Structural Risk Minimization: a robust method for density-dependence detection and model selection =Ecography, 2007 (preprint)=

Giorgio Corani [a],Marino Gatto [a]

[a] *Dipartimento di Elettronica ed Informazione, Politecnico di Milano*
Via Ponzio 34/5
20133 Milano
e-mail:corani[gatto]@elet.polimi.it

Abstract

Statistically distinguishing density-dependent from density-independent populations and selecting the best demographic model for a given population are problems of primary importance. Traditional approaches are PBLR (parametric bootstrapping of likelihood ratios) and Information Criteria (IC), such as the Schwarz Information Criterion (SIC), the Akaike Information Criterion (AIC) or the Final Prediction Error (FPE). While PBLR is suitable for choosing from a couple of models, ICs select the best model from among a set of candidates. In this paper, we use the Structural Risk Minimization (SRM) approach. SRM is the model selection criterion developed within the Statistical Learning Theory (SLT), a theory of great generality for modelling and learning with finite samples. SRM is almost unknown in the ecological literature and has never been used to analyze time series.

First, we compare SRM with PBLR in terms of their ability to discriminate between the Malthusian and the density-dependent Ricker model. We rigorously repeat the experiments described in Dennis and Taper (1994) and find out that SRM is equally powerful in detecting density-independence and much more powerful in detecting density-dependence.

Then, we compare SRM against ICs in terms of their ability to select one of several candidate models; we generate, via stochastic simulation, a huge amount of artificial time series both density-independent and dependent, with and without exogenous covariates, using different dataset sizes, noise levels and parameter values. Our findings show that SRM outperforms traditional ICs, because generally (a) it recognizes the model underlying the data with higher frequency, and (b) it leads to lower errors in out-of-samples predictions. SRM superiority is specially apparent with short time series.

We finally apply SRM to the population records of Alpine ibex (*Capra ibex*) living in the Gran Paradiso National Park (Italy), already investigated by Jacob-

son et. al. (2004) via traditional statistical methods; we both analyze the models cited in Jacobson et. al. (2004) and introduce some novel models. We show that models that are best according to SRM show also the lowest leave-one-out cross-validation error.

Key words: structural risk minimization, density dependence, population dynamics, time series, model selection

---

1 Introduction

A widely addressed problem in ecology is the identification of the basic mechanisms underlying the observed course of population abundances. In particular, statistically distinguishing density-dependent from independent time series, which is of paramount importance to correctly predict future population abundances, stimulated an important research effort over the past decades (Brook and Bradshaw , 2006); well-known methods for density-dependence detection are for instance the reciprocal of von Neumann's ratio (Bulmer , 1975), the randomization test (Pollard et. al. , 1987) and the parametric bootstrapping likelihood ratio test (PBLR) (Dennis and Taper , 1994); the last method was extended to manage multivariate models of jointly fluctuating populations, thus allowing the study of situations in which annual counts are available from a number of sites (Dennis et. al. , 1998).

However, a general weakness of methods based on likelihood ratio testing is that a single density-independent model (usually the Malthusian) is contrasted with a single alternative density-dependent model (usually the Ricker). As a matter of fact, it is difficult to believe that the patterns of population dynamics we observe are generated by just two mechanisms, mutually exclusive. Population regulation may be achieved through a variety of different density-dependence mechanisms (Begon and Mortimer , 1981), which can generate a variety of demographic models. Also, population trends and fluctuations may be affected by exogenous forcings, such as climatic factors. We are thus necessarily led to consider more than two mechanisms as candidate models for explaining demographic time series. On the other hand, managing many models through hierarchical pairwise hypothesis testing does not necessarily lead to the selection of the best model (Strong et. al. , 1999). Therefore, despite their statistical soundness, hypothesis testing frameworks were recognized by the more recent literature as conveying only limited information (Zeng et. al. , 1998; Taper , 2004).

If we want to include density-dependent phenomena and a variety of environmental factors as potential causes of observed population dynamics, we usually end up with a broad suite of candidate models. These are both density-

independent and dependent and may use different sets of covariates. Information Criteria (ICs) are a well-known approach to the problem of choosing the supposedly best model among a set of many alternative candidates. In particular, they provide an expression for the asymptotic value of the expected discrepancy between the true unknown model and the considered candidate model. Such estimates are obtained by correcting the performance of the model on the calibration samples (usually measured through the log-likelihood) by a term containing some measure of the model complexity. Models are then ranked according to the estimated discrepancies, and the model with the best IC value is finally chosen. There exist different ICs, obtained under different hypotheses, and aimed at estimating different discrepancy functions; however, they are based on asymptotic arguments, which therefore hold for large datasets only. Also, some ICs assume that the model underlying the data is contained in the set of candidate models. Nevertheless, ICs are often successfully applied even if such restricting conditions are not met. Well-known Information Criteria are for instance the Schwarz Information Criterion (SIC)(Schwarz , 1978), Akaike's Information Criterion (AIC) (Akaike , 1973) and the Final Prediction Error (FPE) (Akaike , 1970). Since SIC is known to well estimate the true order of the underlying model (Hooten , 1995), several recent papers (Zeng et. al. , 1998; Dennis and Otten , 2000; Peek et. al. , 2002; Taper and Gogan , 2002) dealing with model selection in ecology or density-dependence detection have used SIC as model selection criterion.

In this paper we propose an alternative approach, based on Statistical Learning Theory (SLT). SLT, due to the joint work of Vapnik and Chervonenkis, is also referred to as VC-theory from the name of its authors; an overview of SLT is for instance provided in (Vapnik , 1999). SLT is a general mathematical framework for estimating relationships from a set of finite, empirical observations, using a discrepancy function called *risk*. From among the findings provided by SLT, there is also a method for estimating the risk of a given model, and to then select the best model. Such model selection approach is called Structural Risk Minimization (SRM). Differently from the traditional approaches, SRM (a) is targeted to work on finite datasets, and therefore does not rely on any asymptotic argument, and (b) does not assume any hypothesis about the probability distribution underlying the data, or about the candidate models considered. The core of SLT is the definition of VC-dimension, which is a complexity index for classes of functions. In fact, to use SRM, it is necessary to know the VC-dimension of the considered candidate models. VC-dimension is known by definition for linear models (actually, it corresponds to the number of free parameters), but is generally unknown for nonlinear models. The estimation of the VC-dimension of nonlinear models, which is not dealt with in this paper, is actually a challenge for future research; in fact, little applied work has been done on this topic up to now, although the complex methodology proposed by Vapnik et. al. (1994) allows, in principle, the estimation of the VC-dimension of any model. An attempt to estimate the VC-dimension

of nonlinear demographic models can be found in Corani and Gatto (2006). In the present paper, however, we consider demographic models in which the rate of increase is a linear function of the free parameters, thus avoiding the problem of VC-dimension estimation. Nevertheless, even with reference to linear regression problems only, it has been shown (Cherkassky et. al. , 1999) that SRM can consistently outperform traditional Information Criteria (such as SIC and FPE) for different dataset sizes and noise levels, with stronger advantages for smaller datasets. This indication is specially important in ecology, because only short time series are usually available.

A preliminary application of SRM as model selection criterion for demographic models has been carried out in Corani and Gatto (2005), who showed via simulation that SRM can constitute an alternative to traditional ICs. The present paper provides a thorough analysis of the problem and significant novel contributions. As a first step, SRM is compared with PBLR to test its ability to discriminate between the Malthusian and the Ricker model. We simulate both models with noise under a wide variety of parametric settings (i.e., parameters of the simulated model, simulation length, noise level), creating an ensemble of 500 different stochastic simulations for each parametric setting. On each generated time series, both models are identified and one of the two is selected according to SRM. Having designed experiments that are identical to those of Dennis and Taper (1994), who used parametric bootstrapping (PBLR) to discriminate between the two models, the performances of SRM and PBLR can be rigorously compared. This kind of experiments, where a unique density-dependent model is contrasted with a unique density-independent model, are referred to in the following as *density-dependence detection.*

As a second step, we investigate the problem of choosing a model from among a wide set of candidates; in these experiments we no longer consider the PBLR method, which is not well-suited to select from a large class of models. On the contrary, we compare SRM with the SIC, AIC and FPE criteria. The models are simulated with noise under a wide variety of parametric settings, and the selection criteria performance is assessed with respect to (a) the ability to recognize the model underlying the data and (b) the prediction accuracy of the chosen models on out-of-samples data. These experiments are referred to in the following as *model selection.* As an example, we present the re-analysis of the time series of the ungulate population of Alpine ibex (*Capra ibex*) living in the Gran Paradiso National Park (Italy). In particular, we compare our findings with those of the exhaustive analysis carried out in Jacobson et. al. (2004).

The paper is organized as follows: Section 2 presents the demographic models considered in this work; Section 3 states the model selection problem and describes the different model selection criteria; Section 4 presents the density-detection experiments, and Section 5 the model selection experiments; section 6 reports on the Alpine ibex case study and Section 7 draws the conclusions.

2 Demographic models

We consider models of different complexity without age structure. By N_t we indicate the total population abundance at time t, where t is the time at which counts are available. The simplest model is the Malthusian, namely $N_{t+1} = \lambda N_t$. By taking logarithms, the equation becomes

$$\ln\left(\frac{N_{t+1}}{N_t}\right) = a \tag{1}$$

where $a = \ln(\lambda)$. In this model the rate of increase does not depend on population density; indeed, the main assumption underlying Malthusian models is that environment can provide each individual with the same amount of resources necessary to survival and reproduction, regardless of the population size, or other factors. The Malthusian model can also be seen as a random walk with drift model, and it has received considerable attention in population viability analysis.

If density rises, intraspecific competition can slow down or halt the population increase. In these cases a density-dependent model is necessary to explain the data; the most famous is the logistic equation, introduced by Verhulst (1838) with reference to a time continuous setting. Here, we consider the simple time-discrete demographic model introduced in Ricker and Foerster (1948) and Ricker (1954) which assumes an exponential decrease of the finite rate of demographic increase. It is given by $N_{t+1} = \lambda N_t \exp(bN_t)$ with $b < 0$; by taking logarithms, it becomes:

$$\ln\left(\frac{N_{t+1}}{N_t}\right) = a + bN_t \quad (a > 0,\, b < 0) \tag{2}$$

The Ricker model is probably the most widely used for the analysis of time-discrete demographic data. Its dynamic behavior was lucidly investigated by Beverton and Holt (1957) who remarked that, depending on a, the population can reach a steady state with or without oscillations, undergo permanent oscillations of constant period and amplitude or undergo what they termed irregular and violent oscillations. The nature of this dynamic complexity was not fully understood by ecologists until the basic concepts of bifurcation and chaos were introduced into population ecology (May , 1974; May and Oster , 1976). In particular, the only nontrivial equilibrium of this model, corresponding to $\overline{N} = -a/b$, is stable for $a < 2$, with damped oscillations occurring for $1 < a < 2$. For $a > 2$, the model dynamics is much more complex (i.e., limit cycles and chaos). For this reason, the Ricker model is considered to be quite flexible, as contrasted to other models like for instance the Beverton-Holt

$(N_{t+1} = \lambda N_t/(1+\gamma N_t),\ \lambda, \gamma > 0)$, which does not display cycles or chaos. Obviously, the Malthusian model is a special case of the Ricker model ($b = 0$).

The Ricker model can be further extended to include an exogenous forcing X, which can be for instance a climatic variable that affects the population dynamics. One can assume that the forcing acts linearly on the rate of increase. Denoting as X_t the value of such a covariate at time t, we have:

$$\ln\left(\frac{N_{t+1}}{N_t}\right) = a + bN_t + cX_t \quad (a > 0,\ b < 0) \tag{3}$$

In the same manner, a set of covariates $(X_{1t}, X_{2t}, \ldots, X_{mt})$ can be added to the model structure, to account for several environmental factors that are supposedly relevant for the population dynamics. In the following, we denote the Malthusian model as M, the Ricker model as R, and the Ricker model with i covariates as R followed by the i-th Roman numeral (e.g., RIII is Ricker with 3 covariates).

It is to be remarked that the set of candidate models could be further expanded to include time delays (Turchin , 1990) or inverse density-dependence (Courchamp et. al. , 1999), or else. To keep our analysis reasonably simple, we assume some a priori knowledge that allows us to exclude the occurrence of these phenomena. In any case, some preliminary screening of the candidates is almost unavoidable.

Noise must be included in the models to account for all those exogenous and endogenous factors that affect the population, yet cannot be measured or are not described by the model. They include environmental variability, climatic fluctuations, demographic stochasticity, effects due to age structure variation, etc. We assume stochasticity to act on the finite rate of increase in a multiplicative way, so that the whole set of models we will consider in the sequel can be summarized as:

$$\ln\left(\frac{N_{t+1}}{N_t}\right) = a + bN_t + \sum_{i=1}^{m} c_i X_{it} + \sigma Z_t \quad (a > 0,\ b < 0) \tag{4}$$

where Z_t is a standard white noise and σ^2 is the variance of the logarithm of the noise affecting the system dynamics . In the Malthusian case ($b = 0$) the model is nothing but a random walk with ($a > 0$) or without drift ($a = 0$). Stochastic Malthusian models have often been used for modeling endangered species and for population viability analysis.(Lande et. al. , 2003)

3 Model selection criteria

Model selection aims at choosing the model structure and parameters that, from among a set of candidates, minimize the average discrepancy with the true, unknown, data generation mechanism. Discrepancy can be measured in different ways; in general, its expected value reflects the trade-off between (a) model bias, due to the adoption of a function different in its structure from the true one and (b) model variance, due to the imperfect estimate of the model parameters. Generally, models with too few parameters can be biased, while models having too many parameters can be affected by poor precision of the estimates; therefore, using a more complicated model is convenient only if the decrease of model bias outweighs the increased variance; this is the so-called bias-variance dilemma (Forster , 2000). Since the expected discrepancy could be calculated only if the data generation mechanism were known, model selection criteria are based on estimators of the overall expected discrepancy of a model (Zucchini , 2000); estimates can be computed on the basis of the empirical data and provide a ranking of the candidate models.

In the following, we denote as $\mathbf{x}$ the vector of input variables, and as y the output variable; they can be mapped on the variables of the previously introduced ecological models as follows:

$$\begin{cases} y_{t+1} = ln(\frac{N_{t+1}}{N_t}) \\ \mathbf{x_{t+1}} = [N_t, X_{1t}, \ldots, X_{mt}] \end{cases} \tag{5}$$

where the variables have the meaning already specified in Section 2.

In general terms, a model is a probability distribution relating $\mathbf{x}$ and y (for instance, one might consider the joint probability density $p(\mathbf{x}, y)$). For regression models, such as the ones considered here, the probability distribution is implicitly specified by the equation $y = f(\mathbf{x}, \theta) + \sigma Z_t$, where f is a function of $\mathbf{x}$ and θ, a vector of parameters.

We assume that models are identified on a finite dataset containing q observations $(\mathbf{x}_i, y_i)$, $i = 1, 2 \ldots q$; we denote as d the number of parameters of a model, and as p the parameters-to-data ratio $p = d/q$. In some cases, one would like to estimate also the noise variance σ^2; if so, the parameters include also the variance estimate; in these cases, the total number of parameters of the model is therefore $k = d + 1$. Finally, we denote as ϵ_i the residuals of the estimated models, namely $\epsilon_i = y_i - f\left(\mathbf{x}_i, \hat{\theta}\right)$ where $\hat{\theta}$ is the parameters' estimate.

FPE Akaike's Final Prediction Error (FPE) (Akaike , 1970) is one of the first model selection criteria proposed in the literature; it is an asymptotically unbiased estimator of the expected value of the mean square error of a linear model $y = \mathbf{x}\hat{\beta}$ with respect to the future unknown data, under the assumption that the true data generation mechanism is $y = \mathbf{x}\beta^* + \sigma Z$, where Z is an i.i.d. random noise. Some elements of β^* might be zero, thus determining the dimension of the model. The expectation of the mean square error is taken with reference to all the calibration datasets of size q. FPE is computed as follows:

$$FPE = \frac{1}{q}\sum_{i=1}^{q}\epsilon_i^2\frac{1+d/q}{1-d/q} \tag{6}$$

The model with minimum FPE is selected.

AIC The Akaike Information Criterion (AIC) is aimed at ranking the models according to the Kullback-Leibler discrepancy of each candidate model. A detailed analysis of the AIC theory has been recently carried out by Burnham and Anderson (2003), who also recommend the use of a specific version of AIC, i.e. AIC_c, over the plain AIC; indeed, AIC_c embodies a bias correction factor necessary for small datasets (i.e., $q/k<40$), while AIC_c converges to AIC for larger datasets.

Under the assumption of normally distributed errors, AIC_c can be computed as:

$$\text{AIC}_c = q\log\left(\hat{\sigma}^2\right) + 2k + \frac{2k(k+1)}{q-k-1} \tag{7}$$

where

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{q}\epsilon_i^2}{q} \tag{8}$$

In Burnham and Anderson (2004) it is argumented that AIC and AIC_c can be quite effective even if the true model is not contained in the set of candidates.

Generally, the model with the lowest AIC_c is chosen; however, according to the results of Richards (2005), it is instead convenient to use the following empirical rule based on the principle of parsimony: choose the model with the lowest complexity d within the set of candidates having an AIC_c differential $\Delta AIC < 4$ with respect to the model with the minimum AIC_c. As in our experiments this rule led to a slight yet general improvement of the AIC_c

performance, in the paper we present the results obtained following Richards' prescription.

SIC A further criterion largely used in ecological applications is the Schwarz Information Criterion (SIC), also known as Bayesian Information Criterion (BIC); under the assumption of normally distributed errors, it can be computed as:

$$SIC = q \log \left(\hat{\sigma}^2\right) + k \log q \tag{9}$$

SIC is designed to find the most probable model based on the data: given a fixed set of a priori models $M_1, M_2, \ldots M_r$, and their posterior probabilities $p_1, p_2, \ldots, p_r$ (based on prior probabilities and the likelihood of data), SIC seeks the model M_j such that, for $q \to \infty$, $p_j \to 1$ and $p_{\neq j} \to 0$ for all the models others than M_j. For SIC too, the application of a "rule of thumb" is recommended, which prescribes the consideration of the subset of models having a SIC differential $\Delta SIC < 2$ with respect to the model with the lowest SIC (Raftery , 1995); then, the model with lowest d in such a subset should be chosen. We actually use such a criterion in our simulations.

3.0.0.1 SRM Structural Risk Minimization (SRM) is the model selection criterion derived within the Statistical Learning Theory (SLT). An overview of SLT is given in Vapnik (1999); SLT is based on a novel index of complexity for parameterized functions, both linear and non-linear, known as VC-dimension (denoted as h in the following), where the VC acronym refers to statisticians Vapnik and Chervonenkis, who actually invented it. For the sake of simplicity, we assume to deal with the linear case, in which h coincides with the number of parameters of the model, i.e., $h = d$.

The setting of the model selection problem is as follows (Vapnik , 1999) **:** the random input variable $\mathbf{x}$ is drawn independently from a fixed but unknown distribution and the output y is returned by the system according to a distribution function conditional on $\mathbf{x}$, also fixed but unknown. We denote the joint probability density of $\mathbf{x}$ and y as $p(\mathbf{x}, y)$.

The goal is to select the function $f(\mathbf{x}, \theta)$ (where θ denotes the function parameters) which minimizes the expected value of the loss, expressed by the so-called *risk functional:*

$$R(\theta) = \int_{\mathbf{x}} \int_{y} (y - f(\mathbf{x}, \theta))^2 p(\mathbf{x}, y)\, d\mathbf{x}\, dy \tag{10}$$

However, $p(\mathbf{x}, y)$ is unknown and the only available information is contained in the q input/output samples. Therefore, what can be measured is just the fitting error of the model during calibration, referred to as *empirical risk* (R_{emp}):

$$R(\theta)_{emp} = \frac{1}{q}\sum_{i=1}^{q}(y_i - f_j(\mathbf{x}_i, \theta))^2 \tag{11}$$

One of the key-theorems of SLT provides a distribution-independent upper bound for the risk functional $R(\theta)$, obtained as a function of the empirical risk. Although the original bound formula contains some unknown constant values, for practical regression problems it can be conveniently simplified; according to Cherkassky et. al. (1999), it can be expressed as:

$$R(\theta) \leq R(\theta)_{emp}\left[1 - \sqrt{p - p\ln p + \frac{\ln(q)}{2q}}\right]_{+}^{-1} \tag{12}$$

where $p = h/q = d/q$. The bound is defined only for positive values of the quantity inside square brackets. The inequality of equation (12) holds with probability $\left(1 - \frac{1}{\sqrt{q}}\right)$. The bound value is referred to as *guaranteed risk*.

Actually, Structural Risk Minimization suggests to choose the model that, from among the set of candidates, minimizes the guaranteed risk (12). Looking at the risk functional as a random variable, SRM is a worst-case approach, because it tries to minimize the guaranteed risk, rather than the average risk. However, in Cherkassky et. al. (1999) and Cherkassky and Mulier (1998) it has been shown that minimizing the SRM guaranteed risk generally leads to better performances, in terms of out-of-sample errors, than using traditional model selection criteria.

3.0.0.2 PBLR While the approaches presented up to now are aimed at selecting a model from a broad suite of candidates, traditional tests of hypothesis are apt to choosing between a pair of models. In particular, the parametric bootstrapping of likelihood ratios (PBLR) is a statistical method (Efron and Tibshirani , 1993), which was introduced by Dennis and Taper (1994) in the ecological context. It contrasts model i against model j; the test statistic is the ratio Λ_{ij} of the likelihood function L_i maximized over the parameter values of model i, to the likelihood L_j, also maximized over the parameters of model j. The decision is made in favor of Model i if $\Lambda_{ij} > c$, or in favor of Model j if $\Lambda_{ij} \leq c$, where c is a cutoff value selected so that the probability of wrongly choosing Model j when data arise from Model i is fixed at a small percentage, the *test size*. The PBLR approach estimates such a cutoff value via parametric bootstrapping. Once the test size has been fixed, the efficiency

of the test is evaluated by calculating the *test power*, i.e., the probability of recognizing Model j when it really underlies the data.

Although running PBLR is a matter of a few seconds with modern computers, its implementation and computation is remarkably more complicated than that of ICs or SRM.

4 Experiments on density-dependence detection

In this section we compare the ability of PBLR and SRM in discriminating between a single density-independent model (Malthusian) and a single density-dependent model (Ricker). To rigorously compare the performances of the two methods, we adopt the same experimental design used by Dennis and Taper (1994).

Artificial noisy time series of either the Malthusian or the Ricker model are generated as follows:

$$N_{t+1} = N_t \exp(a + bN_t + \sigma Z_t) \tag{13}$$

where σ defines the noise level and Z_t is a standard normal white noise (mean = 0, standard deviation = 1, $E\left[Z_t Z_{t-\tau}\right] = 0$ for all $\tau \geq 1$). Coefficient b is set to 0 when the Malthusian model is simulated.

An ensemble of stochastic simulations is characterized by a set of parameters, which constitute the *simulation setting:* the initial condition N_0; the model coefficients (a, b); the noise level σ; the simulation length q.

The experimental density-dependence detection methodology, which is repeated 500 times for each simulation setting, is as follows:

(1) *stochastic simulation*: perform a q-steps noisy simulation by means of equation (13), using the current simulation setting;
(2) *identification*: compute the time series $y_{t+1} = \ln \frac{N_{t+1}}{N_t}$ and estimate the parameters of the Ricker and the Malthusian model by means of standard linear regression;
(3) *acceptability check:* use a one-sided test, as done in Dennis and Taper (1994), namely discard the Ricker model if the estimate of b is positive, and in this case select the Malthusian model. This corresponds to assuming a priori that only intraspecific competition affects the population growth rate;
(4) *model selection*: choose the best model according to SRM.

4.1 Malthusian model recognition

The simulation settings for the Malthusian model, coherently with Dennis and Taper (1994), are given by all the possible combinations of the following values: a=[0.05; 0.55; 1.1; 1.6]; σ=[0.05; 0.55; 1.1; 1.6]; q=[10; 20; 40; 60]; N_0=64. We use therefore 64 different simulation settings, for a total of 32,000 simulations. We remark that values of $\sigma > 1$ are rather large. In fact, σ is approximately equal to the coefficient of variation CV of N_t (Lande et. al. , 2003). For many time series of birds and mammals CV rarely exceeds 0.6. Nevertheless, we keep these values of σ to make our results comparable to those of Dennis and Taper (1994).

Within the PBLR hypothesis testing framework, the Malthusian and the Ricker model constitute respectively the null and the alternative hypothesis. The test size, i.e., the probability of rejecting the Malthusian model when it really underlies the data, was set to 5% by Dennis and Taper (1994). By contrast, SRM does not fix a test size a priori, because it selects the Malthusian or the Ricker model according to their performances in terms of guaranteed risk.

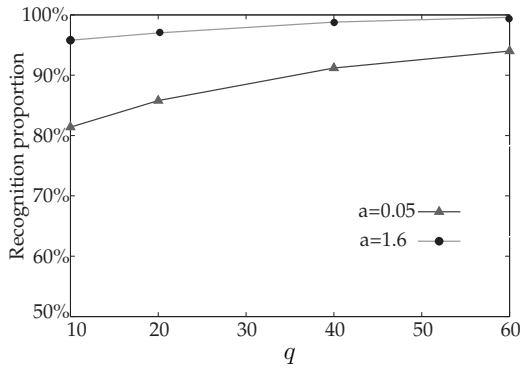
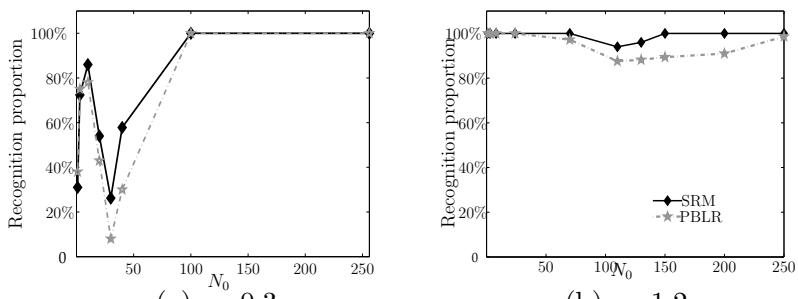


Figure 1. SRM recognition proportion for the Malthusian model as a function of Malthusian parameter a and sample size q. The level of noise is $\sigma = 0.55$. Results obtained for $a = 0.55$ and $a = 1.1$ are intermediate between the curves drawn for $a = 0.05$ and $a = 1.6$.

From our computations we obtain that the correct selection percentage of SRM is 94% on the average, that is very close to that of PBLR, which by definition is 95%; about 30% of the times the Ricker model is discarded because of a positive estimate of parameter b. Figure 1 shows that the SRM recognition proportion increases with both a and q.

4.2 Ricker model recognition

Parameter b of the Ricker model does not influence the probability of recognizing density-dependence as long as it is not zero (Dennis and Taper , 1994). In fact b is a scale parameter, which reflects the units in which the population is measured. More precisely, setting $W_t = bN_t$, we obtain $\ln \frac{W_{t+1}}{W_t} = a + W_t$. In particular, the stability properties of the equilibrium density $\overline{N} = -a/b$ depend on a only; reasonably, the recognition proportion should not depend on the actual abundance around which the population fluctuates, rather on the dynamical characteristics of the population. Therefore, consistently with (Dennis and Taper , 1994), we fix $b = -0.01$ in all the simulations of the Ricker model.



(a) a=0.3 (b) a=1.2

Figure 2. Comparison between SRM and PBLR: proportion of correct detection for the Ricker density-dependent demography as a function of the initial density N_0 and the intrinsic rate of increase a.

Since population growth in the Ricker model is limited, being initially far or close to the equilibrium value $\overline{N}$ is crucial for model recognition. The simulation settings adopted in Dennis and Taper (1994) in order to investigate the role of the size N_0 of the initial population are as follows: $a = [0.3; 1.2]$; $b = [-0.01]$; $\sigma = [0.05]$; $q = [10]$; $N_0 =$[1; 3; 10; 20; 30; 40; 100; 250] for $a = 0.3$ and $N_0 =$[1; 3; 10; 25; 70; 110; 130; 150; 200; 250] for $a = 1.2$.

Figures (2a-2b) report the results. SRM and PBLR show a similar behavior, from a qualitative point of view: in particular, the percentage of correct detection is minimum when the starting condition N_0 is close to $\overline{N} = -a/b$. In fact, for both values of a, $\overline{N}$ is a stable equilibrium and the density-dependent model moves towards the equilibrium. Therefore, small deviations of the initial population from the equilibrium do not allow the exploration of the dynamical characteristics of the model, thus making the recognition more difficult. However, one can easily see that, under any initial condition, SRM performs better than PBLR.

For both SRM and PBLR, correct detection is easier as parameter a increases (compare Fig. 2 a with b). In fact, while a= 0.3 corresponds to a monotonic

approach to equilibrium, $a = 1.2$ implies damped oscillations of the density dependent model around $\overline{N}$. This makes recognition easier.

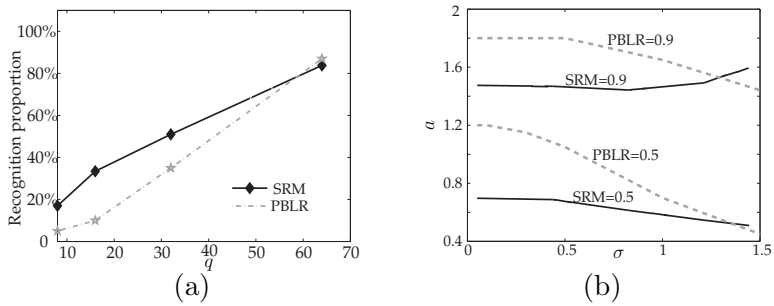


Figure 3. Comparison between SRM and PBLR for Ricker model recognition. (a) Recognition proportion as a function of sample size q (a=0.3); (b): contour plots of the recognition proportion as a function of the intrinsic rate of increase a and the noise level σ; sample size is set to 10.

In a further series of experiments, we investigate the effect of the time series length q. The adopted simulation settings, coherently with Dennis and Taper (1994), are as follows: $a = [0.3; 1.2]$; $b = [-0.01]$; $\sigma = [0.05]$; $q = [8; 16; 32; 64]$; $N_0 = [-a/b]$. As expected, the proportion of correct recognition increases with the time series length; once more, SRM consistently outperforms PBLR, in particular for small q. In the most critical case, i.e., low a and small dataset, SRM outperforms PBLR by 10-20 percent points. Results obtained for a=0.3 are shown in Figure 3a.

In a third series of experiments, we investigate the effect of environmental stochasticity as measured by the noise level σ. Such an issue is investigated jointly with finer variations of parameter a. The simulation settings adopted to this end in Dennis and Taper (1994) are given by all the possible combinations of the following values: a=[0.05; 0.25; 0.45; 0.8; 1.6]; $b = -0.01$; σ =[0.05; 0.25; 0.45; 0.8; 1.6]; $q = 10$; $N_0 = -a/b$.

The results are shown in Figure 3b as contour plots of the detection proportion in the parameter plane ($\sigma - a$). The most striking feature is that the recognition proportion can increase with the noise level σ. This result may seem counterintuitive. However, it can be explained by considering that simulations are started at the model equilibrium; stochastic fluctuations provide deviations from the equilibrium and hence make the correct model selection easier, as already noted by Dennis and Taper (1994) with reference to PBLR. Similar findings are pointed out by Brook and Bradshaw (2006), who analyze the evidence of density dependence in a very large number of population time series. The comparison between PBLR and SRM shows that SRM is more advantageous than PBLR; for instance, the same power (0.5) can be obtained with much lower values of the parameter a. PBLR takes greater advantage from large noise than SRM; for $\sigma > 1.2$ it can reach the same efficiency as

SRM, displaying also a small advantage in the noisiest conditions. However, as we noted above, so high values of σ are not very common.

5 Selection within a suite of candidate models

In this section, we compare the ability of AIC_c, FPE, SIC and SRM to correctly choose from among a suite of different demographic models, both density-dependent and independent, with and without covariates. The considered models are the Malthusian (M), the Ricker (R), and the Ricker with up to 5 covariates (RI, RII, RIII, RIV, RV).

To show evidence of possible tendencies of the criteria to overparameterize, we use just four models (M, R, RI, RII) out of seven to generate artificial time series. In this way, we consider models more complex than necessary (RIII, RIV, RV) to explore whether these models are selected even if the time series is generated by a simpler mechanism.

As usual, models have been simulated using a wide variety of simulation settings, to obtain an ensemble of 500 stochastic simulations for each setting. We use such a huge amount of experiments in order to statistically assess (a) the success of each criterion in choosing the model that really underlies the data and (b) the ability of the chosen models to predict data out of the calibration samples.

The equation used to generate the stochastic simulations is:

$$N_{t+1} = N_t \exp(a + bN_t + c_1 X_{1t} + c_2 X_{2t} + \sigma Z_t) \tag{14}$$

where the symbols have the same meaning as in equation (13), with the addition of covariates X_1 and X_2. Depending on the type of the simulated model, coefficients b and/or c_1 and/or c_2 have been set to 0. Covariates X_1 and X_2 are generated as standard normal white noises with zero cross-correlation.

A simulation is characterized by the following *simulation settings*: the initial condition N_0; the model coefficients (a, b, c_1, c_2); the noise level σ; the simulation length q.

Simulation settings for the different models are obtained by combining the following values in all the possible ways:

- Malthusian model (M): $N_0 = 100$, $a = [.5; 1; 1.5]$, $\sigma = [.05; .1; .25; .5]$, $q = [10; 20; 50; 100]$;

- Ricker model (R): $N_0 = [100; -a/b]$, $a = [.5; 1; 1.5]$, $b = -0.01$, $\sigma = [.05; .1; .25; .5]$, $q = [10; 20; 50; 100]$;
- Ricker model with one covariate (RI): $N_0 = 100$, $a = 1$, $b = -0.01$, $c_1 = [.05; .1; .25; .5]$, $\sigma = [.05; .1; .25; .5]$, $q = [10; 20; 50; 100]$;
- Ricker model with two covariates (RII): $N_0 = 100$, $a = 1$, $b = -0.01$, $c_1 = [0.25]$, $c_2 = [.05; .1; .25; .5]$; , $\sigma = [.05; .1; .25; .5]$;, $q = [10; 20; 50; 100]$.

With respect to the analysis carried out in the previous section, here we use more realistic values of noise, namely $\sigma \leq 0.5$. The values of the intrinsic rate of population increase a are dictated by the consideration that model identifiability is linked to model dynamics. As the eigenvalue of the Ricker model at the nontrivial equilibrium is $1 - a$ we have that $a = 0.5$ corresponds to monotonic stability, $a = 1$ to superstability and $a = 1.5$ to stability with damped oscillations. If the time unit is 1 year, growth rates as big as 1.5 yr^{-1} can be found in insects, smaller rates in vertebrates. Overall, we have 48 different simulation settings for model M, 96 for model R, and 64 for RI and RII; for each simulation setting, 500 different model selection experiments are performed using the following procedure:

(1) *simulation*: perform a $q-$steps simulation according to equation (14), using the given simulation setting;
(2) *identification*: compute the time series $y_{t+1} = \ln \frac{N_{t+1}}{N_t}$ and then estimate the parameters of the seven candidate models by means of standard linear regression;
(3) *acceptability check:* check if there are density-dependent models (R, RI, RII, RIII, RIV, RV) whose estimated parameter b is positive, and remove those models from the set of alternatives considered for the current experiment;
(4) *model selection*: choose the best model according to AIC_c, FPE, SIC and SRM;
(5) *generalization assessment*: run the stochastic equation (14) 20 times, with the simulation setting of step 1, using the last abundance simulated at step 1 to generate the further data N_{q+1} from N_q (i.e., compute 20 stochastic one-step continuations for the simulation), and compute the 20 corresponding rates of increase y_{q+1}. Then, use the model chosen by each different criterion to predict the rate of increase $\hat{y}_{q+1}$, i.e., simulate deterministically the one-step ahead continuation of the model dynamics. Finally, calculate the 20 prediction errors $e_{q+1} = y_{q+1} - \hat{y}_{q+1}$. In this way, $20 \bullet 500 = 10,000$ error samples are collected for each model selection criterion and for each simulation setting. This allows the computation of the frequency distribution of the prediction error, the prediction risk in the sense of eq. (10), i.e., the square of the prediction error, and the absolute value of errors.

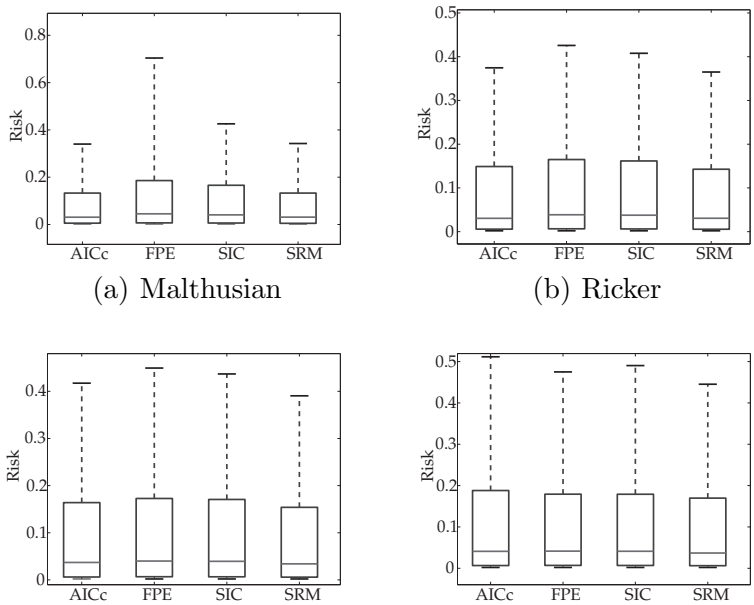
5.1 Results of the experiments

Table 1
Average ability of AIC, FPE, SIC and SRM in recognizing the model underlying the artificially generated time series. For each model selection criterion, the different rows refer to the different models used to generate the artificial time series; the different columns refer instead to the model selected by the different criteria. Percentages in bold (the higher, the better) refer to the cases in which the model underlying the simulated data is chosen by the criterion. M: Malthusian model, R: simple Ricker, RI ... RV : Ricker with1-5 covariates

	AIC$_c$							FPE						
	Selected model							Selected model						
	M	R	RI	RII	$RIII$	RIV	RV	M	R	RI	RII	$RIII$	RIV	RV
Simulated model														
M	**98%**	1%	1%	0%	0%	0%	0%	**83%**	4%	2%	2%	2%	3%	4%
R														
$(N_0 = 100)$	15%	**82%**	1%	1%	1%	0%	0%	1%	**60%**	11%	7%	6%	6%	9%
$(N_0 = -\frac{a}{b})$	19%	**79%**	1%	1%	0%	o%	0%	1%	**59%**	12%	7%	6%	6%	9%
$R\ I$	1 0%	23%	**65%**	1%	1%	0%	0%	0%	11%	**51%**	12%	8%	8%	10%
$R\ II$	7%	9%	19%	**62%**	1%	0%	0%	0%	2%	10%	**52%**	13%	10%	13%

	SIC							SRM						
	Selected model							Selected model						
	M	R	RI	RII	$RIII$	RIV	RV	M	R	RI	RII	$RIII$	RIV	RV
Simulated model														
M	**90%**	2%	1%	1%	1%	2%	4%	**98%**	2%	0%	0%	0%	0%	0%
R														
$(N_0 = 100)$	3%	**80%**	3%	2%	2%	3%	7%	7%	**89%**	3%	1%	0%	0%	0%
$(N_0 = -\frac{a}{b})$	4%	**79%**	3%	2%	2%	3%	7%	9%	**87%**	3%	1%	0%	0%	0%
$R\ I$	1%	20%	**63%**	3%	2%	4%	7%	3%	24%	**69%**	3%	1%	0%	0%
$R\ II$	1%	5%	16%	**62%**	4%	4%	8%	2%	8%	20%	**67%**	2%	1%	0%

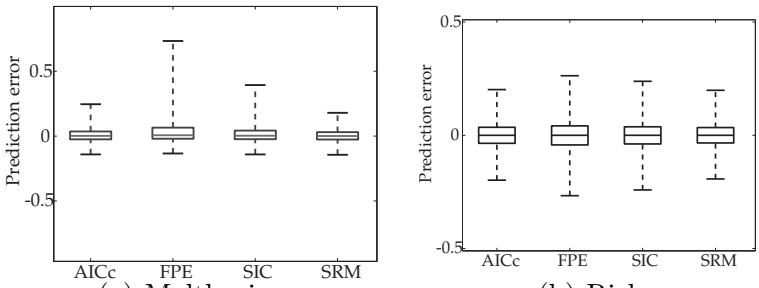
Table 1 summarizes the results by reporting the frequency with which a model is selected on the average using the different model selection criteria. SRM is better or, in a few cases, as good as AIC$_c$ or SIC in detecting the model underlying the data; its recognition percentage is 0-8 points higher than AIC$_c$, 15-22 points higher than FPE, and 0-9 points higher than SIC. SRM and



(a) Malthusian (b) Ricker

(c) Ricker one cov. (d) Ricker two cov.

Figure 4. Box and whiskers plots of prediction risk (e_{q+1}^2) for different model types and selection criteria. The boxes display the lower quartile, the median and the upper quartile, while the whiskers show the 5th and 95th percentiles.



(a) Malthusian (b) Ricker

Figure 5. Box and whiskers plots of the prediction error e_{q+1} for Malthusian and Ricker models. The boxes display the lower quartile, the median and the upper quartile, while the whiskers show the 5th and 95th percentiles.

AIC_c are the most parameter-parsimonious criteria; in fact, they rarely choose models containing more parameters than necessary; on the other hand SIC, and even more FPE, are less parameter-parsimonious; indeed, they select with higher frequency models more parameterized than necessary. In particular, FPE shows the lowest recognition frequency for all the simulated models. On the contrary, SRM and AIC_c have a slightly larger tendency to classify a density-dependent model as density-independent than FPE or SIC.

Figures (4) and (5) display the box and whiskers plots of the prediction risk (e_{q+1}^2) and the prediction error (e_{q+1}). The statistics are estimated from $48 \bullet 500 \bullet 20 = 480,000$ error samples for the Malthusian model, $96 \bullet 500 \bullet 20 =$

$960,000$ for the Ricker and $64 \bullet 500 \bullet 20 = 640,000$ for RI and RII models. The SRM medians of the prediction risk are generally 10-20% lower than SIC and FPE medians, and 10% lower than AIC_c medians. As for the prediction error e_{q+1}, SRM and AIC_c show the most compact plots (Figure 5), i.e., they are less prone than other criteria to underestimate or overestimate y_{q+1}. Boxplots representing the error distributions of all the criteria are centered around 0 and present a symmetrical shape, which means that underestimates and overestimates balance each other. The only exception to this statement is the negative bias of SIC and FPE in the Malthusian case: indeed, in this case they frequently choose a density-dependent demography that results in a systematic underestimation of the population growth rate. FPE is the worst performing criterion for out-of-sample predictions as it corresponds to the widest distributions of both prediction error and risk, as well as the highest risk medians. Boxplots of e_{q+1} referring to the RI and RII models are fairly similar to those obtained for the Ricker model and are therefore omitted.

Figures 6 and 7 display more detailed results for the recognition of the various models. The graphs show the sensitivity of correct model selection to variation of each single parameter. It turns out that the correct selection proportion of the 4 criteria is differentially sensitive to the parameters defining the simulated time series.

The recognition percentages of SRM for the Malthusian model are close to 100% (see Figures 6a,b,c) and practically insensitive to any variation in the simulation settings; they clearly outperform both FPE and SIC. It is worthwhile to highlight the ability of AIC_c and SRM to deal with very small datasets: for instance, they correctly recognize the Malthusian model from time series of just 10 data about 95% of the times, while SIC and FPE are clearly more sensitive to the dataset size (Figure 6c).

In the case of the simple Ricker model, SRM provides the highest recognition proportion under almost all settings (Figures 7a,c,e). Interestingly, for all the model selection criteria, increasing the noise level σ does not worsen the recognition performance (Figure 7c). In fact, noise elicits fluctuations that highlight the nonlinearities of the model. This phenomenon was already described in the previous section, in which, however, we had considered a much wider range of noise levels. Figure 7e shows that, compared to the other methods, SRM is quite reliable even when time series are very short.

Considering the case of the Ricker models with covariates (RI and RII), we find that SRM recognition proportion is almost always higher than those of the other criteria. The shapes of sensitivities are in agreement with what is expected: recognition success increases with the covariate coefficients (Figure 7b), decreases with the noise level σ (Figure 7d), and increases with sample size q (Figure 7f). The effect of noise level is not beneficial to identifiability,

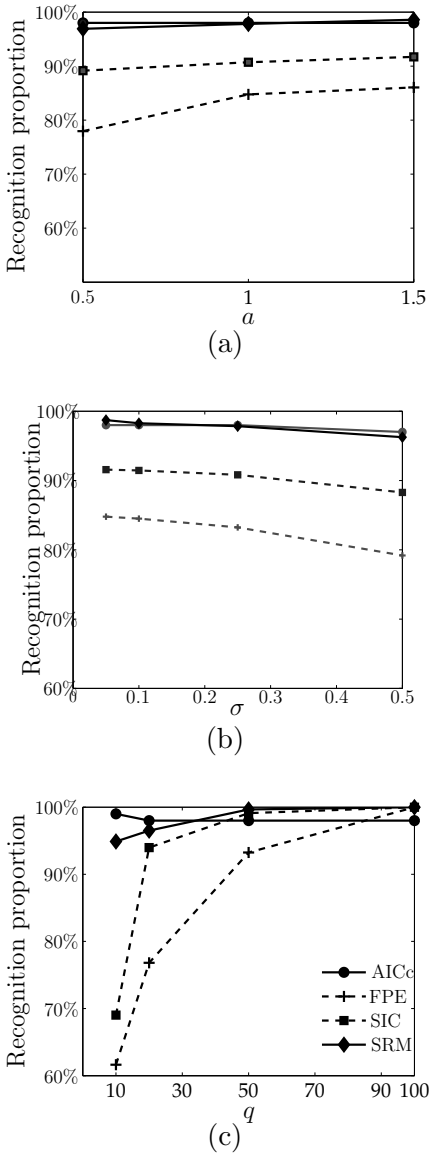


Figure 6. Comparison of 4 model selection methods: sensitivity of the recognition success of the simple Malthusian model to (a) variations of a; (b) variations of σ; (c) variations of q. Percentages are computed by pooling the experiments that share the same value of a given parameter.

as it was in the case of the simple Ricker model. The graphs for model RII have been omitted, as they are very similar to those referring to model RI.

6 The Alpine ibex case study

We will now apply SRM to investigating the demography of the ungulate population of Alpine ibex (*Capra ibex*) living in the Gran Paradiso National

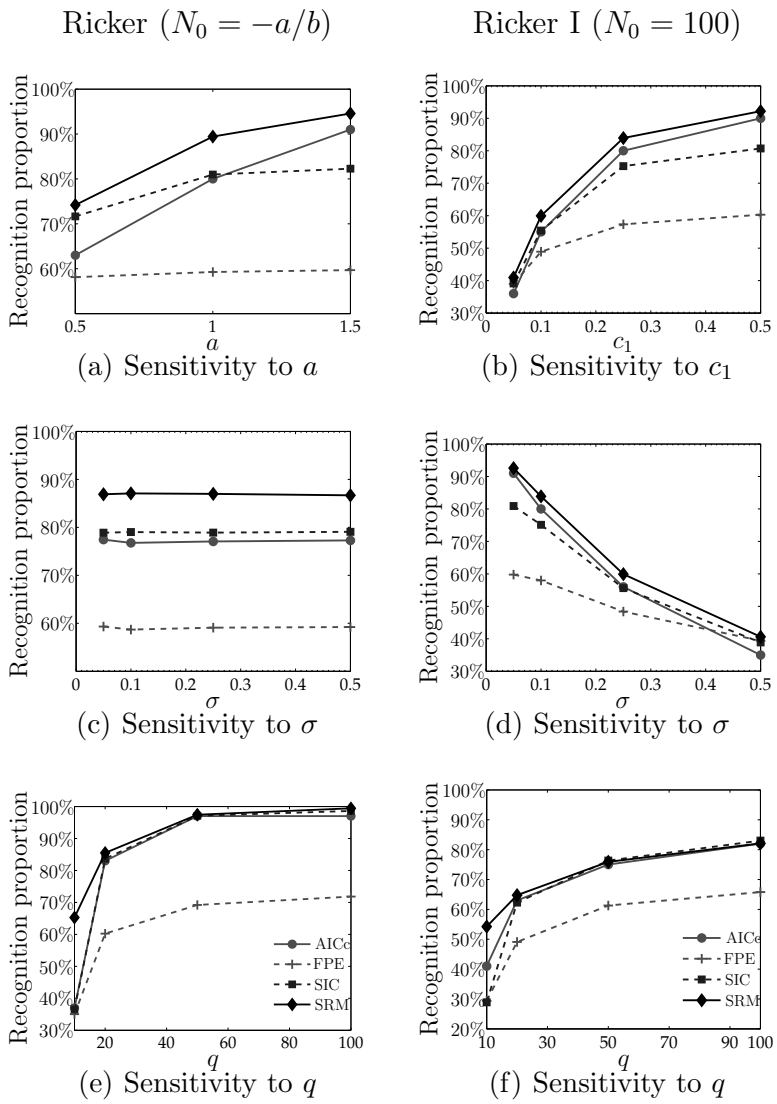


Figure 7. Comparison of 4 model selection methods: sensitivity of the recognition success of the Ricker and RickerI model to variations of each single parameter. Percentages are computed by pooling the experiments that share the same value of a given parameter.

Park (Italy). Recently, the same population has been thoroughly analyzed in Jacobson et. al. (2004), and therefore this study constitutes an important term of comparison for our findings.

The dataset (Jacobson et. al. (2004):Supplement 1) is unusually long, containing a 40-years time series (1960-2000) of both Alpine ibex censuses and meteorological observations (temperature, precipitation and snow depth). The National Park covers an area of about 720 km^2; hunting is not allowed inside or close to the park, and large predators such as lynx and wolf have been absent over the past 100 years. Recently, several studies (see the literature overview in Jacobson et. al. (2004)) suggested that, if large predators are rare or absent,

the changes in ungulate populations can be explained by considering climate forcing and density-dependence only.

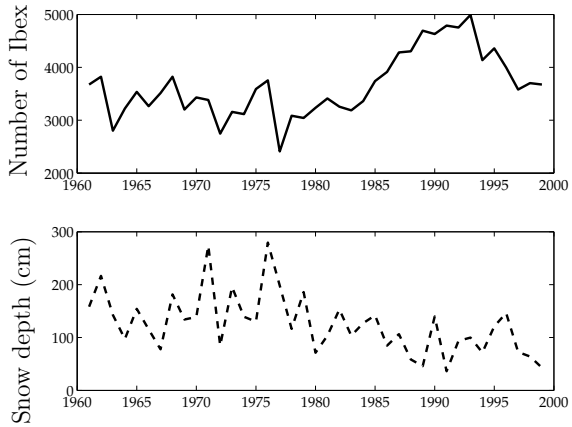


Figure 8. Alpine ibex population abundance and average winter snow depth in the Gran Paradiso National Park.

From Figure 8 one can see that snow depth has a negative impact on population abundance: the three snow peaks of 1962, 1971, 1976 resulted in important decreases of the population in the following year, while the big population increase between 1982 and 1993 corresponds to a period of quite low snow depth. Actually, snow depth has been statistically recognized (Jacobson et. al. , 2004) as the most significant climate driver for the considered population.

6.1 The findings of Jacobson et. al. (2004)

The analysis carried out in Jacobson et. al. (2004) is aimed at developing a simple demographic model to explain the impact of climate forcing and density-dependence on the dynamics of the ibex population. In the first step of their investigation, the authors use three different well-recognized statistical methods (the Bulmer test (Bulmer , 1975), the randomization test (Pollard et. al. , 1987) and PBLR (Dennis and Taper , 1994)) in order to test for the dependence of the population dynamics on density and snow depth. In particular, they first attempt to detect density dependence without considering climate drivers. Using the Gompertz and Ricker models, they conclude that although density plays a role in the population dynamics, it is not a sufficient explanation of the population changes. They reach a similar conclusion when testing a model with snow-depth dependence alone. Therefore, they decide to test models including density-dependence, snow-depth dependence and further terms representing their interactions. In particular, the authors include as variables the density N_t, the snow depth S_t and the product N_tS_t representing the interaction of these two quantities. The first class of models they

consider is a Ricker model with two covariates, whose equation is as follows:

$$ln\left(\frac{N_{t+1}}{N_t}\right) = a + bN_t + cS_t + dN_tS_t \tag{15}$$

Within this class, they analyze several "subset" models obtained by dropping individual terms from the complete model.

A further family of models analyzed by Jacobson et. al. (2004) is obtained by letting the population dynamics depend on the logarithm of abundances $L_t = ln(N_t)$ (Gompertz model) rather than abundances themselves:

$$ln\left(\frac{N_{t+1}}{N_t}\right) = a + bL_t + cS_t + dL_tS_t \tag{16}$$

In this case too, subset models derived from equation 16 are considered.

In a further modeling attempt, Jacobson et. al. (2004) consider snow-threshold models, which have two different sets of parameters for low and high snow conditions. They set the snow threshold $\overline{S}$ equal to the mean snow depth plus 1/2 the standard deviation. In particular, the parameters b, c, d of these models are estimated twice (i.e., for low and high snow depth), while the intrinsic rate of increase a is assumed to be independent of snow being low or high, so the estimate of a is unique. At this point, having generated a wide set of candidate models, the authors use information criteria to choose from among them.

Among all the considered models, the authors select a set containing models 2, 3, 4, 5 (their structure is reported in Table 2) on the basis of the Akaike Information Criterion (AIC). Then, they perform an out-of-samples simulation analysis and finally choose model 4.

6.2 Analysis via SRM

The paper of Jacobson et. al. (2004) constitutes a relevant term of comparison, because of its methodological rigor and the cross-checking of several statistical approaches. It is therefore of interest to check whether we reach similar conclusions by our approach. This is specially important because we now use SRM considering a set of candidate models which is much wider than the set considered in the extensive testing phase described in the previous section 5. In our analysis, we firstly evaluate via SRM all the models proposed by Jacobson et. al. (2004); according to SRM, the best models are ranked as 2, 3, 1, 4, 5 (see last column of Table 2). This result is partially different

Table 2
Comparison of the best models for the Alpine ibex population as chosen according to SRM. Each model includes the variables marked with an asterisk (see eqs.15 and 16). Column DoF displays the degrees of freedom of each model; column SRM shows the penalized empirical risk calculated according to SRM. As for the model types: LIN denotes linear models; THRESH denotes snow-threshold models with a threshold independent of parameter a; THRESH(a_1, a_2) denotes snow-threshold models with two a's, respectively below and above the threshold. N_t is the population abundance at time t; $L_t = log(N_t)$; S_t is the snow-depth at time t.

model #	model type	DoF	const	N_t	L_t	S_t	N_tS_t	L_tS_t	SRM
1	LIN	3	*			*		*	0.0130
2	THRESH	7	*	*		*	*		0.0120
3	THRESH	7	*		*	*		*	0.0121
4	THRESH	5	*			*	*		0.0140
5	THRESH	5	*			*		*	0.0140
6	THRESH (a_1, a_2)	6	*			*	*		0.0107
7	THRESH (a_1, a_2)	6	*			*		*	0.0108
8	THRESH (a_1, a_2)	6	*	*			*		0.0110

from Jacobson et. al. (2004), because SRM singles out also a model without threshold (i.e., model 1) as a reasonable candidate for the ibex case. It is to be noted that models 2 and 3, which stand out in terms of the estimated prediction risk, are much less parsimonious than model 1, being characterized by 7, instead of 3, parameters.

To improve the robustness of our analysis, we further investigate the five best models via cross-validation (CV). The idea underlying CV is to split the dataset into a calibration set of size $(q - v)$ and a validation set of size v; then, (a) each model is identified on the calibration set, and (b) its square errors are measured on the validation set. Steps (a) and (b) are iterated for all the possible calibration sets of size $(q - v)$; the cross-validation criterion is the average, over these repetitions, of the square error measured at step b. For a comprehensive discussion on cross-validation, see Turchin (2003). In particular, we set $v = 1$, thus adopting the so-called leave-one-out cross-validation criterion (LOO-CV). We assess the model errors in terms of both predicted rates of increase $\hat{y}_{t+1}$ and predicted population abundances, computed as $\hat{N}_{t+1} = N_t \exp(\hat{y}_{t+1})$.

The results reported in Table 3 show that indeed models 2 and 3 have a better performance than the more parsimonious models 1 (3 parameters) and 4, 5 (5 parameters). Actually, LOO-CV analysis confirms the SRM ranking, with the only exception of model 1, which now performs worst. On the basis of both

SRM and leave-one-out analysis we are thus led to choose model 2 or 3.

Table 3
Leave-one-out performances of models considered in Table 2.

Model #	$\frac{1}{q}\sum_t (y_t - \hat{y}_t)^2$	$\frac{1}{q}\sqrt{\sum_t (N_t - \hat{N}_t)^2}$
1	.0072	51
2	.0046	43
3	.0047	43
4	.0062	46
5	.0064	47
6	.0044	42
7	.0045	42
8	.0045	41

However, we must honestly admit that the differences among the performances of the 5 models in terms of both SRM and LOO-CV are not very large, whereas the differences in terms of degrees of freedom are quite important. One is thus led to think that perhaps the set of models considered by Jacobson et. al. (2004) might not include the one most appropriate for the ibex data. As a matter of fact, there are no particular reasons for the intrinsic rate of increase a to be independent of the snow depth threshold. Therefore, we introduce and analyze an additional subset of models, in which parameter a is estimated twice, like the remaining parameters. Namely, we estimate a value a_1 for $S < \overline{S}$ and a value a_2 for $S > \overline{S}$. The value of the snow threshold $\overline{S}$ is left unchanged. Widening the model base does not pose any problem as the calculations via SRM are quite simple and quick.

Considering further models has indeed a positive effect, since 6 new models out of 9 display an improvement of the SRM score over model 2, i.e., the best model according to the previous analysis. All of them have a smaller number (6) of parameters than model 2; so, considering a as dependent upon snow depth leads also to a simplification of the model. The structure and the SRM scores of the three best models with a double estimate of a are reported in Table 2 (model IDs: 6,7,8) .

This result is confirmed by LOO-CV analysis which is reported in Table 3. We thus finally propose model 6, best according to SRM. It is worthwhile to remark that model 6 is a modification of model 4 (the one chosen by Jacobson et. al. (2004)), having the same structure with the only difference that a depends upon snow depth. The entire LOO-CV simulation computed via model 6 is provided in Figure 9.

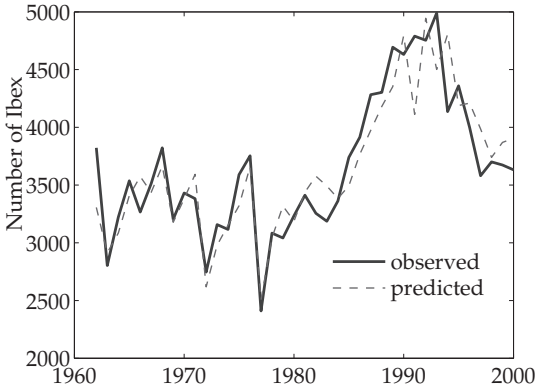


Figure 9. Leave-one-out cross validation of model 6, best according to SRM. Each predicted value at year t is obtained by using model 6 calibrated on the dataset from which the data of year t were excluded.

7 Conclusions

In the history of ecology one of the most debated issues has been the importance of density dependence, as opposed to exogenous factors, for population regulation. Both can actually play an important role in determining the abundances of populations through time (Begon and Mortimer , 1981). It is thus of great importance to have methods that can extract the maximum of information from the existing time series without a priori advocating a precise mechanism of population regulation. Many possible alternative models must be considered and tested by the population ecologist. The developments of statistical ecology in the past decade have exactly aimed at achieving this goal. In the present work, we have further gone into that direction. In fact, we have proposed the use of Structural Risk Minimization (SRM) as a simple and robust alternative to the traditional model selection criteria. It is simple, because in practice it requires that one makes a few regressions, penalizes the resulting fitting errors in the suitable way and compares the penalized errors between each other. It is robust, as it proves to work well under a wide variety of different conditions, characterized by different noise levels, dataset sizes, complexity of the model generating the time series.

As a first step, we have compared SRM and parametric bootstrapping (PBLR) repeating the simulation experiments described in Dennis and Taper (1994); our findings show that SRM recognizes the Malthusian model with the same, very high, probability of PBLR, but SRM outperforms PBLR in detecting the density-dependent Ricker model. Both conclusions are confirmed under almost all the many simulation settings investigated. In particular, the gap between SRM and PBLR in density-dependence detection is especially strong for short time series, and for simulations started close to the population equilibrium.

The subsequent comparison between SRM, FPE, AIC_c and SIC addresses the problem of choosing the best demographic model for a given population time series among a wide set of candidates. Our experiments show that SRM generally recognizes the model underlying the data with uniformly higher success than FPE, AIC_c and SIC; only in a few cases, its performance is neared by AIC_c and SIC. The advantage in using SRM is especially remarkable with short time series (10-20 samples), while for bigger sample sizes the difference between the criteria becomes tighter. This is important, because small datasets are the norm more than the exception in ecology: for instance, in a recent analysis (Sibly et. al. , 2005) involving more than 3200 populations, the reported average length of time series is 16.5 years, with standard deviation of 14.2 years.

On the basis of the presented results, we can moreover conclude that SRM leads fairly consistently to low out-of-sample prediction errors. It must be noted that we have assessed the performance of the various methods using one-step-ahead predictions; we hypothesize that several-steps-ahead predictions, which heavily penalize wrong model choices, would show an even greater advantage of SRM compared to traditional criteria. Robust multi-step-ahead predictions are of great practical interest in the design of population management policies and in the analysis of population viability for endangered species. Also, assessing the long-term response of populations to density and environmental forcing is of paramount importance, specially when we consider that global climate is ineluctably changing (Wigley , 2005) and affecting the distribution and persistence of many species (Parmesan et. al. , 1999, Thomas et. al. 2004).

Of course, our statements about SRM reliability for the analysis of ecological populations should be considered with some caution, as they are predicated on the assumption that data are generated by models that assume no population structure. More research should be devoted in the future to investigating the performances of the different model selection methods when age, stage and sex structure are important determinants of population dynamics. Time delays (Turchin , 1990) can also make a difference, although we do not expect that SRM can be at disadvantage in this case. Moreover, populations are often estimated rather than censused. Thus sampling error can be a source of variability. Dennis and Taper (1994) investigated the effect of census error on PBLR and concluded that it has little influence on test size and power. This result was subsequently challenged by Shenk et. al. (1998) and later by Freckleton et. al. (2006) . They used a different way of introducing sampling error in the model and showed that in this case PBLR can detect density dependence when the true dynamics is actually density independent. Certainly, the problem of the influence of census error on SRM performance is important and worth being explored in the future.

The greatest limitation in using SRM is at the moment the ignorance of the VC-dimensions of nonlinear models; although in principle SRM does not require candidate models to be linear, this ignorance prevents the analysis of nonlinear models via SRM. Therefore, a future research challenge will be the estimation of the VC-dimensions of the commonest nonlinear demographic models; it would allow the detection of density dependence other than the usual Ricker or Gompertz dependence. The problem can be solved using the methodology proposed in Vapnik et. al. (1994), and then improved by Shao and Cherkassky (2000). Our preliminary results based on this methodology show that, given two non-linear demographic models with the same number of parameters d, but different dynamic complexity, the model with the more complex dynamics has also a higher VC-dimension estimate. For instance, the Beverton-Holt model (2 parameters), which always tends monotonically towards an equilibrium, has a lower VC-dimension than the Ricker model, which can generate permanent oscillations or chaotic fluctuations. Remarkably, since both models have 2 parameters, their correction terms for model complexity would be the same with traditional model selection criteria such as SIC or AIC. So the ranking of the two models would be based only on the fit. On the contrary, SRM would allow one to trade-off model fit against complexity, because the penalization factor would be smaller for the Beverton-Holt model. Should the problem of VC-dimension estimation be solved for a large class of nonlinear demographies, SRM could be straightforwardly applied to a very wide variety of candidate models. As the Statistical Learning Theory, of which SRM is part, has been explicitly conceived for nonlinear models, the approach proposed in this paper seems apt to effectively dealing with the daunting complexity of ecological time series.

A further direction of future research is multi-model inference (MMI). Indeed, multi-model inference Burnham and Anderson (2004) can lead via model averaging to better predictive performance, as it avoids the model selection uncertainty inherent in the choice of any single, supposedly best model. MMI makes it unnecessary to defend the choice of a single model, as it incorporates several competing models, though with different weights. Model-averaged estimators have more reliable measures of precision and reduced bias when compared to estimators from a single "best" model. Recently, Brook and Bradshaw (2006) have used MMI (based on AIC and Akaike's weights) to analyze density-dependence in 1098 time series. They have shown that through such an approach one can also quantify the relative empirical support for a set of working hypotheses that encompass a wide range of real populations dynamical behaviors. Unfortunately, we are currently unaware of a model averaging theory based on SRM. Nevertheless, in some experiments we performed, we found the weighted average of the predictions issued by a set of models (the weights being proportional to the inverse of the SRM scores) to have a statistically lower error than the predictions returned by any of the single models in the set. However, this topic needs more careful analysis, including both

an investigation of the statistical meaning of the resulting estimator, and an experimental assessment of its performance, similar to that carried out in the present paper.

Acknowledgments

We thank A. Provenzale and J. von Hardenberg for helpful comments and discussion on the manuscript. B. Kendall's suggestions helped improve the quality of the manuscript. This work was developed under the international mobility program "Consequences and impacts of global climate change on the management and conservation of natural resources", Italian Ministry of University and Research.

References

Akaike, H. 1970. Statistical predictor identification. - Ann. Inst. Stat. Math. 22: 203-217.

Akaike, H. 1973. Information Theory as an extension of the maximum likelihood principle. - In: B.N.Petrov, and Csaki, F. (ed.), Second International Symposium on Information Theory. , pp. 267-281.

Begon, M. and Mortimer, M. 1981. Population ecology. - Sinauer Assoc..

Beverton, R. and Holt, S. 1957. On the Dynamics of Exploited Fish Populations. - H.M. Stationery Office.

Brook, B. and Bradshaw, C. 2006. Strength of evidence for density dependence in abundance time series of 1198 species. - Ecology 87: 1445-1451.

Bulmer, M. 1975. The statistical analysis of density dependence. - Biometrics 31: 901-911.

Burnham, K. and Anderson, D. 2003. Model Selection and Multi-Model Inference: a Practical Information-Theoretic Approach. - Springer.

Burnham, K. and Anderson, D. 2004. Multimodel inference - understanding AIC and BIC in model selection. - Sociological Methods & Research 33: 261-304.

Cherkassky, V., Shao, X., Mulier, F. and Vapnik, V. 1999. Model complexity control for regression using VC generalization bounds. - IEEE Trans. on Neural Networks 10: 1075-1089.

Cherkassky, V. and Mulier, F. 1998. Learning from data. - Wiley-Interscience.

Corani, G. and Gatto, M. 2005. Model selection in demographic time series using VC-bounds. - Ecological Modelling 191: 186-195.

Corani, G. and Gatto, M. 2006. VC-dimension and Structural Risk Minimization for the analysis of nonlinear ecological models and time series. - Appl. Math. and Comp. 176: 166-176.

Courchamp, F., Clutton-Brock, T. and Grenfell, B. 1999. Inverse density dependence and the Allee effect. - Trends in ecology & evolution 14: 405-410.

Dennis, B., Taper, M. and Kemp, W. 1998. Joint density dependence. - Ecology 79: 426-441.

Dennis, B. and Otten, M. 2000. Joint effects of density dependence and rainfall on abundance on san joaquin kit fox. - J. Wild. Management 64: 388-400.

Dennis, B. and Taper, M. 1994. Density dependence in time series observation of natural populations: estimation and testing. - Ecological Monographs 64: 205-244.

Efron, B. and Tibshirani, J. T. 1993. An Introduction to the Bootstrap. - Chapman & Hall.

Forster, M. 2000. Key concepts in model selection: Performance and generalizability. - J. Math. Psychology 44: 205-231.

Freckleton, R., Watkinson, R., Green, R. E. and Sutherland, W. 2006. Census error and the detection of density dependence. - J. of Animal Ecology 75: 837-851.

Hooten, M. 1995. Distinguishing forms of statistical density dependence and independence in animal time series data using Information Criterion. - .

Jacobson, A., Provenzale, A., Von Hardenberg, A., Bassano, B. and Festa-Bianchet, M. 2004. Climate forcing and density dependence in a mountain ungulate population. - Ecology 85: 1598-1610.

Lande, R., Engen, S. and Saether, B. 2003. Stochastic population dynamics in ecology and conservation. - Oxford University Press..

May, R. 1974. Biological populations with nonoverlapping generations: stable points, stable cycles and chaos. - Science 186: 645-647.

May, R. and Oster, G. 1976. Bifurcations and dynamic complexity in simple ecological models. - American Naturalist 110: 573-599.

Parmesan, C., Ryrholm, N., Stefanescu, C., Hillk, J. K., Thomas, C. D., Descimon, H., Huntley, B., Kaila, L., Kullberg, J., Tammaru, T., Tennent, W. J., Thomas, J. A. and Warren, M. 1999. Poleward shifts in geographical ranges of butterfly species associated with regional warming. - Nature 399: 579-583.

Peek, J., Dennis, B. and Hershey, T. 2002. Predicting population trends of mule deer. - J. Wild. Management 66: 729-736.

Pollard, E., Lakhani, K. and Rothery, P. 1987. The detection of density dependence from a series of annual censuses. - Ecology 68: 2046-2055.

Raftery, A. 1995. Bayesian model selection in social research (with Discussion). - Sociological Methodology 25: 111-196.

Richards, S. A. 2005. Testing ecological theory using the information-theoretic approach: examples and cautionary results. - Ecology 86: 2805-2814.

Ricker, W. 1954. Stock and recruitment. - J. of the Fisheries Research Board of Canada 11: 559-623.

Ricker, W. and Foerster, R. 1948. Computation of fish production. - Bull. Bingham Oceanogra. Coll 11: 173-211.

Schwarz, G. 1978. Estimating the dimension of a model. - Ann. Stat. 6: 461-

464.
Shao, X. and Cherkassky, V. 2000. Measuring the VC-dimension using Optimized Experimental Design. - Neural Computation 12: 1969-1986.
Shenk, M., White, G. and Burnham, K. P. 1998. Sampling-variance effects on detecting density dependence from temporal trends in natural populations. - Ecological Monographs 68: 445-463.
Sibly, R. M., Barker, D., Denham, M., Hone, J. and Pagel, M. 2005. On the Regulation of Populations of Mammals, Birds, Fish, and Insects. - Science 309: 607-610.
Strong, D., Whippler, A., Child, A. and Dennis, B. 1999. Model selection for a subterranean trophic cascade: root-feeding caterpillars and entomopathogenic nematodes. - Ecology 80: 2750-2761.
Taper, M. 2004. Model identification from many candidates. - In: Taper, M. L. and Lele, S. R. (ed.), The nature of scientific evidence. University of Chicago Press, pp. .
Taper, M. and Gogan, P. 2002. The northern Yellowstone elk: density dependence and climatic conditions. - J. Wild. Management 66: 106-122.
Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., Erasmus, B. F., De Siqueira, M. F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., Van Jaarsveld, A. S., Midgley, G. F., Miles, L., Ortega-Huerta, M. A., Peterson, A. T., Phillips, O. L. and Williams, S. E. 2004. Extinction risk from climate change. - Nature 427: 145-148.
Turchin, P. 1990. Rarity of density dependence or population regulation with lags. - Nature 344: 660-663.
Turchin, P. 2003. Complex population dynamics: a theoretical/empirical synthesis. - Princeton University Press.
Vapnik, V., Levin, E. and Cun, Y. 1994. Measuring the VC-dimension of a learning machine. - Neural Computation 6: 851-876.
Vapnik, V. 1999. An Overview of Statistical Learning Theory. - IEEE Trans. on Neural Networks 10: 988-999.
Verhulst, P. 1838. Recherches mathematiques sur la loi d'accrossement de la population. - Memoirs de l'Academie Royal Bruxelles 18: 1-38.
Wigley, T. M. L. 2005. The climate change commitment. - Science 307: 1766-1769.
Zeng, Z., Nowiersky, R., Taper, M., Dennis, B. and Kemp, W. 1998. Complex population dynamics in the real world: modeling the influence of time-varying parameters and time lags. - Ecology 79: 2193-2209.
Zucchini, W. 2000. An introduction to model selection. - J. Math. Psychology 44: 41-61.