

# 1 Dealing with many correlated covariates in capture-recapture models

2 Olivier Gimenez<sup>1</sup>, Christophe Barbraud<sup>2</sup>

3  
4 <sup>1</sup>CEFE UMR 5175, CNRS, Université de Montpellier, Université Paul-Valéry Montpellier,  
5 EPHE, 1919 Route de Mende, 34293 Montpellier Cedex 5, France.

6 <sup>2</sup>CEBC UMR 7372, CNRS – Université de La Rochelle, 79360 Villiers en Bois, France.

7 Corresponding author: Olivier Gimenez (olivier.gimenez@cefe.cnrs.fr)

8  
9 *Word count: 2412*

10 *Summary:* Capture-recapture models for estimating demographic parameters allow covariates  
11 to be incorporated to better understand population dynamics. However, high-dimensionality  
12 and multicollinearity can hamper estimation and inference. Principal component analysis is  
13 incorporated within capture-recapture models and used to reduce the number of predictors  
14 into uncorrelated synthetic new variables. Principal components are selected by sequentially  
15 assessing their statistical significance. We provide an example on seabird survival to illustrate  
16 our approach. Our method requires standard statistical tools, which permits an efficient and  
17 easy implementation using standard software.

18 *Key words:* Animal demography, Population dynamics, Principal-component capture-  
19 recapture model, Survival estimation.  
20

## INTRODUCTION

Capture-recapture (CR) methods (e.g. Lebreton *et al.* 1992) are widely used for assessing the effect of explanatory variables on demographic parameters such as survival (Pollock 2002). Generally however, complex situations arise where multiple covariates are required to capture patterns in survival. In such situations, one usually favors a multiple regression-like CR modeling framework that is however hampered by two issues: first, because it increases the number of parameters to be estimated, incorporating many covariates results in a loss of statistical power and a decrease in the precision of parameter estimates; second, correlation among the set of predictors – aka multicollinearity – may alter interpretation (see below).

To overcome these two issues, Grosbois *et al.* (2008) recommended to perform a principal component analysis (PCA) on the set of explanatory variables before fitting CR models. PCA is a multivariate technique that explains the variability of a set of variables in terms of a *reduced* set of *uncorrelated* linear combinations of such variables – aka principal components (PCs) – while maximizing the variance (Jolliffe 2002). Grosbois *et al.* (2008) then expressed survival as a function of the PCs that explained most of the variance in the set of original covariates, typically the first one or the first two ones.

However, the main drawback of this approach is that the PCs are selected based on covariates variation pattern alone, regardless of the response variable, and without guarantee that survival is most related to these PCs (Graham 2003). To deal with this issue in the context of logistic regression, Aguilera *et al.* (2006) proposed to test the significance of *all* PCs to decide which ones should be retained, instead of a priori relying on the PCs that explain most of the variation in the covariates.

In this paper, we implement the algorithm proposed by Aguilera *et al.* (2006) to deal with many possibly correlated covariates in CR models, a method we refer to as principal

46 component capture-recapture (P2CR). We apply this new approach to a case study on  
47 survival of Snow petrels (*Pagodroma nivea*) that is possibly affected by climatic conditions.  
48 In this example, the issue of multicollinearity occurs, and summarizing the set of covariates  
49 in a subset of lower dimension is also crucial to get precise survival estimates. Overall, P2CR  
50 models can be fitted with statistical programs that perform PCA and CR data analysis. The  
51 data and R code are available from GitHub at <https://github.com/oliviergimenez/p2cr>.

## 53 METHODS

54 We used capture-recapture (CR) models to study open populations over  $K$  capture  
55 occasions to estimate the probability  $\phi_i$  ( $i = 1, \dots, K - 1$ ) that an individual survives to  
56 occasion  $i + 1$  given that it is alive at time  $i$ , along with the probability  $p_j$  ( $j = 2, \dots, K$ ) that  
57 an individual is recaptured at time  $j$  – aka as the Cormack-Jolly-Seber (CJS) model (Lebreton  
58 *et al.* 1992). Covariates were incorporated in survival probabilities using a linear-logistic  
59 function:

$$\text{logit}(\phi_i) = \log\left(\frac{\phi_i}{1-\phi_i}\right) = \alpha + \sum_{j=1}^p \beta_j X_{ij} \quad (1)$$

60 where  $\alpha$  is the intercept parameter,  $X_{ij}$  is the value of covariate  $j$  ( $j = 1, \dots, p$ ) in year  $i$  ( $i =$   
61  $1, \dots, K - 1$ ), and  $\beta_j$  is its associated slope parameter. Covariates were standardized to avoid  
62 numerical instabilities. To assess the significance of a covariate in CR models, we used the  
63 analysis of deviance (ANODEV; Skalski, Hoff & Smith 1993) that compares the amount of  
64 deviance explained by this covariate with the amount of deviance not explained by this  
65 covariate, the CR model with fully time-dependent survival serving as a reference. The  
66 ANODEV test statistic is given by:

$$\text{ANODEV} = \frac{\text{Dev}(X) - \text{Dev}(\text{constant})}{1} / \frac{\text{Dev}(\text{time}) - \text{Dev}(X)}{K - 1} \quad (2)$$

67 where  $\text{Dev}(\text{constant})$ ,  $\text{Dev}(X)$  and  $\text{Dev}(\text{time})$  stand for the deviance of models with constant,  
68 covariate-dependent and time-dependent survival probabilities. To obtain the associated p-  
69 value, the value of the ANODEV is compared with the quantile of Fisher-Snedecor  
70 distribution with 1 and K-1 degrees of freedom.

71 To reduce the dimension of the set of covariates  $(X_1, \dots, X_p)$ , we used PCA which  
72 aims at finding a small number of linear combinations of the original variables – the principal  
73 components (PCs) – while maximizing the variance in  $(X_1, \dots, X_p)$ . Because the variables  
74 measurement units often differ, we performed the PCA on the correlation matrix (Jolliffe  
75 2002). To select PCs, we used a forward model selection algorithm as proposed by Aguilera  
76 *et al.* (2006) for the logistic regression. The forward algorithm begins with no covariates in  
77 the model. Each PC is incorporated in simple linear regression-like CR models and the  
78 ANODEV p-value calculated. The PC that has the lowest p-value is added to the null model,  
79 say  $\text{PC}_k$ . Then the PCs that were not retained are incorporated along with  $\text{PC}_k$  in multiple  
80 regression-like CR models, and ANODEV p-values are calculated. In other words, we need  
81 to assess the effect of  $\text{PC}_j$  for  $j \neq k$  in the presence of  $\text{PC}_k$  to decide whether  $\text{PC}_j$  should be  
82 retained. To do so,  $\text{Dev}(\text{constant})$  and  $\text{Dev}(X)$  are replaced by  $\text{Dev}(\text{PC}_k)$  and  $\text{Dev}(\text{PC}_k + \text{PC}_j)$   
83 in Equation 2, where  $\text{Dev}(\text{PC}_k + \text{PC}_j)$  is the deviance of the model with survival as a function  
84 of both principal components  $\text{PC}_k$  and  $\text{PC}_j$ . We repeat the process until no remaining PC is  
85 selected.

86 All models were fitted using the maximum-likelihood method using MARK (White &  
87 Burnham 1999) called with R (Laake 2013).

88

89

90

91

92

## CASE STUDY

93

The Snow petrel is a medium sized Procellariiform species endemic to Antarctica that breeds

94

in summer. Birds start to occupy breeding sites in early November, laying occurs in early

95

December and chicks fledge in early March. This highly specialized species only forages

96

within the pack-ice on crustaceans and fishes. Data on survival were obtained from a long-

97

term CR study on Ile des Pétrels, Pointe Géologie Archipelago, Terre Adélie, Antarctica. We

98

refer to Barbraud *et al.* (2000) for more details about data collection. We removed the first

99

capture to limit heterogeneity among individuals, and worked with a total of 604 female

100

capture histories from 1973 to 2002.

101

The following covariates were included to assess the effect of climatic conditions

102

upon survival variation: sea ice extent (SIE; [http://nsidc.org/data/seaice\\_index/](http://nsidc.org/data/seaice_index/)); air

103

temperature, which was obtained from the Météo France weather station at Dumont

104

d'Urville, as a proxy for sea surface temperature; southern Oscillation Index (SOI) as a proxy

105

for the overall climate condition (<https://crudata.uea.ac.uk/cru/data/soi/>). These

106

environmental variables were averaged over seasonal time periods corresponding to the chick

107

rearing period (January to March: summer period), the non-breeding period (April to June:

108

autumn and July to September: winter), and the laying and incubation period of the same year

109

(October to December: spring). In total, 9 covariates were included in the analysis: sea ice

110

extent in summer (SIEsummer), in autumn (SIEautumn), in winter (SIEwinter), in spring

111

(SIEspring), annual SOI, air temperature in summer (Tsummer), in autumn (Tautumn), in

112

winter (Twinter) and in spring (Tspring).

113

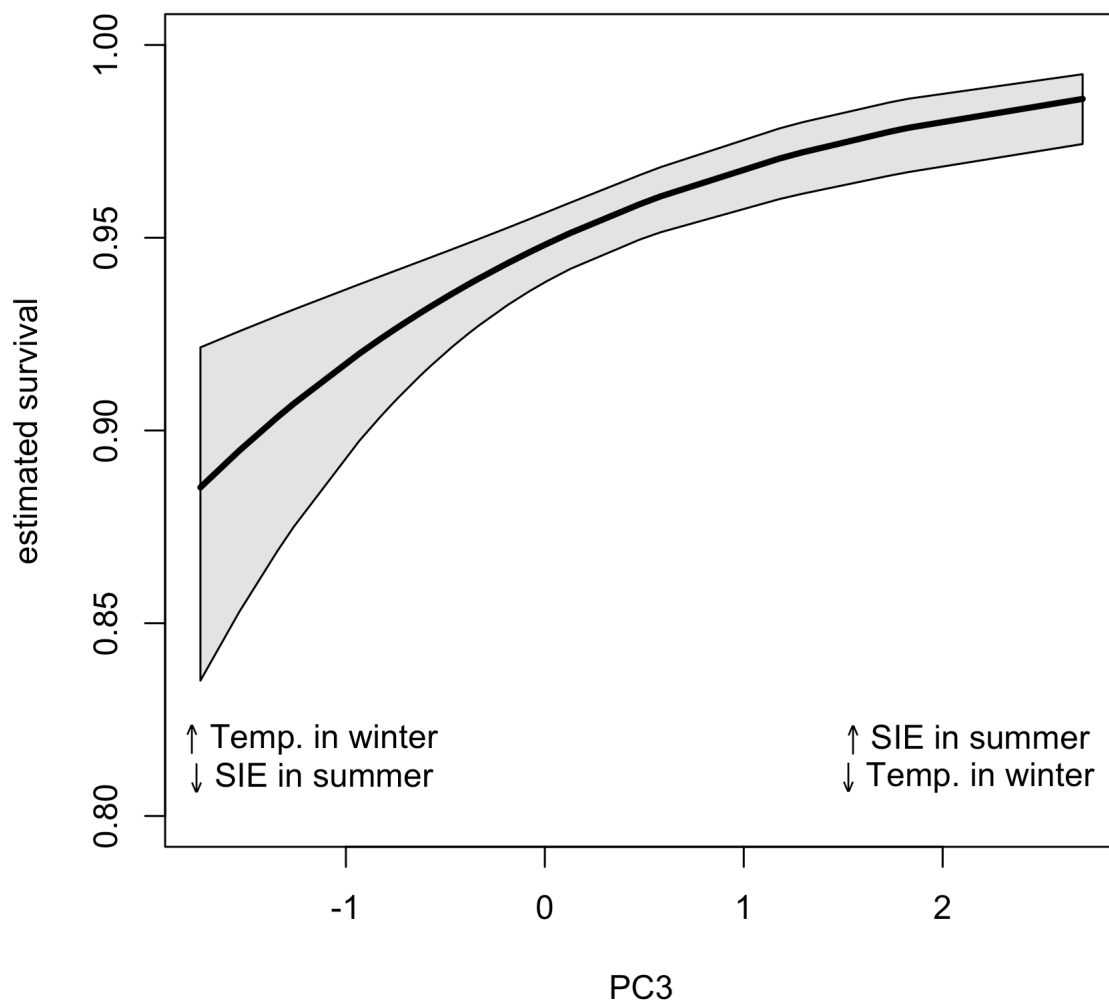
114

## RESULTS

115 The CJS model poorly fitted the data ( $\chi^2 = 221.2$ ,  $df = 127$ ,  $p \ll 0.01$ ), and a closer  
116 inspection of the results revealed that the lack of fit was explained by a trap-dependence  
117 effect (Test2CT,  $\chi^2 = 103.1$ ,  $df = 27$ ,  $p \ll 0.01$ ). Consequently, we estimated two recapture  
118 probabilities that differed according to whether or not a recapture occurred the occasion  
119 before (Pradel 1993). By first attempting to simplify the structure of recapture probabilities,  
120 we were led to consider an additive effect of time and a trap effect (Supplementary material).  
121 Estimates of recapture probabilities ranged from 0.14 (standard error [SE] = 0.07) to 0.79 (SE  
122 = 0.09) when no recapture occurred the occasion before and from 0.25 (SE = 0.18) to 0.89  
123 (SE = 0.09) when a recapture occurred the occasion before (Supplementary material).

124 Because of multicollinearity, we were led to counterintuitive estimates of regression  
125 parameters in the CR model including all covariates (Supplementary material): the coefficient  
126 of SIE in autumn was estimated at 0.5 (SE = 0.24) and that of SIE in winter was estimated at  
127 -0.5 (SE = 0.21) while these two covariates were significantly positively correlated ( $r = 0.67$ ,  
128  $p < 0.01$ ).

129 When we applied the P2CR approach, the algorithm selected two PCs, namely PC3  
130 ( $F_{1,27} = 7.34$ ,  $p = 0.01$ ) at step 1 and PC4 ( $F_{1,26} = 4.63$ ,  $p = 0.04$ ) at step 2 (Supplementary  
131 material), but never did we pick PC1 as we would have done using a classical approach  
132 (Grosbois et al. 2008). PC3 was positively correlated to SIE in summer and negatively  
133 correlated to temperature in winter, while PC4 was positively correlated to temperature in  
134 spring and negatively correlated to SIE in summer (Supplementary material). Survival  
135 increased with increasing values of PC3 (Figure 1), with high values of SIE in summer and  
136 low values of temperature in winter (resp. low values of SIE in summer and high values of  
137 temperature in winter) corresponding to high (resp. low) survival.



138

139

**Figure 1: Survival of Snow petrel as a function of PC3.**

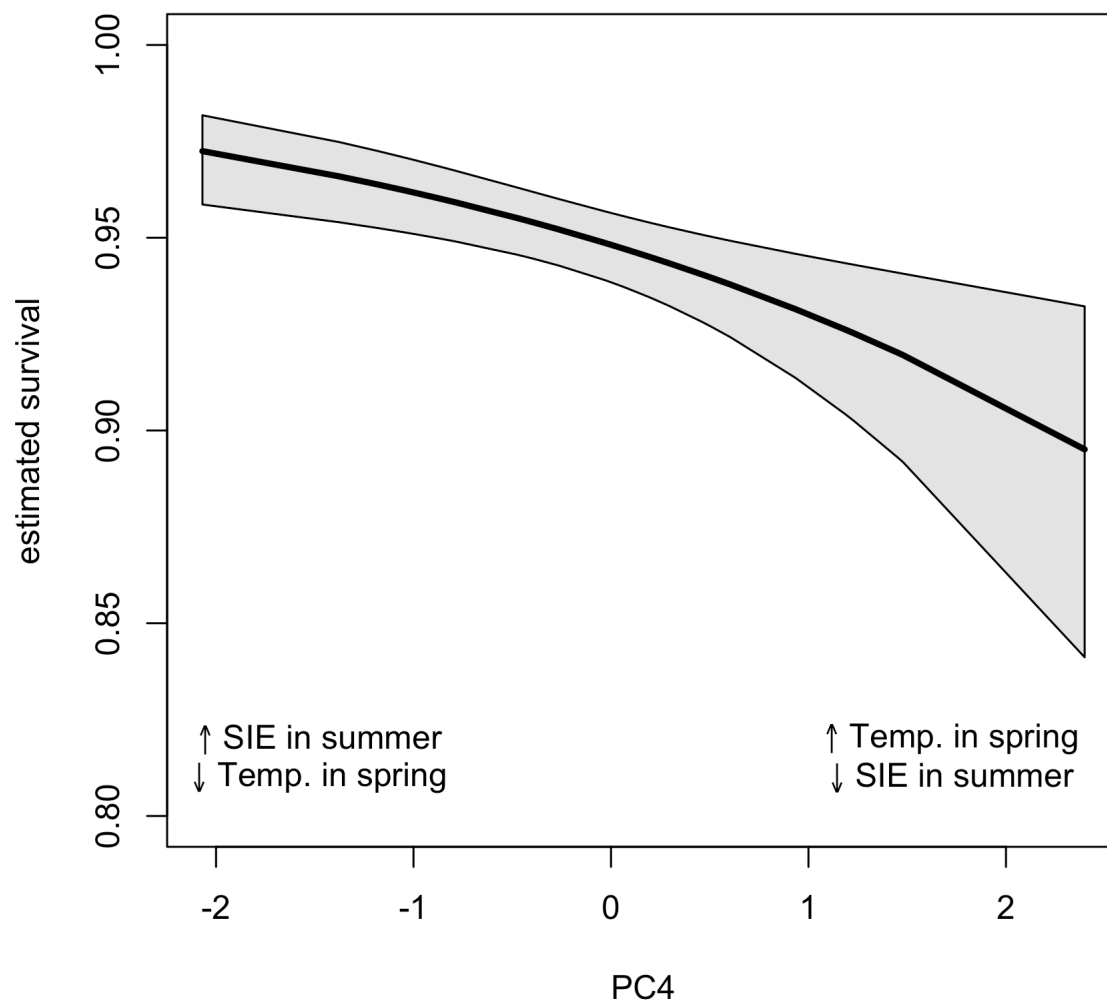
140

141 Survival decreased with increasing values of PC4 (Figure 2), with high values of temperature

142 in spring and low values of SIE in summer (resp. low values of temperature in spring and

143 high values of SIE in summer) corresponding to low (resp. high) survival.

144



145

146

**Figure 2: Survival of Snow petrel as a function of PC4.**

147

148

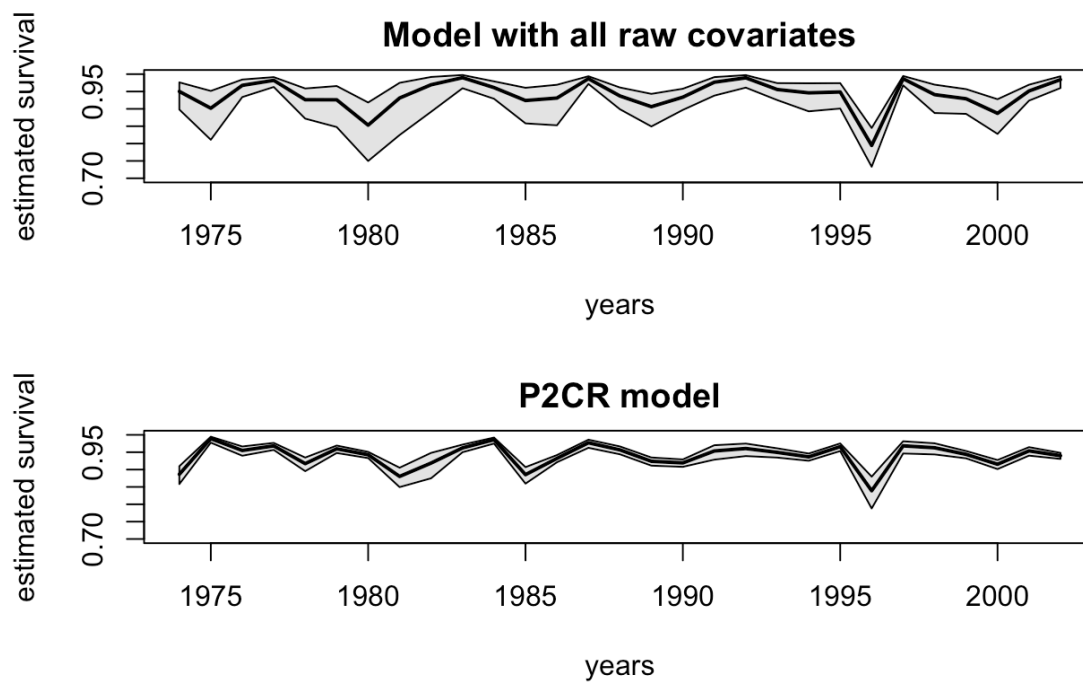
The P2CR approach also led to more precise survival estimates when compared to the model

149

incorporating all original covariates (Figure 3).

150





151

152 **Figure 3: Survival of Snow petrel over time as estimated from the model with all**  
153 **original covariates (top panel) vs. the PC2R model (bottom panel).**

154

155

156

## DISCUSSION

157

158 We introduce a new approach combining principal component analysis and capture-recapture  
159 models to deal with many possibly correlated explanatory covariates. Our approach requires  
160 standard statistical tools, which allows an efficient and easy implementation using standard  
161 software.

162

### *Snow petrels and climatic conditions*

164 In summer, snow petrels exclusively forage within the pack-ice tens to hundreds of  
165 kilometers from the colony where they catch sea ice-associated species, such as Antarctic  
166 silverfish (*Pleuragramma antarcticum*) and Euphausiids, to feed their chick (Ridoux &  
167 Offredo 1989). This is an energetically demanding period for breeding adults and, during

168 years with reduced sea-ice extent, food resources may be less abundant and snow petrels may  
169 be forced to cover larger distances to find suitable foraging habitats, with potential survival  
170 costs. Assuming air temperature was a proxy of sea surface temperature variations, the  
171 negative effect of warmer temperatures on survival is coherent with general patterns found  
172 between sea surface temperature and demographic parameters in seabirds (Barbraud *et al.*  
173 2012). In many marine ecosystems warmer temperatures are associated with decreased  
174 primary production and food resources for top predators. Although the low survival in 1996  
175 corresponded to a year with reduced sea-ice extent in summer, the drop in survival was high  
176 and remains unexplained at the moment.

177

#### 178 *Principal component CR models*

179 When multiple covariates have to be considered to estimate survival, both issues of  
180 dimensionality and multicollinearity can lead to biased estimates, inflated precision as well as  
181 lack of statistical power. In such a context, the P2CR modeling framework has proved  
182 particularly useful in our example, mainly because few PCs were selected which were easily  
183 interpretable. We acknowledge that PCs with little interpretability might have been picked up  
184 by our method. To make the interpretation easier, PCA results can be post-processed by  
185 rotating axes to improve correlations between raw variables and PCs like in the varimax  
186 method (Kaiser 1958). Recent developments in the field of multivariate analyses could also  
187 be useful, like methods to handle with missing values in PCA (Dray & Josse 2015).

188 In statistical ecology, one of our objectives is to try and explain variation in state  
189 variables such as abundance, survival and the distribution of species. Dimension-reduction  
190 methods are promising to deal with many correlated covariates for the analysis of CR or  
191 occupancy data.

192

193

## ACKNOWLEDGEMENTS

194

We greatly acknowledge all of the wintering fieldworkers involved in the monitoring programs in Terre Adélie since 1962, and Dominique Besson for the management of the database. The study on petrels was supported by Expéditions Polaires Françaises, Institut Paul Emile Victor (program IPEV 109, resp. H. Weimerskirch) and Terres Australes et Antarctiques Françaises.

199

200

## LITERATURE CITED

201

Aguilera, A.M., Escabias, M. & Valderrama, M.J. (2006) Using principal components for

202

estimating logistic regression with high-dimensional multicollinear data. *Computational*

203

*Statistics and Data Analysis*, **50**, 1905–1924.

204

Barbraud, C., Rolland, V., Jenouvrier, S., Nevoux, M., Delord, K. & Weimerskirch, H.

205

(2012) Effects of climate change and fisheries bycatch on Southern Ocean seabirds: A

206

review. *Marine Ecology Progress Series*, **454**, 285–307.

207

Barbraud, C., Weimerskirch, H., Guinet, C. & Jouventin, P. (2000) Effect of sea-ice extent on

208

adult survival of an Antarctic top predator: the snow petrel *Pagodroma nivea*.

209

*Oecologia*, **125**, 483-488

210

Dray, S. & Josse, J. (2015) Principal component analysis with missing values: a comparative

211

survey of methods. *Plant Ecology*, **216**, 657–667.

212

Graham, M.H. (2003) Confronting multicollinearity in ecological multiple regression.

213

*Ecology*, **84**, 2809–2815.

214

Grosbois, V., Gimenez, O., Gaillard, J.M., Pradel, R., Barbraud, C., Clobert, J., Møller, A.P.

215

& Weimerskirch, H. (2008) Assessing the impact of climate variation on survival in

216

vertebrate populations. *Biological Reviews*, **83**, 357–99.

217

Jolliffe, I.T. (2002) Principal Component Analysis, Second Edition. Springer-Verlag, New

- 218           York.
- 219    Kaiser, H.F. (1958) The varimax criterion for analytic rotation in factor analysis.
- 220           *Psychometrika*, **23**, 187–200.
- 221    Laake, J.L. (2013) RMark: An R Interface for Analysis of Capture-Recapture Data with
- 222           MARK. AFSC Processed Rep 2013-01, 25p. Alaska Fish. Sci. Cent., NOAA, Natl. Mar.
- 223           Fish. Serv., 7600 Sand Point Way NE, Seattle WA 98115.
- 224    Lebreton, J.-D., Burnham, K.P., Clobert, J. & Anderson, D.R. (1992) Modeling survival and
- 225           testing biological hypotheses using marked animals: A unified approach with case
- 226           studies. *Ecological Monographs*, **62**, 67–118.
- 227    Pollock, K.H. (2002) The use of auxiliary variables in capture-recapture modelling: an
- 228           overview. *Journal of Applied Statistics*, **29**, 85–102.
- 229    Ridoux, V. & Offredo, C. (1989) The diets of five summer breeding seabirds in Adélie Land,
- 230           Antarctica. *Polar Biology*, **9**, 137–145.
- 231    Skalski, J.R., Hoff, A. & Smith, S.G. (1993) Testing the significance of individual- and
- 232           cohort-level covariates in animal survival studies. *Marked Individuals in the Study of*
- 233           *Bird Population*, eds J.D. Lebreton & P.M. North, pp. 9–28. Birkäuser Verlag, Basel.
- 234    White, G.C. & Burnham, K.P. (1999) Program MARK: survival estimation from populations
- 235           of marked animals. *Bird Study*, **46**, 120–139.
- 236