Ecography

E7585
Li, W. and Guo, Q. 2013. How to assess the prediction accuracy of species presence–absence models without absence data? – Ecography 36: xxx–xxx.

Supplementary material

Appendix 1 Derivations of equations (8) and (9)

p and r are estimated from the presence–absence data, which are a simple random sample. In contrast, p' and r' are estimated from the presence–background data, which are a case-control sample. To differentiate from the presence–absence sample, here we use $\eta = 1$ to denote the case-control presence–background sample.

Note that p, r, p', and r' have the following probabilistic interpretation:

$$p = \Pr(y = 1 \mid y' = 1) \tag{A1}$$

$$r = \Pr(y' = 1 \mid y = 1) \tag{A2}$$

$$p' = \Pr(s = 1 \mid y' = 1, \eta = 1) \tag{A3}$$

$$r' = \Pr(y' = 1 \mid s = 1, \eta = 1) \tag{A4}$$

Because the observed presence data ($s = 1$) are random samples of the presence data ($y = 1$), r' is an estimate of r:

$$\Pr(y' = 1 \mid s = 1, \eta = 1) = \Pr(y' = 1 \mid y = 1) \tag{A5}$$

Equation (A5) is equivalent to equation (8) of the paper.

According to the definition of p'' in equation (7) of the paper, we have

$$p'' = \frac{p'}{1 - p'}$$

$$= \frac{\Pr(s = 1 \mid y' = 1, \eta = 1)}{1 - \Pr(s = 1 \mid y' = 1, \eta = 1)}$$

$$= \frac{\Pr(s = 1, y' = 1, \eta = 1)/\Pr(y' = 1, \eta = 1)}{1 - \Pr(s = 1, y' = 1, \eta = 1)/\Pr(y' = 1, \eta = 1)}$$

$$= \frac{\Pr(s = 1, y' = 1, \eta = 1)}{\Pr(y' = 1, \eta = 1) - \Pr(s = 1, y' = 1, \eta = 1)}$$

$$= \frac{\Pr(s=1, y'=1, \eta=1)}{\Pr(s=1, y'=1, \eta=1) + \Pr(s=0, y'=1, \eta=1) - \Pr(s=1, y'=1, \eta=1)}$$

$$= \frac{\Pr(s=1, y'=1, \eta=1)}{\Pr(s=0, y'=1, \eta=1)}$$

$$= \frac{\Pr(y'=1 \mid s=1, \eta=1) \times \Pr(s=1, \eta=1)}{\Pr(y'=1 \mid s=0, \eta=1) \times \Pr(s=0, \eta=1)} \quad \text{(A6)}$$

Note that the background data (s = 0) are random samples of the population. Hence, we have

$$\Pr(y'=1 \mid s=0, \eta=1) = \Pr(y'=1) \quad \text{(A7)}$$

Substituting (A5) and (A7) into (A6), we obtain

$$p'' = \frac{\Pr(y'=1 \mid y=1) \times \Pr(s=1, \eta=1)}{\Pr(y'=1) \times \Pr(s=0, \eta=1)}$$

$$= \frac{\Pr(y'=1, y=1)}{\Pr(y'=1)} \times \frac{\Pr(s=1, \eta=1)}{\Pr(y=1) \times \Pr(s=0, \eta=1)}$$

$$= \Pr(y=1 \mid y'=1) \times \frac{\Pr(s=1, \eta=1)}{\Pr(y=1) \times \Pr(s=0, \eta=1)}$$

$$= p \times c \quad \text{(A8)}$$

where $c = \frac{\Pr(s=1, \eta=1)}{\Pr(y=1) \times \Pr(s=0, \eta=1)}$; c is a constant whose value depends on the species prevalence $\pi = \Pr(y = 1)$ and the ratio of the number of observed presence data to the number of background data $\frac{\Pr(s=1, \eta=1)}{\Pr(s=0, \eta=1)}$. Equation (A8) is equivalent to equation (9) of the paper.

Appendix 2 Confusion matrices and statistics

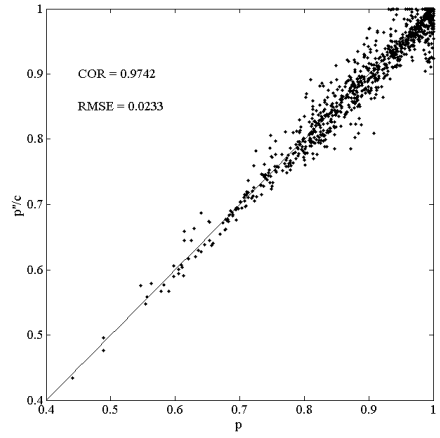
Table A1. Confusion matrices and statistics generated by DOMAIN, GLM, and MAXENT.

	Reference																	
	DOMAIN						GLM						MAXENT					
				s = 0						s = 0						s = 0		
Prediction	s = 1	s = 0	Total	y = 1	y = 0	Total	s = 1	s = 0	Total	y = 1	y = 0	Total	s = 1	s = 0	Total	y = 1	y = 0	Total
y' = 1	1000	1987	2987	875	1112	1987	1000	1465	2465	875	590	1465	975	933	1908	852	81	933
y' = 0	0	3013	3013	0	3013	3013	0	3535	3535	0	3535	3535	25	4067	4092	23	4044	4067
Total	1000	5000	6000	875	4125	5000	1000	5000	6000	875	4125	5000	1000	5000	6000	875	4125	5000
							n_1 = 1000; n_0 = 5000; π = 875 / 5000 = 0.1750; c = n_1 / (π * n_0) = 1.1429											
Statistics	r'=1000/1000 =1			r=875/875=1			r'=1000/1000 =1			r=875/875=1			r'=975/1000=0.9750			r=852/875=0.9737		
	p'=1000/2987=0.3348			p=875/1987=0.4404			p'=1000/2465=0.4057			p=875/1465=0.5973			p'=975/1908=0.5110			p=852/933=0.9132		
	p"/c=0.3348/(1-0.3348)/1.1429=0.4404						p''/c=0.4057/(1-0.4057)/1.1429=0.5973						p''/c=0.5110/(1-0.5110)/1.1429=0.9144					
	F_{pb}=2*1000/(1000+0+1987)=0.6696			F=2*875/(2*875+0+1112)=0.6115			F_{pb}=2*1000/(1000+0+1465)=0.8114			F=2*875/(2*875+0+590)=0.7479			F_{pb}=2*975/(975+25+933)=1.0088			F=2*852/(2*852+23+81)=0.9425		
	F_{cpb}=2*1000/(1000+0+1.1429*1987)=0.6115						F_{cpb}=2*1000/(1000+0+1.1429*1465)=0.7479						F_{cpb}=2*975/(975+25+1.1429*933)=0.9437					

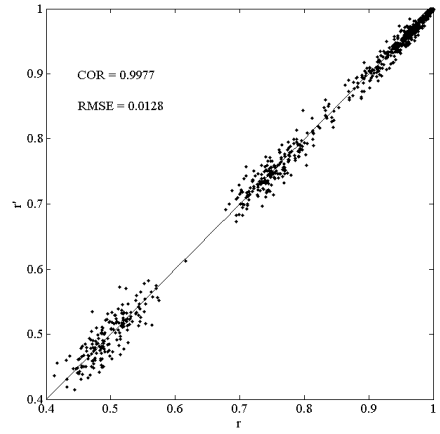
y = 1: observed presence; y = 0: observed absence; s = 1: observed presence; s = 0: background data; y' = 1: predicted presence; y' = 0: predicted absence; n_1: the number of observed presence data; n_0: the number of background data; π: species prevalence; c: a constant.

Please see text for equations of the statistics p, r, F, p', p''/c, r', F_{pb}, and F_{cpb}.

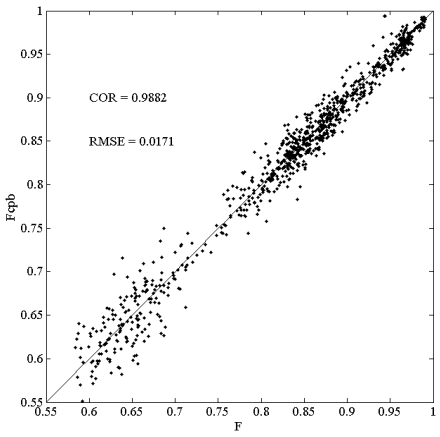
Appendix 3 Scatter plots (true species prevalence is provided)



(a)



(b)



(c)

Figure A1 Scatter plots. (a) p''/c vs. p, (b) r' vs. r, and (c) F_{cpb} vs. F. COR: the Pearson's correlation coefficient; RMSE: root-mean-square error. Prior information on species prevalence is provided. Please see text for equations of the statistics p, r, F, p''/c, r', and F_{cpb}.

Appendix 4 Scatter plots (the species prevalence is estimated from model)

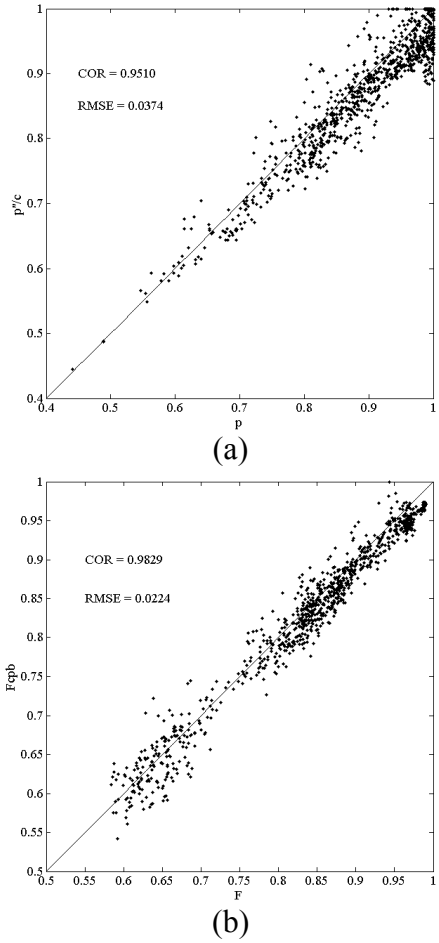


Figure A2 Scatter plots. (a) p''/c vs. p, and (b) F_{cpb} vs. F. COR: the Pearson's correlation coefficient; RMSE: root-mean-square error. The information on species prevalence is estimated by MAXENT with maximizing F_{pb} as the thresholding method. Please see text for equations of the statistics p, F, p''/c, and F_{cpb}.