# Theory-driven analysis of ecological data

Summary of the course

## 1 Introduction to theoretical models and link to data (Day 1)

### 1.1 Why mathematical models?

In this introduction, we first discuss if we still need theory at the age of big data. We define what is theory and what is a model in a general way, and more specifically mathematical models as used in ecology. We develop on the necessary trade-offs involved in the construction of a model, in a relation of its specific goals in science: understanding or predicting. We then discuss to which extent we can reach these goal by instead using artificial intelligence informed by big data. Then we come back to the principles of the scientific method and elaborate on how mathematical models participate in this cycle of scientific knowledge production involving data and theory. We finally advocate for the necessity of deeper interactions between data and theory despite some identified current weakness in this link.

### 1.2 What types of theoretical models in ecology?

In this second part of the first day, we first consider that a model is necessarily defined in relation to a given question. The biological system and the associated research question jointly define the scale of interest. Depending on our question and the scientific goal underlying this question (understand vs predict), we can use a range of different model types which features vary along two axes: from simple to complex and from phenomenological to process-based. We give the main reasons why focussing on process-based simple models and stressed out that the different types of assumptions we make while building models constrain the interpretation of model predictions. In a second time we present the different model formalisms regarding stochasticity (deterministic vs stochastic models), time (static models vs dynamical models, with either discrete or continuous time) and space (non-spatial vs spatially implicit vs explicit models). We discuss the situations in which some formalisms are more appropriate than others. Finally we present main technical choices to make: using either agent-based modelling or dynamical equations, highlighting their pros and cons. Furthermore, we discuss tractability of dynamical models in relation to their complexity.

### 1.3 How to build a model?

In this third part, we will first try to apply what we've just discuss on a spcific system and question in an interactive way, that is to define the variables,the processes to be included in the model, the formalism we can choose. We then will identify together the main assumptions underlying a classical simple model (Rosenzweig & McArthur revisited) from the mathematical formulation of the processes. Finally, we will show the principle of numerical integration to simulate a system of differential equations.

### 1.4 How to analyze a model?

In the fourth part, we see how we classically analyse a mathematical model (ODEs): search for equilibria, from analytical expression to graphical representations; run a local stability analysis of this equilibria to study the behaviour of the model; look for dependence to initial conditions. In a second time we see the different ways to use a theoretical model to answer different types of research questions (parameter exploration, change the modelstructure, in sillico experiments), and we finally present how to assess the robustness of conclusions (sensitivity analysis).

### 1.5 Making a link between models and data

On the last part of the day, we discuss how one can make a connection between models (of any type) and data. This is an overview of a broad field in statistics. We first discuss how to make a qualitative comparison of model predictions and data, and then consider the ways to make the comparison quantitative. First, finding the best parameter values (model fitting) with least squares, maximum likelihood, or Bayesian approaches. Second, quantifying the uncertainty in parameter estimates (intervals), and the goodness of fit. Finally,comparing different models and selecting the 'best' (model selection), with the possibility of using alternative models rather than just one (multimodel inference).

## 2 Temporal data (Day 2)

### 2.1 Introduction

In ecology and evolution we may encounter data collected over time, such as population density time-series. Faced with this kind of data, we have fundamentally two options for analysis: First, a purely statistical approch with methods such as ARIMA. Second, we could use the large body of theory from population and community ecology that allows us to formulate relationships between densities and time from theoretically established principles.

Here, we will focus on the latter and an example of such an approach can be found, for example in Sibly et al. 2005 who study the shape of density-dendependence confronting a theta-logistic model with time-series data from the Global Population Density Database.

### 2.2 Building a population dynamics model

As we have seen in the first day, a population dynamics model fundamentally is a bookkeeping exercise that foamlizes inputs (e.g., births) and outputs (e.g., deaths). We will see this using the logistic model in its r-K- (Pianka) and in its r-alpha (Verhulst) formulation (see Mallet 2012). We will dive more deeply into the logistc by trying to derive it assuming either competition for food resources or for space.

### 2.3 Confronting model with data

After this exercise of model building, we will learn how to confront such models with data focusing on the statistical interface using a Bayesian approach (stan, rstan). Besides diving into code we will explore error types (observation error, process error) and data structure complexity.

The afternoon will be dedicated to a hands-on exercise for fitting time-series data to different dynamics models. Details can be found e.g. in Rosenbaum & Fronhofer (2023).

### 2.4 Two papers for student project on Friday

- Yoshida et al. 2003
- Gause 1934

### 2.5 References

@Article{Gause1934, author = {Gause, G. F.}, title = {Experimental analysis of Vito Volterra's mathematical theory of the struggle for existence}, journal = {Science}, year = {1934}, volume = {79}, pages = {16–17}, }

@Article{Mallet2012, author = {Mallet, J.}, title = {The struggle for existence: how the notion of carrying capacity, {K}, obscures the links between demography, {D}arwinian evolution, and speciation}, journal = {Evol. Ecol. Res.}, year = {2012}, volume = {14}, number = {5}, pages = {627–665}, }

@Article{Rosenbaum2023, author = {Benjamin Rosenbaum and Emanuel A. Fronhofer}, journal = {Ecosphere}, title = {Confronting population models with experimental microcosm data: from trajectory matching to state-space models}, year = {2023}, number = {4}, pages = {e4503}, volume = {14}, data_doi = {https://doi.org/10.5281/zenodo.7702324}, doi = {10.1002/ecs2.4503}, }

@Article{Sibly2005, author = {Richard M. Sibly and Daniel Barker and Michael C. Denham and Jim Hone and Mark Pagel}, title = {On the Regulation of Populations of Mammals, Birds, Fish, and Insects}, journal = {Science}, year = {2005}, volume = {309}, number = {5734}, pages = {607–610}, month = {Jul}, issn = {1095-9203}, doi = {10.1126/science.1110760}, url = {http://dx.doi.org/10.1126/science.1110760}, }

@Article{Yoshida2003, author = {Yoshida, Takehito and Jones, Laura E. and Ellner, Stephen P. and Fussmann, Gregor F. and Hairston, Nelson G.}, title = {Rapid evolution drives ecological dynamics in a predator–prey system}, journal = {Nature}, year = {2003}, volume = {424}, number = {6946}, pages = {303–306}, issn = {0028-0836}, doi = {10.1038/nature01767}, url = {http://dx.doi.org/10.1038/nature01767}, }

# 3 Spatial data (Day 3)

## 3.1 Objectives and introduction

Compared to Day 2, we will consider a slighly larger scale, considering communities of several species, distributed over a large spatial extent. The typical spatial dataset would consist in a community matrix (sites*species), with either presence/absence data, detection/non detection data (if there is imperfect detection), or abundance data (density or number of individuals). Possibly, one may have presence only data, which causes specific issues. Ideally, this community matrix is observed repeatedly at several points in time (time series), which significantly expands the range of methods and models we can mobilize, comapred to static (snapshot) datasets. This basic data set is often complemented with other relavent information describing species (traits, phylogenetic relationships…) and/or locations (coordinates, environental descriptors…). The typical questions we want to address are: where are species distributed? What is the relative importance of environmental attributes versus dispersal in shaping communities? Can we infer interactions (competition, facilitation…) between species?

We briefly present three general frameworks used in this context: -1- The habitat filtering approach; -2- The metacommunity framework; -3- Vellend's community ecology theory framework, and discuss their strengths/weaknesses. We refer to this cool website presenting reflections from Vellend and Leibold on the fundamental publications which introduced frameworks (2) and (3).

We then go over a repertoire of methods and models that can be used to describe and analyze such datasets. We make a distinction between (1) data-driven approaches and (2) process-driven approaches.

## 3.2 Data-driven approaches

We first consider community-level approaches, i.e. the ones that consider community composition/diversity as a patch metric Unconstrained methods do not directly use environmental variables: the associations with environmental descriptors is done a *posteriori* (e.g. PCA). Constrained methods simultaneously consider species composition and environmental descriptors (e.g. RDA). A short focus is made on Cottenie's approach of variance partitioning (partialRDA + PCNM) with examples from mollusk communities in the French Antilles. We then turn to species-level methods, starting with permutation approaches, the Cscore, with some connection here to networks and Day 4. We mention the fact that even without interactions, several processes may yield spurious species associations (e.g. Calcagno et al. 2022). After that, we go to Species Distribution Models, and finally Occupancy models, whose hallmark is to add a detection layer (probability to detect a species when it is present). These will be discussed more in the next section (mechanistic approaches).

## 3.3 Process-driven (mechanistic) approaches

We start by introducing patch-occupancy models, and the two simplest and historically most important models: Levins (1969) metapopulation model and McArthur & Wilson (1967) IB model. We highlight the mathematical connection between Levins' model and the logistic model that was studied on the previous day. We show how an intrinsic covariation of r and K shows up in this formulation. We walk-through the computation of equilibrium occupancies in the two models and their natural generalization. We discuss the IB model and its predictions, reflecting on why the curves are non linear. We go through the the example of fitting this model to island presence/absence data, inspired from data in Manne et al. (1998).

We discuss possible extensions of patch-occupancy models: -1- explicit space (IFM models from Hanski and followers); -2- competition/colonization trade-offs; -3- spatial networks; -4- trophic webs and others. A more profound extension is to include an explicit description of population dynamics within patches. Models quickly become complicated and need to be simulated. One exception is Hubbel's neutral model, which has a dynamics for species abundances in patches and remains reasonably tractable. An example of simulation framework for metacommunities is Thompson et al. (2020) Ecology Letters. This is the one we will use to simulate data in the afternoon practical.

### 3.4 Conclusion

We conclude by reviewing he different approaches possible and compare their advantages and drawbacks. We discuss how one approach is not universally better than the others, but rather a set of approaches can yield complementary insights, and different approaches are more papropriate depending on the underlying question and the type/amount of data available.

## 4 Interaction networks, food webs, and complexity (Day 4)

During this day on interaction networks, we will first (1) introduce the concepts and definitions associated with networks of interactions, and then tackle two general questions regarding interaction networks: (2) how to simplify complexity by considering random matrices to model interactions and (3) how to assess the link between complexity and dynamical stability in empirical data. As a conclusion to the thematic day, we will (4) delve into multiplex networks, i.e. networks combining interactions of different types.

The introductory course covers a variety of topics including networks and their definition, their representation as matrix and their interpretation in dynamical model. Notable theoretical results based on network properties are then reviewed, including results on stability and feasibility, the search for invariant properties structuring the topology of empirical networks, and the general framework of robustness studies, i.e. studies aimed at understanding how species removal affect the rest of the network through cascading extinctions. The introductory course is concluded by a primer on network statistics commonly used to characterize interaction networks, including degree distributions, the randomizations used to generate null predictions, and the models used to assess groups of species interacting in the same way or interacting in a nested fashion.

The simplification of complexity through large ecosystems and random matrices is then presented in a series of steps. The emerging simplicity of systems comprising many species is explained, with a particular focus on some fingerprint properties that could characterize the predictions of competing models. The use of random interaction matrix can be considered as a shortcut to obtaining such properties, as well as a path to analytical tractability of complex disordered systems.

The link between complexity and stability of empirical food webs is the focus of the third part of the day. By using simple relationships between interaction coefficients and biomasses, and then computing Jacobian matrices, which characterize system stability, the complexity-stability relationship of empirical food webs can be compared with that of their randomized counterparts, which in turn allows for a close examination of ecological hypotheses explaining food web stability.

The last presentation of the day introduces multiplex networks of interactions, i.e. networks in which several types of interactions are represented together. Following on the idea of keystone predators, as introduced by the work of R. Paine on intertidal communities, the modulation of trophic interactions and species parameters by non-trophic interactions is explored through dynamical models confronted with existing data on intertidal systems with both competition and resource consumption. The parallel topologies of trophic and non-trophic interaction networks are then gauged through block models and random network models in order to assess the effect of network structure on food web feasibility, stability and total biomass.