1   **Efficient abstracting of dive profiles using a broken-stick model**

2   Short running title: Efficient abstracting of dive profiles

3

4   Theoni Photopoulou (corresponding author)

5   Centre for Statistics in Ecology, Environment and Conservation, Department of

6   Statistical Sciences, University of Cape Town, Rondebosch 7701, Cape Town, South

7   Africa

8   Email: theoni.photopoulou@uct.ac.za

9   Philip Lovell

10   Sea Mammal Research Unit, Scottish Oceans Institute, University of St Andrews,

11   Scotland KY16 8LB, UK

12   Michael A Fedak

13   Sea Mammal Research Unit, Scottish Oceans Institute, University of St Andrews,

14   Scotland KY16 8LB, UK

15   Len Thomas

16   Centre for Research into Ecological and Environmental Modelling, The Observatory,

17   University of St Andrews, Scotland KY16 9LZ, UK

18   Jason Matthiopoulos

19   Institute of Biodiversity, Animal Health and Comparative Medicine, Graham Kerr

20   Building University of Glasgow, Glasgow, Scotland G12 8QQ, UK

21

22

23

24

25

26   **Summary**

27   1. For diving animals, animal-borne sensors are used to collect time-depth

28   information for studying behaviour, ranging patterns and foraging ecology. Often, this

29   information needs to be compressed for purposes of storage or transmission. Widely

30   used devices called Conductivity-Temperature-Depth Satellite Relay Data Loggers

31   (CTD-SRDLs) sample time and depth at high resolution during a dive and then

32   abstract the time-depth trajectory using a broken-stick model (BSM). This

33   approximation method can summarise efficiently the curvilinear shape of a dive,

34   using a piecewise linear shape with a small, fixed number of vertices.

35   2. We present the process of abstracting dives using the BSM and quantify its

36   performance, by measuring the uncertainty associated with the profiles it produces.

37   We develop a method for obtaining a confidence zone and an index for the

38   goodness-of-fit (dive zone index, DZI) for abstracted dive profiles. We validate our

39   results with a case study using dives from elephant seals (*Mirounga* spp.). We use

40   Generalised Additive Models (GAMs) to determine if the DZI can be used as a proxy

41   for an absolute measure of fit, and investigate the relationship between the DZI and

42   dive shape.

43   3. We found a strong correlation between the residual sum of squares (RSS) for the

44   difference between the detailed and abstracted profiles and the DZI and maximum

45   residual (R4), for dives resulting from CTD-SRDLs (69% deviance explained). On it's

46   own the DZI explained a lower percentage of deviance which was variable for

47   abstracted dives with different numbers of points. We also found evidence for

48   systematic differences in the DZI for different dive shapes (65% deviance explained).

49   4. Although the proportional loss of information in the abstraction of time-depth dive

50   profiles by BSM is high, what remains is sufficient to infer goodness-of-fit of the

51   abstracted profile by reversing the abstraction process. Our results suggest that

52   together the DZI and R4 can be used as a proxy for the RSS, and we present the

53   method for obtaining these metrics for BSM-abstracted profiles.

54

55   Keywords: animal telemetry, broken-stick model, CTD-SRDL, dive profile, dive type,

56   dive zone index, elephant seal, data abstraction

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76   INTRODUCTION

77   **Data abstraction in animal telemetry: needs and consequences**

78   One of the most effective ways to remotely study movement and behaviour in marine

79   animals is to use animal-borne sensors. Satellite-linked and archival animal

80   telemetry devices have developed rapidly, driven by questions about the behaviour

81   and movement of large vertebrates at sea. A range of purpose-built hardware and

82   software is widely available for deployment on animals. Although animal telemetry

83   devices are able to record information at high temporal and spatial resolution, in

84   many cases devices cannot be recovered, which means the data have to be

85   transmitted. Additionally, it is seldom possible to transmit all data that are collected

86   during a deployment, because the quantity and resolution of the received telemetry

87   data are constrained by several factors: the desired observation time, battery life of

88   the device, bandwidth of the communication system used to relay data, behaviour of

89   the animal, and the software specifications. This means that not all information that

90   is recorded can be recovered. Consequently, data abstraction (defined here as

91   reduction in volume to a simplified representation of the original) is unavoidable for

92   many types of telemetry device.

93

94   The trade-off between temporal data resolution (i.e., the rate of data sampling and

95   delivery) and the operational longevity of the telemetry device has driven the

96   development of efficient software and memory-saving processing algorithms. One of

97   these is a broken-stick model (BSM), used for the abstraction of two-dimensional

98   dive trajectories (time-depth dive profiles) on-board telemetry devices prior to

99   transmission (Fedak *et al.* 2002). BSMs are change-point models, falling under

100   piecewise linear approximation, and are used to identify points of abrupt change in

101   time-series. The piecewise linear profile of a time-series, generated by an efficient

102   linear abstraction method, should have low average deviation from the detailed dive

103   profile. When processing takes place on a small device with limited power for

104   processing and transmissions, it is advantageous to represent piecewise linear

105   profiles using a fixed and small number of bits of information, particularly when using

106   CLS Argos, where message size is fixed (Argos 2011). The algorithm should also be

107   time-efficient, scale linearly in execution time with dive duration, and consistently

108   encode biologically relevant information that might enable inferences on dive

109   function. The BSM fulfils these criteria and was adopted as the default dive

110   abstraction algorithm on CTD-SRDLs in 2006 (pers. comm. Phil Lovell). The

111   predecessor of this model placed breakpoints at locations of maximum inflection in

112   the detailed profile, and while it performed equally well, processing required more

113   time and energy (Fedak, Lovell & Grant 2001). The BSM was chosen empirically,

114   because it was found to provide a highly satisfactory compromise between the

115   priorities described above. However, its performance has not been formally tested,

116   nor have the consequences of its performance on the biological and ecological

117   conclusions that are drawn in studies using dive profiles abstracted with the BSM.

118

119   In ecology, change-point models have a long history (MacArthur 1957; MacArthur &

120   MacArthur 1961) and have seen application in a range of fields, e.g., in

121   oceanography to reduce data volume (Rual 1989), to identify edge effects in plant

122   communities (Toms & Lesperance 2003), to locate ontogenetic shifts in southern

123   elephant seal diet using stable isotopes (Authier *et al.* 2012) and in a Bayesian

124   context applied to allometric relationships between tree height and diameter

125   (Beckage *et al.* 2007). We illustrate the use of the BSM for dive abstraction with a

126   time-depth dive profile as an example (Figure 1). We call *abstracted* those dive

127   profiles that have been processed and reduced in resolution using this algorithm. We

128   call *detailed* those dive profiles that are recorded at regular and frequent time

129   intervals, at the sampling resolution of the device.

130

131   **Assessing uncertainty in abstracted dive profiles from elephant seals**

132   **(*Mirounga* sp.) as a case study**

133   The need for abstraction becomes critical for deployments on wide-ranging marine

134   species, such as seals and turtles, when geographic and temporal data coverage is

135   of interest, and when devices cannot be recovered. For elephant seals and other

136   phocid seals for example, complete time-series of year-round locations and

137   behaviour may be more biologically interesting than detailed information over short

138   periods, and more useful for understanding their life-histories (McConnell, Chambers

139   & Fedak 1992; Hebblewhite & Haydon 2010). Until now, the uncertainty associated

140   with abstracted profiles has not been quantified and the implications for ecological

141   studies that use abstracted profiles have not been assessed.

142

143   Abstracted profiles are, by construction, information-poor versions of the detailed

144   trajectories, but since the abstraction process is known in the case of CTD-SRDLs, it

145   is possible to reverse the deterministic steps and retrieve some of the information.

146   Historically, once the abstraction was completed, the high-resolution time-depth

147   profile was overwritten, but current tags store all information that they record, and

148   this can be accessed if the tag is retrieved. Here, we show that it is possible to

149   construct a 100% confidence zone around an abstracted profile (i.e., upper and

150   lower depth limits at each time point within which the true depth must lie) and

151   compare the zone for different dives. This confidence zone is hereafter referred to as

152   the *dive zone*, and the relative measure of maximum deviation of the abstracted

153   profile, from the detailed profile, is referred to as the *dive zone index* (DZI).

154

155   The consequences of the abstraction regime by BSM are investigated here using

156   detailed and abstracted dive profiles, from northern (*M. angustirostris*) and southern

157   elephant seals (*M. leonina*), as a case study. Elephant seals are large-bodied, long-

158   lived and abundant marine mammals. They spend many months at sea in the open

159   ocean and in coastal or marginal ice zones (McConnell *et al.* 1992; Jonker & Bester

160   1998; Campagna *et al.* 2007; Bailleul *et al.* 2007; Biuw *et al.* 2010). They frequently

161   visit high latitudes for extended periods, diving deeply and almost continually,

162   returning to land twice a year to breed and moult. CTD-SRDLs are regularly used in

163   studies of their movement and diving behaviour, and that of other wide-ranging

164   phocid seals.

165

166   The characterisation of dives into types, based on dive parameters, has been a

167   popular approach to the study of diving behaviour (Hindell, Slip & Burton 1991;

168   Schreer & Testa 1996; Schreer, Kovacs & O'Hara Hines 2001; Baechler 2002). In

169   general, the identification of types or groups of behaviour is useful for comparing

170   behavioural patterns and activity budgets between individuals and in different spatial

171   and temporal contexts, and is carried out using a wide range of methods including

172   empirical methods, machine learning, and state-space methods, to name a few (e.g.,

173   Fauchald & Tveraa 2003; Thums, Bradshaw & Hindell 2008; McKellar *et al.* 2014).

174   BSM dives are used widely in studies of diving behaviour and physiology (e.g.,

175   McConnell *et al.* 1999; Biuw *et al.* 2007, 2010; Bailleul *et al.* 2007), without

176   considering or accounting for uncertainty in the abstracted dive profiles. Ignoring the

177   uncertainty in BSM-derived time-depth profiles may lead to incorrect inferences if the

178   BSM output has substantial error associated with it, and if dives with different shape

179   characteristics differ systematically in the amount of error associated with them. This

180   computationally expedient method has been thought to perform well at capturing

181   biologically relevant aspects of time-depth dive profiles, but this impression has, to

182   date, remained anecdotal.

183

184   **Aims and questions**

185   In this paper we aim to explain the BSM for dive profile abstraction, provide a

186   method that extracts as much information as possible from abstracted dive data, and

187   improve the interpretability of abstracted dive profiles. To do this we 1) present an

188   overview of the process by which the BSM generates abstracted dive profiles; 2)

189   assess the performance of the BSM for dive profile representation, by comparing

190   detailed and abstracted time-depth dive profiles from elephant seals, as a case

191   study; 3) present a three-step method for obtaining, *post hoc*, the depth limits on the

192   detailed dive (i.e., the dive zone) based on its BSM abstracted profile; 4) develop an

193   index of goodness-of-fit of abstracted dives (i.e., DZI) and use detailed dive profiles

194   to validate it; and 5) use this index to determine if there are systematic differences in

195   the amount of error associated with different dive types, following (Hindell *et al.*

196   1991).

197

198   We recast these five aims as research questions: 1) How does the BSM work for

199   dive profile abstraction? How does the representation of the detailed dive change

200   with increasing BSM points? 2) Is the sample of study dives representative?

201   3) What can we learn from abstracted dives? 4) How is the DZI derived? And, 5) Can

202   the DZI be used as a proxy for the RSS? Does the DZI vary systematically between

203   dive types?

204

205   MATERIALS AND METHODS

206   **How does the BSM work for dive profile abstraction?**

207   The BSM is an iterative process. For time-depth dive profiles, it is based on

208   minimising the vertical distance (i.e., difference in depth) between the detailed

209   trajectory recorded by the tag and the abstracted dive profile being proposed, at the

210   sampling resolution of the dive (for CTD-SRDLs 4s, 8s, 16s or 32s). We call these

211   vertical distances *residuals* (Figure 1). The basic principle of the model is that at

212   each iteration, the residuals are calculated, and the point with the biggest residual is

213   added to the abstracted profile. At the first iteration, which we call step zero, the

214   abstracted profile consists only of the start and end points of the dive, forming a

215   straight line at what the tag perceives to be zero depth. This corresponds to a depth

216   buffer at the surface (0-6 m), which is intended to exclude any less interesting

217   shallow undulations from the dive record, that would compete with regular deep

218   dives for transmission. The distance from this straight line to the detailed profile is

219   measured at each time point and the point at which the piecewise linear abstracted

220   profile deviates most from the detailed profile is added to the abstracted profile,

221   creating two new line segments. This is called a breakpoint. This step creates a new

222   piecewise linear profile comprising $I+2$ points, connected by linear segments, and

223   completes one iteration of the model (Figure 1).

224

225   The maximum residual (R_I) is calculated for each of the resulting line segments and

226   the point with the greatest departure is selected as the next breakpoint and added to

227   the profile. This process is repeated until the desired number of breakpoints is

228   reached, and the resulting piecewise linear abstracted profile has been constructed.

229   When the abstraction process is complete, the abstracted time-depth dive profile

230   includes $I+2$ time points (T_1 to T_I), and the corresponding I depth points (D_1 to D_I).

231   The first and last points, (T_0, T_{I+1}, D_0 and D_{I+1}) are not transmitted. At T_0 time is

232   considered to be zero and at T_{I+1} the time elapsed since the beginning of the dive

233   will equal the dive duration. Similarly, at D_0 and D_{i+1}, the depth will both be 0-6 m.

234   The order in which the time-depth points were selected is not stored or transmitted.

235

236   **How does the representation of the detailed dive change with increasing BSM**

237   **points?**

238   Using detailed dives from four high-resolution datasets, each representing a

239   continuous dive record from one individual (10943, 12454, 12453 and 12451 in

240   Table 1), we estimated the proportion of high-resolution samples that was

241   represented by the number of BSM points in the corresponding abstracted dive

242   profile (Table 2). We generated an abstracted profile with 3-12 BSM points for each

243   study dive. This resulted in a dataset of 2400 proportions, from 240 dives.

244

245   **Is the sample of study dives representative?**

246   Four iterations of the BSM are carried out on-board CTD-SRDLs, resulting in a dive

247   profile consisting of six time-depth points: two at the surface, and four at depth, at

248   irregular times, which vary from dive to dive. The number of iterations of the BSM

249   algorithm to be carried out on CTD-SRDLs was chosen as the minimum sufficient

250   number to convey the geometric shape of the dive profile, while keeping the number

251   of computations low (Fedak *et al.* 2001).

252

253   We confirmed that the sample of detailed dives from the four individuals was

254   consistent with elephant seal diving behaviour in general. To do this, the DZI was

255   calculated for a sample of 4,000 abstracted dives from 45 southern elephant seals

256   instrumented in four different field seasons (1,000 dives each from two post-moult

257   deployments and two post-breeding deployments; ct40, ct45, ct49, ct58) over two

258   years at the island of South Georgia, South Atlantic (Table 1). The resulting

259   distribution of DZI was visually compared with the distribution of DZI for the detailed

260   dives (Appendix S2, Figures S2.1 and S2.2).

261

262   **What can we learn from abstracted dives?**

263   The process by which abstracted dive profiles arise when they are collected by CTD-

264   SRDL, is known, therefore it is possible to reverse the deterministic steps and obtain

265   limits to the depth at which the trajectory could have passed, before it was

266   abstracted. The information required to build the dive zone (the 100% confidence

267   zone for depth) includes the 1) temporal resolution at which the dive was recorded

268   by the tag, the $I+2$ locations, in time and depth (including maximum dive depth), 2)

269   the residual associated with the final, $I+1^{th}$ breakpoint, and critically, 3) the order in

270   which breakpoints were selected during abstraction. The temporal resolution of the

271   detailed dive data is known from the duration of the dive, and the locations are

272   received in the satellite transmissions, but the residuals and the order in which the

273   breakpoints were added need to be determined. Constructing the dive zone involves

274   three steps.

275

276   First, to find the order in which points were added to the profile, the BSM must be

277   applied to the already abstracted dive trajectory (Figure 2). At each of the I iterations,

278   one of the breakpoints is selected as the point of greatest deviation hence retrieving

279   the order in which the breakpoints were added.

280

281   Second, to determine the limits of the zone, the residuals corresponding to the

282   breakpoints need to be calculated. It is tempting to assume that the residual

283   associated with the Ith breakpoint, R_I, applies to all segments in the final profile,

284   determining the limits of the dive zone. This is not the case. The limits of the dive

285   zone at each iteration of the model interact with those from previous iterations. As

286   breakpoints are added to the profile, the dive zone changes shape and size.

287   Although the dive zone will always get smaller with subsequent iterations of the

288   algorithm, this geometric effect means that the dive zone needs to be constructed

289   based on all iterations of the model, up to the last one. Furthermore, the resulting

290   dive zone at the final iteration is not symmetric around any segments in a profile and

291   all breakpoints touch the limit of the dive zone (Figure 3).

292

293   The depth points selected by the BSM are coded before transmission according to a

294   pseudo-logarithmic mantissa and exponent representation. With this representation,

295   resolution can be made proportional to the scale of the number being represented,

296   making it useful for depth. Data are then truncated during the decoding process,

297   once they are received. As a consequence, the received depth measurements are

298   binned (i.e. each reported depth has an upper and lower bound) and bin width

299   increases with depth. Bin width is usually smaller than the dive zone height at each

300   breakpoint. As a result, the dive zone is truncated near the deepest point(s) of the

301   dive where the depths along the line segment near the bottom of the dive exceed the

302   known accuracy of the greatest depth reached during the dive. This is justified since

303   we know that no point in the true profile can be deeper than the maximum depth

304   recorded by the tag. Depth is also truncated at each breakpoint, for the same

305   reason, so the dive zone appears "pinched" at each breakpoint (Figure 3).

306

307   When the Ith iteration is complete, it is known with certainty that there were no points

308   in the detailed trajectory that had a greater residual than the one corresponding to

309   the last point (Figure 2). All other depth points in the true trajectory will now have a

310   smaller vertical distance to the abstracted profile.

311

312   Third, to construct the dive zone, a number of equally spaced time points need to be

313   selected at which to sample vertical sections of the time-depth space. The resolution

314   of time points should not exceed the resolution at which depth data were collected by

315   the tag. At each of these time points ($t, ..., tmax$) the estimated lower, L_t, and upper,

316   U_t, depth bounds define the depth interval through which the true trajectory passed

317   with 100% confidence (Figure 3). Computer code for all algorithms described was

318   written in R (R Core Team 2013) and can be found in the supplementary material

319   (Appendix S1).

320

321   **How is the DZI derived?**

322   The approximation of a non-linear path will improve as the number of points (N) that

323   are used to approximate it increases. This should be reflected in a reduction in the

324   size of the maximum residual at the last iteration, as the number of BSM iterations

325   increases. In absolute terms, a small value for R_I means that the biggest vertical

326   outlier in the true path was a small distance away from the abstracted path, which

327   might suggest a better fit. However, R_I is affected by the depth, duration and

328   sinuosity of the detailed dive trajectory, and does not follow a strictly decreasing

329   relationship with the number of iterations of the model. When detailed dive data are

330   not available, a reliable and unbiased way of measuring goodness-of-fit is required to

331   assess the accuracy of abstracted profiles, irrespective of depth and duration. On its

332   own R_I does not provide an objective way of assessing goodness-of-fit over the

333   whole dive, because it depends on the maximum depth of the dive and the slopes

334   and lengths of the segments that make up the abstracted profile. The mean dive

335   zone height should be a better measure of fit, because it does contain information on

336   the whole dive, and includes the geometric effects that result from the slopes and

337   lengths of the segments that make up the abstracted dive, though importantly, it also

338   does not include information on sinuosity. However, to be comparable between dives

339   it needs to be standardised by dive depth and dive duration. This is the basis for the

340   construction of the DZI.

341

342   The DZI is calculated using the sum of the differences between the upper, U, and

343   lower, L, limits of the dive zone at each time step, t, in a dive. The sum of these

344   heights is divided by the product of the maximum dive depth, $maxdep$, and the

345   number of depth points that were recorded by the tag for the dive prior to abstraction,

346    $tmax$. This quotient ranges between 0 and 1, where values close to 0 indicate a

347   narrow zone around an abstracted dive and are desirable, and values close to 1

348   indicate a wide zone around an abstracted dive and relatively low confidence in the

349   abstracted dive profile.

$$DZI = \frac{\sum_{t=1}^{tmax} U_t - L_t}{maxdep * tmax}$$

350   Equation 1

351

352   **Can the DZI be used as a proxy for the RSS?**

353   When detailed dive data are available, goodness-of-fit can also be assessed using

354   the sum of squared residuals (RSS) between the detailed and abstracted depths,

$$RSS = \sum_{t=1}^{tmax} (T_t - A_t)^2$$

355   Equation 2

356   where T_t is the tth depth in the detailed profile and A_t is the tth depth in the linearly

357   interpolated abstracted profile. The RSS, an absolute measure of fit for detailed dive

358   profiles, was compared to the DZI, the relative measure of fit developed here for

359   abstracted dives, and the biggest residual at the last iteration of the BSM, R_I to

360   describe the relationship between them and determine if they could be used as a

361   proxy for the RSS.

362

363   To do this, the RSS, DZI and R_I were calculated for abstracted profiles with 3-12

364   BSM points for each of the 240 study dives (n=2400). RSS is a non-zero real

365   number, so a Generalized Additive Model (GAM) with a Gamma distribution and a

366   log link function were used, fitted with `mgcv` in R (Wood 2000, 2011). We wanted to

367   know if the DZI could be used as a proxy for the RSS in already abstracted dives

368   received by CTD-SRDL, but we were also interested to know if the amount of

369   deviance explained by the DZI changed with increasing breakpoints. Hence, we

370   fitted a model to a subset of the data representing six breakpoints, and did model

371   selection to find the best model, but also fitted a model to a subset of the data

372   representing each number of breakpoints explored in this study (three to twelve)

373   including only DZI as a covariate without doing model selection.

374

375   In the first case, RSS was modelled with DZI and R_4 as explanatory variables, and

376   individual seal as a random effect. In the second case, RSS was modelled with DZI

377   as the only explanatory variable, and individual seal as a random effect, as above. In

378   both cases, the relationship between the RSS and the covariates was non-linear, so

379   they were fitted as smooth functions with a shrinkage smoother ("cs") as the basis

380   function and k=4 knots. The number of knots was found to be sufficient using

381   standard `mgcv` checks. This basis allows for the smooth coefficients to be shrunk to

382   zero and effectively removed from the model when there is no relationship with the

383   response. We specified a gamma parameter value of 1.4 to reduce the chance of

384   overfitting. The random effect was fitted using the "re" smooth, as described above.

385   We used restricted maximum likelihood (REML) as the fitting method (Wood 2011).

386

387   **Does the DZI vary systematically between dive types?**

388   The availability of detailed dive data (depth sampled every 4s) made it possible to

389   investigate the effects of the abstraction process, develop methods to reverse the

390   abstraction and quantify goodness-of-fit. Detailed data were taken from a high

391   temporal resolution dataset recorded by a specially configured archival SRDL

392   deployed on a northern elephant seal after the moult at Año Nuevo, California, and

393   recovered six months later (Table 1). These dives were chosen by eye, as being

394   representative of the six functional or behavioural characterisations sometimes used

395   to classify elephant seal dives into types; U-shaped dives (U), V-shaped dives (V),

396   square-bottom dives (SQ), wiggle dives (W), root-shaped dives (R) and drift dives

397   (DR), after Hindell, Slip, and Burton (1991). Ten dives of each type from each

398　　individual were included in the case study. No detailed dive data from southern

399　　elephant seals were used, but several datasets of abstracted dives were available. A

400　　sample of over 22,000 abstracted dives that had already been classified into types

401　　were used to investigate systematic differences in goodness-of-fit between dive

402　　types (Table 1).

403

404　　The classification of abstracted dives into types was done using the random forest

405　　tree-building method (Breiman 2001). Random forest is a machine learning tool; we

406　　used an implementation in the `randomForest` library in R (Liaw & Wiener 2002). A

407　　supervised version of the method was used to classify dives, whereby 3,000 dives,

408　　14% of the dataset, were classified based on visual cues, and used to train the

409　　remaining dives. The overall "out-of-bag" error, an unbiased estimate of classification

410　　error, was 3.6%. This represents the aggregate of the prediction error rate at each

411　　bootstrap iteration (Liaw & Wiener 2002). The variables supplied to the function for

412　　classification were maximum dive depth, bathymetry and fifteen dive parameters

413　　(Photopoulos 2007). This method has been found to work well for dive classification,

414　　using both detailed and abstracted dive data (Thums *et al.* 2008).

415

416　　The DZI and R_4 were also calculated for the each dive in the dataset. The DZI used

417　　as the response variable in GAM with a quasibinomial distribution and a logit link

418　　function, dive type as a factor variable, R_4 as a smooth covariate using "cs" basis

419　　function (k=4 knots, checked as above) and individual animal as a simple random

420　　effect using the "re" smooth. The model was fitted using the REML method and

421　　gamma was specified as 1.4, as above.

422

423 Model assessment was done based on inspection of the residuals, the relationship

424 between the observations and the fitted values for each model, and the percentage

425 of deviance explained.

426

427 RESULTS

428 In the introduction, we outlined five questions relating to the BSM for dive profile

429 abstraction, the usefulness of the DZI as a goodness-of-fit measure and the validity

430 of our assessment of it as such: 1) How does the BSM work for dive profile

431 abstraction? How does the representation of the detailed dive change with

432 increasing BSM points? 2) Is the sample of study dives representative? 3) What can

433 we learn from abstracted dives? 4) How is the DZI derived? And, 5) Can the DZI be

434 used as a proxy for the RSS? Does the DZI vary systematically between dive types?

435

436 1) The BSM algorithm for dive profile abstraction is illustrated in Figure 1 and

437 animated in Appendix S2 of the supplementary material. The proportion of detailed

438 samples in a dive increases with the number of breakpoints in the abstracted profile,

439 and depends on dive duration. Overall, the proportion of a dive represented by its

440 abstracted profile is very low, starting at 1% with three breakpoints and reaching 4%

441 with twelve, and increases by a constant 0.33% for each breakpoints added (Table

442 2). The mean relationship was similar for all animals but there were differences in the

443 variability in the relationship (Figure 4).

444

445 2) Variability between the samples of abstracted and detailed dives was expected,

446 due to individual variability and the different regions where data were collected.

447 However, visual comparison of the distributions of DZI from the different samples did

448   not suggest any striking differences (Appendix S2, Figures S2.1 and S2.2,

449   Supplementary material).

450

451   3) Even though the detailed trajectory cannot be recovered unless the device is

452   physically recovered, the order in which breakpoints were added to the profile and

453   the 100% confidence limits to the detailed profile, which we call the dive zone, can

454   be calculated from abstracted dives. This makes it possible to derive the DZI (Figure

455   3).

456

457   4) The derivation of the DZI is based on the maximum depth of the dive, its duration

458   and the upper and lower limits to the dives zone.

459

460   5) There was a strong, positive relationship between the RSS and the DZI together

461   with R_4. We found that the DZI, R_4 and a random effect for individual explained 69%

462   of the variability in the RSS for abstracted profiles with six breakpoints (deviance

463   explained) (Figure 5). On its own, the DZI explained a variable proportion of

464   deviance for abstracted dives with differing numbers of breakpoints, but there was an

465   overall positive relationship with increasing breakpoints for dives with four or more

466   breakpoints (Figure 6). The DZI varied substantially between dive type and had an

467   increasing relationship with R_4 (Figure 7). The dive type associated with the biggest

468   DZI values were square dives (SQ) and both V-shaped (V) and drift dives (DR) had

469   the smallest DZI, under the model (Figure 8). Dive type and R_4 together explained

470   65% of the variability in the DZI (deviance explained), having accounted for

471   individual variability by fitting a random effect for individual.

472

473   DISCUSSION

474   Change-points models are used in many fields, from neuroscience and epidemiology

475   to genetics and finance, to identify changes in time-series. One of these, the BSM,

476   was adopted on board CTD-SRDLs, as a working solution to the problem of linearly

477   approximating a non-linear path in the vertical dimension with as little information as

478   possible, while aiming to retain biologically relevant content. Until now, the

479   approximation error associated with abstracted profiles had not been investigated.

480   We provide a way of calculating and summarizing the error associated with dive

481   profiles derived using the BSM, in order to assess the information content of

482   abstracted dives. Our results suggest that BSM-abstracted dive profiles do, in fact,

483   retain enough information to estimate goodness-of-fit of the abstracted profile to the

484   detailed profile. The strong, positive relationship between the DZI and R_4, and the

485   RSS is evidence for this. It means that researchers using BSM-abstracted dive

486   profiles to make inference about animal behaviour and diving ecology can now

487   calculate the DZI and R_4 and incorporate a relative measure of error into their

488   analyses.

489

490   With the BSM, as with other linear approximation methods, the number of iterations

491   is critical to the quality of the abstracted dive. Our results confirm the findings of a

492   preliminary investigation during tag development, regarding the adequacy of different

493   numbers of iterations, that resulted in the standard use of four iterations of the

494   algorithm for dives collected by CTD-SRDLs. That study found that the information

495   gained in the transmission of a fifth breakpoint is relatively small, and that it would be

496   more useful to receive a measure of the variance of the detailed dive, either for the

497   whole dive, or each segment, which is supported by the results presented in this

498   paper.

499

500   It is worth noting that although the BSM can efficiently summarise a curvilinear

501   trajectory, using a piecewise-linear shape, profiles resulting from low iteration

502   numbers, may mask important biological features even if they closely approximate

503   the detailed trajectory. Our methods do not provide means for assessing the

504   biological content of abstracted dives, since the detail lost, however small, might be

505   the most biologically interesting. For example, a useful feature of the BSM is that it is

506   efficient at identifying long sections where the trajectory has low variability. In the

507   case of dives, these are often the descent and ascent phases, leaving only two

508   points to confer information about the bottom phase of the dive, which is arguably

509   the most interesting biologically. Dives with low variability in change in depth in the

510   bottom phase, also have the lowest DZI, as we found here for drift dives (DR type).

511   However, when classifying abstracted dives with a method like a random forest

512   algorithm, ancillary behavioural data are necessary for validating classes as being

513   functionally distinct, in addition to being phenomenologically distinct.

514

515   Large numbers of dive profiles are collected using CTD-SRDLs from many different

516   species (over 21 million profiles since 1991, SMRU 2012, unpublished data) and

517   used to make inferences about the biology and behaviour of the instrumented

518   animals. It seems essential that a method for assessing the accuracy of these

519   abstracted dives, at least statistically if not biologically, is made widely available. The

520   methods we have presented here make that possible. They also provide a way of

521   carrying out a "pilot" analysis when detailed dive data are available. The differences

522   in diving behaviour between species, and habitats, may render different numbers of

523 breakpoints appropriate for the questions being asked. When detailed dive data are

524 available, the result of dive abstraction with different number of iterations can be

525 investigated to achieve the best result for a specific study, prior to deployment.

526 Together, these uses for our methods may help make more robust the behavioural

527 conclusions we can draw from telemetry data.

528

529 More generally, through this work we have developed a method for quantifying

530 uncertainty in the fitted values for BSMs, when the original time-series is no longer

531 available. This method could be applied to any situation where the original time-

532 series data are not longer available. This could be useful in situations where large

533 amounts of data are being generated and cannot be stored or transmitted at the

534 original resolution. We demonstrate that measures of fit like the DZI and R_4 have a

535 strong correspondence to the RSS and could therefore be used instead.

536

537 ACKNOWLEDGEMENTS

553

554   DATA ACCESSIBILITY

555   • R scripts: uploaded as supporting material

556   • Example data of detailed and abstracted dive data: uploaded as supporting

557   material

558   • For access to the full datasets used in this study please contact the data

559   owners directly. South Georgia Island and Año Nuevo deployments: Mike

560   Fedak (maf3@st-andrews.ac.uk) and Lars Boehme (lb284@st-

561   andrews.ac.uk). Kerguelen Islands deployments: Christophe Guinet

562   (Christophe.GUINET@cebc.cnrs.fr).

563

564   REFERENCES

565   Aoki, K., Watanabe, Y., Crocker, D., Robinson, P.W., Biuw, M., Costa, D., Miyazaki,

566   N., Fedak, M.A. & Miller, P.J.O. (2011) Northern elephant seals adjust gliding

567   and stroking patterns with changes in buoyancy: validation of at-sea metrics of

568   body density. *The Journal of experimental biology*, **214**, 2973–87.

569   Argos. (2011) Argos User's Manual. URL http://wwwargos-systemorg/files/pmedia/p

570   ublic/r363_9_argos_manual_en.pdf

571 Authier, M., Martin, C., Ponchon, A., Steelandt, S., Bentaleb, I. & Guinet, C. (2012)

572   Breaking the sticks: a hierarchical change-point model for estimating

573   ontogenetic shifts with stable isotope data. *Methods in Ecology and Evolution*, **3**,

574   281–290.

575 Baechler, J. (2002) Dive shapes reveal temporal changes in the foraging behaviour

576   of different age and sex classes of harbour seals (Phoca vitulina). *Canadian*

577   *Journal of Zoology*, **80**, 1569–1577.

578 Bailleul, F., Charrassin, J.-B., Monestiez, P., Roquet, F., Biuw, M. & Guinet, C.

579   (2007) Successful foraging zones of southern elephant seals from the

580   Kerguelen Islands in relation to oceanographic conditions. *Philosophical*

581   *transactions of the Royal Society of London. Series B, Biological sciences*, **362**,

582   2169–81.

583 Beckage, B., Joseph, L., Belisle, P., Wolfson, D.B. & Platt, W.J. (2007) Bayesian

584   change-point analyses in ecology. *The New Phytologist*, **174**, 456–67.

585 Biuw, M., Boehme, L., Guinet, C., Hindell, M., Costa, D., Charrassin, J.-B., Roquet,

586   F., Bailleul, F., Meredith, M., Thorpe, S.E., Tremblay, Y., McDonald, B., Park,

587   Y.-H., Rintoul, S.R., Bindoff, N., Goebel, M., Crocker, D., Lovell, P., Nicholson,

588   J., Monks, F. & Fedak, M.A. (2007) Variations in behavior and condition of a

589   Southern Ocean top predator in relation to in situ oceanographic conditions.

590   *Proceedings of the National Academy of Sciences of the United States of*

591   *America*, **104**, 13705–10.

592   Biuw, M., Nøst, O., Stien, A., Zhou, Q., Lydersen, C. & Kovacs, K. (2010) Effects of

593   hydrographic variability on the spatial, seasonal and diel diving patterns of

594   southern elephant seals in the eastern Weddell Sea. *PloS one*, **5**, 1–14.

595   Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5–32.

596   Campagna, C., Piola, A.R., Marin, M.R., Lewis, M., Zajaczkovski, U. & Fernández, T.

597   (2007) Deep divers in shallow seas: Southern elephant seals on the Patagonian

598   shelf. *Deep Sea Research Part I: Oceanographic Research Papers*, **54**, 1792–

599   1814.

600   Fauchald, P. & Tveraa, T. (2003) Using first-passage time in the analysis of area-

601   restricted search and habitat selection. *Ecology*, **84**, 282–288.

602   Fedak, M.A., Lovell, P. & Grant, S. (2001) Two approaches to compressing and

603   interpreting time-depth information as collected by time-depth recorders and

604   satellite-linked data recorders. *Marine Mammal Science*, **17**, 94–110.

605   Fedak, M.A., Lovell, P., McConnell, B. & Hunter, C. (2002) Overcoming the

606   constraints of long range radio telemetry from animals: getting more useful data

607   from smaller packages. *Integrative and comparative biology*, **42**, 3–10.

608   Hebblewhite, M. & Haydon, D.T. (2010) Distinguishing technology from biology: a

609   critical review of the use of GPS telemetry data in ecology. *Philosophical

610   transactions of the Royal Society of London. Series B, Biological sciences*, **365**,

611   2303–12.

612   Hindell, M., Slip, D. & Burton, H. (1991) The diving behaviour of adult male and

613   female southern elephant seals, Mirounga leonina (Pinnipedia : Phocidae).

614   *Australian Journal of Zoology*, **39**, 595–619.

615   Jonker, F. & Bester, M. (1998) Seasonal movements and foraging areas of adult

616   southern female elephant seals, Mirounga leonina, from Marion Island. *Antarctic*

617   *Science*, **10**, 21–30.

618   Liaw, A. & Wiener, M. (2002) Classification and regression by randomForest. *R*

619   *News*, **2**, 18–22.

620   MacArthur, R. (1957) On the relative abundance of bird species. *Proceedings of the*

621   *National Academy of Sciences of the United States of America*, **43**, 293–295.

622   MacArthur, R. & MacArthur, J. (1961) On bird species diversity. *Ecology*, **42**, 594–

623   598.

624   McConnell, B., Chambers, C. & Fedak, M.A. (1992) Foraging ecology of southern

625   elephant seals in relation to the bathymetry and productivity of the Southern

626   Ocean. *Antarctic Science*, **4**, 393–398.

627   McConnell, B., Fedak, M.A., Lovell, P. & Hammond, P. (1999) Movements and

628   foraging areas of grey seals in the North Sea. *Journal of Applied Ecology*, **36**,

629   573–590.

630   McKellar, a. E., Langrock, R., Walters, J.R. & Kesler, D.C. (2014) Using mixed

631   hidden Markov models to examine behavioral states in a cooperatively breeding

632   bird. *Behavioral Ecology*, **00**, 1–10.

633   Photopoulos, T. (2007) *Behavioural Changes of a Long-Ranging Diver in Response*

634   *to Oceanographic Conditions*. University of St Andrews.

635   R Core Team. (2013) R: A language and environment for statistical computing. R

636   Foundation for Statistical Computing.

637   Rual, P. (1989) For a better XBT bathy-message on board quality control, plus a new

638   data reduction method. *Western Pacific International Meeting and Workshop on*

639   *Toga Coare* (eds J. Picaut, R. Lukas & T. Delcroix), pp. 823–833. Noumea, New

640   Caledonia.

641   Schreer, J., Kovacs, K. & O'Hara Hines, R. (2001) Comparative diving patterns of

642   pinnipeds and seabirds. *Ecological Monographs*, **71**, 137–162.

643   Schreer, J. & Testa, J. (1996) Classification of Weddell seal diving behavior. *Marine*

644   *Mammal Science*, **12**, 227–250.

645   Thums, M., Bradshaw, C.J.A. & Hindell, M.A. (2008) A validated approach for

646   supervised dive classification in diving vertebrates. *Journal of Experimental*

647   *Marine Biology and Ecology*, **363**, 75–83.

648   Toms, J. & Lesperance, M. (2003) Piecewise regression: a tool for identifying

649   ecological thresholds. *Ecology*, **84**, 2034–2041.

650   Wood, S.N. (2000) Modelling and smoothing parameter estimation with multiple

651   quadratic penalties. *Journal of the Royal Statistical Society: Series B (Statistical*

652   *Methodology)*, **62**, 413–428.

653   Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood

654   estimation of semiparametric generalized linear models. *Journal of the Royal*

655   *Statistical Society: Series B (Statistical Methodology)*, **73**, 3–36.

656

Figures and Tables

Table 1. Deployment information and morphometrics for the detailed datasets from a northern and three southern elephant seal (a sample of 60 study dives was taken from the dive record of each seal), and the abstracted dataset(s) from 45 southern elephant seals. For "ct" deployments, length is given as the mean (standard error) of all animals of each sex. For deployments 12454, 12453, 12451, two length measurements were available for each animal, so we present the mean (standard error) of the two measurements. Only one length measurement was made during deployment 10943.

Deployment	Species	Deployment location	Period (UTC)	Sampling regime	Morphometrics	Sex	Dive duration
10943	*M. angustirostris* 1 adult male	Año Nuevo CA, USA	23/08/2008 to 16/02/2009	Time and depth every 4s **	Length: 350.0 cm Axial Girth: 295.0 cm	M	24.95 min (0.77)
12454	*M.leonina* 1 adult female	Kerguelen Islands	30/10/2012 to 15/02/2013	Time and depth every 4s	Length: 235.5.0 cm (6.5) Weight: 258 kg	F	19.18 min (0.63)
12453	*M.leonina* 1 adult female	Kerguelen Islands	20/10/2012 to 11/02/2013	Time and depth every 4s	Length: 271.5 cm (6.5) Weight: 425 kg	F	23.48 cm (0.67)
12451	*M.leonina* 1 adult female	Kerguelen Islands	30/10/2012 to 15/02/2013	Time and depth every 4s	Length: 234.0 cm (1.0) Weight: 275 kg	F	18.52 min (0.64)
ct1	*M. leonina* 6 adult females	Husvik, South Georgia Island	06/01/2004 to 25/08/2004	CTD_GEN_07B*	Length: 242.5 cm (6.7)	F	23.96 min (0.07)
ct8	*M. leonina* 7 adult females 4 adult males	Husvik, South Georgia Island	13/01/2005 to 31/10/2005	CTD_GEN_07B	Length: 255.7 cm (2.0) Length: 301.3 cm (8.5)	F M	36.59 min (0.25)
ct40	*M. leonina* 5 adult females 5 adult males	Husvik, South Georgia Island	28/01/2008 to 09/12/2008	CTD_GEN_07B	Length: 250.5 cm (13.0) Length: 343.6 cm (24.4)	F M	27.47 min (0.38)
ct45	*M. leonina* 10 adult females	Husvik, South Georgia Island	17/10/2008 to 01/02/2009	CTD_GEN_07B	Length: 248.7 cm (2.9)	F	19.77 min (0.18)
ct49	*M. leonina* 11 adult females 1 adult male	Husvik, South Georgia Island	28/01/2009 to 12/01/2010	CTD_GEN_07B	Length: 240.5 cm (4.0) Length: 237.00 cm	F M	25.92 min (0.35)

Photopoulou et al. Efficient abstracting of dive profiles

29

| ct58 | *M. leonina*
13 adult females | Husvik,
South Georgia Island | 22/10/2009 to
05/02/2010 | CTD_GEN_07B | Length: 249.5 cm (3.2) | F | 18.97 min (0.19) |

* CTD_GEN_07B is the parameter specification with which the instruments were programmed.

** This tag sampled time and depth at 1Hz but resolution was reduced to one sample every 4s for consistency other time-depth data. The tag operated on a three-day duty cycle (3 days on, 3 days off).

Table 2. Observed mean proportion of detailed depth samples from individuals 10943, 12454, 12453, 12451 represented by an abstracted profile of a given number of BSM points (standard error, SE), and mean DZI (standard error, SE). The grey row highlights the case of a dive profile with 6 points, produced by 4 iterations of the BSM.

Number of BSM points	Mean proportion of depth samples represented by abstracted profile (SE)	Mean DZI (SE)
3	0.010 (0.0002)	0.993 (0.0002)
4	0.013 (0.0003)	0.549 (0.0156)
5	0.017 (0.0003)	0.303 (0.0103)
6	0.020 (0.0004)	0.158 (0.0048)
7	0.023 (0.0005)	0.110 (0.0036)
8	0.027 (0.0005)	0.085 (0.0030)
9	0.030 (0.0006)	0.071 (0.0026)
10	0.033 (0.0007)	0.060 (0.0023)
11	0.037 (0.0007)	0.051 (0.0019)
12	0.040 (0.0008)	0.044 (0.0017)

Photopoulou et al. Efficient abstracting of dive profiles

31
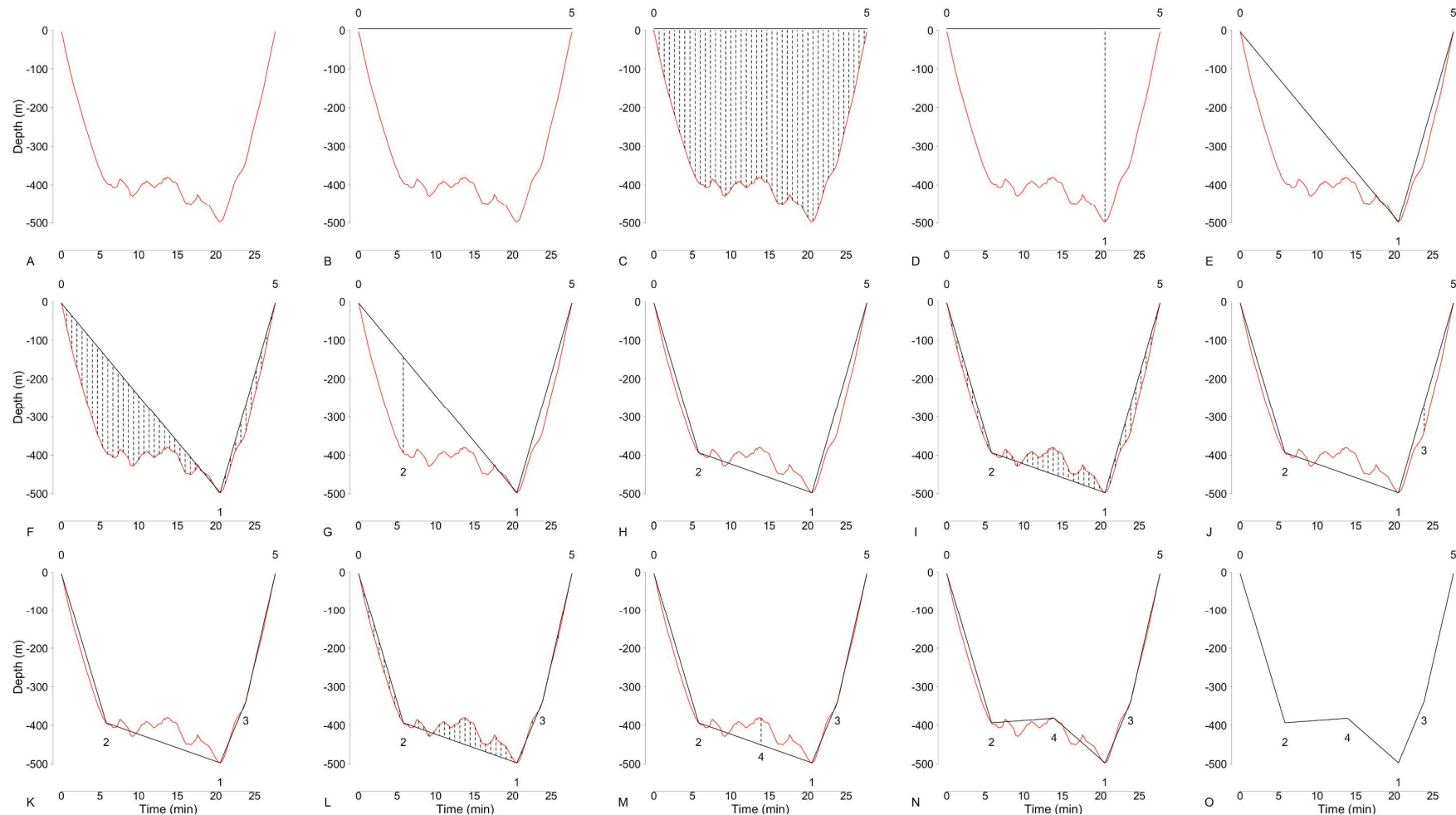
Figure 1. Step-wise illustration of time-depth dive profile abstraction for a dive from elephant seal 10943 (see Table 1) using a broken-stick model. With this abstraction regime, highly resolved time-depth measurements are processed to generate abstracted dive profiles made up of $I+1$ consecutive line segments that approximate the true, non-linear time-depth trajectory travelled. This is achieved through I (=4, here) iterations of the algorithm. The red lines represent the true time-depth dive path, the solid black lines represent the abstracted dive path and the dashed lines represent the residuals for the abstracted dive

Photopoulou et al. Efficient abstracting of dive profiles

path, at each iteration of the algorithm. The numbers represent the order in which points are added to the abstracted profile. Points 0 and 5 mark the beginning and end of the dive. The dashed black lines represent the residuals that are calculated at each time step between the true and abstracted dive.
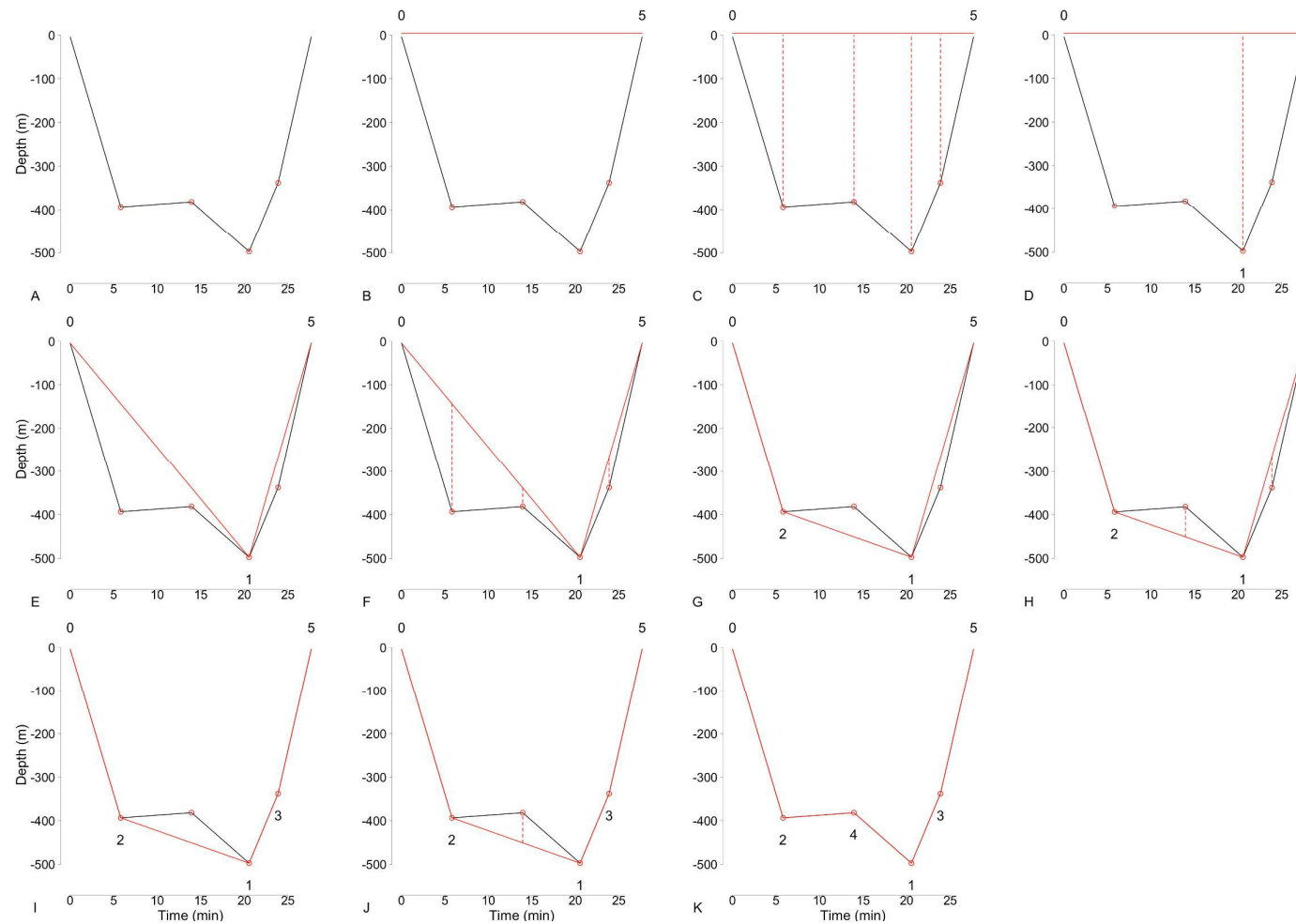
Figure 2. Step-wise illustration of the calculation of the order in which breakpoints were added to a time-depth dive profile abstracted using the BSM, received from CTD-SRDLs telemetry devices. The black line represents the abstracted time-depth dive profile received from the device. The red points represent known breakpoints. The solid red lines the proposed BSM abstracted dive path at each iteration of the model, and the dashed red lines represent the residuals that correspond to known breakpoints.
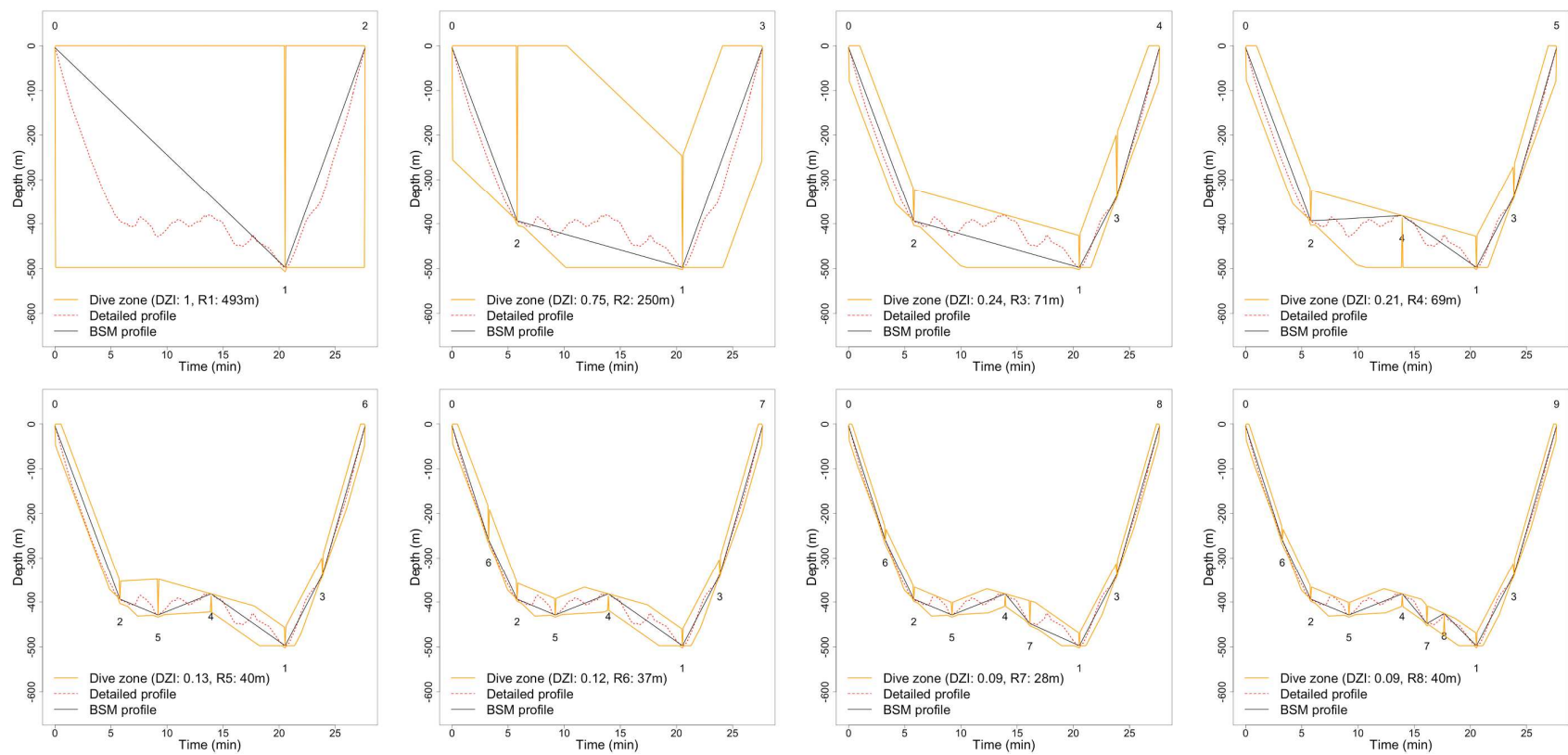
Figure 3. Step-wise illustration of the construction of the dive zone and its evolution as points are added to the abstracted dive profile. We use an example dive from individual 10943 (see Table 1), with up to ten iterations of the broken-stick model, i.e., using 3-12 points, in total.

Photopoulou et al. Efficient abstracting of dive profiles

1
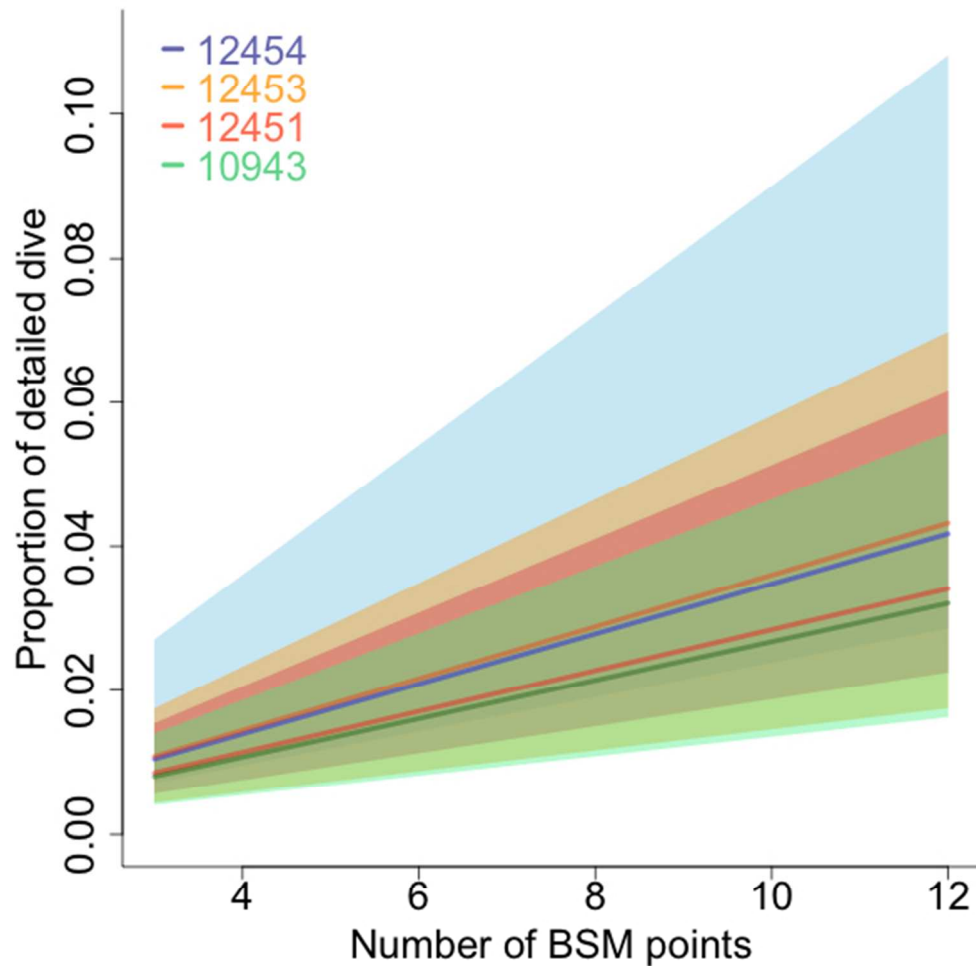
Figure 4. The relationship between the proportion of high-resolution time-depth samples and the number of BSM points in the abstracted dive profiles of 240 case study dives from a northern elephant seal and three southern elephant seals (see Table 1). Abstracted profiles with 3-12 points in total were generated for each of the study dives and compared with the full resolution profile. The coloured areas incudes minimum and maximum observed range of the relationship for each individual.
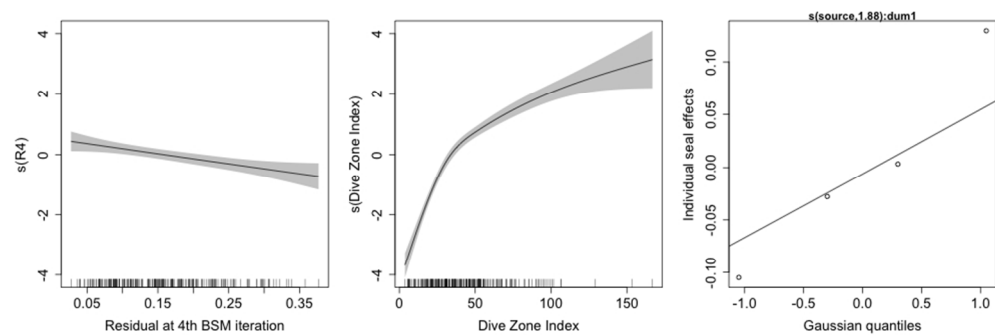211x211mm (72 x 72 DPI)

Figure 5. Smooth functions of the covariates in a Generalized Additive Model with log(RSS) as the response, DZI a smooth covariate, and individual dive as a random effect. The data used to fit this model were abstracted dive profiles of 4,000 study dives from 45 southern elephant seals instrumented at South Georgia Island in 2008 and 2009. The grey area includes two standard errors for the fitted relationship.
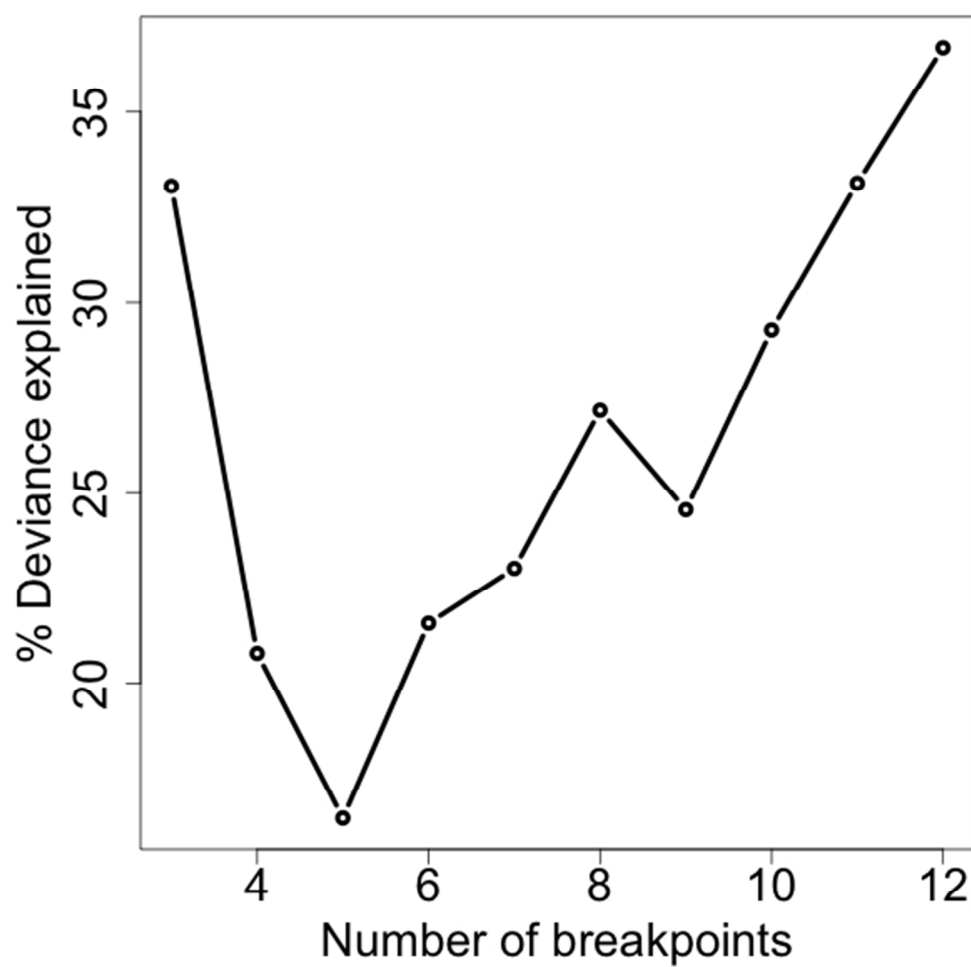352x117mm (72 x 72 DPI)

Figure 6. The relationship between the percentage of deviance explained by a model for RSS with DZI as the explanatory variable and individual seal as a random effect, and the number of breakpoints in the dives being considered.
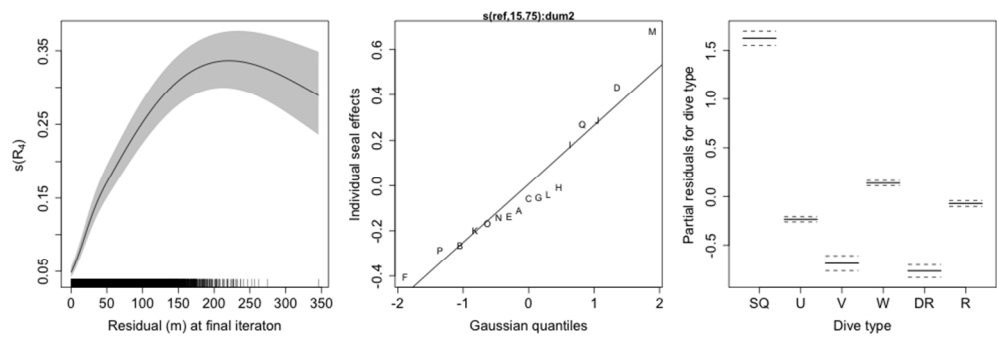211x211mm (72 x 72 DPI)

Figure 7. Smooth functions of the covariates in a Generalized Additive Model with dive zone index as the response, dive type as a factor variable, the residual as the final iteration of the BSM as a smooth covariate, and individual seal as a random effect. The data used to fit this model were abstracted dive profiles of 22,305 study dives from 17 southern elephant seals instrumented at South Georgia Island in 2004 and 2005. The grey area includes two standard errors for the fitted relationship.
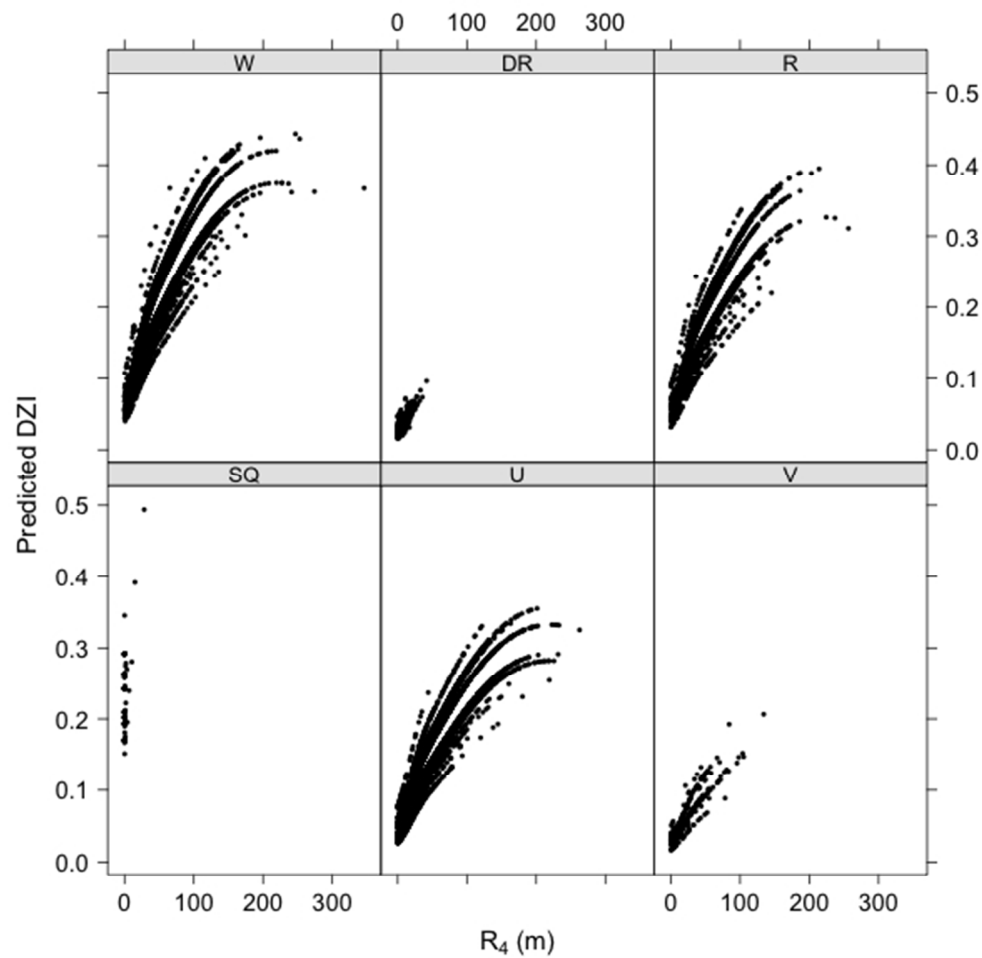352x117mm (72 x 72 DPI)

Figure 8. The fitted relationship between the dive zone index (DZI) and the residual at the final interaction of the broken stick model (R4) for each dive type. These are the predictions based on a Generalized Additive Model, with R4 as smooth covariate, dive type as a factor variable, and individual seal as a random effect. The data used to fit this model were abstracted dive profiles of 22,305 study dives from 17 southern elephant seals instrumented at South Georgia Island in 2004 and 2005.

211x211mm (72 x 72 DPI)