SORT 34 (2) July-December 2010, 239-248

Application of receiver operating characteristic (ROC) methodology in biological studies on marine resources: sex determination of *Paracentrotus lividus* (Lamarck, 1816)

Vicente Lustres-Pérez[1], María Xosé Rodríguez-Álvarez[2,3], María P. Pata[1]
Eugenio Fernández-Pulpeiro[1], Carmen Cadarso-Suárez[2,3]
Universidade de Santiago de Compostela (USC)

Abstract

The receiver operating characteristic (ROC) curve is usually used in biomedicine as an indicator of the accuracy of diagnostic tests. However, this measure of discrimination has been little used in other areas, such as animal biology or ecology. We present a novel application of an ROC analysis in which gonad colour was used to determine the sex of *Paracentrotus lividus* (Lamarck, 1816), a sea urchin of considerable commercial interest. A better classifier than gonad colour was obtained by transforming these colours through flexible logistic generalized additive models.

MSC: 6207, 62G08, 62G09, 62H30

Keywords: ROC, GAM, *Paracentrotus lividus*, bootstrap

1. Introduction

Paracentrotus lividus (Lamarck, 1816) is an echinoderm of high commercial value that is found along the coasts of Europe and North Africa. As this species is particularly abundant on the coast of Galicia (NW Spain), commercial harvesting began in the early 1980s. Although the reported annual average catch is in the order of 750 T (as per official

[1] Departamento de Zoología y Antropología Física, Universidade de Santiago de Compostela (USC), Spain.
[2] Unidad de Bioestadística, Departamento de Estadística e Investigación Operativa, USC, Spain.
[3] Instituto de Investigación Sanitaria de Santiago (IDIS), Santiago de Compostela, Spain.
Received: November 2009
Accepted: June 2010

data published by the Galician Regional Authority/*Xunta de Galicia*: www.pescadegalicia.com), real production is in fact higher, as a large volume of the catch goes undeclared. Currently, the *P. lividus* harvesting period in Galicia lasts from October to April, coinciding with the time of year when this species reaches sexual maturity. The spawning period on this stretch of coast usually begins at some point between the start and middle of spring (Catoira, 1995; Monteiro-Torreiro and Garcia-Martinez, 2003; Lustres-Pérez, 2006).

Although *P. lividus* is a dioecious species, with separate sexes, studies conducted by other authors report no method of determining the creature's sex externally, despite its outward display of a great variety of colours (Tortonese, 1965). Nevertheless, gonad colour has been linked to sex by some authors, though the majority of such studies have been of a descriptive nature (see e.g. Sellem and Guillou, 2007).

Commercial interest in this species lies in exploitation of the reproductive organs, the gonads. Female gonads are of higher quality than those of males, inasmuch as the former are reputed to have a better flavour and so be more palatable. This, in turn, means that clear criteria for sexual selection are vital for proper commercial and biological management.

From a statistical point of view, the discriminatory capacity of a given continuous or ordinal classifier, *Y* (gonad colour in our case), in terms of distinguishing between two alternative states, S1 and S2 (i.e., sex), is usually based on receiver operating characteristic (ROC) curve analysis (Metz, 1978; Swets and Pickett, 1982; Hanley and McNeil, 1982). The ROC curve is based on dichotomisation of the classifier *Y* by choosing a cut-off, such that values above this value will classify an individual as belonging to one of the states (say S1), and values below it as belonging to the alternative state (S2).

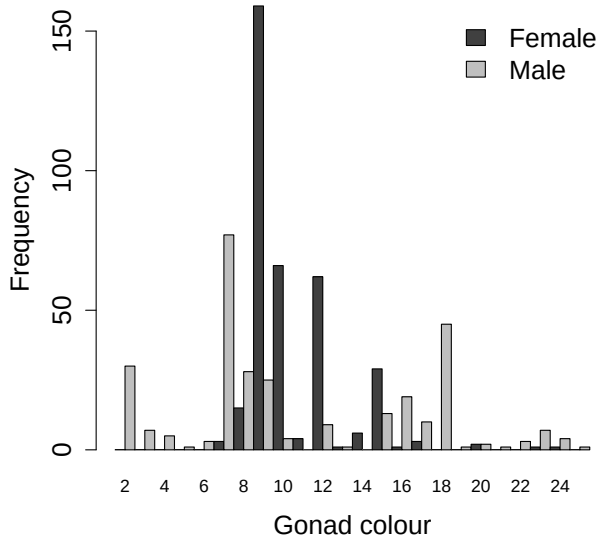


Figure 1: *Frequency of the different gonad colours selected according to sex.*

In many situations, however, the classification rule based on Y values that minimise the overall misclassification error is not necessarily the criterion used in ROC analysis. Figure 1 shows gonad colour frequency (codified numerically) for each sex. An irregular distribution of gonad colour can be observed, with a dominance of sexes in non-contiguous regions. Consequently, sex classification by means of a cut-off value is neither feasible nor logical. A modification of the classification rule is thus necessary. Indeed, if the discriminatory capacity of gonad colour is evaluated by means of the ROC curve, a lower discriminatory capacity will be obtained than that to be expected from a visual examination of Figure 1 (see Results, Section 3). Hence, use of such an analysis would lead to erroneous conclusions.

An intuitive solution to this problem, is to estimate the probability of belonging to one of the states as a function of the values of the marker Y (e.g., $P[S1|Y]$), and to base the classification on these probabilities, i.e., to transform the marker in such a way that the classification rules can be based on cut-off values.

This study proposes to model $P[S1|Y]$ by means of a generalised additive model (GAM, Hastie and Tibshirani, 1990) for binary data. GAMs are flexible non-parametric regression models that allow for much more accurate fitting of real data than do the parametric linear models usually used. Furthermore, in the case of the *P. lividus* data shown above, the use of a generalised linear model (GLM, McCullagh and Nelder, 1989) would not enable this probability to be correct modelled (see Section 3 for more details).

This paper is structured as follows: Section 2 outlines the statistical methodology; Section 3 reports the results of applying the proposed methodology to *P. lividus* data; and lastly, Section 4 concludes with a discussion.

2. Statistical methodology

Let Y be a continuous or ordinal classifier. Classification on the basis of Y of an individual as belonging to state S1 or S2 can be made by choosing a cut-off value, c, such that if $Y \geq c$ the observation is classified as S1, and if $Y < c$ it is classified as S2. Hence, each cut-off value chosen, c, will give rise to a true positive fraction (or sensitivity), $TPF(c) = P[Y \geq c|S1]$, and a false positive fraction (or 1-specificity), $FPF(c) = P[Y < c|S2]$. In such a situation, the ROC curve is defined as the set of all TPF-FPF pairs that can be obtained by a varying cut-off value c, $\{(FPF(c),\ TPF(c)),\ c \in (-\infty, \infty)\}$, or, equivalently, as the function of the form $ROC(t) = S_{S1}\left(S_{S2}^{-1}(t)\right)\ t \in (0,1)$, where S_{S2} and S_{S1} denote the survival functions of Y in the groups defined by states S1 and S2 respectively. Several indices can be used as summaries of the discriminatory capacity of the ROC curve. The area under the ROC curve (AUC) is most commonly used, taking values from 0.5 (no power of discrimination) to 1 (perfect power of discrimination). Generally, discrimination is deemed accurate where AUC exceeds 0.8.

As was illustrated in the Introduction with real *P. lividus* data, in many situations the classification rule based on the classifier Y that minimises the overall misclassification

error, is not the criterion used in ROC analysis. Moreover, the best Y-based classifier with a cut-off as the classification decision is that which is based on the conditional probability of one of the states (e.g., S1) given the values of Y (Neyman and Pearson, 1933; McIntosh and Pepe, 2002). Therefore, the best classifier, $\tilde{Y}$, can be expressed as:

$$\tilde{Y} \equiv f(Y) = P[S1|Y] \subset (0,1). \tag{1}$$

In practice, however, the function $f(\cdot)$ of (1) is not known, and its estimation would be required. In this study, we propose to model the function $f(\cdot)$ using a logistic GAM regression model as follows:

$$f(Y) = P[S1|Y] = g^{-1}(\alpha + h(Y)) = \frac{\exp(\alpha + h(Y))}{1 + \exp(\alpha + h(Y))}, \tag{2}$$

where $g(\cdot)$ is the logit link function (known) and $h(\cdot)$ is a smooth unknown function.

To date, several approaches to estimating the model (2) have been suggested in the statistical literature, e.g., methods based on penalised regression splines (Eilers and Marx, 1996; Wood, 2003) or the Bayesian versions of these (Lang and Brezger, 2004). Alternatively, the local scoring algorithm with kernel-type smoothers can be also used (McCullagh and Nelder, 1989; Wand and Jones, 1995).

In this paper, penalised regression combined with thin plate splines as smoothers (Wood, 2003) are proposed for the purpose of estimating the function $h(\cdot)$. A crucial step in estimating $h(\cdot)$ is choosing the smoothing parameter that controls the smoothness of the resultant estimate. In this paper, the optimal smoothing parameter is chosen automatically by use of the Un-Biased Risk Estimator criterion (UBRE) (Wood, 2004).

Once the model (2) is fitted, the estimated probabilities are used as the new classifier, and the ROC curve and the corresponding AUC are obtained. In addition, bootstrap regression techniques (Efron and Tibshirani, 1993) are used to construct a 95% bootstrap confidence interval (CI) for the AUC.

3. Results

3.1. Materials and methods

The study was undertaken at the following two sites along Galicia's Atlantic seaboard (NW Spain): Punto Area das Vacas (42°06′54″ N; 008°54′30″ W) situated on the Vigo estuary (*Ría de Vigo*); and Lago (42°19′25″ N; 008°49′37″ W) located on Aldán Bay (*Ensenada de Aldán*), on the southern edge of the Pontevedra estuary (*Ría de Pontevedra*). Both sites are located in fishing area of *P. lividus* and feature extensive rocky areas with a high abundance of this specie. However, the sampling areas are exploited only occasionally.

Table 1: *List of colours selected (C, cyan; M, magenta; Y, yellow; and K, black).*

Code	Colour	Pantone CVC	%CMYK			
			C	M	Y	K
1	Bright yellow	*Yellow C*	0	0	100	0
2	Yellow	107	0	0	79	0
3	Pale yellow	100	0	0	51	0
4	Dark yellow	110	0	11	94	6
5	Yellow+Black	1405	0	38	100	65
6	Bright yellow orange	137	0	34	91	0
7	Orange yellow	136	0	27	79	0
8	Pale orange	1495	0	30	69	0
9	Orange	1505	0	38	76	0
10	Bright orange	*Orange* 021C	0	51	87	0
11	Orange pink	1485	0	23	56	0
12	Orange red	172	0	65	83	0
13	Red	*Warm red*	0	79	91	0
14	Dark red	1795	0	94	100	0
15	Dark orange	1595	0	65	100	9
16	Orange light brown	167	0	60	100	18
17	Orange brown	1605	0	56	100	30
18	Light brown orange	160	0	60	100	34
19	Light brown red	1815	0	91	100	51
20	Light brown	724	0	51	100	43
21	Brown orange	1615	0	56	100	43
22	Brown red	181	0	72	79	47
23	Brown	168	0	56	100	60
24	Dark brown	1545	0	51	100	83
25	Black	*Black C*	0	0	0	100

The species of algae present in the intertidal zone at both sites included *Lithophyllum incrustans* Philippi 1837, *Corallina officinalis* Linnaeus 1758, *Corallina elongata* J. Ellis & Solander 1786, *Chondrus crispus* Stackhouse 1797, *Bifurcaria bifurcata* R. Ross 1958, *Ulva rigida* C. Agardh 1823, etc., which are all very common along intertidal areas on the Galicia coast. For its part, there is a high abundance of *Saccorhiza polyschides* (Lightfoot) Batters 1902, in the sublittoral zone studied.

Samples were collected monthly from January 2002 to February 2003 along the lower intertidal zone of both sites, and along the sublittoral zone of the latter. Samples were collected randomly, with each comprising 25 individuals of *P. lividus*. A total of 750 specimens were finally studied.

The sex was determined according to the colour of gametic fluid. Histological examination of the gonads showed that male gonads emit white gametes, while female emission was orange in colour although some cases were red. These observations are in agreement with the findings of other studies (e.g. Crapp and Willis, 1975). Samples were disregarded where it was not possible to collect gametic fluid.

A colour table (Pantone CVC, Pantone Inc) was used to determine the gonad colour of the samples. The table breaks the colours down into four component parts, namely, cyan (C), magenta (M), yellow (Y) and black (K) (collectively, CMYK). A total of 25 colours were observed in the samples collected, and codified using Table 1. The observations were made by three researchers under constant low light conditions across the study.

3.2. Statistical modelling

The discriminatory capacity of gonad colour for distinguishing male from female individuals was assessed by using the following two different classifiers: (a) raw gonad colour (without transformation); and, (b) gonad colour transformed through equation (2). In the latter case, the following logistic GAM regression model was fitted:

$$f(Colour) = P[Sex = 1|Colour] = g^{-1}(\alpha + h(Colour)), \tag{3}$$

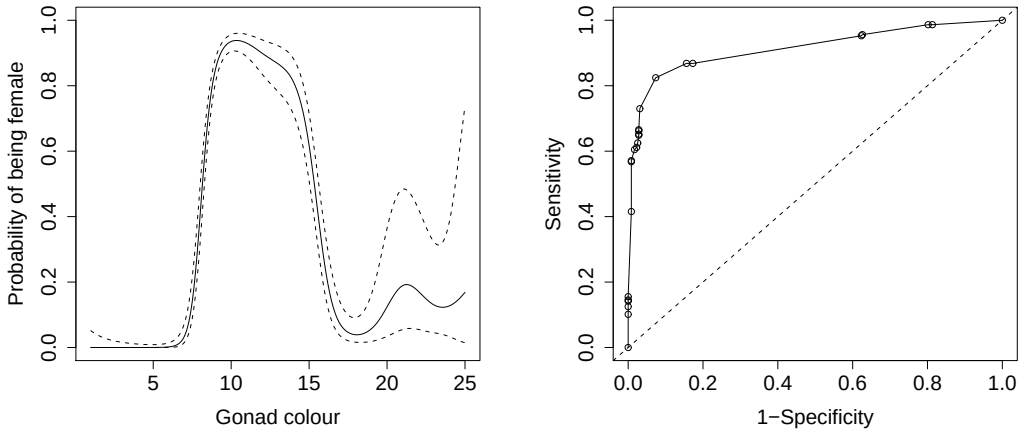
where *Colour* denotes gonad colour, *Sex* is a binary variable taking the value 1 for female and 0 for male, $g(\cdot)$ is the logit function (known), and $h(\cdot)$ is a smooth unknown function.

The logistic GAM regression model (3) was fitted by using the `gam` function of the `mgcv` package (Wood, 2006). The R software package, `ROCR` (Sing *et al.*, 2005), was used to estimate the ROC curve and AUC.

3.3. Results

The results shown below are based on the global data collected in the two sites of the study: Punto Area das Vacas and Lago. For the analysis of the discriminatory capacity of raw gonad colour (where the darkest values were assumed to be indicators of male gender), the estimated AUC was 0.586 with a 95% bootstrap confidence interval (CI) of (0.542, 0.638). Based on this result, gonad colour would not seem to be reliable for accurate classification of the sex of *P. lividus*.

With respect to the analysis performed with the transformed data, Figure 2a shows the estimated probability of being female according to gonad colour. Intermediate



(a) (b)

Figure 2: *(a) Estimated probability of being female according to gonad colour, together with its 95% confidence interval. Shown in (b) is the estimation of the ROC curve for gonad colour transformation.*

Table 2: *Range of colours for classifying an individual as female, along with the probabilities of reaching a correct decision.*

Range of female gonad colour	Probability of correct classification of a female individual (TPF)	Probability of correct classification of a male individual (TNF)
$[9, 14]$	0.84	0.87
$[9, 15]$	0.82	0.92

colours corresponded to a high probability of being female, while colours at the extremes of the palette corresponded to a low probability of being female and an ensuing higher probability of being male. The ROC curve associated with the above probability (employed as the classifier) is shown in Figure 2b. The corresponding AUC was 0.914 with a 95% bootstrap CI of (0.891, 0.936) ($n = 649$), with similar results being observed for all three populations after they had been separately analysed. Use of the logistic GAM regression model led to optimum predictive capacity. In contrast to the analysis of raw data, these results confirm that gonad colour can afford a high degree of accuracy in classifying the sex of *P. lividus*.

It is important to note that, on working with transformed data, the TPF-FPF pairs which give rise to the ROC curve are obtained on the basis of the probabilities estimated by the model (3). For this example, however, it is easy to obtain the gonad colour ranges that yield the said TPF-FPF pairs. Table 2 shows some possible colour ranges for classifying an individual as female, together with the corresponding probability of reaching a correct decision (in ROC terminology, the true positive, TPF, and the true negative, TNF, fractions).

4. Conclusions

In this paper, a new flexible alternative for evaluating the discriminatory capacity of a continuous or ordinal classifier is suggested. The proposed methodology is based on: (a) transformation of the classifier by means of a logistic GAM regression model; (b) use of the probabilities estimated by this model as a new classifier; and (c) evaluation of the discriminatory capacity of this new classifier by the ROC curve. This transformation makes it possible to obtain better cut-off values (or intervals) on which to base the classification.

The methodology presented in this paper was applied to the task of assessing the discriminatory capacity (accuracy) of gonad colour in terms of determining the sex of *P. lividus*. The results obtained with crude gonad colour indicated that this classifier had little accuracy. Yet when transformed gonad colour was used, this same discriminatory capacity proved to be high. Similarly, using transformed gonad colour, this paper furnishes two possible colour ranges on which to base *P. lividus* sex-classification in wild populations. Sexual determination of this species according to gonad colour serves both to enhance knowledge of its biology and to improve its commercial exploitation by enabling a better quality product to be obtained.

According to many authors, the diet of *P. lividus* greatly affects gonad colour, being particularly influenced by the accumulation of carotenoids (e.g. Shpigel *et al.*, 2005; Shpigel *et al.*, 2006). Many studies have investigated the influence of distinct diets (natural or artificial), in increasing gonad yield and improving gonad quality (from a commercial perspective). While the samples collected in this study originated from distinct habitats (intertidal/sublittoral), in which the diet of the urchins could be different, important changes were not observed neither in the distribution of gonad colours, nor in the ability of the gonad colour in discriminating the sex of *P. lividus*.

ROC methodology has an infinity of possibilities in the field of ecology and biology. In our opinion, the alternative presented in this paper could be of great utility in aspects relating to improvement in marine resource management, e.g., for determining the size or age at which examples of a species reach sexual maturity, or the periods during which a given resource reaches specific sexual stages. Likewise, it offers great possibilities in spatial distribution studies (presence/absence), among others. Application of this methodology will allow for solid results, based on appropriate statistical models, to be obtained.

Acknowledgements

The authors would like to express their gratitude for the support received in the form of the Spanish MEC Grant MTM2008-01603 and the Galician Regional Authority (Xunta de Galicia) projects INCITE08PXIB208113PR and 07MMA001200PR. We are also grateful to the referee for her/his valuable comments and suggestions, which served to make a substancial improvement to this paper.

References

Catoira, J. L. (1995). Spatial and temporal evolution of the gonad index of the sea urchin *Paracentrotus lividus* (Lamarck) in Galicia, Spain. In: Emson, R., Smith, A. and Campbell, A. (eds.). *Echinoderm Research*. Balkema, Rotterdam, 295-298.

Crapp, G. B. and Willis, M. E. (1975). Age determination in the sea urchin *Paracentrotus lividus* (Lamarck), with notes on the reproductive cycle. *Journal of Experimental Marine Biology and Ecology*, 20, 157-178.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRP Press, New York.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89-121.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29-36.

Hastie, T. J and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall, London.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13, 183-212.

Lustres-Pérez, V. (2006). El erizo de mar: *Paracentrotus lividus* (Lamarck, 1816) en las costas de Galicia. PhD thesis. Universidad de Santiago de Compostela.

McIntosh, M. W. and Pepe, M. S. (2002). Combining several screening tests: optimality of the risk score. *Biometrics*, 58, 657-664.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Second Edition. Chapman and Hall, London.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8, 283-298.

Monteiro-Torreiro, M. F. and Garcia-Martinez, P. (2003). Seasonal changes in the biochemical composition of body components of the sea urchin, *Paracentrotus lividus*, in Lorbé (Galicia-north-western Spain). *Journal of the Marine Biological Association of the United Kingdom*, 83, 575-581.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypothesis. *Philosophical Transactions of the Royal Society of London, Series A*, 231, 289-337.

Sellem, F. and Guillou, M. (2007). Reproductive biology of *Paracentrotus lividus* (Echinodermata: Echinoidea) in two contrasting habitats of northern Tunisia (south-east Mediterranean). *Journal of the Marine Biological Association of the United Kingdom*, 87, 763-767.

Shpigel, M., McBride, S. C., Marciano, S., Ron, S. and Ben-Amotz, A. (2005). Improving gonad colour and somatic index in the European sea urchin *Paracentrotus lividus*. *Aquaculture*, 245, 101-109.

Shpigel, M., Schlosser, S. C., Ben-Amotz, A., Lawrence, A. L. and Lawrence, J. M. (2006). Effects of dietary carotenoid on the gut and the gonad of the sea urchin *Paracentrotus lividus*. *Aquaculture*, 261, 1269-1280.

Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20), 3940-3941.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York.

Tortonese, E. (1965). *Fauna d'Italia. Echinodermata*. Calderini, Bologna.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Chapman & Hall, London.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B* 65, 95-114.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99, 673-686.

Wood, S. N. (2006). *Generalized Additive Models, An Introduction with R.* Chapman & Hall/CRC, Boca Raton, Florida.

Xunta de Galicia. Plataforma tecnolóxica da pesca. Consellería do mar. http://www.pescadegalicia.com/default.htm (accessed: 23 November, 2010).