uses two basic rules:

R1: Always endeavour to keep the value of a nonlinear measure as steady as possible during prediction (Figure 6-1).

R2: The new value must be chosen from a set of potential values generated from a probability distribution.

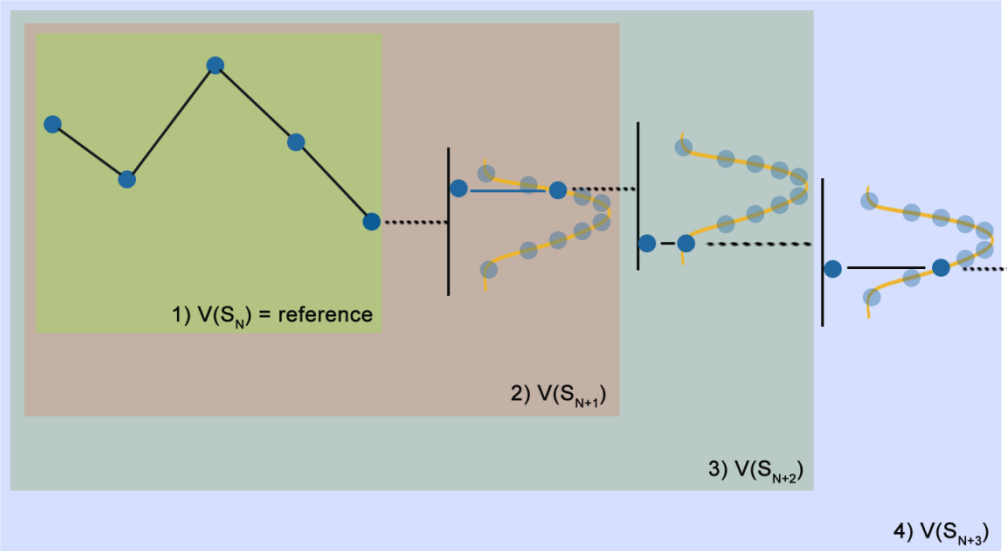


Figure 6-1. Successive steps of the GenericPred method for time series prediction

The prediction has to be pursued one step at a time because the predicted value in the current step is needed for determining the valid range of change for the next step. For those problems for which a binary prediction ('yes' or 'no') is required, (e.g., the epileptic seizure prediction), a threshold t is computed from the learning data. Whenever the value y_{N+i} is greater than the threshold t, the prediction is positive. For example, yes there is an epileptic seizure at time $N+i$ if $y_{N+i} > t$; otherwise, there will be no seizure at time $N+i$.

Classical model-based prediction approaches consider a unique value for the next step, whereas in the GenericPred method, several points are considered simultaneously. Our method is also able to constantly adjust the information regarding the current time series, whereas classical predictive methods apply the model without taking into account the concordance between the original time series and the predicted ones. Technically, any nonlinear measure could be used for the time series characterisation. However, here, we used the P&H method [149] because it has been shown that this method can efficiently discriminate between different types of nonlinear behaviour [32], [150].

6.2. Results

6.2.1. Prediction of Dow Jones Industrial Average Stock Index

To evaluate our method, three financial time series were considered. For each time series, 1500 time steps (about six years) were analyzed to predict the next unseen 500 time steps (about two years). It is therefore an out-sample prediction. We examined the Dow-Jones Industrial index (DJIA) time series with respect to the daily closing values of the DJIA for three periods of time: 1) September 1993- September 2001, 2) July 2001-July 2009, and 3) August 2004- August 2012. For the first period, our goal was to evaluate the GenericPred method when markets are stable with no major change and no financial crisis. In the second period, we evaluated the performance of the GenericPred method with respect to the prediction of financial crises. In the third period, the recession was set to the mid-range to determine how a financial crisis in the mid-range can affect the prediction of the market index. The GenericPred method prediction errors were significantly less than other methods in all three periods with any length of prediction (Table 6-1). Moreover, the GenericPred predictions were more stable with a constant lower standard deviation regardless of whether the target data lies before the recession, during the recession, or after the recession. The prediction error for the first 200 steps is especially smaller than that of the other methods.

Table 6-1. Comparison of mean absolute percentage error (MAPE) [300] **between several methods and the GenericPred method for the prediction of DJIA time series.**

	1	**10**	**50**	**100**	**200**	**300**	**400**	**500**	**Mean (1-500)**	**Std (1-500)**
DJIA 1993-2001										
ARIMA	0.22%	0.5%	4%	5%	10%	11%	10%	18%	10%	5%
GARCH	0.22%	0.7%	6%	8%	15%	18%	20%	28%	16%	7%
VAR	0.25%	0.46%	5%	7%	15%	17%	19%	28%	15%	7%
GenericPred	0.14%	0.23%	1%	3%	3%	0.1%	3%	2%	3%	2%
L-FABS	0.57%	-	-	-	-	-	-	-	-	-
MLP	1.06%	-	-	-	-	-	-	-	-	-
DJIA 2001-2009										
ARIMA	0.15%	2.5%	3%	4%	7%	13%	41%	40%	19%	17%
GARCH	0.02%	1.5%	6%	9%	17%	25%	52%	54%	27%	20%
VAR	0.02%	2%	5%	8%	14%	23%	50%	52%	26%	19%
GenericPred	0.03%	0.8%	1.5%	3%	0.27%	7%	24%	15%	10%	8%

DJIA 2004-2012										
ARIMA	0.93%	3.5%	7%	12%	17%	8%	19%	17%	12%	5%
GARCH	0.65%	2%	12%	19%	37%	44%	95%	125%	47%	32%
VAR	0.93%	3%	9%	16%	26%	20%	40%	46%	24%	12%
GenericPred	0.43%	1.5%	0.3%	2%	4%	3%	13%	8%	7%	4%

The first period corresponds to the DJIA time series between 1993 and 2001. The GenericPred method also outperformed F-FABS and MLP, two methods dedicated to short term prediction, for the first step prediction on this data. The GenericPred method still clearly outperformed the three others with an overall error rate of 2% (Table 6-1). Moreover, the GenericPred method predicts the trend with very high accuracy whereas the predictions of the other methods strongly and rapidly diverge from the real data (Figure 6-2).

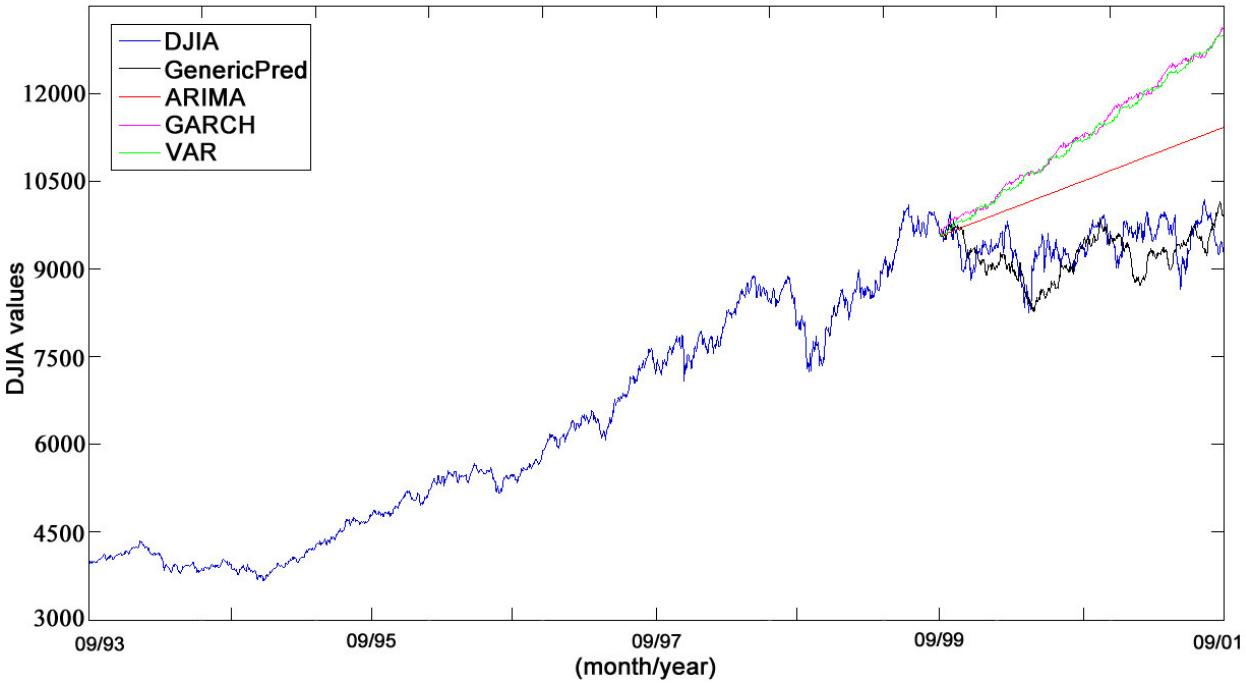


Figure 6-2. Prediction of DJIA time for first period by four different methods including proposed method. Time period between September 1993 and September 1999 has been used for prediction of time period between September 1999 until September 2001.

In the second considered period, the US stock market peaked in October 2007 but by March 2009, the Dow Jones average had reached its minimum, which reflects the worst effects of a financial crisis. The prediction for the first 300 steps of the GenericPred method are still high (less than 3% error), whereas the accuracy decreases significantly for the last 200 steps at the

peak of the financial crisis (Table 6-1). The ARIMA method still outperformed the GARCH and VAR methods, although its performance is significantly worse than that of the GenericPred method for the 500 time steps. Moreover, GenericPred is the only method able to discover the decreasing trend corresponding to the financial crisis, while the three other methods predicted a growth in the stock market (Figure 6-3).

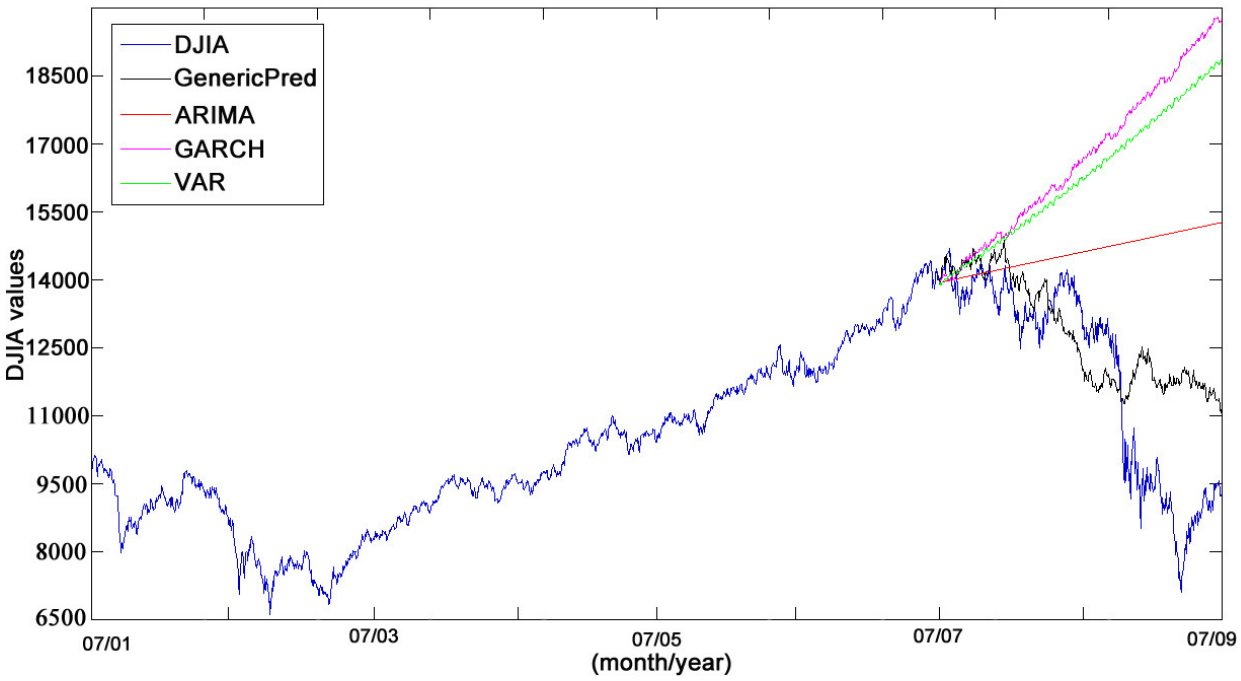


Figure 6-3. Comparison of prediction for DJIA time series for second period by four different methods including proposed method. Time period between July 2001 and July 2007 has been used for prediction of time period between July 2007 until July 2009.

In the third period (August 2004- August 2012), the recession is in the middle of the considered range. The GenericPred method has high overall predictive accuracy (4% errors in average) (Table 6-1). For the same data, with respect to prediction accuracy, ARIMA outperformed GARCH and VAR although it performed significantly worse than our method, GenericPred (12% error on average for ARIMA). Even though the 2009 financial crisis data are used for training in this experiment, the GenericPred method successfully discovers the general trends for the next 500 steps, with a particularly high accuracy for the first 300 steps, effectively predicting the increase in the stock market (Figure 6-4). The other three methods failed to predict the trend.

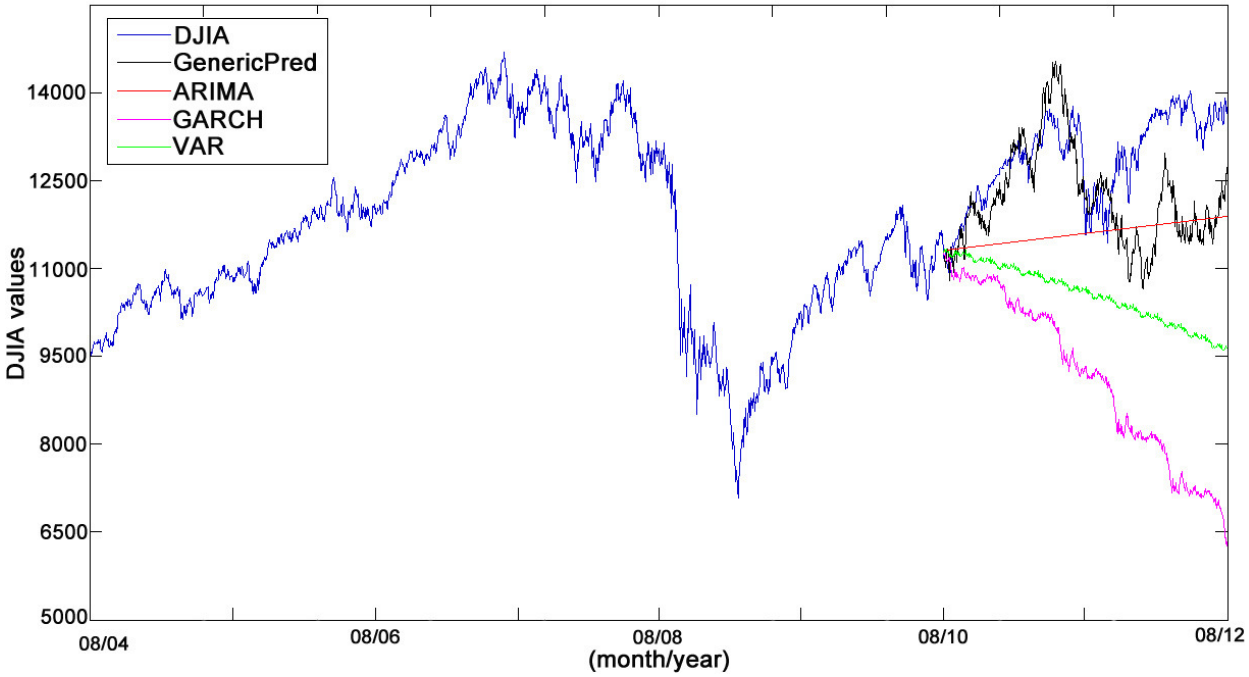


Figure 6-4. Comparison of the prediction for DJIA time series for third period by four different methods including proposed method. Time period between August 2004 and August 2010 has been used for prediction of time period between August 2010 until August 2012.

6.2.2. Prediction of Epileptic Seizure

Another important domain of time series prediction is in medical science. The detection of seizures inside an EEG even for a trained neurologist is extremely hard since there is no obvious change during epileptic seizure (Figure 6-5). The GenericPred method has been applied to all three stages (before seizure, during seizure and after seizure).

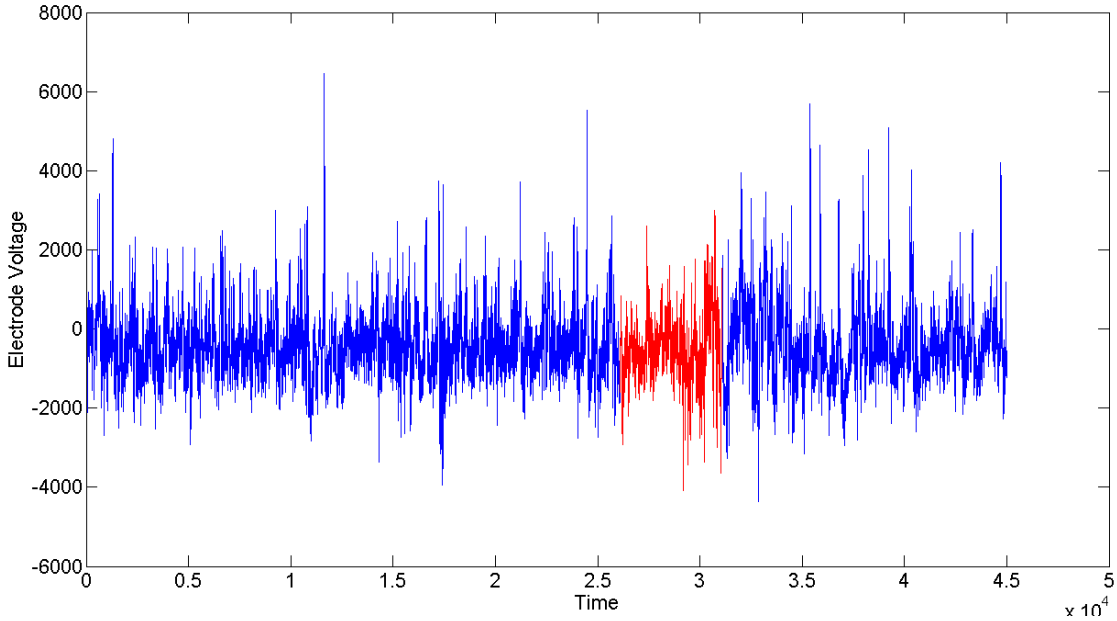


Figure 6-5. The recorded EEG time series from electrode #5 of patient #1for about three hours. The red color shows the ictal part (during seizure) of EEG.

The P&H chaoticity values [30] have been predicted using GenericPred (Figure 6-5 and Figure 6-6) on a constant-length (20 minutes) sliding window (the window moves every 20 seconds) of the five EEG time series for all 21 patients was subject to analysis. During seizure, a peak in P&H values obtained from EEG time series appears. Based on the analysis of all patients, a threshold for prediction of seizure (P&H value equal to 2.4) has been determined from the 21 patients' data. Using this threshold, the GenericPred method can predict the epileptic seizure with a 100% sensitivity and specificity up to 17 minutes in advance with a precision of few seconds. This represent a considerable improvement compared to the current state of art, which reaches a 73% sensitivity and 67% specificity accuracy for 10 patients within a 1-10 minutes range [292]. The same results have been obtained by considering data of any five electrodes independently.

Table 6-2. Sensitivity and specificity of epileptic seizure prediction for 21 patients for different length of prediction. For each patient one positive and 10 negative samples have been built. The positive sample contains one epileptic seizure event and the ten negative samples are seizure-free. Therefore, there are in total 21 positive and 210 negative samples that were used to compute the specificity and the sensitivity accuracy.

Length of prediction before seizure	Sensitivity	Specificity
16 minutes ± 7 seconds	100%	100%

17 minutes ± 7 seconds	100%	100%
18 minutes ±13 seconds	85%	100%
19 minutes ± 13 seconds	57%	100%
20 minutes ± 43 seconds	43%	100%

To illustrate the methodology we designed with GenericPred, we present here an example of prediction of epileptic seizure. For this example, the EEG time series recorded by the electrode #5 for the first patient has been considered. The EEG time series we considered, including before seizure, during seizure and after seizure, has the length of 920000 time steps (about 24 hours). According to the database, the seizure starts from the time 91100 and last until 96090 (around 8 minutes). In order to evaluate the GenericPred method for the false detection, we considered the EEG time series before time step 91100 and after time step 96090 (before and after seizure) when there is no seizure. We considered 10 ranges of EEG time series before and after seizure. No peak were predicted by GenericPred method in any of these cases (Figure 6-6 for an example of one of these cases), while GenericPred predicted the peak of P&H measure during seizure 17 minutes before it happens (Figure 6-7). The average P&H value during seizure-free part of EEG time series, is -0.3 (±0.7) while the average P&H value during seizure is 2.8 (±0.05). The same methodology has been applied to the data of the 21 patients to compute the sensitivity and specificity of GenericPred (Table 6-2).

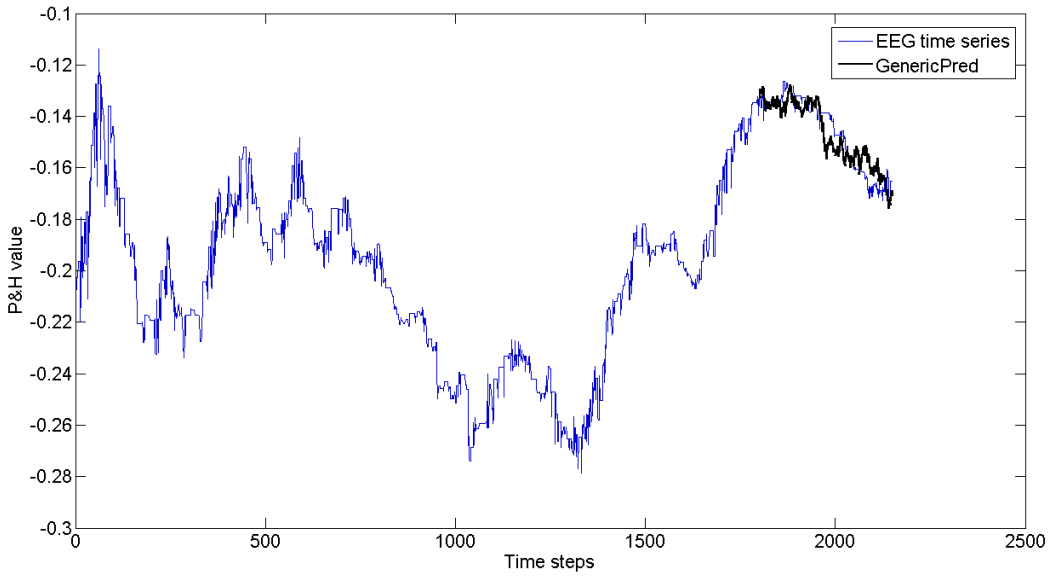


Figure 6-6. No peak up to 2.8 in P&H value is predicted by GenericPred during the seizure-free part of EEG.

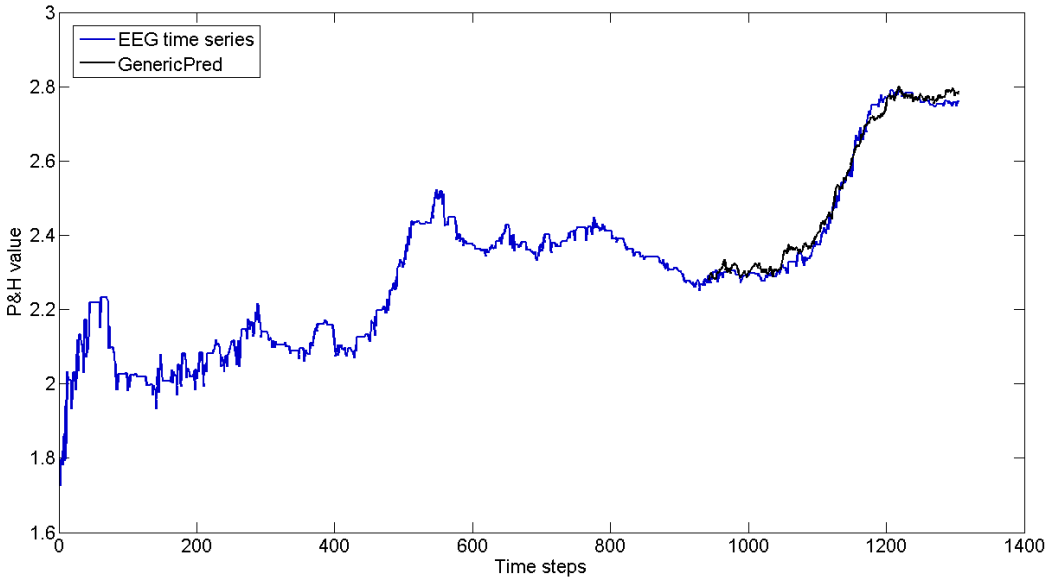


Figure 6-7. The seizure starts when the P&H value of EEG reaches to a value close to 2.8. The GenericPred method can predict the peak in P&H value (epileptic seizure) 17 minutes in advance.

6.2.3. Prediction of global temperature anomaly

We used the global temperature anomaly data from 1880 to 1983 to train for prediction of global temperature during 1983-2013. Unlike ARIMA, GenericPred accurately predicts the increasing trend in the last 30 years (Figure 6-8A). Moreover, most of the successive peaks and depressions are predicted with a high precision. The mean square error for GenericPred is 0.64 and for ARIMA is 1.6. GARCH and VAR methods cannot make prediction because of insufficient data.

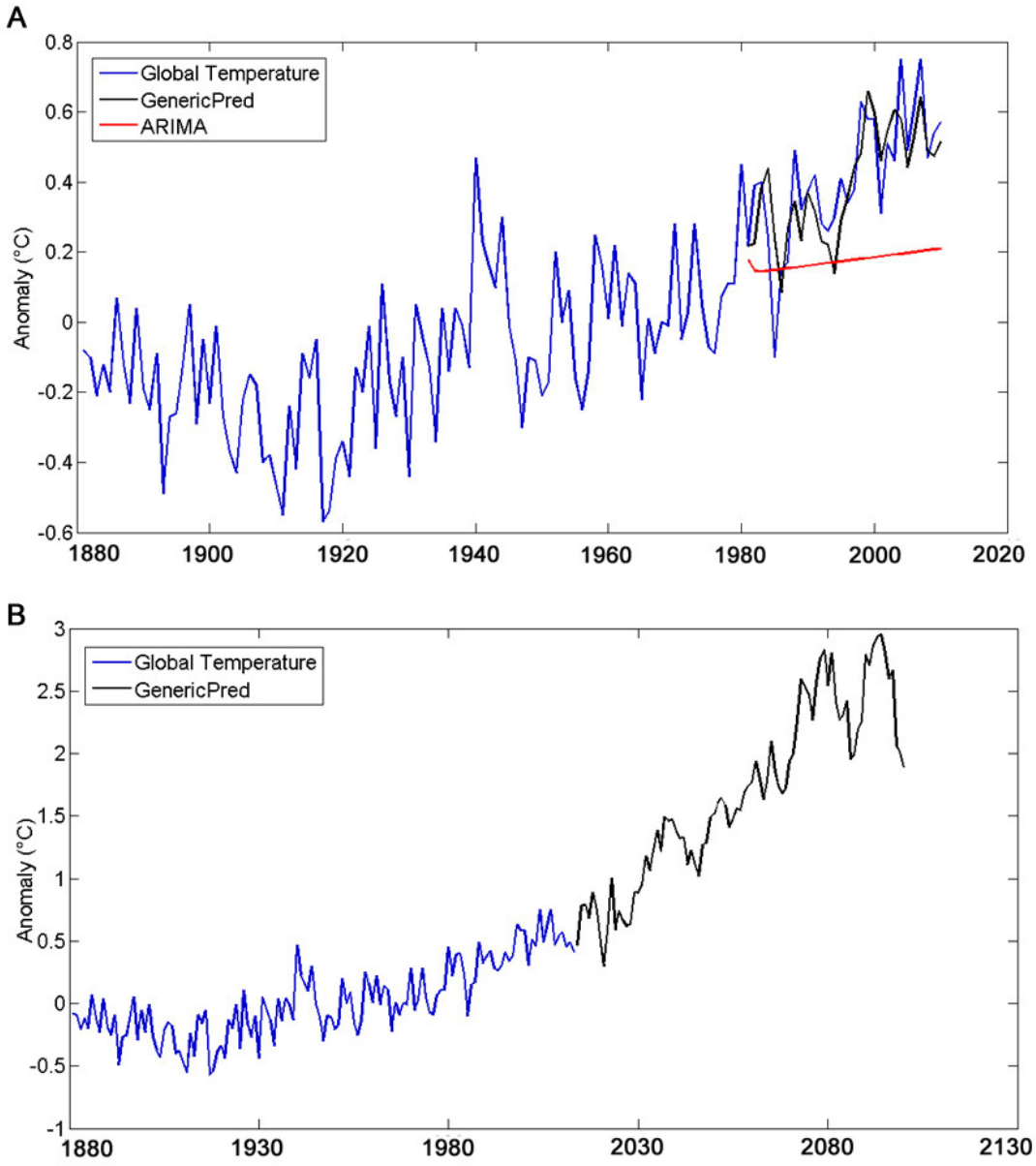


Figure 6-8. Predicting the annual records of global temperature anomaly (A) for 30 years (1983-2013) (B) until end of 21st century (2014-2100).

The existing dedicated forecasting models for global temperature anomaly predict that, as the world consumes more fossil fuel, greenhouse gas concentrations will continue to rise, and Earth's average surface temperature will continue to rise [301]. Based on recent prediction, average surface temperatures could rise between 2°C and 6°C by the end of the 21st century [175], [302]. We predict the global temperature anomaly until end of 21st century (2014-2100) (Figure 6-8B). The new method predicted an average anomaly of 2.5°C for the years 2085-2100, which is in accordance with the predictions of the dedicated models.

6.3. Conclusion

In our approach, predictions of the next previously unseen points are always performed using the complete time series, including the previous predicted points, whereas in traditional approaches, after generating the model, predictions are performed using only the model with the original time series being discarded. Therefore, our method is able to constantly adjust the information regarding the current time series, whereas classical predictive methods apply the model without taking into account the concordance between the original time series and the predicted ones.

Our approach demonstrates a significant gain over traditional methods with respect to different DJIA time series in terms of accuracy for both short and long term prediction and its ability to predict the evolutionary trend of the stock market is vastly superior to the predictive ability of existing methods. The capability of GenericPred method to predict epileptic seizure could be a major breakthrough in the fight against epilepsy providing for the first time a robust method able to predict an epileptic seizure long before its occurrence. The successful GenericPred's prediction of the increasing trend in global temperature shows that this method could also be a powerful tool for controlling global warming phenomenon. Another advantage of our approach is that it does not rely on a complex model of the original time series and it is therefore very general and very computationally efficient.

This method provides a first step towards accurate and comprehensive time series long-term predictions. Although with respect to long-term predictions it is impossible to predict the exact values, GenericPred's performance shows great potential for predicting the time series' trend.

Chapter 7

7. Conclusion

The understanding of many complex natural phenomena is still at the level of hypothesis, which in some cases can even be contradictory. Investigating those questions is important for scientists to have a better understanding about the world around us. Among those phenomena those related to biology and biological systems are among the most amazing and the most difficult to understand. However, in the last two decades new theoretical approaches, such as individual-based modeling and chaos analysis, have emerged that bring new possibilities to investigate them. With increases in computational power, it is possible to make complex individual-based models to simulate natural phenomena. However, due to their complexity, emerging from the multiple interactions between individuals, all these systems are expected to have chaotic behaviours. Chaos analysis should therefore be considered to reach a full understanding of the resulting systems. Chaos analysis can, for example, extract the history of complexity of a system to monitor how multiple components of the system affect the overall complexity as the system evolves with time. Being chaotic systems, and thus being strongly dependent on the initial conditions, it can not be expected also that two independent experiments will show the same behaviour. Chaos analysis can help to characterize the similar properties between those experiments by bringing a higher level point of view.

Biologists can hardly investigate many difficult evolutionary or ecological questions only by studying real ecosystems since in most cases there is not enough information available and even if there is, it is very time consuming and expensive to run an experiment. We employed EcoSim, a complex simulation platform, to investigate several ecological questions, as well as long-term evolutionary patterns and processes such as speciation and macroevolution. The main difference between EcoSim and the classic modeling approaches is that classic ecological modeling are based on pre-defined fitness functions. This causes a strong bias because what is “good” is predetermined and is therefore not an emerging property. For removing pre-defined fitness functions a complex system in which fitness emerges from the multiple interactions between numerous individuals is needed.

During my doctoral study, two main hypothesizes concerning species have been investigated. It is widely accepted that, if a part of a population becomes fully genetically isolated from the other part of a population, the two subpopulations may be subjected to their own unique mutations and genetic drift effects; thus, they will follow their own separated evolutionary path leading to

speciation. Once gene flow between the two groups of individuals is disrupted, speciation becomes a possibility. The first hypothesis we investigated shows that the reduction of gene flow between populations due to partial geographic barriers isolation can increase the speciation rate. Moreover, the extent to which various degrees of habitat heterogeneity influences speciation rates, which is not well understood, was also considered. We investigated how small, randomly distributed physical obstacles influence the distribution of populations and species, the level of population connectivity (e.g., gene flow) as well as the mode and tempo of speciation in a virtual ecosystem. We observed a direct and continuous increase in the speed of evolution with the increasing number of obstacles in the world, bringing a first insight to intermediate speciation mechanisms that do not relay on complete isolation. Our second hypothesis investigatetion is about Darwin's theory, the origin of species. It is a difficult problem that can hardly been studied in nature because – in most cases – the process is rare, protracted, and unreplicated. We investigated the fact that species are an inevitable byproduct of evolution. Our results confirmed the role of natural selection in speciation by showing its importance. Although abiotic conditions can certainly drive speciation, our results support assertions that biotic interactions could be particularly important drivers of the selection that causes the formation of new species.

Complex systems, such as an individual based ecosystem simulation, generate a huge amount of data. For a simulation approach to be useful for answering theoretical question, efficient methods for data analysis and knowledge extraction are also vital. We use several machine learning techniques, including feature selection, classification and rule extraction to analyze the data generated by the EcoSim experiments. Our objective was to conduct a robust test of the effectiveness of our framework for identifying important features for different theoretical ecological questions. By interpreting the obtained models we have been able to extract meaningful rules to enrich our knowledge about the kind of features involved in several biological problems and how their combination can be used to predict a biological event, such as species richness variation. By all these studies, we contributed to a deeper understanding of concepts such as: the evolutionary process and the emergence of species. We also showed that machine learning techniques are particularly efficient to analyze such data bringing semantically interpretable rules with high predictive accuracy and therefore these techniques should be considered as important tools for future theoretical or empirical studies. All of these studies could have some significance in ecological resource management, epidemiology, or in studying the impact of human behaviour on ecosystems.

Any attempt to model a real complex system without knowing how realistic the model or simulation is, can lead to an inaccurate and unacceptable result. In order to make sure that the behaviour of our simulation can reach the same level of complexity as natural phenomena, we applied different chaos and nonlinear analyses to the results of EcoSim. We used four different methods: Higuchi fractal dimension, correlation dimension, largest Lyapunov exponent, P&H method to investigate the behaviour of population time series in EcoSim. According to the results obtained, we can conclude that behaviour of population time series in EcoSim is deterministic. Also among various cases of deterministic behaviour, we show that the behaviour of population time series in EcoSim is chaotic. Since it has been shown the multifractal property is a common feature in the spatial distribution of different animal communities, we also applied multifractal analysis to the data generated by the EcoSim simulation. Multifractal analysis of EcoSim's results demonstrated self-similarity characteristics in the spatial distribution of individuals as it has been observed in real ecosystems. We analyzed different parameters of the simulation to detect which ones cause the multifractal behaviour. More importantly, we showed that the combination of the predation pressure associated with the distribution of food is an important factor for the emergence of multifractal phenomena. These results also s thehows capacity of EcoSim to generate data with complex characteristics generally observed in real ecosystem studies. Moreover, we have shown that complex systems, having chaotic behaviour, can easily be created to realistically model biological phenomena. We have also shown that, even if these system are complex, it is still possible to accurately analyze them and to extract some general and pertinent properties from them that allows to describe and characterize them.

During our studies we have had experiences with: nonlinear analysis of different complex time series, design and implementation of accurate models (simulation) for complex systems, and study of the relative influence of different parameters on the emerging patterns of such systems. These studies showed us the possibility and interest of making prediction for complex systems, in particular chaotic ones. However an intensive investigation of the state of the art revealed that, even focusing only on nonlinear time series, all the existing methods cannot perform correct long-term prediction. We came up with a new approach, which could bring a whole new perspective on this topic. A vast number of applications could be addressed by this method. The new method's ability in predicting the evolutionary trend of the stock market is vastly superior to the predictive ability of existing methods (including the successful prediction of 2009 financial crisis two years before it happened). Such predictions could be useful to improve businesses investment decisions or to help governments to make better fiscal and monetary policy decisions. In medical

science, there are also many applications for which an efficient prediction algorithm could save lives; for example the capability of the new method to predict epileptic seizure could be a major breakthrough in the fight against epilepsy providing for the first time a robust method able to predict an epileptic seizure long before its occurrence. The successful new method's prediction of the increasing trend in global temperature shows that this method could also be a powerful tool for controlling global warming phenomenon. Another advantage of our approach is that it does not rely on a complex model of the original time series and it is therefore very general and very computationally efficient. Time series analysis of earth's seismic waves can be used for earthquake prediction.

All the methods and tools we designed to model, analyze and predict complex system have shown us that there is a strong interest to investigate more deeply such approaches. However, complex system design and analysis is still at an early stage of development. It should be particularly interesting to study how different techniques could be combined and integrated in an unique system. For example, up to now, modeling natural phenomenon and chaos analysis are two separate steps. To guaranty that an individual-based model is conceived respecting the true level of complexity of targeted system, these two steps could be combined. Involving the chaos computations inside the behaviour of model could help to control the chaotic behaviour of model, although more research is still needed for finding the true level of chaos of a system. Having the possibility to simulate a complex system, such as a financial market or brain generating signals, targeting a particular level of chaos, or the variation of level of chaos, monitoring the ones of the corresponding real system, open a completely new promising field of research.

Appendix A

Glossary

Allopatry	Allopatric speciation is speciation that occurs when biological populations of the same species become vicariant, or isolated from each other to an extent that prevents or interferes with genetic interchange.
Chaotic behaviour	Deterministic behaviour extremely sensitive to initial conditions.
Coexistence	Organisms exist together, at the same time and in the same place
Competitive exclusion	The inevitable elimination from a habitat of one of two different species with identical needs for resources.
Complex systems	Systems that consist of many diverse and autonomous but interrelated and interdependent components or parts linked through many interconnections.
Extrinsic adaptation	Not using a pre-defined fitness function.
Fitness function	The fitness function evaluates how good a potential solution, or a population of organisms, or a set of physiological traits is relative to others.
Fuzzification	The fuzzification comprises the process of transforming values into grades (probabilities) of membership for fuzzy sets.
Non-chaotic behavior	The behavior of signal which has not all of these properties: 1) Nonlinear 2) Aperiodic 3) Extremely sensitive to initial condition.
Species	A set of naturally interbreeding organisms that are genetically reproductively isolated from other sets of organisms.
Sympatry	Two species or populations are considered sympatric when they exist in the same geographic area and thus regularly encounter one another. An initially

	interbreeding population that splits into two or more distinct species sharing a common range exemplifies sympatric speciation.

Appendix B

Copyright Permissions

1- Robin Gras

I give permission to include materials for the papers presented in chapters 2, 4, 5 and 6 of Abbas Golestani's dissertation.

2- Morteza Mashayekhi, Marwa Khater, Yasaman Farahani

I do give permission to include materials for the papers presented in chapters 2 and 4 of Abbas Golestani's dissertation.

REFERENCES / BIBLIOGRAPHY

[1] J. A. Coyne, *Why Evolution is True*. Viking, 2009, p. 282.

[2] C. Adami, *Introduction to Artificial Life [Hardcover]*. Springer; Corrected edition, 1997, p. 374.

[3] T. J. Taylor, "From Artificial Evolution to Artificial Life." University of Edinburgh. College of Science and Engineering. School of Informatics., Jul-1999.

[4] I. L. Boyd, "Ecology. The art of ecological modeling.," *Science*, vol. 337, no. 6092, pp. 306–7, Jul. 2012.

[5] D. L. DeAngelis and L. J. Gross, *Individual-based Models and Approaches in Ecology: Populations, Communities, and Ecosystems*. Chapman & Hall, 1992, p. 525.

[6] M. Bithell and J. Brasington, "Coupling agent-based models of subsistence farming with individual-based forest models and dynamic models of water distribution," *Environ. Model. Softw.*, vol. 24, no. 2, pp. 173–190, Feb. 2009.

[7] F. L. Hellweger and V. Bucci, "A bunch of tiny individuals—Individual-based modeling for microbes," *Ecol. Modell.*, vol. 220, no. 1, pp. 8–22, Jan. 2009.

[8] T. Filatova, P. H. Verburg, D. C. Parker, and C. A. Stannard, "Spatial agent-based models for socio-ecological systems: challenges and prospects," *Environmental modelling & software*. 27-Jul-2013.

[9] C. Ricotta, "From theoretical ecology to statistical physics and back: self-similar landscape metrics as a synthesis of ecological diversity and geometrical complexity," *Ecol. Modell.*, vol. 125, no. 2–3, pp. 245–253, 2000.

[10] R. Sneyers, "Climate Chaotic Instability: Statistical Determination and Theoretical Background," *Environmetrics*, vol. 8, no. February, pp. 517–532, 1997.

[11] K. T. Alligood, T. D. Sauer, and J. A. Yorke, *Chaos: An Introduction to Dynamical Systems (Textbooks in Mathematical Sciences)*. Springer, 2000, p. 604.

[12] J. Kurths and H. Herzel, "Can a solar pulsation event be characterized by a low-dimensional chaotic attractor?," *Sol. Phys.*, vol. 107, no. 1, pp. 39–45, 1986.

[13] G. P. Pavlos, L. P. Karakatsanis, and M. N. Xenakis, "Tsallis non-extensive statistics, intermittent turbulence, SOC and chaos in the solar plasma, Part one: Sunspot dynamics," *Phys. A Stat. Mech. its Appl.*, vol. 391, no. 24, pp. 6287–6319, 2012.

[14] J. Kurths and M. Karlicky, "The route to chaos during a pulsation event," *Sol. Phys.*, vol. 119, no. 2, pp. 399–411, 1989.

[15] C. Pahl-Wostl, *The Dynamic Nature of Ecosystems: Chaos and Order Entwined*. Wiley, 1995, p. 267.

[16] "An elegant chaos," *Nature*, vol. 507, no. 7491, pp. 139–140, Mar. 2014.

[17] A. M. Stomp, "Genetic information and ecosystem health: arguments for the application of chaos theory to identify boundary conditions for ecosystem management.," *Environ. Health Perspect.*, vol. 102 Suppl , pp. 71–4, Dec. 1994.

[18] E. E. Popova, M. J. R. Fasham, A. V Osipov, and V. A. Ryabchenko, "Chaotic behaviour of an ocean ecosystem model under seasonal external forcing," *J. Plankton Res.*, vol. 19, no. 10, pp. 1495–1515, 1997.

[19] J. Kurths and U. Schwarz, "Chaos theory and radio emission," *Space Sci. Rev.*, vol. 68, no. 1–4, pp. 171–184, May 1994.

[20] C. Yi, Z. Jinkui, and H. Jiaxin, "Research on application of earthquake prediction based on chaos theory," in *Intelligent Computing and Integrated Systems (ICISS), 2010 International Conference on*, 2010, pp. 753–756.

[21] I. P. Janecka, "Cancer control through principles of systems science, complexity, and chaos theory: a model.," *Int. J. Med. Sci.*, vol. 4, no. 3, pp. 164–73, Jan. 2007.

[22] D. N. Chorafas, *Chaos Theory in the Financial Markets*. McGraw Hill Professional, 1994, p. 382.

[23] L. Romanelli, M. A. Figliola, and F. A. Hirsch, "Deterministic chaos and natural phenomena," *J. Stat. Phys.*, vol. 53, no. 3, pp. 991–994, 1988.

[24] A. Accardo, M. Affinito, M. Carrozzi, and F. Bouquet, "Use of the fractal dimension for the analysis of electroencephalographic time series," *Biol. Cybern.*, vol. 77, no. 5, pp. 339–350, 1997.

[25] G. Kubin, "What is a chaotic signal?," in *in Proceedings of the 1995 IEEE Workshop on Nonlinear Signal and Image Processing, Ed. I. Pitas*, 1995.

[26] S. Basu and E. Foufoula-Georgiou, "Detection of nonlinearity and chaoticity in time series using the transportation distance function," *Phys. Lett. A*, vol. 301, no. 5–6, pp. 413–423, 2002.

[27] Z. J. Kowalik and T. Elbert, "Changes of chaoticness in spontaneous EEG/MEG," *Integr. Psychol. Behav. Sci.*, vol. 29, no. 3, pp. 270–282, 1994.

[28] J. Fell, G. Fernández, and C. E. Elger, "More than synchrony: EEG chaoticity may be necessary for conscious brain functioning," *Med. Hypotheses*, vol. 61, no. 1, pp. 158–160, 2003.

[29] R. Gras, D. Devaurs, A. Wozniak, and A. Aspinall, “An individual-based evolving predator-prey ecosystem simulation using a fuzzy cognitive map as the behavior model,” *Artif. Life*, vol. 15, no. 4, pp. 423–463, 2009.

[30] A. Golestani, R. Gras, and M. Cristescu, “Speciation with gene flow in a heterogeneous virtual world: can physical obstacles accelerate speciation?,” *Proc. R. Soc. B Biol. Sci.*, 2012.

[31] G. Robin, A. Golestani, C. Melania, and A. P. Hendry, “Speciation without pre-defined fitness functions,” *PLoS One*, 2014.

[32] A. Golestani and R. Gras, “Regularity analysis of an individual-based ecosystem simulation,” *Chaos*, vol. 20, no. 4, p. 3120, 2010.

[33] A. Golestani and R. Gras, “Multifractal phenomena in EcoSim, a large scale individual-based ecosystem simulation,” in *Int. Conf. Artificial Intelligence, Las Vegas, 2011*, 2011, pp. 991–999.

[34] A. Golestani and R. Gras, “Identifying Origin of Self-Similarity in EcoSim, an Individual-Based Ecosystem Simulation, using Wavelet-based Multifractal Analysis,” in *Proceedings of the World Congress on Engineering and Computer Science*, 2012, vol. 2, pp. 1275–1282.

[35] A. Golestani and R. Gras, “Can we predict the unpredictable?,” *Sci. Rep.*, vol. 4, p. 6834, 2014.

[36] G. W. S. II, *Ecological Risk Assessment, Second Edition*. CRC Press, 2006, p. 680.

[37] M. Gillman, *Introduction to Ecological Modelling*. Wiley, 1997, p. 202.

[38] S. E. Jørgensen and B. D. Fath, *Fundamentals of Ecological Modelling: Applications in Environmental Management and Research*. Elsevier Science & Technology Books, 2011, p. 399.

[39] C. A. Drew, Y. F. Wiersma, and F. Huettmann, *Predictive Species and Habitat Modeling in Landscape Ecology: Concepts and Applications (Google eBook)*, vol. 2010. Springer, 2010, p. 328.

[40] J. H. Holland, *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press, 1975.

[41] M. Mitchell, “An introduction to genetic algorithms,” Mar. 1996.

[42] L. Kallel, B. Naudts, and A. Rogers, *Theoretical aspects of evolutionary computing*. Springer, 2001.

[43] N. H. Packard, “Intrinsic adaptation in a simple model for evolution,” *Artif. Life*, vol. 141, 1989.

[44] A. D. Channon and R. I. Damper, “Perpetuating evolutionary emergence,” 1998.

[45] D. E. Goldberg and J. Richardson, “Genetic algorithms with sharing for multimodal function optimization,” pp. 41–49, Oct. 1987.

[46] S. W. Mahfoud, “Simple Analytical Models of Genetic Algorithms for Multimodal Function Optimization.,” in *ICGA*, 1993, p. 643.

[47] J. J. Grefenstette, “Evolvability in dynamic fitness landscapes: a genetic algorithm approach,” in *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)*, 1999, vol. 3, pp. 2031–2038.

[48] Y. Chen, *Extending the Scalability of Linkage Learning Genetic Algorithms: Theory & Practice (Google eBook)*. Springer, 2006, p. 120.

[49] M. Pelikan, D. E. Goldberg, and F. G. Lobo, “A Survey of Optimization by Building and Using Probabilistic Models,” *Comput. Optim. Appl.*, vol. 21, no. 1, pp. 5–20, Jan. 2002.

[50] M. Pelikan, *Hierarchical Bayesian Optimization Algorithm: Toward a New Generation of Evolutionary Algorithms*. Springer, 2005, p. 166.

[51] A. D. Channon and R. I. Damper, “Towards the evolutionary emergence of increasingly complex advantageous behaviours,” *Int. J. Syst. Sci.*, vol. 31, no. 7, pp. 843–860, 2000.

[52] N. H. P. C. Langton, C. Taylor, D. Farmer, Mark A. Bedau, “Measurement of Evolutionary Activity, Teleology, and Life,” 1996.

[53] L. Yaeger, “Computational genetics, physiology, metabolism, neural systems, learning, vision, and behavior or PolyWorld: Life in a new context,” in *Artificial Life III, Vol. XVII of SFI Studies in the Sciences of Complexity, Santa Fe Institute*, 1993, pp. 263–298.

[54] S. Gavrilets, *Fitness landscapes and the origin of species (MPB-41)*. Princeton University Press, 2004.

[55] S. Gavrilets, A. Vose, M. Barluenga, W. Salzburger, and A. Meyer, “Case studies and mathematical models of ecological speciation. 1. Cichlids in a crater lake.,” *Mol. Ecol.*, vol. 16, no. 14, pp. 2893–909, Jul. 2007.

[56] U. Dieckmann and M. Doebeli, “On the origin of species by sympatric speciation,” *Nature*, vol. 400, no. 6742, pp. 354–357, 1999.

[57] M. Kirkpatrick and S. L. Nuismer, “Sexual selection can constrain sympatric speciation.,” *Proc. Biol. Sci.*, vol. 271, no. 1540, pp. 687–93, Apr. 2004.

[58] D. I. Bolnick, “Multi-species outcomes in a common model of sympatric speciation.,” *J. Theor. Biol.*, vol. 241, no. 4, pp. 734–44, Aug. 2006.

[59] B. Drossel and A. Mckane, “Competitive speciation in quantitative genetic models.,” *J. Theor. Biol.*, vol. 204, no. 3, pp. 467–78, Jun. 2000.

[60] M. Doebeli, H. J. Blok, O. Leimar, and U. Dieckmann, “Multimodal pattern formation in phenotype distributions of sexual populations.,” *Proc. Biol. Sci.*, vol. 274, no. 1608, pp. 347–57, Feb. 2007.

[61] M. Doebeli and U. Dieckmann, “Speciation along environmental gradients,” *Nature*, vol. 421, no. 6920, pp. 259–264, 2003.

[62] M. Higashi, G. Takimoto, and N. Yamamura, “Sympatric speciation by sexual selection,” *Nature*, vol. 402, no. 6761, pp. 523–526, 1999.

[63] G. Takimoto, M. Higashi, and N. Yamamura, “A deterministic genetic model for sympatric speciation by sexual selection.,” *Evolution*, vol. 54, no. 6, pp. 1870–81, Dec. 2000.

[64] S. Gavrilets and A. Vose, “Case studies and mathematical models of ecological speciation. 2. Palms on an oceanic island.,” *Mol. Ecol.*, vol. 16, no. 14, pp. 2910–21, Jul. 2007.

[65] C. Seven, “Dynamic patterns of adaptive radiation : evolution of mating preferences sergey gavrilets and aaron vose,” pp. 102–126, 2009.

[66] X. Thibert-Plante and A. P. Hendry, “Factors influencing progress toward sympatric speciation.,” *J. Evol. Biol.*, vol. 24, no. 10, pp. 2186–96, Oct. 2011.

[67] X. Thibert-Plante and A. P. Hendry, “The consequences of phenotypic plasticity for ecological speciation.,” *J. Evol. Biol.*, vol. 24, no. 2, pp. 326–42, Feb. 2011.

[68] F. Débarre, “Refining the conditions for sympatric ecological speciation.,” *J. Evol. Biol.*, vol. 25, no. 12, pp. 2651–60, Dec. 2012.

[69] B. Allen, M. A. Nowak, and U. Dieckmann, “Adaptive dynamics with interaction structure.,” *Am. Nat.*, vol. 181, no. 6, pp. E139–63, Jun. 2013.

[70] K. Thearling and T. Ray, “Evolving multi-cellular artificial life,” *Artif. Life IV*, vol. IV, no. July, pp. 283–288, 1994.

[71] R. K. Standish, “Open-Ended Artificial Evolution,” Oct. 2002.

[72] C. Ofria and C. O. Wilke, “Avida: a software platform for research in computational evolutionary biology.,” *Artif. Life*, vol. 10, no. 2, pp. 191–229, Jan. 2004.

[73] R. E. Lenski, C. Ofria, R. T. Pennock, and C. Adami, “The evolutionary origin of complex features.,” *Nature*, vol. 423, no. 6936, pp. 139–44, May 2003.

[74] C. O. Wilke, J. L. Wang, C. Ofria, R. E. Lenski, and C. Adami, “Evolution of digital organisms at high mutation rates leads to survival of the flattest,” *Nature*, vol. 412, no. 6844, pp. 331–333, 2001.

[75] M. Huston, D. DeAngelis, and W. Post, “New Computer Models Unify Ecological Theory,” *Bioscience*, vol. 38, no. 10, pp. 682–691, Nov. 1988.

[76] P. T. Hraber, T. Jones, and S. Forrest, “The ecology of echo.,” *Artif. Life*, vol. 3, no. 3, pp. 165–90, Jan. 1997.

[77] S. Forrest and T. Jones, “Modeling complex adaptive systems with Echo,” *Complex Syst. Mech. Adapt. RJ Stonier XH Yu), IOS Press*, pp. 3–21, 1994.

[78] L. Yaeger, “Poly world: Life in a new context,” *Proc. Artif. Life*, vol. 3, p. 263, 1994.

[79] J. T. Lizier, M. Piraveenan, D. Pradhana, M. Prokopenko, and L. S. Yaeger, “Functional and structural topologies in evolved neural networks,” pp. 140–147, Sep. 2009.

[80] V. Grimm and S. F. Railsback, *Individual-based Modeling and Ecology*. Princeton University Press, 2005, p. 428.

[81] M. Smith, “Using massively-parallel supercomputers to model stochastic spatial predator-prey systems,” *Ecol. Modell.*, vol. 58, no. 1–4, pp. 347–367, Nov. 1991.

[82] V. Volterra, “Variations and fluctuations of the number of individuals in animal species living together,” *J. Cons. Int. Explor. Mer*, vol. 3, no. 1, pp. 3–51, 1928.

[83] G. Bell, “The evolution of trophic structure.,” *Heredity (Edinb).*, vol. 99, no. 5, pp. 494–505, Nov. 2007.

[84] W. Yamaguchi, M. Kondoh, and M. Kawata, “Effects of evolutionary changes in prey use on the relationship between food web complexity and stability,” *Popul. Ecol.*, vol. 53, no. 1, pp. 59–72, Apr. 2010.

[85] C. J. Scogings, K. A. Hawick, and H. A. James, “Tools and techniques for optimisation of microscopic artificial life simulation models,” in *Proceedings of the Sixth IASTED International Conference on Modelling, Simulation, and Optimization, Gabarone, Botswana*, 2006, pp. 90–95.

[86] C. J. Scogings and K. A. Hawick, “Modelling Predator Camouflage Behaviour and Tradeoffs in an Agent-Based,” in *Proc. IASTED International Conference on Modelling and Simulation*, pp. 32–802.

[87] C. J. Scogings and K. A. Hawick, “Introducing a gestation period of time-delayed benefit into an animat-based artificial life model,” in *Proc. 12th IASTED Int. Conf. on Artificial Intelligence and Applications (AIA□13), Innsbruck, Austria, IASTED*, 2013, pp. 43–50.

[88] V. Grimm, U. Berger, F. Bastiansen, S. Eliassen, V. Ginot, J. Giske, J. Goss-Custard, T. Grand, S. K. Heinz, G. Huse, A. Huth, J. U. Jepsen, C. Jørgensen, W. M. Mooij, B. Müller, G. Pe’er, C. Piou, S. F. Railsback, A. M. Robbins, M. M. Robbins, E. Rossmanith, N. Rüger, E. Strand, S. Souissi, R. A. Stillman, R. Vabø, U. Visser, and D. L. DeAngelis, “A standard protocol for describing individual-based and agent-based models,” *Ecol. Modell.*, vol. 198, no. 1–2, pp. 115–126, Sep. 2006.

[89] M. Mashayekhi, A. Golestani, Y. M. F. Farahani, and R. Gras, "An enhanced artificial ecosystem: Investigating emergence of ecological niches," in *ALIFE 14: The Fourteenth Conference on the Synthesis and Simulation of Living Systems*, vol. 14, pp. 693–700.

[90] B. Kosko, "Fuzzy cognitive maps," *Int. J. Man. Mach. Stud.*, vol. 24, no. 1, pp. 65–75, 1986.

[91] J. Tisseau, "R{é}alit{é} virtuelle: autonomie in virtuo," *Habilit. Dir. des Rech. Univ. Rennes*, vol. 1, 2001.

[92] M. Khater, D. Murariu, and R. Gras, "Contemporary Evolution and Genetic Change of Prey as a Response to Predator Removal," *Ecol. Inform.*, vol. 22, pp. 13–22, Feb. 2014.

[93] L. Seuront, F. Schmitt, Y. Lagadeuc, D. Schertzer, S. Lovejoy, and S. Frontier, "Multifractal analysis of phytoplankton biomass and temperature in the ocean," *Geophys. Res. Lett.*, vol. 23, no. 24, pp. 3591–3594, 1996.

[94] V. N. Biktashev, J. Brindley, a V Holden, and M. a Tsyganov, "Pursuit-evasion predator-prey waves in two spatial dimensions.," *Chaos*, vol. 14, no. 4, pp. 988–94, Dec. 2004.

[95] N. F. Otani, A. Mo, S. Mannava, F. H. Fenton, E. M. Cherry, S. Luther, and R. F. Gilmour Jr, "Characterization of multiple spiral wave dynamics as a stochastic predator-prey system," *Phys. Rev. E*, vol. 78, no. 2, p. 21913, 2008.

[96] J. Mallet, "A species definition for the modern synthesis," *Trends Ecol. Evol.*, vol. 10, no. 7, pp. 294–299, 1995.

[97] A. Aspinall and R. Gras, "K-means clustering as a speciation mechanism within an individual-based evolving predator-prey ecosystem simulation," *Act. Media Technol.*, pp. 318–329, 2010.

[98] D. Devaurs and R. Gras, "Species abundance patterns in an ecosystem simulation studied through Fisher's logseries," *Simul. Model. Pract. Theory*, vol. 18, no. 1, pp. 100–123, 2010.

[99] R. A. Fisher, A. S. Corbet, and C. B. Williams, "The relation between the number of species and the number of individuals in a random sample of an animal population," *J. Anim. Ecol.*, pp. 42–58, 1943.

[100] M. Mashayekhi, B. MacPherson, and R. Gras, "A machine learning approach to investigate the reasons behind species extinction," *Ecol. Inform.*, vol. 20, pp. 58–66, Feb. 2014.

[101] B. D. Hughes, "Random Walks and Random Environments Clarendon." Oxford, 1995.

[102] S. P. Hubbell, *The unified neutral theory of biodiversity and biogeography (MPB-32)*, vol. 32. Princeton University Press, 2001.

[103] A. J. Lotka, “Contribution to the theory of periodic reactions,” *J. Phys. Chem.*, vol. 14, no. 3, pp. 271–274, 1910.

[104] V. Volterra, *Variazioni e fluttuazioni del numero d’individui in specie animali conviventi*. C. Ferrari, 1927.

[105] J. C. Sprott, *Chaos and time-series analysis*, vol. 69. Oxford University Press Oxford, UK:, 2003.

[106] C. Werndl, “What Are the New Implications of Chaos for Unpredictability?,” *Br. J. Philos. Sci.*, vol. 60, no. 1, pp. 195–220, Jan. 2009.

[107] C. S. Bertuglia and F. Vaio, *Nonlinearity, Chaos, And Complexity: The Dynamics Of Natural And Social Systems*. Oxford University Press, Incorporated, 2005, p. 387.

[108] *Complex Nonlinearity: Chaos, Phase Transitions, Topology Change and Path Integrals*. Springer; 2008 edition, 2008, p. 844.

[109] S. N. Elaydi, *Discrete Chaos, Second Edition: With Applications in Science and Engineering*, vol. 2007. CRC Press, 2007, p. 440.

[110] V. G. Ivancevic, *Complex nonlinearity chaos, phase transitions, topology change, and path integrals /*. Berlin : Springer, 2008.

[111] A. Medio and M. Lines, *Nonlinear Dynamics: A Primer*. Cambridge University Press, 2001, p. 300.

[112] R. Cohen and S. Havlin, *Complex Networks: Structure, Robustness and Function (Google eBook)*. Cambridge University Press, 2010, p. 238.

[113] B. J. Carr and A. A. Coley, “Self-similarity in general relativity,” *Class. Quantum Gravity*, vol. 16, no. 7, pp. R31–R71, Jul. 1999.

[114] D. C. Parker, A. Hessl, and S. C. Davis, “Complexity, land-use modeling, and the human dimension: Fundamental challenges for mapping unknown outcome spaces,” *Geoforum*, vol. 39, no. 2, pp. 789–804, Mar. 2008.

[115] D. C. Parker, S. M. Manson, M. A. Janssen, M. J. Hoffmann, and P. Deadman, “Multi-Agent Systems for the Simulation of Land-Use and Land-Cover Change: A Review,” *Ann. Assoc. Am. Geogr.*, vol. 93, no. 2, pp. 314–337, Jun. 2003.

[116] T. E. V. M. Dawn C. Parker, “Measuring Emergent Properties of Agent-Based Landcover/Landuse Models using Spatial Metrics,” *Comput. Econ. Financ. 2001*, Apr. 2001.

[117] D. C. Parker, T. Berger, S. M. Manson, and L. Use, *Agent-based models of land-use and land-cover change: report and review of an international workshop, October 4-7, 2001, Irvine, California, USA*. LUCC Focus 1 Office, 2002.

[118] D. T. Robinson, D. G. Brown, D. C. Parker, P. Schreinemachers, M. A. Janssen, M. Huigen, H. Wittmer, N. Gotts, P. Promburom, E. Irwin, and others, "Comparison of empirical methods for building agent-based models in land use science," *J. Land Use Sci.*, vol. 2, no. 1, pp. 31–55, 2007.

[119] J. H. Brown, V. K. Gupta, B.-L. Li, B. T. Milne, C. Restrepo, and G. B. West, "The fractal nature of nature: power laws, ecological complexity and biodiversity.," *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 357, no. 1421, pp. 619–26, May 2002.

[120] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*. Wiley, 2003, p. 368.

[121] G. A. Edgar, *Classics on fractals*. Addison-Wesley, 1993, p. 366.

[122] N. A. Salingaros and B. J. West, "A universal rule for the distribution of sizes," *Environ. Plan. B*, vol. 26, pp. 909–924, 1999.

[123] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, 2009.

[124] A.-L. Barabási, "Emergence of scaling in complex networks," *Handb. graphs networks from genome to internet*, pp. 69–84, 2002.

[125] J. D. Skufca, J. A. Yorke, and B. Eckhardt, "Edge of chaos in a parallel shear flow," *Phys. Rev. Lett.*, vol. 96, no. 17, p. 174101, 2006.

[126] K. SANDAU and H. KURZ, "Measuring fractal dimension and complexity - an alternative approach with an application," *J. Microsc.*, vol. 186, no. 2, pp. 164–176, May 1997.

[127] A. Brandstäter, J. Swift, H. L. Swinney, A. Wolf, J. D. Farmer, E. Jen, and P. J. Crutchfield, "Low-dimensional chaos in a hydrodynamic system," *Phys. Rev. Lett.*, vol. 51, no. 16, p. 1442, 1983.

[128] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Phys. D Nonlinear Phenom.*, vol. 31, no. 2, pp. 277–283, 1988.

[129] T. Higuchi, "Relationship between the fractal dimension and the power law index for a time series a numerical investigation," *Phys. D Nonlinear Phenom.*, vol. 46, no. 2, pp. 254–264, 1990.

[130] L. Telesca, G. Colangelo, V. Lapenna, and M. Macchiato, "Monofractal and multifractal characterization of geoelectrical signals measured in southern Italy," *Chaos, Solitons & Fractals*, vol. 18, no. 2, pp. 385–399, 2003.

[131] W. Klonowski, "Everything you wanted to ask about EEG but were afraid to get the right answer," *Nonlinear Biomed. Phys.*, vol. 3, no. 2, 2009.

[132] W. Klonowski, "Chaotic dynamics applied to signal complexity in phase space and in time domain," *Chaos, Solitons & Fractals*, vol. 14, no. 9, pp. 1379–1387, 2002.

[133] B. S. Raghavendra and D. N. Dutt, "Nonlinear dynamical characterization of heart rate variability time series of meditation," *Health (Irvine. Calif).*, vol. 3, p. 10, 2010.

[134] M. Small, *Applied Nonlinear Time Series Analysis: Applications in Physics, Physiology and Finance (World Scientific Series on Nonlinear Science Series a)*. World Scientific Pub Co Inc, 2005, p. 245.

[135] D. Yu, M. Small, R. Harrison, and C. Diks, "Efficient implementation of the gaussian kernel algorithm in estimating invariants and noise level from noisy time series data," *Phys. Rev. E. Stat. Phys. Plasmas. Fluids. Relat. Interdiscip. Topics*, vol. 61, no. 4 Pt A, pp. 3750–6, Apr. 2000.

[136] C. Diks, "Estimating invariants of noisy attractors," *Phys. Rev. E*, vol. 53, no. 5, pp. R4263–R4266, May 1996.

[137] R. Lopes and N. Betrouni, "Fractal and multifractal analysis: a review.," *Med. Image Anal.*, vol. 13, no. 4, pp. 634–49, Aug. 2009.

[138] P. A. Moreno, P. E. Vélez, E. Martínez, L. E. Garreta, N. Díaz, S. Amador, I. Tischer, J. M. Gutiérrez, A. K. Naik, F. Tobar, and F. García, "The human genome: a multifractal analysis.," *BMC Genomics*, vol. 12, no. 1, p. 506, Jan. 2011.

[139] C. Atupelage, H. Nagahashi, M. Yamaguchi, M. Sakamoto, and A. Hashiguchi, "Multifractal feature descriptor for histopathology.," *Anal. Cell. Pathol. (Amst).*, vol. 35, no. 2, pp. 123–6, Jan. 2012.

[140] P. S. Addison, *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*. Taylor & Francis, 2002.

[141] B. Enescu, K. Ito, and Z. R. Struzik, "Wavelet-Based Multifractal Analysis of real and simulated time series of earthquakes," *Annu. Disas. Prev. Res. Inst., Kyoto Univ*, no. 47, 2004.

[142] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *Inf. Theory, IEEE Trans.*, vol. 38, no. 2, pp. 617–643, 1992.

[143] J. F. Muzy, E. Bacry, and A. Arneodo, "The multifractal formalism revisited with wavelets," *Int. J. Bifurcat. Chaos*, vol. 4, no. 2, pp. 245–302, 1994.

[144] A. Arneodo, E. Bacry, and J. F. Muzy, "The thermodynamics of fractals revisited with wavelets," *Phys. A Stat. Mech. its Appl.*, vol. 213, no. 1, pp. 232–275, 1995.

[145] S. Sanei and J. Chambers, *EEG signal processing*. Wiley-Interscience, 2007.

[146] M. Baranger and N. E. C. S. Institute, *Chaos, Complexity, and Entropy: A Physics Talk for Non-physicists*. New England Complex Systems Institute.

[147] Z. J. Kowalik, “A PRACTICAL METHOD FOR THE MEASlJREMENTS OF THE CHAOTICITY OF ELECTRIC AND MAGNETIC BRAIN ACTIVITY,” *Int. J. Bifurc. Chaos*, vol. 5, no. 2, pp. 475–490, 1995.

[148] M. Mashayekhi and R. Gras, “Investigating the Effect of Spatial Distribution and Spatiotemporal Information on Speciation using Individual-Based Ecosystem Simulation,” *J. Comput.*, vol. 2, no. 1, p. in press, 2012.

[149] A. Golestani, M. R. Jahed Motlagh, K. Ahmadian, A. H. Omidvarnia, and N. Mozayani, “A new criterion to distinguish stochastic and deterministic time series with the Poincar{é} section and fractal dimension,” *Chaos An Interdiscip. J. Nonlinear Sci.*, vol. 19, no. 1, p. 13137, 2009.

[150] A. Golestani, A. Ashouri, K. Ahmadian, M. Jahed-Motlagh, and M. Doostari, “Irregularity Analysis of Iris patterns,” *IPCV08*, 2008.

[151] N. B. Tufillaro, “Poincare Map,” 1997.

[152] S. H. Strogatz, *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*. Westview Pr, 1994.

[153] A. A. Tsonis, *Chaos: from theory to applications*. Plenum Press New York, 1992.

[154] F. Takens and others, “Dynamical systems and turbulence,” *Lect. notes Math.*, vol. 898, no. 9, p. 366, 1981.

[155] A. H. Omidvarnia and A. M. Nasrabadi, “A new irregularity criterion for discrimination of stochastic and deterministic time series,” *FRACTALS-LONDON-*, vol. 16, no. 2, p. 129, 2008.

[156] P. Grassberger and I. Procaccia, “Measuring the strangeness of strange attractors,” *Phys. D Nonlinear Phenom.*, vol. 9, no. 1, pp. 189–208, 1983.

[157] M. T. Rosenstein, J. J. Collins, and C. J. De Luca, “A practical method for calculating largest Lyapunov exponents from small data sets,” *Phys. D Nonlinear Phenom.*, vol. 65, no. 1, pp. 117–134, 1993.

[158] U. Parlitz, “Nonlinear Time-Series Analysis,” *Nonlinear Model. Black-Box Tech.*, pp. 209–239, 1998.

[159] S. Sato, M. Sano, Y. Sawada, and others, “Practical methods of measuring the generalized dimension and the largest Lyapunov exponent in high dimensional chaotic systems,” *Prog. Theor. Phys*, vol. 77, no. 1, pp. 1–5, 1987.

[160] J. Kurths and H. Herzel, “An attractor in a solar time series,” *Phys. D Nonlinear Phenom.*, vol. 25, no. 1–3, pp. 165–172, 1987.

[161] M. Dämmig and F. Mitschke, “Estimation of Lyapunov exponents from time series: the stochastic case,” *Phys. Lett. A*, vol. 178, no. 5–6, pp. 385–394, 1993.

[162] A. Ray and A. Roy Chowdhury, "On the characterization of non-stationary chaotic systems: Autonomous and non-autonomous cases," *Phys. A Stat. Mech. its Appl.*, vol. 389, no. 21, pp. 5077–5083, Nov. 2010.

[163] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. Doyne Farmer, "Testing for nonlinearity in time series: the method of surrogate data," *Phys. D Nonlinear Phenom.*, vol. 58, no. 1, pp. 77–94, 1992.

[164] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: forecasting and control*. Wiley. com, 2013.

[165] D. C. Montgomery, L. A. Johnson, and J. S. Gardiner, *Forecasting and Time Series Analysis*. Mcgraw-Hill (Tx), 1990, p. 381.

[166] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control (Google eBook)*. John Wiley & Sons, 2013, p. 746.

[167] J. M. Gottman, *Time-Series Analysis: A Comprehensive Introduction for Social Scientists*. Cambridge University Press, 1982, p. 416.

[168] *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications (Google eBook)*. John Wiley & Sons, 2006, p. 514.

[169] K. J. Arrow, R. Forsythe, M. Gorham, R. Hahn, R. Hanson, J. O. Ledyard, S. Levmore, R. Litan, P. Milgrom, F. D. Nelson, G. R. Neumann, M. Ottaviani, T. C. Schelling, R. J. Shiller, V. L. Smith, E. Snowberg, C. R. Sunstein, P. C. Tetlock, P. E. Tetlock, H. R. Varian, J. Wolfers, and E. Zitzewitz, "Economics. The promise of prediction markets.," *Science*, vol. 320, no. 5878, pp. 877–8, May 2008.

[170] K. Lehnertz and C. Elger, "Can Epileptic Seizures be Predicted? Evidence from Nonlinear Time Series Analysis of Brain Electrical Activity," *Phys. Rev. Lett.*, vol. 80, no. 22, pp. 5019–5022, Jun. 1998.

[171] J. Martinerie, C. Adam, M. Le Van Quyen, M. Baulac, S. Clemenceau, B. Renault, and F. J. Varela, "Epileptic seizures can be anticipated by non-linear analysis.," *Nat. Med.*, vol. 4, no. 10, pp. 1173–6, Oct. 1998.

[172] J. Jeong, "Nonlinear dynamics of EEG in Alzheimer's disease," *Drug Dev. Res.*, vol. 56, no. 2, pp. 57–66, Jun. 2002.

[173] J. Dauwels, F. Vialatte, C. Latchoumane, J. Jeong, and A. Cichocki, "EEG synchrony analysis for early diagnosis of Alzheimer's disease: a study with several synchrony measures and EEG data sets.," *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, vol. 2009, pp. 2224–7, Jan. 2009.

[174] *Nonlinear Dynamics and Predictability of Geophysical Phenomena (Geophysical Monograph Series)*. American Geophysical Union, 1994, p. 107.

[175] P. A. Stott and J. A. Kettleborough, “Origins and estimates of uncertainty in predictions of twenty-first century temperature rise.,” *Nature*, vol. 416, no. 6882, pp. 723–6, Apr. 2002.

[176] G. Sugihara, R. May, H. Ye, C. Hsieh, E. Deyle, M. Fogarty, and S. Munch, “Detecting causality in complex ecosystems.,” *Science*, vol. 338, no. 6106, pp. 496–500, Oct. 2012.

[177] S. A. Levin, “Ecosystems and the Biosphere as Complex Adaptive Systems,” *Ecosystems*, vol. 1, no. 5, pp. 431–436, Sep. 1998.

[178] F. Neri, “Learning and predicting financial time series by combining natural computation and agent simulation,” pp. 111–119, Apr. 2011.

[179] J. S. Zirilli, *Financial Prediction Using Neural Networks*. Intl Thomson Computer Pr (Sd), 1996, p. 135.

[180] T. Bollerslev, “Generalized autoregressive conditional heteroskedasticity,” *J. Econom.*, vol. 31, no. 3, pp. 307–327, 1986.

[181] M. P. Clements, P. H. Franses, and N. R. Swanson, “Forecasting economic and financial time-series with non-linear models,” *Dep. Work. Pap.*, Oct. 2003.

[182] G. P. Zhang and D. M. Kline, “Quarterly Time-Series Forecasting With Neural Networks,” *IEEE Trans. Neural Networks*, vol. 18, no. 6, pp. 1800–1814, Nov. 2007.

[183] M. Zanin, “Forbidden patterns in financial time series.,” *Chaos*, vol. 18, no. 1, p. 013119, Mar. 2008.

[184] R. Hyndman, J. K. Ord, J. G. De Gooijer, and R. J. Hyndman, “25 years of time series forecasting,” *Int. J. Forecast.*, vol. 22, no. 3, pp. 443–473, 2006.

[185] C. C. Holt, “Forecasting seasonals and trends by exponentially weighted moving averages,” *Int. J. Forecast.*, vol. 20, no. 1, pp. 5–10, Jan. 2004.

[186] R. G. Brown, *Statistical Forecasting for Inventory Control*. McGraw-Hill, 1959, p. 232.

[187] R. G. Brown, *Smoothing, Forecasting and Prediction of Discrete Time Series*. Courier Dover Publications, 2004, p. 468.

[188] R. F. Engle, “Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation,” *Econometrica*, vol. 50, no. 4, pp. 987–1007, 1982.

[189] S. J. Taylor, “Forecasting the volatility of currency exchange rates,” *Int. J. Forecast.*, vol. 3, no. 1, pp. 159–170, Jan. 1987.

[190] H. Tong, *Threshold Models in Non-linear Time Series Analysis*. Springer; Softcover reprint of the original 1st ed. 1983 edition, 1983, p. 336.

[191] H. Tong, *Non-linear Time Series: A Dynamical System Approach*. Clarendon Press, 1993, p. 564.

[192] M. P. Clements and J. (Jeremy P. . Smith, "The performance of alternative forecasting methods for SETAR models." University of Warwick, Department of Economics, 01-Jun-1996.

[193] J. G. De Gooijer and K. Kumar, "Some recent developments in non-linear time series modelling, testing, and forecasting," *Int. J. Forecast.*, vol. 8, no. 2, pp. 135–156, Oct. 1992.

[194] J. A. Coyne and H. A. Orr, "Speciation. Sunderland, MA." Sinauer Associates, Inc, 2004.

[195] A. S. Kondrashov and F. A. Kondrashov, "Interactions among quantitative traits in the course of sympatric speciation," *Nature*, vol. 400, no. 6742, pp. 351–354, 1999.

[196] D. I. Bolnick and B. M. Fitzpatrick, "Sympatric speciation: models and empirical evidence," *Annu. Rev. Ecol. Evol. Syst.*, vol. 38, pp. 459–487, 2007.

[197] B. M. Fitzpatrick, J. A. Fordyce, and S. Gavrilets, "What, if anything, is sympatric speciation?," *J. Evol. Biol.*, vol. 21, no. 6, pp. 1452–1459, 2008.

[198] J. C. Avise, J. Arnold, R. M. Ball, E. Bermingham, T. Lamb, J. E. Neigel, C. A. Reeb, and N. C. Saunders, "Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics," *Annu. Rev. Ecol. Syst.*, vol. 18, pp. 489–522, 1987.

[199] S. K. Lyons, "A quantitative assessment of the range shifts of Pleistocene mammals," *J. Mammal.*, vol. 84, no. 2, pp. 385–402, 2003.

[200] N. J. Savill, P. Rohandi, and P. Hogeweg, "Self-reinforcing spatial patterns enslave evolution in a host-parasitoid system," *J. theor. Bio.*, vol. 188, pp. 11–20, 1997.

[201] V. I. Krinsky and K. I. Agladze, "Interaction of rotating waves in an active chemical medium," *Phys. D Nonlinear Phenom.*, vol. 8, no. 1, pp. 50–56, 1983.

[202] J. Bascompte, R. V Solé, and N. Mart\i□nez, "Population cycles and spatial patterns in snowshoe hares: an individual-oriented simulation," *J. Theor. Biol.*, vol. 187, no. 2, pp. 213–222, 1997.

[203] C. J. Krebs, M. S. Gaines, B. L. Keller, J. H. Myers, and R. H. Tamarin, "Population Cycles in Small Rodents Demographic and genetic events are closely coupled in fluctuating populations of field mice," *Science (80-.).*, vol. 179, no. 4068, pp. 35–41, 1973.

[204] M. Turelli, N. H. Barton, and J. A. Coyne, "Theory and speciation.," *Trends Ecol. Evol.*, vol. 16, no. 7, pp. 330–343, Jul. 2001.

[205] M. Kirkpatrick and V. Ravigné, "Speciation by natural and sexual selection: models and experiments.," *Am. Nat.*, vol. 159 Suppl , pp. S22–35, Mar. 2002.

[206] X. Thibert-Plante and A. P. Hendry, "Five questions on ecological speciation addressed with individual-based simulations.," *J. Evol. Biol.*, vol. 22, no. 1, pp. 109–23, Jan. 2009.

[207] G. S. van Doorn, P. Edelaar, and F. J. Weissing, "On the origin of species by natural and sexual selection.," *Science*, vol. 326, no. 5960, pp. 1704–7, Dec. 2009.

[208] P. Nosil and S. M. Flaxman, "Conditions for mutation-order speciation.," *Proc. Biol. Sci.*, vol. 278, no. 1704, pp. 399–407, Feb. 2011.

[209] R. Lande, "Models of speciation by sexual selection on polygenic traits," *Proc. Natl. Acad. Sci.*, vol. 78, no. 6, pp. 3721–3725, Jun. 1981.

[210] S. Gavrilets, "Rapid evolution of reproductive barriers driven by sexual conflict.," *Nature*, vol. 403, no. 6772, pp. 886–9, Feb. 2000.

[211] L. Parrott, R. Proulx, and X. Thibert-Plante, "Three-dimensional metrics for the analysis of spatiotemporal data in ecology," *Ecol. Inform.*, vol. 3, no. 6, pp. 343–353, Dec. 2008.

[212] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003, p. 640.

[213] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[214] D. Schluter, "Ecological character displacement in adaptive radiation," *Am. Nat.*, vol. 156, no. S4, pp. S4–S16, 2000.

[215] P. Nosil and J. L. Feder, "Genomic divergence during speciation: causes and consequences," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 367, no. 1587, pp. 332–342, 2012.

[216] U. Dieckmann, *Adaptive Speciation*. Cambridge University Press, 2004, p. 460.

[217] T. Price, *Speciation in birds*. Roberts and Co., 2008, p. 470.

[218] R. A. Duckworth and L. E. B. Kruuk, "Evolution of genetic integration between dispersal and colonization ability in a bird," *Evolution (N. Y).*, vol. 63, no. 4, pp. 968–977, 2009.

[219] B. McGill, R. Etienne, J. Gray, D. Alonso, M. Anderson, H. Benecha, M. Dornelas, B. Enquist, J. Green, F. He, A. Hurlbert, A. Magurran, P. Marquet, B. Maurer, A. Ostling, C. Soykan, K. Ugland, and E. White, *Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework*, vol. 10. 2007, pp. 995 – 1015.

[220] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004, p. 376.

[221] M. J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data (Springer Series in Statistics)*. Springer, 2011, p. 698.

[222] A. Golestani and R. Gras, “A new species abundance distribution model based on model combination.,” *Int. J. Biostat.*, vol. 9, no. 1, Jan. 2013.

[223] A. E. MAGURRAN, “Species abundance distributions: pattern or process?,” *Funct. Ecol.*, vol. 19, no. 1, pp. 177–181, Feb. 2005.

[224] H. Irie and K. Tokita, “Species-area relationship for power-law species abundance distribution,” *Int. J. Biomath.*, vol. 5, no. 03, 2012.

[225] J. Harte, “Self-Similarity in the Distribution and Abundance of Species,” *Science (80-.).*, vol. 284, no. 5412, pp. 334–336, Apr. 1999.

[226] L. Borda-de-Água, S. P. Hubbell, and M. McAllister, “Species-area curves, diversity indices, and species abundance distributions: a multifractal analysis,” *Am. Nat.*, vol. 159, no. 2, pp. 138–155, 2002.

[227] A. K. Dewdney, “A dynamical model of communities and a new species-abundance distribution.,” *Biol. Bull.*, vol. 198, no. 1, pp. 152–65, Feb. 2000.

[228] G. Bell, “Neutral macroecology.,” *Science*, vol. 293, no. 5539, pp. 2413–8, Sep. 2001.

[229] B. J. McGill, B. A. Maurer, and M. D. Weiser, “Empirical evaluation of neutral theory.,” *Ecology*, vol. 87, no. 6, pp. 1411–23, Jun. 2006.

[230] M. G. Bulmer, “On fitting the Poisson lognormal distribution to species-abundance data,” *Biometrics*, pp. 101–110, 1974.

[231] R. MacArthur, “On the relative abundance of species,” *Am. Nat.*, pp. 25–36, 1960.

[232] G. Sugihara, “Minimal Community Structure: An Explanation of Species Abundance Patterns,” *Am. Nat.*, vol. 116, no. 6, p. 770, Dec. 1980.

[233] J. Chave, “Neutral theory and community ecology,” *Ecol. Lett.*, vol. 7, no. 3, pp. 241–253, Feb. 2004.

[234] J. M. Potts and J. Elith, “Comparing species abundance models,” *Ecol. Modell.*, vol. 199, no. 2, pp. 153–163, Nov. 2006.

[235] R. A. Kempton and L. R. Taylor, “Log-series and log-normal parameters as diversity discriminants for the Lepidoptera,” *J. Anim. Ecol.*, pp. 381–399, 1974.

[236] M. Dornelas and S. R. Connolly, “Multiple modes in a coral species abundance distribution,” *Ecol. Lett.*, vol. 11, no. 10, pp. 1008–1016, 2008.

[237] J. Periaux, G. Winter, and others, “Genetic algorithms in engineering and computer science,” 1995.

[238] J. C. Spall, *Introduction to Stochastic Search and Optimization*. Wiley-Interscience, 2003, p. 618.

[239] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach (3rd Edition)*. Prentice Hall, 2009, p. 1152.

[240] R. Izsák, "Maximum likelihood fitting of the Poisson lognormal distribution," *Environ. Ecol. Stat.*, vol. 15, no. 2, pp. 143–156, Oct. 2007.

[241] R. Condit, S. Aguilar, A. Hernandez, R. Perez, S. Lao, G. Angehr, S. P. Hubbell, and R. B. Foster, "Tropical forest dynamics across a rainfall gradient and the impact of an El Nio dry season," *J. Trop. Ecol.*, vol. 20, no. 1, pp. 51–72, Jan. 2004.

[242] P. W. Frank, "Patterns in the Balance of Nature. And related problems in quantitative ecology. C. B. Williams. Academic Press, New York, 1964. viii + 324 pp. Illus. $9.50," *Science (80-.).*, vol. 144, no. 3625, pp. 1439–1440, Jun. 1964.

[243] H. Bell, "A bird community of lowland rainforest in New Guinea. I. Composition and Density of the Avifauna," *Emu*, vol. 82, no. 1, p. 24, 1982.

[244] J. M. Thiollay, "Structure compar{é}e du peuplement avien dans trois sites de for{ê}t primaire en Guyane," *Rev. d'{é}cologie*, vol. 41, no. 1, pp. 59–105, 1986.

[245] R. Sukumar, H. S. Suresh, H. S. Dattaraja, R. John, and N. V Joshi, "Mudumalai forest dynamics plot, India," *Trop. For. Divers. dynamism Find. from a large-scale plot Netw.*, pp. 551–563, 2004.

[246] L. Kish, *Statistical Design for Research (Wiley Classics Library)*. Wiley, 2004, p. 267.

[247] S. L. Pimm, M. Ayres, A. Balmford, G. Branch, K. Brandon, T. Brooks, R. Bustamante, R. Costanza, R. Cowling, L. M. Curran, A. Dobson, S. Farber, G. A. da Fonseca, C. Gascon, R. Kitching, J. McNeely, T. Lovejoy, R. A. Mittermeier, N. Myers, J. A. Patz, B. Raffle, D. Rapport, P. Raven, C. Roberts, J. P. Rodriguez, A. B. Rylands, C. Tucker, C. Safina, C. Samper, M. L. Stiassny, J. Supriatna, D. H. Wall, and D. Wilcove, "Environment. Can we defy nature's end?," *Science*, vol. 293, no. 5538, pp. 2207–8, Sep. 2001.

[248] C. M. Roberts, C. J. McClean, J. E. N. Veron, J. P. Hawkins, G. R. Allen, D. E. McAllister, C. G. Mittermeier, F. W. Schueler, M. Spalding, F. Wells, C. Vynne, and T. B. Werner, "Marine biodiversity hotspots and conservation priorities for tropical reefs.," *Science*, vol. 295, no. 5558, pp. 1280–4, Mar. 2002.

[249] M. Austin, "Species distribution models and ecological theory: A critical assessment and some possible new approaches," *Ecol. Modell.*, vol. 200, no. 1–2, pp. 1–19, Jan. 2007.

[250] A. Guisan and N. E. Zimmermann, "Predictive habitat distribution models in ecology," *Ecol. Modell.*, vol. 135, no. 2–3, pp. 147–186, Dec. 2000.

[251] M. . Austin, "Spatial prediction of species distribution: an interface between ecological theory and statistical modelling," *Ecol. Modell.*, vol. 157, no. 2–3, pp. 101–118, Nov. 2002.

[252] D. J. Currie, G. G. Mittelbach, H. V. Cornell, R. Field, J.-F. Guegan, B. A. Hawkins, D. M. Kaufman, J. T. Kerr, T. Oberdorff, E. O'Brien, and J. R. G. Turner, "Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness," *Ecol. Lett.*, vol. 7, no. 12, pp. 1121–1134, Dec. 2004.

[253] C. Rahbek, N. J. Gotelli, R. K. Colwell, G. L. Entsminger, T. F. L. V. B. Rangel, and G. R. Graves, "Predicting continental-scale patterns of bird species richness with spatially explicit models.," *Proc. Biol. Sci.*, vol. 274, no. 1607, pp. 165–74, Jan. 2007.

[254] V. Gewin, "Ecosystem health: the state of the planet.," *Nature*, vol. 417, no. 6885, pp. 112–3, May 2002.

[255] R. L. Pressey, T. C. Hager, K. M. Ryan, J. Schwarz, S. Wall, S. Ferrier, and P. M. Creaser, "Using abiotic data for conservation assessments over extensive regions: quantitative methods applied across New South Wales, Australia," *Biol. Conserv.*, vol. 96, no. 1, pp. 55–82, Nov. 2000.

[256] G. C. Reese, K. R. Wilson, J. A. Hoeting, and C. H. Flather, "Factors Affecting Species Distribution Predictions: A Simulation Modeling Experiment," *Ecol. Appl.*, vol. 15, no. 2, pp. 554–564, Apr. 2005.

[257] N. J. Gotelli, M. J. Anderson, H. T. Arita, A. Chao, R. K. Colwell, S. R. Connolly, D. J. Currie, R. R. Dunn, G. R. Graves, J. L. Green, J.-A. Grytnes, Y.-H. Jiang, W. Jetz, S. Kathleen Lyons, C. M. McCain, A. E. Magurran, C. Rahbek, T. F. L. V. B. Rangel, J. Soberón, C. O. Webb, and M. R. Willig, "Patterns and causes of species richness: a general simulation model for macroecology.," *Ecol. Lett.*, vol. 12, no. 9, pp. 873–86, Sep. 2009.

[258] S. J. Pittman, J. D. Christensen, C. Caldow, C. Menza, and M. E. Monaco, "Predictive mapping of fish species richness across shallow-water seascapes in the Caribbean," *Ecol. Modell.*, vol. 204, no. 1–2, pp. 9–21, May 2007.

[259] W. B. Sherwin, "Entropy and information approaches to genetic diversity and its expression: Genomic geography," *Entropy*, vol. 12, no. 7, pp. 1765–1798, 2010.

[260] J. R. Quinlan, *C4. 5: programs for machine learning*, vol. 1. Morgan kaufmann, 1993.

[261] L. Rokach, *Data Mining with Decision Trees: Theory and Applications*. World Scientific, 2008, p. 244.

[262] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001, p. 533.

[263] C. Strobl, J. Malley, and G. Tutz, "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests.," *Psychol. Methods*, vol. 14, no. 4, pp. 323–48, Dec. 2009.

[264] W. D. Kissling, C. Rahbek, and K. Böhning-Gaese, “Food plant diversity as broad-scale determinant of avian frugivore richness,” *Proc. R. Soc. B Biol. Sci.*, vol. 274, no. 1611, pp. 799–808, 2007.

[265] D. Oro, E. Cam, R. Pradel, and A. Martínez-Abraín, “Influence of food availability on demography and local population dynamics in a long-lived seabird,” *Proc. R. Soc. London. Ser. B Biol. Sci.*, vol. 271, no. 1537, pp. 387–396, 2004.

[266] C. Devaux and R. Lande, “Incipient allochronic speciation due to non-selective assortative mating by flowering time, mutation and genetic drift,” *Proc. R. Soc. B Biol. Sci.*, vol. 275, no. 1652, pp. 2723–2732, 2008.

[267] K. L. Evans, J. J. D. Greenwood, and K. J. Gaston, “Dissecting the species--energy relationship,” *Proc. R. Soc. B Biol. Sci.*, vol. 272, no. 1577, pp. 2155–2163, 2005.

[268] D. J. Currie, “Energy and large-scale patterns of animal-and plant-species richness,” *Am. Nat.*, pp. 27–49, 1991.

[269] K. Roy, D. Jablonski, J. W. Valentine, and G. Rosenberg, “Marine latitudinal diversity gradients: tests of causal hypotheses,” *Proc. Natl. Acad. Sci.*, vol. 95, no. 7, pp. 3699–3702, 1998.

[270] J. A. Crame, “Taxonomic diversity gradients through geological time,” *Divers. Distrib.*, vol. 7, no. 4, pp. 175–189, 2001.

[271] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, vol. 7, no. 3. Cambridge University Press, 1997, p. 304.

[272] C. J. Stam, “Nonlinear dynamical analysis of EEG and MEG: review of an emerging field.,” *Clin. Neurophysiol.*, vol. 116, no. 10, pp. 2266–2301, 2005.

[273] A. H. Meghdadi, W. Kinsner, and R. Fazel-Rezai, “Characterization of healthy and epileptic brain EEG signals by monofractal and multifractal analysis,” *2008 Can. Conf. Electr. Comput. Eng.*, pp. 001407–001412, May 2008.

[274] A. Bershadskii, “Some universal properties of multifractal chaos at nuclear giant resonance,” *Phys. Rev. C*, vol. 59, pp. 3469–3472, 1999.

[275] R. V Sole and S. C. Manrubia, “Self-similarity in rain forests: Evidence for a critical state,” *Phys. Rev. E-Statistical Phys. Plasma Fluids Relat. Interdiscpl Top.*, vol. 51, no. 6, pp. 6250–6253, 1995.

[276] B. B. Mandelbrot, “Self-affine fractal sets, I: the basic fractal dimensions,” in *Fractals in physics*, 1986, vol. 1, p. 3.

[277] M. G. Turner and R. H. Gardner, *Quantitative methods in landscape ecology: the analysis and interpretation of landscape heterogeneity*, vol. 82. Springer Verlag, 1991.

[278] B. T. Milne, “Applications of fractal geometry in wildlife biology,” in *Wildlife and Landscape Ecology: Effects of Pattern on Scale*, J. A. Bissonette, Ed. New York: Springer and Verlag, 1997, pp. 32–69.

[279] L. Seuront, *Fractals and multifractals in ecology and aquatic science*. CRC Press, 2010.

[280] I. Scheuring and R. H. Riedi, “Application of multifractals to the analysis of vegetation pattern,” *J. Veg. Sci.*, vol. 5, no. 4, pp. 489–496, 1994.

[281] J. B. Drake and J. F. Weishampel, “Multifractal analysis of canopy height measures in a longleaf pine savanna,” *For. Ecol. Manage.*, vol. 128, no. 1–2, pp. 121–127, 2000.

[282] J. Ozik, B. R. Hunt, and E. Ott, “Formation of multifractal population patterns from reproductive growth and local resettlement,” *Phys. Rev. E*, vol. 72, no. 4, p. 46213, 2005.

[283] D. I. Iudin and D. B. Gelashvily, “Multifractality in ecological monitoring,” *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 502, no. 2, pp. 799–801, 2003.

[284] M. Pascual, F. A. Ascioti, and H. Caswell, “Intermittency in the plankton: a multifractal analysis of zooplankton biomass variability,” *J. Plankton Res.*, vol. 17, no. 6, p. 1209, 1995.

[285] L. Seuront and N. Spilmont, “Self-organized criticality in intertidal microphytobenthos patch patterns,” *Phys. A Stat. Mech. its Appl.*, vol. 313, no. 3, pp. 513–539, 2002.

[286] S. G. Mallat, *A wavelet tour of signal processing*. Academic Pr, 1999.

[287] J. W. Crawford, K. Ritz, and I. M. Young, “Quantification of fungal morphology, gaseous transport and microbial dynamics in soil: an integrated framework utilising fractal geometry,” *Geoderma*, vol. 56, no. 1–4, pp. 157–172, 1993.

[288] R. W. Hahn and P. C. Tetlock, “Using Information Markets to Improve Public Decision Making,” *Work. Pap.*

[289] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Springer, 2007, p. 764.

[290] H. Meinardi, R. A. Scott, R. Reis, and J. W. Sander, “The treatment gap in epilepsy: the current situation and ways forward.,” *Epilepsia*, vol. 42, no. 1, pp. 136–49, Jan. 2001.

[291] R. S. Fisher, W. van Emde Boas, W. Blume, C. Elger, P. Genton, P. Lee, and J. Engel, “Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE).,” *Epilepsia*, vol. 46, no. 4, pp. 470–2, Apr. 2005.

[292] S. Wang, W. A. Chaovalitwongse, and S. Wong, “Online Seizure Prediction Using an Adaptive Learning Approach,” *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 12, pp. 2854–2866, Dec. 2013.

[293] Y. Park, L. Luo, K. K. Parhi, and T. Netoff, "Seizure prediction with spectral power of EEG using cost-sensitive support vector machines.," *Epilepsia*, vol. 52, no. 10, pp. 1761–70, Oct. 2011.

[294] H. Feldwisch-Drentrup, B. Schelter, M. Jachan, J. Nawrath, J. Timmer, and A. Schulze-Bonhage, "Joining the benefits: combining epileptic seizure prediction methods.," *Epilepsia*, vol. 51, no. 8, pp. 1598–606, Aug. 2010.

[295] B. Schelter, M. Winterhalder, T. Maiwald, A. Brandt, A. Schad, J. Timmer, and A. Schulze-Bonhage, "Do false predictions of seizures depend on the state of vigilance? A report from two seizure-prediction methods and proposed remedies.," *Epilepsia*, vol. 47, no. 12, pp. 2058–70, Dec. 2006.

[296] A. K. Tafreshi, A. M. Nasrabadi, and A. H. Omidvarnia, "Epileptic Seizure Detection Using Empirical Mode Decomposition," in *2008 IEEE International Symposium on Signal Processing and Information Technology*, 2008, pp. 238–242.

[297] H. Adeli, S. Ghosh-Dastidar, and N. Dadmehr, "A wavelet-chaos methodology for analysis of EEGs and EEG subbands to detect seizure and epilepsy.," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 2, pp. 205–11, Feb. 2007.

[298] L. Ye, G. Yang, E. Ranst, and H. Tang, "Time-series modeling and prediction of global monthly absolute temperature for environmental decision making," *Adv. Atmos. Sci.*, vol. 30, no. 2, pp. 382–396, Feb. 2013.

[299] M. T. Rosenstein, J. J. Collins, and C. J. De Luca, "A practical method for calculating largest Lyapunov exponents from small data sets," *Phys. D Nonlinear Phenom.*, vol. 65, no. 1–2, pp. 117–134, 1993.

[300] H. Y. YAMIN and S. M. SHAHIDEHPOUR, "Bidding Strategies Using Price Based Unit Commitment in a Deregulated Power Market," *Electr. Power Components Syst.*, vol. 32, no. 3, pp. 229–245, Mar. 2004.

[301] S. I. Seneviratne, M. G. Donat, B. Mueller, and L. V. Alexander, "No pause in the increase of hot temperature extremes," *Nat. Clim. Chang.*, vol. 4, no. 3, pp. 161–163, Feb. 2014.

[302] D. M. Smith, S. Cusack, A. W. Colman, C. K. Folland, G. R. Harris, and J. M. Murphy, "Improved surface temperature prediction for the coming decade from a global climate model.," *Science*, vol. 317, no. 5839, pp. 796–9, Aug. 2007.

VITA AUCTORIS

Abbas Golestani received his Bachelor's and Master's Degrees from Shahed University and Iran University of Science and Technology, Iran, in 2005, and 2008, respectively. He studies in the School of Computer Science, University of Windsor, Canada, from 2009 to 2014 for a Degree of Doctoral Philosophy. His research interests include Machine learning, Medical time series analysis, Multi-agent systems, Evolutionary algorithms, Time series forecasting and Chaos theory.