

Model selection in demographic time series using VC-bounds

Giorgio Corani*, Marino Gatto

Dipartimento di Elettronica ed Informazione, Politecnico di Milano, Via Ponzio 34/5, 20133 Milano, Italy

Available online 22 September 2005

Abstract

The problem of distinguishing density-independent (DI) from density-dependent (DD) demographic time series is important for understanding the mechanisms that regulate populations of animals and plants. We address this problem in a novel way by means of Statistical Learning Theory (SLT); SLT is built around the idea of VC-dimension, a complexity index for classes of parameterized functions. Though VC-dimensions of nonlinear models are generally unknown, in the linear case VC-dimension actually corresponds to the number of free parameters; this allows one to straightforwardly apply the model selection framework developed within SLT, and called Structural Risk Minimization (SRM). We generate noisy artificial time series, both DI and DD, and use SRM to recognize the model underlying the data, choosing among a suite of both density-dependent and independent demographics. We show that SRM significantly outperforms traditional model selection approaches, such as the Schwartz Information Criterion and Final Prediction Error in recognizing both density-dependence and independence.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Model selection; Density-dependence; VC-dimension; Structural risk minimization; Time series analysis

1. Introduction

To recognize whether a population is growing in a density-dependent or independent way is of great practical importance in the design of proper policies for sustainable management and exploitation of natural populations. In fact, statistically distinguishing density-dependent from independent time series is of paramount interest for predicting future population abundances and understanding the mechanisms that regulate the species demography. Therefore, this topic

stimulated a great research effort over the past three decades (Bulmer, 1975; Pollard et al., 1987; Dennis and Taper, 1994).

Earlier works were based on hypothesis-testing approaches and contrasted a single density-independent model (usually Malthusian) with a single alternative density-dependent one (often the Ricker model). A milestone in this context is for instance the work of Dennis and Taper (1994), who proposed a powerful hypothesis testing framework based on parametric bootstrapping of likelihoods ratios. Despite their statistical soundness, hypothesis testing frameworks often suffered from the problem of low power, and were therefore recognized as conveying only limited information (Zeng et al., 1998). In fact, the diversity in pat-

* Corresponding author. Tel.: +39 02 2399 3562; fax: +39 02 2399 3412.

E-mail address: corani@elet.polimi.it (G. Corani).

terns of natural population regulation can be hardly addressed by comparing just a couple of models and, 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).

To overcome such limitations, several authors (Zeng et al., 1998; Strong et al., 1999; Dennis and Otten, 2000; Taper and Gogan, 2002) have proposed the use of information criteria (IC) to choose the best among a suite of alternative models including both density-independent and density-dependent demographies. Generally speaking, ICs address the model selection by choosing the model that minimizes the product of the squared deviation from data and a penalization factor, which is an increasing function of the ratio of free parameters d to the size q of the dataset; the rationale for such a penalization is that an optimal trade-off should be found between the quality of data fitting and model complexity. Different ICs provide different expressions of the penalization factor, obtained under different hypotheses; however, they are all (a) based on asymptotic arguments – which therefore hold just for large datasets – underlying a set of common assumptions such as (b) the linearity of the target function, which (c) must be contained in the set of the candidate approximating functions. As a consequence of these restricting hypotheses, ICs are very often applied even if their constitutive assumptions are not strictly met. In the literature, the Final Prediction Error (FPE) and in particular the Schwartz Information Criterion (SIC) appear to be the most widely used by ecologists (Strong et al., 1999; Dennis and Otten, 2000; Zeng et al., 1998; Taper and Gogan, 2002).

As a viable alternative to classical ICs, we propose the use of the model selection framework developed within Statistical Learning Theory (SLT). SLT (Vapnik, 1995) is derived under very general hypotheses, such as finite sample settings and nonlinear estimation, and is built around the idea of VC-dimension h , a complexity index for classes of functions. In the case of linear estimators, VC-dimension is known and actually corresponds to the number of free parameters, i.e., $h \equiv d$; on the contrary, VC-dimensions of nonlinear models are generally unknown a priori and this constitutes a major obstacle to a wide application of SLT findings. The challenge of estimating the VC-dimension of nonlin-

ear models can be addressed through the methodology originally proposed by Vapnik et al. (1994), although little applied work has been done on this topic up to now.

The model selection framework developed within SLT is called Structural Risk Minimization (SRM); it provides an analytical upper bound on the future error of a given model. The VC-bound is the product of the training error and a penalization factor, which depends on the VC-dimension h and on the sample size q . The model which minimizes such a bound is finally selected.

With reference to linear regression problems, it has been shown (Cherkassky et al., 1999a) that SRM can consistently outperform traditional Information Criteria for different dataset sizes and noise levels; the performance gap in favor of SRM is especially strong on small datasets, which are usual in ecological modelling.

As far as we know, this paper uses for the first time SRM to detect density-dependence, comparing its effectiveness against SIC and FPE. We simulate different demographic models – with different parametric settings – and then corrupt the simulated data with different levels of noise. For each noisy simulation, we identify different candidate models (including the one underlying the time series), and then select the best one according to the different model selection criteria. This way, we statistically assess the skill of each model selection criterion in recognizing the model which really underlies the data.

Our experimental results show that SIC may sometimes favor too parsimonious models (which implies that the Malthusian demography can be chosen even if the time series is a density-dependent one) while, on the contrary, FPE may favor overparameterized models (i.e., it may select a density-dependent demography even if it is not the case, or include useless covariates in the chosen model). SRM is well-balanced, and in fact it shows the best ability in choosing the appropriate model complexity.

The paper is organized as follows: Section 2 presents the suite of demographic models considered, Section 3 details the different model selection approaches, Section 4 explains the experimental methodology, and Section 5 reports the obtained results. In Appendix A, we sketch the theoretical definition of VC-dimension.

2. The demographic models of the suite

The considered demographic models are linear with respect to the parameters, i.e. the output variable is given by a weighted sum of the input variables. In particular, the models return as output the instantaneous rate of population increase between years, i.e., $Y_{t+1} = \ln(N_{t+1}/N_t)$, where N_t denotes the population abundance at time t . Input variables are, for instance, the abundance in the previous year and exogenous driving forces, such as temperature, precipitation, etc.

We introduce the models in increasing order of complexity, the next being obtained by adding a linear term in the previous. The suite is as follows:

- the Malthusian model (M):

$$\ln(N_{t+1}/N_t) = a \tag{1}$$

which is the only density-independent model;

- the Ricker model (R):

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

whose only nontrivial equilibrium corresponds to $\bar{N} = -a/b$. Also, note that b is a scale parameter. In fact, setting $Z_t = bN_t$, we obtain $\ln \frac{Z_{t+1}}{Z_t} = a + Z_t$.

- the Ricker model with one covariate (RI):

$$\ln(N_{t+1}/N_t) = a + bN_t + cX_1(t) \tag{3}$$

where $X_1(t)$ is an exogenous forcing variable, such as a climatic indicator, which may affect the population dynamics;

- the Ricker model with two covariates (RII):

$$\ln(N_{t+1}/N_t) = a + bN_t + cX_1(t) + dX_2(t) \tag{4}$$

where $X_2(t)$ is a further exogenous variable.

The above presented models have VC-dimension h (see Appendix A) that corresponds to the number of free parameters d , because all the equations are linear. These models are often used as alternative candidates when one has to identify the demographic mechanism underlying the available field data, and are usually compared by SIC (Zeng et al., 1998; Dennis and Otten, 2000; Taper and Gogan, 2002; Peek et al., 2002).

3. The model selection problem

From an abstract viewpoint we can think of the model selection problem as the problem of approximating the functioning of a true system; such a system receives an input vector \mathbf{x} , characterized by a probability distribution $P(\mathbf{x})$ and correspondingly returns an output y , according to the conditional distribution $P(y|\mathbf{x})$. Both $P(\mathbf{x})$ and $P(y|\mathbf{x})$ are unknown. We assume that the system is represented by the unknown relationship:

$$y = g(\mathbf{x}) + \epsilon \tag{5}$$

where ϵ is an independent identically distributed zero mean random noise.

A model selection procedure is aimed at choosing the best approximating function among a set of several candidates $f_j(\mathbf{x}, \omega)$, where ω denotes the parameters specifying the function, and the subscript j refers to one of different classes of functions. For example, class j might be a polynomial of degree j .

The choice is based on a finite number q of samples (\mathbf{x}_i, y_i) , $i = 1, \dots, q$. If, as usual, the quality of the approximation is measured through the squared error, the optimal approximating function should in principle minimize the following prediction risk functional:

$$R_j(\omega) = \int (y - f_j(\mathbf{x}, \omega))^2 dP(\mathbf{x}, y) \tag{6}$$

which is however unknown because the joint probability distribution function $P(\mathbf{x}, y) = P(y|\mathbf{x})P(\mathbf{x})$ is unknown.

On the other hand, what can be experimentally measured by using the q samples is the empirical risk:

$$R_j(\omega)_{\text{emp}} = \frac{1}{q} \sum_{i=1}^q (y_i - f_j(\mathbf{x}_i, \omega))^2 \tag{7}$$

Information criteria attempt to estimate the unknown prediction risk (6) as the known empirical risk (7), penalized by some measure of the model complexity. Once an accurate estimate of the prediction risk is found, the model that minimizes the estimated prediction risk with respect to both the class j of functions and the parameters defining each function inside the class is chosen. In general, for a function f_j having d_j

free parameters, ICs take the form:

$$\text{estimated risk}(f_j) = R_j(\omega)_{\text{emp}} r(p_j) \quad (8)$$

where $r(p)$ is the penalization function and p_j denotes the ratio d_j/q . In this paper we consider the Final Prediction Error (Akaike, 1970):

$$\begin{aligned} \text{FPE estimated risk}(f_j) \\ = R_j(\omega)_{\text{emp}} [(1 + p_j)/(1 - p_j)] \end{aligned} \quad (9)$$

and the Schwartz Information Criterion (Cherkassky and Mulier, 1998; Cherkassky et al., 1999a)

$$\begin{aligned} \text{SIC estimated risk}(f_j) \\ = R_j(\omega)_{\text{emp}} \left[1 + (\ln(q)/2) p_j (1 - p_j)^{-1} \right] \end{aligned} \quad (10)$$

These classical approaches are motivated by asymptotic arguments ($q \rightarrow \infty$) for linear models and indeed risk estimates provided by FPE and SIC are asymptotically equivalent. They also assume that the target function $g(\mathbf{x})$ is contained in the set of candidate approximating functions $f_j(\mathbf{x}, \omega)$. It is worthwhile to note that the experiments carried out in this paper will actually meet such an assumption.

These classical approaches can be contrasted to the VC-theory approach where, for a sample of finite length q , one can calculate a bound for the risk function (6). For “practical” regression problems, the following Structural Risk Minimization holds with probability $(1 - (1/\sqrt{q}))$ (Cherkassky et al., 1999a):

$$\text{SRM estimated risk}(f_j) \leq R_{\text{emp}} \left[1 - \sqrt{p_j - p_j \ln p_j + \ln(q)/2q} \right]_+^{-1} \quad (11)$$

where $p_j = h_j/q$ (h_j is the VC-dimension of the j -th class of functions). If the models are linear, h_j coincides with the number of free parameters, so $p_j = d_j/q$. The model that minimizes the right-hand-side of (11) is finally selected. Therefore, in practice SRM is yet another way of penalizing the empirical risk R_{emp} .

With reference to our application, the problem of predicting the rate of demographic increase between year t and year $t + 1$ can be obtained by setting

$$y = \ln(N_{t+1}/N_t), \quad \mathbf{x} = [N_t, X_1(t), X_2(t)] \quad (12)$$

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

Actually, VC-theory has been formalized under the assumption of \mathbf{x} and y being independent identically independent distributed, which is not true for our problem. Nevertheless, the effectiveness of the approach also in this unusual context is clearly shown by the experimental findings reported later.

4. The experimental methodology

To test the methodology, we generate artificial time series, simulating the models and adding noise to the simulations. By adopting a log-normal multiplicative noise, the equations of a noisy simulation can be written as:

$$N_{t+1} = N_t \exp(a + bN_t + cX_1(t) + dX_2(t) + n\text{WN}) \quad (13)$$

where n is the noise level and WN a standard normal white noise ($\mu = 0, \sigma^2 = 1$). Depending on the type of the simulated model, coefficients b, c or d can be set to 0. Covariates X_1 and X_2 are also generated as standard normal white noises.

A simulation is therefore characterized by the following *simulation settings*:

- the initial condition N_0 ;
- the model coefficients (a, b, c, d) ;
- the noise level n ;
- the simulation length q .

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

- Malthusian model: $N_0 = 100; a = [0.5; 1; 1.5]; n = [0.05; 0.1; 0.25; 0.5]; q = [10; 20; 50; 100]$;
- Ricker model: since parameter $b < 0$ does not influence the results (see e.g. Dennis and Taper (1994)) being just a scaling parameter, we fix $b = -0.01$. On the other hand, simulations started in correspondence of the model equilibrium ($N_0 = -a/b$) make the model recognition especially difficult (Dennis and Taper, 1994), and therefore deserve particular consideration. $N_0 = 100$ and $-a/b; a = [.5; 1; 1.5]; b = -0.01; n = [.05; .1; .25; .5]; q = [10; 20; 50; 100]$;

- Ricker I model: $N_0 = 100; a = 1; b = -0.01; c = [0.01; 0.1; 0.5]; n = [.05; .1; .25; .5]; q = [10; 20; 50; 100];$
- Ricker II model: $N_0 = 100; a = 1; b = -0.01; c = 0.01; d = [0.01; 0.1; 0.5]; n = [0.05; 0.1; 0.25; 0.5]; q = [10; 20; 50; 100].$

For each simulation setting, we perform 500 model recognition experiments. The model selection procedure can be summarized as follows:

- (1) *simulation*: 500 noisy simulations are performed according to Eq. (13), using the given simulation setting;
- (2) *identification*: the time series $Y_{t+1} = \ln(N_{t+1}/N_t)$ is calculated and the parameters of the four candidate models are estimated by means of linear least squares;

- (3) *acceptability check*: density-dependent models (i.e., all but the Malthusian) should have a negative parameter b , meaning that intra-specific competition negatively affects the population growth rate; thus, models with a positive estimate of b are discarded;
- (4) *model selection*: FPE, SIC and SRM are used in order to choose the model from among the set of candidates.

5. Results

Table 1 summarizes the results by reporting the average proportion a model is selected using the different model selection criteria. The average is taken over all the different simulation settings (model parameters, simulation length, noise level). More detailed results

Table 1
Average ability of FPE, SIC and SRM in recognizing the model underlying the artificially generated time series

Simulated model	Selected model (FPE)				Selected model (SIC)				Selected model (SRM)			
	M (%)	R (%)	RI (%)	RII (%)	M (%)	R (%)	RI (%)	RII (%)	M (%)	R (%)	RI (%)	RII (%)
M	81	9	5	5	99	1	0	0	98	2	0	0
R($N_0 = 100$)	0	72	15	13	19	78	2	1	4	92	3	1
R($N_0 = -\frac{a}{b}$)	2	71	15	12	49	49	1	1	10	86	3	1
RI	0	23	58	18	19	29	49	3	3	36	58	3
RII	0	7	21	72	6	24	18	51	2	17	24	58

Percentages in bold refer to the model that really generated the data (the higher, the better). See text for symbols.

Table 2
Detailed results for the recognition of the simple Malthusian model with different settings

	Recognized models (FPE)				Recognized models (SIC)				Recognized models (SRM)			
	M (%)	R (%)	RI (%)	RII (%)	M (%)	R (%)	RI (%)	RII (%)	M (%)	R (%)	RI (%)	RII (%)
<i>a</i>												
0.50	79	10	6	5	99	1	0	0	97	3	0	0
1.00	81	8	6	5	99	1	0	0	98	2	0	0
1.50	84	7	5	4	99	1	0	0	98	2	0	0
<i>n</i>												
0.05	84	7	5	4	99	1	0	0	99	1	0	0
0.10	83	7	5	5	100	0	0	0	98	2	0	0
0.25	81	9	5	5	99	1	0	0	98	2	0	0
0.50	76	12	6	6	99	1	0	0	97	3	0	0
<i>q</i>												
9	76	8	7	9	97	1	1	1	95	4	1	0
19	81	9	5	5	100	0	0	0	97	3	0	0
49	84	9	4	3	100	0	0	0	99	1	0	0
99	84	8	5	3	100	0	0	0	100	0	0	0

Percentages in bold refer to the model that really generated the data (the higher, the better).

Table 3
Detailed results for the recognition of the simple Ricker model ($N_0 = 100$) with different settings

	Recognized models (FPE)				Recognized models (SIC)				Recognized models (SRM)			
	M (%)	R (%)	RI (%)	RII (%)	M (%)	R (%)	RI (%)	RII (%)	M (%)	R (%)	RI (%)	RII (%)
<i>a</i>												
0.50	1	72	15	12	55	42	2	1	7	89	3	1
1.50	0	72	15	13	1	95	2	2	1	95	3	1
<i>n</i>												
0.05	0	72	15	13	0	96	2	2	0	96	3	1
0.10	0	71	16	13	23	74	2	1	0	95	4	1
0.25	0	72	15	13	43	54	1	2	5	92	3	0
0.50	1	72	16	11	48	50	2	0	10	86	3	1
<i>q</i>												
9	2	61	17	20	18	70	6	6	10	82	6	2
19	0	71	17	12	25	75	0	0	4	90	5	1
49	0	77	13	10	33	67	0	0	0	98	2	0
99	0	78	13	9	0	100	0	0	0	99	1	0

Percentages in bold refer to the model that really generated the data (the higher, the better).

are reported in Tables 2–6 that evidence the sensitivity of the recognition proportion to variations of each single parameter, by pooling the experiments that share the same value of a given parameter. For instance, the row ($a = 0.5$) in Table 2 refers to the average result obtained on $8000 = 500 \times 4$ (different values of n) \times 4 (different values of q) experiments.

As for the Malthusian model (Table 1), it is almost always recognized both by SIC and SRM (99 and 98%), while FPE fails about 20% of times, selecting a density-dependent demography. Looking at the detailed results

(Table 2), one notes that SIC and SRM recognize correctly the Malthusian demography, as they are in practice insensitive to any variation of the noise level n , the simulation length q or the drift parameter a ; however, FPE too shows little sensitivity of its performances to changes in one of these parameters.

As for the ability to recognize the Ricker model (Table 3), SRM (92%) strongly outperforms FPE and SIC (78 and 72%, respectively); for whatever value of a , n and q , a consistent advantage of SRM over both FPE and SIC is found. Remarkably, FPE and SRM ap-

Table 4
Detailed results for the recognition of the simple Ricker model ($N_0 = -a/b$) with different settings (simulations started at the model equilibrium)

	Recognized models (FPE)				Recognized models (SIC)				Recognized models (SRM)			
	M (%)	R (%)	RI (%)	RII (%)	M (%)	R (%)	RI (%)	RII (%)	M (%)	R (%)	RI (%)	RII (%)
<i>a</i>												
0.50	4	70	14	12	94	4	1	1	23	74	2	1
1.00	1	72	15	12	47	50	2	1	6	90	3	1
1.50	0	72	16	12	5	92	2	1	2	94	3	1
<i>n</i>												
0.05	2	71	15	12	51	47	1	1	10	86	3	1
0.10	2	71	15	12	51	47	1	1	10	86	3	1
0.25	1	71	15	13	49	49	1	1	10	86	3	1
0.50	2	71	15	12	46	52	1	1	10	86	3	1
<i>q</i>												
9	5	59	17	19	50	42	4	4	31	63	5	1
19	1	72	15	12	51	49	0	0	10	84	5	1
49	0	75	15	10	48	52	0	0	0	97	2	1
99	0	78	14	8	47	53	0	0	0	99	1	0

Percentages in bold refer to the model that really generated the data (the higher, the better).

Table 5

Detailed results for the recognition of the Ricker (I) model with one covariate and with different settings

	Recognized models (FPE)				Recognized models (SIC)				Recognized models (SRM)			
	D (%)	R (%)	RI (%)	RII (%)	D (%)	R (%)	RI (%)	RII (%)	D (%)	R (%)	RI (%)	RII (%)
<i>c</i>												
0.01	1	57	28	14	45	52	2	1	6	82	11	1
0.10	0	12	68	20	12	33	52	3	2	25	70	3
0.50	0	0	79	21	0	2	94	4	0	1	95	4
<i>n</i>												
0.05	0	11	69	20	16	15	65	4	1	22	74	3
0.10	0	20	62	18	18	17	62	3	3	30	64	3
0.25	0	28	54	18	23	36	39	2	3	39	55	3
0.50	0	35	48	17	20	46	32	2	4	54	41	1
<i>q</i>												
9	0	27	48	25	23	23	45	9	11	36	50	3
19	0	27	55	18	20	27	51	2	0	39	56	5
49	0	22	63	15	18	31	51	0	0	36	62	2
99	0	18	67	15	15	34	51	0	0	33	66	1

Percentages in bold refer to the model that really generated the data (the higher, the better).

pear much more robust than SIC, whose performance strongly worsens when high noise level or low values of the demographic parameter *a* are used, because it wrongly tends to choose the Malthusian model. On the other hand, all the three criteria take advantage in a similar way (improvement between 20 and 30% points) from increasing the simulation length.

If the Ricker simulation is started at the equilibrium (Table 4), FPE and SRM behave in a quite robust way,

because successfully recognition decrease just a few points. On the contrary, SIC displays a strong worsening (about 30 points) with a disappointing behavior in particular when *a* is low (4% average selection success for *a* = 0.5). Interestingly, none of the criteria shows in this case any worsening as the noise level increases. The explanation is as follows: noise elicits fluctuations around the stable equilibrium of the Ricker thus making the model recognition easier; such

Table 6

Detailed results for the recognition of the Ricker model (II) with two covariates and with different settings

	Recognized models (FPE)				Recognized models (SIC)				Recognized models (SRM)			
	D (%)	R (%)	RI (%)	RII (%)	D (%)	R (%)	RI (%)	RII (%)	D (%)	R (%)	RI (%)	RII (%)
<i>d</i>												
0.01	0	13	54	33	11	34	50	5	2	24	63	11
0.1	0	9	7	84	7	36	4	53	2	23	8	67
0.5	0	0	1	99	0	3	0	97	0	4	1	95
<i>n</i>												
0.05	0	0	13	87	0	0	30	70	0	0	24	76
0.10	0	1	22	77	2	2	30	66	0	3	31	66
0.25	0	8	24	68	11	40	9	40	2	18	26	54
0.50	0	21	22	57	13	55	2	30	3	47	15	35
<i>q</i>												
9	0	15	20	65	12	18	17	53	6	30	22	42
19	0	10	22	68	8	22	19	51	0	19	23	58
49	0	4	21	75	4	27	18	51	0	11	25	64
99	0	1	19	80	1	30	18	51	0	7	26	67

Percentages in bold refer to the model that really generated the data (the higher, the better).

an effect balances the usual negative impact of the noise.

As concerns the sensitivity of the recognition of the Ricker I model (Table 5), SRM and FPE provide the best overall result (58%), while SIC is significantly worse (49%); moreover, while FPE and SRM choose a density-dependent model almost always, SIC chooses the Malthusian model about 20% of times. As can be expected, the value of the covariate coefficient in the original model causes the largest variations of the percentage recognition of all the criteria. In particular, as the covariate coefficient is very small ($c = 0.01$), FPE is able to recognize the model more than SRM (which chooses often the simple Ricker model), and SIC (which chooses often the Ricker or the Malthusian model). As c increases, SRM becomes however the best performing approach.

The recognition of the Ricker II model (Table 6) shows a prevalence of FPE (about 70%) over SRM (about 56%) and SIC (about 52%). Also in this case all the model selection criteria show a higher sensitivity to the covariate coefficient than to the any other parameter. In this case, the underlying tendency of FPE in choosing complex models is favorable, and the advantage is especially significant when the covariate coefficient is low (22% points more than SRM and 28 than SIC for $d = 0.01$).

6. Conclusions

In this work, we address the density-dependence detection problem by comparing the performances provided by the traditional and well-known SIC and FPE model selection criteria with SRM, the model selection criterion developed within the Statistical Learning Theory framework. As far as we know, this is the first time that SRM is used to tackle the problem of model selection in population ecology.

Since VC-dimensions of linear estimators are known to simply correspond to the number of free parameters, the SRM application with linear models is straightforward. In future works, it would be however of great interest to estimate also the VC-dimension of non-linear ecological models, by using the methodology proposed by Vapnik et al. (1994). This would allow the inclusion of further models in the suite that is tested for density-dependence.

The philosophy underlying our experiments can be summarized as follows: we simulate different demographic models, both density-dependent and independent, investigating for each model a wide variety of simulation settings (i.e., parameters of the simulated model, simulation length, noise level used to corrupt the data). We perform 500 different simulations for each simulation setting and on each simulation we identify a suite of different models. Then, we select one of them, using as choice criterion the lowest risk according to the FPE, SIC, SRM risk estimates. Finally, we assess the skill of each model selection criterion in recognizing the model which underlies the time series.

The overall outcome of our experiments can be summarized as follows: SIC appears prone to parsimonious model and, therefore, it is very effective in detecting density-independence; on the other hand, it may encounter significant failures when it has to detect density-dependence, or a covariate which impacts on the demography trend. On the contrary, FPE seems prone to parameterized models: in fact, it is less effective than SRM or SIC in detecting density-independence, but outperforms them when the most complex density-dependent model has to be recognized. SRM appear to be the best balanced criterion and indeed it provides the overall best performance. In fact, it performs as well as SIC in recognizing density-independence, and at the same time it is very effective in recognizing density-dependent models, as it is outperformed by FPE only in the recognition of the Ricker II model.

Such results allow one to conclude that SRM is a really viable approach for model selection.

Appendix A. The VC-dimension definition

Statistical learning theory is built around the idea of VC-dimension as a measure of the complexity of classes¹ of models. In this section we clarify the general definition of VC-dimension, and then analyze then the particular case of linear classifiers and regressors.

¹ By *class of models* we mean a set of models that have an identical mathematical expression, and that can differ only in the parameters value. For example, the models $y = 4x_1 + 2x_2$ and $y = x_1 + 3x_2$ belong to the same class, while $y = 4x_1 + 2x_2 + 3$ or $y = 4x_1 + 2x_2 + 2x_3$ do not.

The theoretical definition of VC-dimension is clearly stated for instance in (Burges, 1998):

The VC-dimension is a property of a specific class of functions $f(\mathbf{x}, \theta)$ (θ is a generic set of parameters). [...] If we consider the two-class pattern recognition-case, $f(\mathbf{x}, \theta) \in \{-1, 1\} \forall (\mathbf{x}, \theta)$. Now, if a given set of l points can be labeled in all the possible 2^l ways, and for each labeling, a member of the set $f(\mathbf{x}, \theta)$ can be found which correctly assigns those labels, we say that that set of points is shattered by that set of functions. The VC dimension for the set of functions $f(\mathbf{x}, \theta)$ is defined as the maximum number of training points that can be shattered by $f(\mathbf{x}, \theta)$. Note that, if the VC dimension is h , then there exists at least one set of h points that can be shattered, but in general it will not be true that every set of h points can be shattered.

Now, let us analyze how the theoretical definition applies to the case of linear classifiers. A linear classifier in the space $\mathbf{x} = (x_1, x_2, \dots, x_m)$ is given by:

$$g(\mathbf{x}, \mathbf{w}, b) = \text{sign}(\mathbf{x} \cdot \mathbf{w} + b)$$

with $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{w} \in \mathbb{R}^m$, $b \in \mathbb{R}$. Hence, each classifier contains $(m + 1)$ parameters (m weights w_i and the bias b); it is an *indicator function* since it labels a given point in the space \mathbf{X} in a binary way. For the sake of simplicity, we consider now a bi-dimensional space $\mathbf{X} = \mathbb{R}^2$; the class of classifiers is therefore given by

Table A.1

Possible binary classifications for a set of three points

Points	Classifications							
	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>
A	-1	-1	-1	-1	1	1	1	1
B	-1	-1	1	1	-1	-1	1	1
C	-1	1	-1	1	-1	1	-1	1

$\text{sign}(x_1 w_1 + x_2 w_2 + b)$. A couple of points (A, B) in \mathbf{X}^2 can be labeled as:

- ($A = 1; B = 1$)
- ($A = -1; B = -1$)
- ($A = 1; B = -1$)
- ($A = -1; B = 1$).

One can easily figure out a linear classifier that realizes all these four classifications on a set of two points. We say out that it *shatters* two points, since there exist *at least* one set of two points that can be separated in all the possible ways. Actually, all the set of two points can be separated in all the possible ways, since the classifier has two degrees of freedom.

A set of three points (A, B, C) can be labeled in $2^3 = 8$ possible ways, as listed in Table A.1. If we refer for instance to the points represented in Fig. 1(a–h), the class of linear classifiers can realize all the classifications. Hence, it shatters also three points. However,

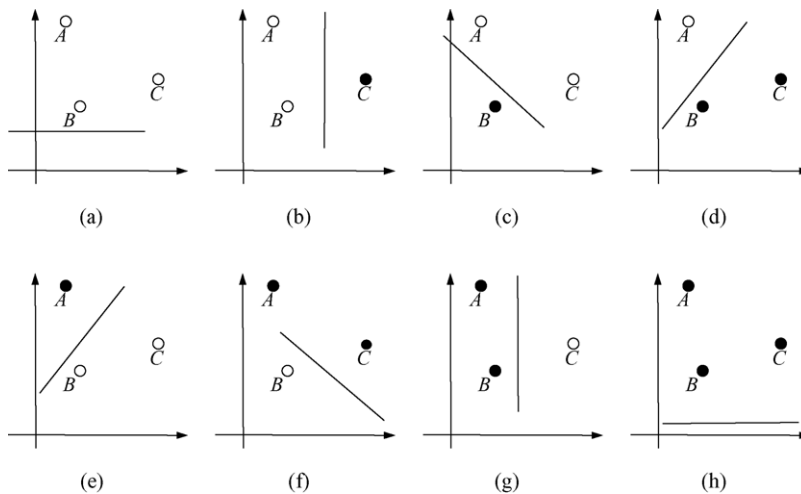


Fig. 1. Graphical representation of the classifications of Table A.1: linear classifiers in the space (x_1, x_2) shatter three points. Black circles correspond to values -1 , open circles to values $+1$.

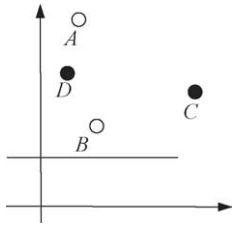


Fig. A.1. Linear classifiers in the space (x_1, x_2) cannot shatter any combination of four points.

not all the possible sets of three points can be separated; for instance, three points lying on the same straight line cannot.

However, a linear classifier in \mathbb{R}^2 does not shatter four points. For instance, no linear classifier can separate the points of Fig. A.1. More generally, it is possible to prove (Cherkassky et al., 1999) that a linear classifier in \mathbb{R}^m shatters $m + 1$ points. Therefore, linear classifiers have VC-dimension equal to $m + 1$. Similarly, the VC-dimension of a class of linear real valued regressors of type $g(\mathbf{x}, \mathbf{w}, b) = (\mathbf{w} \cdot \mathbf{x} + b)$, $\mathbf{x} = (x_1, x_2, \dots, x_m)$, is $(m + 1)$ (Cherkassky and Mulier, 1998).

Only in the linear case, the VC-dimension is guaranteed to actually equal the number of free parameters; indeed, the VC-dimension of a nonlinear function can be either larger or smaller than the number of parameters.

References

- Akaike, H., 1970. Statistical predictor identification. *Ann. Inst. Stat. Math.* 22, 203–217.
- Bulmer, M.G., 1975. The statistical analysis of density dependence. *Biometrics* 31, 901–911.
- Burges, J.C., 1998. A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discov.* 2, 121–167.
- Cherkassky, V., Mulier, F., 1998. *Learning from Data*. Wiley-Interscience.
- Cherkassky, V., Shao, X., Mulier, F., Vapnik, V., 1999. Model complexity control for regression using VC generalization bounds. *IEEE Trans. Neural Netw.* 10 (5), 1075–1089.
- Cherkassky, V., Shao, X., Mulier, F., Vapnik, V., 1999. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* 10 (5), 988–999.
- Dennis, B., Otten, M.R., 2000. Joint effects of density dependence and rainfall on abundance on san joaquin kit fox. *J. Wildl. Manage.* 64 (2), 388–400.
- Dennis, B., Taper, M.L., 1994. Density dependence in time series observation of natural populations: estimation and testing. *Ecol. Monogr.* 64 (2), 205–244.
- Peek, J.M., Dennis, B., Hershey, T., 2002. Predicting population trends of mule deer. *J. Wildl. Manage.* 66 (3), 729–736.
- Pollard, E., Lakhani, K., Rothery, P., 1987. The detection of density dependence from a series of annual censuses. *Ecology* 68, 2046–2055.
- Strong, D.R., Whippler, A.V., Child, A.L., Dennis, B., 1999. Model selection for a subterranean trophic cascade: root-feeding caterpillars and entomopathogenic nematodes. *Ecology* 80 (8), 2750–2761.
- Taper, M.L., Gogan, P.J., 2002. The northern yellowstone elk: density dependence and climatic conditions. *J. Wildl. Manage.* 66 (1), 106–122.
- Vapnik, V., Levin, E., Cun, Y., 1994. Measuring the VC-dimension of a learning machine. *Neural Comput.* 6, 851–876.
- Vapnik, V.N., 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag.
- Zeng, Z., Nowiersky, R., Taper, M.L., Dennis, B., Kemp, W.P., 1998. Complex population dynamics in the real world: modeling the influence of time-varying parameters and time lags. *Ecology* 79 (6), 2193–2209.