

1 03 August 2009
2 Bram Van Moorter
3 Department of Biology
4 Norwegian Institute of Technology in Trondheim (NTNU)
5 Realfagbygget, Høgskoleringen 5, NO-7491 Trondheim (Norway)
6 +47 73596266; fax: +47 73596100
7 Bram.Van.Moorter@gmail.com

8
9 RH: Van Moorter et al. · From Behavior to Movement States through Clustering

10 **Identifying Movement States from Location Data using Cluster Analysis**

11 BRAM VAN MOORTER¹, *Department of Biology*, Norwegian Institute of Technology in
12 Trondheim, Høgskoleringen 5, NO-7491 Trondheim, Norway

13 DARCY R. VISSCHER, *Department of Biological Sciences*, University of Alberta, Edmonton,
14 AB T6G 2E9, Canada

15 CHRISTOPHER L. JERDE², *Department of Biological Sciences*, University of Alberta,
16 Edmonton, AB T6G 2E9, Canada

17 JACQUELINE L. FRAIR³, *Department of Biological Sciences*, University of Alberta,
18 Edmonton, AB T6G 2E9, Canada

19 EVELYN H. MERRILL, *Department of Biological Sciences*, University of Alberta, Edmonton,
20 AB T6G 2E9, Canada

21 **ABSTRACT** Animal movement studies regularly use movement states (such as slow and fast)
22 derived from remotely sensed locations to make inferences about strategies of resource use.
23 However, the number of movement state categories used is often arbitrary and rarely evaluated
24 from the data. Identifying groups with similar movement characteristics (i.e., movement states)
25 is a statistical problem. We present a k-means clustering procedure along with the gap statistic as

¹ E-mail: Bram.Van.Moorter@gmail.com

² Present address: *Department of Biological Sciences*, University of Notre Dame, P.O. Box 369, Notre Dame, Indiana 46556-0369, USA

³ Present address: *Department of Environmental and Forest Biology*, 250 Illick Hall, State University of New York

26 a framework for evaluating the number of potential movement states without making a priori
27 assumptions about the number of clusters present. We illustrate the use of this methodology for
28 distinguishing movement states of free-ranging elk (*Cervus elaphus*) in west-central Alberta
29 using turning angle and step length derived from GPS locations, and horizontal and vertical head
30 movements derived from tip switches in a neck collar. We suggest how researchers could
31 biologically interpret the clusters by linking them to landscape features.

32 **KEY WORDS** Alberta, *Cervus elaphus*, gap statistic, GPS, K-means, movement state.

33 *Journal of Wildlife Management: 00(0): 000-000, 200x*

34 **INTRODUCTION**

35 Linking animal behavior to processes, such as dispersal, population dynamics, and habitat
36 selection is an important theme in ecology (Lima and Zollner 1996, Nathan et al. 2008). A major
37 focus of the behavior-oriented approach has been to understand how landscape heterogeneity
38 influences animal movements (Wiens 1989, Crist et al. 1992, Johnson et al. 2002). While early
39 animal movement studies primarily relied on directly observing animals in controlled arenas or
40 animals that had relatively narrow ranges of movement (Jones 1977, Roitberg and Mangel 1997),
41 advances in the technology of global positioning systems (GPS) now allows fine-scale
42 monitoring of movements with high accuracy for free-ranging animals. This has led to a surge in
43 methods for characterizing movement patterns and identifying movement states (Johnson et al.
44 1992b, Fauchald and Tveraa 2003, Nams 2005, Morales et al. 2004, Luque and Guinet 2007,
45 Barraquand and Benhamou 2008).

46 One common approach, with roots in random walk theory (Kareiva and Shigesada 1983),
47 uses sequential GPS locations to describe the animal's movement path in terms of distance
48 moved between points (referred to as step lengths) and the turning angle between points (Turchin

49 1998). Based on changes in characteristics of these measures (e.g. differences in the distribution
50 of step lengths and turning angles), several studies have successfully identified multiple
51 movement “states” (Patterson et al. 2008), such as slow and fast states thought to reflect within
52 and between patch movements or foraging and exploratory movements (Morales et al. 2004).
53 Such state-space modeling provides a powerful means to model movement processes given
54 multiple movement states (e.g. Morales et al. 2004), and can also be used to explore the number
55 of identifiable states using information criteria (Burnham and Anderson 2003) to select among
56 models based on different numbers of states. However, the computational challenges associated
57 with large numbers of parameters, which are needed for a realistic movement model, and the
58 assumptions involved in the modeling of multi-state movement processes are not trivial,
59 especially for data with many variables and many states. Ideally, a technique for identifying the
60 number of movement states would evaluate the evidence in multivariate movement data for how
61 many groupings are distinguishable without strong assumptions about the movement process
62 itself. Such an exploratory analysis may be complementary to state-space modeling by providing
63 initial parameter estimates associated with a set number of states.

64 In this paper, we present a multivariate clustering approach based on k-means for
65 statistically identifying patterns of different groupings of movement characteristics, and we
66 illustrate its use with movement data taken from elk (*Cervus elaphus*) in west central Alberta,
67 Canada. We assume no a priori underlying joint distributions, nor number of movement states,
68 and we use the gap statistic (Tibshirani et al. 2001) to evaluate the evidence for the number of
69 movement clusters. The gap statistic compares the within-cluster variance from the observed data
70 to that expected from a reference distribution; the smallest number of clusters is selected where
71 this gap statistic shows a large increase. We also demonstrate the importance of preprocessing

72 and standardization of data in cluster analyses. Finally, we suggest how the movement states can
73 be linked back to landscape patterns, when a movement model is not the research aim (if this
74 were the case, state-space modeling may provide a useful framework to model these patterns
75 with movement processes, e.g. Morales et al. 2004).

76

77 **METHODS**

78 **K-means Clustering and Gap Statistic**

79 The purpose of cluster algorithms is to group observations with similar characteristics using
80 multivariate information. Fisher (1958) stated the problem as: given n observations with d
81 variables, find k similar groups. When there are 2 variables, the clusters are identifiable using a
82 scatter plot and by inspecting the density of points (Fig. 1A). Clusters of similarly related
83 observations have a high density and the separation between clusters shows low or no density.
84 For the problem of identifying different movement states, we would have n observations (animal
85 locations) each with d associated variables (such as step length and turning angle) where we
86 attempt to identify k groups (movement behaviors, modes, or states) based on the movement
87 characteristics.

88 One commonly used algorithm for clustering data is the k-means procedure (MacQueen
89 1967). This procedure starts by initializing the number of clusters, k , identified within the data.
90 Once k is designated, k random, initial nodes, often referred to as seed points, are generated. Fig.
91 1A shows simulated data with two randomly initialized seed points. Once the seeds are
92 initialized, a multivariate distance measure (most often a squared Euclidian distance) is
93 calculated from each data point to each seed. Observations are assigned to the seed with the
94 shortest distance, which reflects similarity in the observations. The within-cluster means of all

95 variables are then calculated. The vector of variable means then becomes the new seed point for
96 each cluster. The process is iterated until membership of observations within clusters stabilizes.
97 Fig. 1B shows the final seed points with the surrounding observations assigned to the cluster.
98 The k-means algorithm can be sensitive to the initial seed points. One approach to solve this
99 problem is to run the clustering routine multiple times with different initial seed points and
100 compare the results (both the number of clusters and the cluster each datum is assigned to).
101 Steinley (2006a) provided a comprehensive review of k-means clustering, and we refer the
102 reader to Johnson and Wichern (1998) or Legendre and Legendre (1998) for a thorough
103 treatment of k-means analysis.

104 Determining the variables important to identifying clusters is an important area of current
105 research. Kaufman and Rousseauw (1990) noted the addition of non-informative variables will
106 confound the underlying cluster structure. Building upon this observation, Lleti et al. (2004)
107 developed a procedure to eliminate non-informative variables to investigate clustering when
108 there are a large number of variables available. However, as Johnson and Wichern (1998)
109 suggested that the selection of variables to be included in the clustering is best justified by the
110 process, in our case, the biology of movement variables related to between-patch and within-
111 patch movements.

112 Studies have shown that the outcome of cluster analyses, including methods like k-means
113 are sensitive to the distribution and range in values of the input (x) variables (Steinley 2006b,
114 Yingqiu et al. 2007). Heavily skewed data can hide the cluster structure present in these data,
115 whereas range differences among variables affect contributions from different variables to the
116 clustering result. Clustering relies on multi-dimensional measurement of (Euclidean) distance;
117 variables with a larger range will have a larger contribution to this distance measure (Kaufman

118 and Rousseeuw 1990). Therefore, variables with a larger range will have an increased weight in
 119 cluster analyses. As a result both data transformation and standardization have been
 120 recommended (Yingqiu et al. 2007). Range standardization is the recommended standardization
 121 procedure for clustering (Steinley 2006a):

$$122 \quad z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

123 Due to the role of the minimum and maximum value of a variable in range standardization, it is
 124 crucial to inspect the data for outliers (i.e. atypical data that are distant from the rest of the data)
 125 before performing this standardization.

126 It is critical in clustering observations into groups to determine the number of groups
 127 present in the data. There are a few suggested methods for evaluating the evidence for the
 128 number of clusters (Calinski and Harabasz 1974, Milligan and Cooper 1985). Recently,
 129 Tibshirani et al. (2001) proposed the gap statistic and found it correctly identified the true
 130 number of clusters more regularly than other approaches when compared in simulation. The gap
 131 statistic (Equation 2) provides a hypothesis testing framework that requires 4 steps. First, using
 132 the observed data, cluster analysis is performed for $k = 1, 2, 3 \dots n$. With the data clusters defined,
 133 a measure of the dispersion of the observations is calculated by summing the within cluster
 134 squared Euclidean distances (i.e., $SS_{k,data}$ for a given k). Second, a number of reference data sets
 135 (β) are created by Monte Carlo simulation using the uniform distribution on the principal
 136 components. Fig. 1B shows the distribution of observations on the principal components, and
 137 Fig. 1C shows 1 random reference data set on the same principal components. Principal
 138 components are used to ensure the reference data span the same space as the observed data. The
 139 reference data sets are then clustered for $k = 1, 2, 3 \dots n$, and the $SS_{k,ref}$ is calculated for each
 140 reference data set.

$$141 \quad \text{gap}(k) = E[\log(\text{SS}_{k,\text{ref}})] - \log(\text{SS}_{k,\text{data}}) \quad (2)$$

$$142 \quad \text{SE}_k = \sqrt{\frac{1}{\beta} \sum_{b=1}^{\beta} \left(\log(\text{SS}_{k,\text{ref}(b)}) - \frac{1}{\beta} \sum_{b=1}^{\beta} \log(\text{SS}_{k,\text{ref}(b)}) \right)^2} \sqrt{1 + \frac{1}{\beta}} \quad (3)$$

143 Third, the standard error, SE_k (Equation 3), from the reference data set is required to
 144 create a decision rule. Finally, the decision rule in the gap framework is to choose the smallest
 145 number of clusters, k , such that the gap statistic of the data with k groups is greater than or equal
 146 to the gap statistic of $k + 1$ clusters less the standard error of $k + 1$ clusters (Equation 4).

$$147 \quad \hat{k} = \min(k) : \text{GAP}(k) \geq \text{GAP}(k + 1) - T \times \text{SE}_{k+1} \quad (4)$$

148 The tolerance T is analogous to setting the alpha level in the standard hypothesis testing
 149 framework, where increased tolerance is similar to selecting a smaller alpha rejection region.
 150 Tibshirani et al. (2001) used a tolerance of 1, but larger values of tolerance increase the strength
 151 of evidence required to include additional clusters (see Tibshirani et al. 2001 for full details and
 152 formulations). Tibshirani et al. (2001) demonstrated that the gap statistic performed well in
 153 detecting the number of clusters when clusters are well separated, however it was sensitive to the
 154 amount of overlap between clusters. Fortunately, the bias in sensitivity is such that it is likely to
 155 identify fewer clusters than there are in truth.

156

157 **Example: Analysis of Elk Movement Data**

158 *Elk movement data.* — We illustrated this approach with data from elk in the Rocky Mountain
 159 foothills of west–central Alberta, adjacent to Jasper and Banff National Parks (Frair et al. 2005,
 160 Frair et al. 2007). The area is topographically diverse and predominately forested (> 60 %),
 161 making direct observation of an individual elk’s movement behavior impractical. Forest
 162 harvesting in the area has created small clear-cut patches of regenerating forest (hereafter,

163 cutblocks) throughout the forest matrix. These patches provide considerably more foraging
164 opportunities than the surrounding forest, making them attractive to elk (Visscher and Merrill
165 2009).

166 We used locations sampled every 2-hr from 15 June - 15 September 2002 for a resident
167 female elk captured by aerial netgunning (University of Alberta Animal Care and Use Protocol #
168 300201) and equipped with a GPS collar (GPS 2200, Lotek Wireless, New Market, Ontario,
169 Canada). For more details on the data collection see Frair et al. (2005). A total of 930 locations
170 were recorded throughout this 3-month period, with an 83% fix success rate. Using consecutive
171 locations stored in the collar, we calculated the step length and turning angle for each movement
172 segment (Turchin 1998, Jerde and Visscher 2005). All statistics were calculated only for 3
173 consecutive successful fixes to ensure comparable time intervals. Turning angles ranged from 0°
174 to 180°, so they were always positive. The neck collar also contained an activity sensor that
175 recorded the number of times a tip switch was activated by an up and down (vertical) or side to
176 side (horizontal) motion of the head during a preset 4-min period directly prior to when the collar
177 acquired an animal's location. Frequency of vertical or horizontal head movements ranged from
178 zero to 255 (maximum count stored over 4 minutes). Step lengths and turning angles described
179 movement trajectories (Turchin 1998); head movements have been shown to reflect feeding
180 behavior (Adrados et al. 2003). We used both the measures of path trajectories (step length [SL],
181 turning angle [TA]) and head movements (vertical activity [ACT1] and horizontal activity
182 [ACT2]) as variables in the cluster analysis because we expected they had the potential to
183 distinguish behaviors reflecting within and between feeding patch movements.

184 *Data preprocessing.* — Both activity measures and step length were log-transformed to reduce
185 positively skew. Further, we standardized all variable values on their range (Steinley 2006a).

186 Because we found no outliers (no data seemed outlying by visual inspection of the distribution,
187 nor were any observations farther than 3.3 standard deviations from the mean, which
188 corresponds to 1‰ in a normal distribution), we did not remove any data prior to data
189 standardization.

190 *K-means analysis.* — To demonstrate the importance of data preprocessing, we first performed a
191 cluster analysis on the raw data, these data after log-transformation, and then after range
192 standardization. In subsequent analyses, we used the transformed and standardized data, which
193 has been the recommended strategy. Second, we replicated the k-means analysis 3 times to
194 investigate the effect of variable inclusion on choice of number of groups using: (1) all 4
195 movement measures, (2) only trajectory measures (i.e., SL and TA), and (3) only activity
196 measures (i.e., ACT1 and ACT2). Finally, we performed 1 more analysis using all 4
197 measurements, but with unequal weighting of the different variables. The use of differential
198 weighting for further analysis of these data was suggested by the results from our previous
199 analyses on the effect of variable inclusion. For each of these analyses, we used the gap statistic
200 with 50 reference data sets (β in Equation 3) and tolerance levels of 1 and 2. The k-means cluster
201 analyses were conducted 100 times with random seed points. All procedures for the gap
202 framework we describe are available in the open-source software for statistical computing and
203 graphics R (Development Core Team 2008), using package clusterSim (Walesiak and Dudek
204 2008). A modified version of the function `index.Gap` is available at [http://ase-](http://ase-research.org/moorter)
205 [research.org/moorter](http://ase-research.org/moorter), together with the example data used in this paper.

206

207 **RESULTS**

208 **Untransformed and Non-standardized variables**

209 The raw data without any data preprocessing did not reveal any cluster structure; the gap statistic
210 was maximum for 1 cluster (i.e., no cluster structure). After log-transforming both activity
211 measures and step lengths, we found a structure with 2 clusters. The 2 movement states differed
212 only in their degree of directional persistence (mean TA: 43° versus 136° ; $t_{897} = -55$, $P < 0.001$).
213 The larger range of the turning angles (0° to 180°) than the other measures (on a log-scale: 0 to
214 8.22 for the step lengths and 0 to 5.55 for both activity measures) explained the clustering based
215 on turning angles. Fig. 2A, B shows how range standardization of the different variables led to a
216 distinctly different cluster structure (instead of on turning angles, the clustering is now mainly
217 based on both activity measures). These results illustrated the importance of proper data
218 preprocessing before any cluster analysis.

219

220 **Equally weighted variables**

221 Irrespective of whether only trajectory measures (SL, TA), only activity measures (ACT1,
222 ACT2) or all 4 variables were included in the k-means analysis, we found evidence for only 2
223 movement states at both levels of tolerance (Fig. 2A, 2C and 2E). However, when using all 4
224 measures, the gap statistic continued to increase for more clusters than 2 (Fig. 2A). Such an
225 increase in gap values could indicate a more complex or even hierarchical structure in the data
226 (Tibshirani et al. 2001). We used variable weighting to further explore the cluster structure in our
227 data (see below).

228 Although we found agreement about the number of movement states among the analyses
229 with different input variables, we also found important differences in cluster assignment of the
230 data. First, the agreement between cluster membership based on trajectory measures and on all 4
231 measures were barely different from random (proportion agreement = 0.54; $\chi_1^2 = 3.6$, $P = 0.06$),

232 whereas the agreement between clusters of movement states from activity measures and from all
233 input variables was nearly the same (proportion agreement = 1.00, $\chi_1^2 = 909.0$, $P < 0.001$). This
234 indicated that when all variables were analyzed together, cluster assignment was driven primarily
235 by the activity measures and that movement behaviors related to head movement were much
236 more distinct than trajectory variables. Nonetheless, the pattern of turning angles among
237 movement states also differed among movement states based solely on trajectory data (Fig. 2D).
238 One movement state ($n = 421$) was characterized by strong directional persistence, whereas the
239 other ($n = 503$) indexed frequent direction reversals by elk.

240

241 **Differentially weighted variables**

242 Two aspects of the previous analyses suggested that the cluster structure present in the trajectory
243 measures (notably, TA) was lost in the structure presented by both activity measures: first, the
244 clusters found using only the trajectory measures did not show up in the clustering based on all
245 variables, and second, the gap statistic for clustering based on all 4 variables continued to
246 increase after 2 clusters (Fig. 1A). This dominant effect of the activity measures can be explained
247 by the high correlation between both activity measures ($r = 0.51$), which is similar to pseudo-
248 replication. Therefore, we continued our investigation of the cluster structure by decreasing the
249 weight of both activity measures. To compensate for the variance shared by both variables ($r^2 =$
250 0.26), we decreased the weight of each activity measures by 0.13 (i.e. $r^2/2$).

251 The reduced weight of the activity measures led to a more stable gap statistic pattern (see
252 Fig. 2G). The value of the gap statistic increased until 4 or 5 clusters were identified, after which
253 it became rather stable. Such pattern indicated that most of the cluster structure was accounted
254 for with 4 or 5 clusters. Whether 4 or 5 clusters were selected depended on the choice of

255 tolerance value (respectively 2 or 1), with 4 clusters a more conservative and parsimonious
256 choice. Fig. 2H and Table 1 summarize the movement characteristics of the 4 movement state
257 clusters. These 4 clusters corresponded roughly to a subdivision of both activity-based clusters
258 into 2 sub-clusters based on the turning angle. Thus, the 4 movement states were the 4
259 combinations of high and low activity with high and low directional persistence.

260

261 **DISCUSSION**

262 The strength of the k-means approach with the gap statistic as a framework for identifying
263 movement states is that it allows the movement data to be explored in a hypothesis testing
264 framework, using a well formulated statistical criterion (i.e., gap statistic) to identify the number
265 of movement states. Unlike previous approaches to investigate movement states (e.g. Morales et
266 al. 2004), k-means clustering investigates data patterns without the requirement to model the
267 processes underlying these data. Therefore, this approach provides a simple way to obtain a
268 better understanding of the cluster structure and how the different variables are related to this
269 structure, as is illustrated by our analysis (see Fig. 2).

270 From our results it also was clear that the outcome of the k-means was sensitive to the
271 distribution of the data, because in the absence of any data preprocessing no clusters were
272 identified in our data. Data transformations, such as a logarithmic transformation, can decrease
273 the skew in data, which can aid in revealing cluster structure (for another example see Yingqiu et
274 al. 2007). It is, therefore, important to carefully inspect the data before undertaking a cluster
275 analysis. Equally important is to use a comparable measurement scale for all variables. When
276 large scale-differences exist, those variables with the largest range will contribute more to the
277 Euclidean distances among points, resulting in an unintended weighting of the variables. Range

278 standardization has been demonstrated as one of the better approaches to remove scale
279 differences in k-means clustering (Steinley 2006a). Intentional variable weighting can be useful
280 when there are statistical (e.g. Steinley and Brusco 2008) or biological reasons to assign more
281 weight in the clustering approach to specific variables. We demonstrated this by decreasing the
282 weight of the head movement data to counter the effect of having included 2 such highly
283 correlated measures in the analysis. More generally, we would recommend in the case of
284 multiple measurements of the same underlying dimension, like our measures of head movement,
285 to decrease the weight of these variables proportional to their common variance.

286 Using the approach we presented, the researcher can explicitly test for the number of
287 distinctive states and can adjust the strength of evidence required to add clusters by changing the
288 tolerance used in the gap statistic. The tolerance T in equation 4 determines the number of
289 standard errors used to assess the change in gap statistic with increasing number of clusters.
290 Similar to setting the alpha level in conventional statistics, choosing a tolerance T leads to more
291 or less conservative choices in the number of clusters (increasing T leads to a more conservative
292 choice). In most of our analyses, we did not observe a change in the number of clusters with
293 different levels of tolerance (i.e., 1 or 2), indicating low uncertainty regarding the number of
294 clusters present in the data. However, in the analysis with weighted activity measures we found
295 some uncertainty regarding the number of clusters. The more conservative choice of requiring
296 changes of at least 2 standard errors led to the selection of 4 clusters, whereas the more relaxed
297 criterion of 1 standard error increase would have led to the selection of 5 clusters. Our analysis
298 illustrated the value of careful inspection of the changes in the gap statistic with the number of
299 clusters in the development of the analysis, instead of exclusively relying on an arbitrary
300 tolerance value (similar to the arbitrary nature of the alpha level). If there is a clear peak (as in

301 Fig. 2D), then low uncertainty in the number of clusters exists, with all clusters being clearly
302 separated. However, when the gap statistic increases beyond the selected number of clusters
303 (e.g., Fig. 2B), this would suggest a more complex cluster structure (Tibshirani et al. 2001). In
304 this situation, the researcher should investigate the cluster structure further. Variable weighting,
305 as we did, can be used with the k-means cluster analysis. Alternatively, hierarchical clustering
306 can be employed to investigate hierarchical cluster organization; for such hierarchical clustering,
307 other tools are available, including agglomerative nesting or divisive analysis clustering
308 (Kaufman and Rousseeuw 1990).

309 In the context of the analysis of GPS data, two important issues must be discussed:
310 sampling interval and missing fixes. The sampling interval determines the temporal scale of the
311 data (Nathan et al. 2008), therefore, it is important to select an interval appropriate for the
312 research question. The sample interval should match the temporal patterns of the focal process
313 and the error structure of the data. When travel distances become too short at short sampling
314 intervals, inaccurate movement assessment may arise due to GPS error (Hurford 2009). With
315 increasing sample intervals behaviors become more and more aggregated. For instance, elk
316 movement sampled at a daily interval aggregates rest and movement states (found in our data at
317 2-h intervals), instead, these data showed behaviors at a larger temporal scale, like encamped and
318 exploratory states (Morales et al. 2004). Therefore, the error in the data and the temporal scale of
319 the behavior of interest will dictate the sampling interval.

320 Similar to studies in habitat selection (e.g. Frair et al. 2004, Bourgoin et al. 2009)
321 selective fix success could lead to the underestimation of certain movement states. Unfortunately,
322 it is not possible to estimate movement-related fix-success bias using stationary GPS receivers,
323 like is done for habitat-related fix-success bias (Frair et al. 2004). In an alternative approach,

324 Bourgoin et al. (2009) derived the habitat-related bias from interpolated locations. When
325 movement states show sufficient temporal autocorrelation, then information regarding the
326 behavioral state contained in successful fixes can be used to interpolate movement states for
327 missing fixes. Similar to Bourgoin et al. (2009), these interpolated movement states allow the
328 estimation of fix bias in relation to movement state. The presence of this fix-success bias can be
329 expected, when animals select different habitats depending on their state, due to habitat-related
330 fix success (e.g. Frair et al. 2004).

331 One of the interesting possibilities offered by distinguishing movement states of animals
332 in ecology is to investigate animal responses to the environment. As a simple illustration we
333 compared the occurrence of the 4 movement states in cutblocks versus non-cutblocks, and during
334 crepuscule (am/pm) versus the rest of the 24-h period. As discussed above, cutblocks offer
335 important foraging opportunities for elk, and the major activity peaks for elk occur at dusk and
336 dawn (Merrill 1991, Olsson et al. 2007). We found (see Table 1) important differences among
337 movement states with respect to their occurrence within cutblocks ($\chi_3^2 = 18.64, P < 0.001$) and
338 during twilight ($\chi_3^2 = 25.58, P < 0.001$). The first movement state seemed to correspond to
339 between-patch movements, characterized by high activity and high directional persistence. This
340 movement state occurred during crepuscule, outside the cutblocks. Whereas, the second
341 movement state was most likely composed of foraging movements, which combined frequent
342 head movements with frequent direction changes. Not surprisingly, these foraging movements
343 occurred primarily during twilight within the cutblocks. The third movement state was
344 characterized by low activity and low directional persistence. The movement state was unlikely
345 to be feeding, because of few up and down head movements; this may be the result of GPS-error
346 during an animal's resting phase (Hurford 2009). In further support for this interpretation, we

347 found this movement state occurred less during the main activity periods (i.e., twilight) and more
348 within cutblocks, where high visibility may aid the early detection of approaching predators. In
349 contrast, little head movements and high directional persistence were typical of the final
350 movement state; a clear biological interpretation for this state is not straightforward. This final
351 movement state occurred less during twilight and less within cutblocks. We could speculatively
352 propose an interpretation of this movement state as consisting of between-patch movements
353 without foraging whereas in the first movement state the between-patch movements were
354 accompanied by occasional foraging, the presence or absence of foraging during these between
355 patch movements would then explain the amount of head movements. Linking our states to
356 landscape features thus helped corroborate that biologically plausible states were identified using
357 our technique, and can further help inform modeling of the movement process such as with a
358 state-based approach.

359 Ultimately, the usefulness of groupings of movement characteristics derived from the k-
360 means approach comes from their correspondence to expected biological mechanisms and
361 processes that produce the clusters. Indeed, the interpretation of the different movement states
362 follows from the descriptive characteristics of these states. Cluster analysis investigates the
363 structure and patterns in the data. Subsequently, the hypothesized processes underlying these
364 patterns could be tested with simultaneous direct observations of animal behavior. If behavioral
365 observations were taken while the GPS collar was deployed, then once the data were
366 downloaded, clustered, and temporally matched to the observations, the researcher could explore
367 the correspondence of observed behaviors to the different clusters. However, this may prove
368 difficult in the field given the constraints on observing behavior in free-ranging individuals,
369 particularly in forested conditions. We limited our illustration to a few explanatory variables; it is

370 possible that other variables could provide better predictions of the movement states we
371 identified.

372 An important asset of the k-means approach combined with the gap statistic is the ability
373 to include multivariate information. While previous methods (Johnson et al.1992a, Johnson et
374 al.1992b, Morales et al., 2004) have included turning angles and step lengths, they are not
375 conducive to using a range of variables, such as the head movements used in our elk example.
376 Although, it is possible to include other variables in state-space models, doing so results in an
377 increasing model complexity and computational challenges due to the definition of functional
378 relationships between model variables in this approach. With advancements in animal collar
379 technology, we expect additional variables, such as physiological responses (e.g. heart rate),
380 abiotic conditions (ambient temperature), or duration of head movements, to become particularly
381 useful in remotely identifying animal behaviors such as inactivity/thermogenesis or predatory
382 flight responses. However, care should be taken in the selection of the variables that are
383 included, as non-discriminating variables can clutter and hide cluster structure present in the
384 data.

385 The k-means approach with the gap statistic allows the researcher to explicitly investigate
386 the number of movement states in a transparent hypothesis testing framework without
387 assumptions about the underlying movement process. In previous methods the number of states
388 was not an explicit part of the research hypothesis (e.g. Fauchald and Tveraa 2003) or were
389 identified using a parametric approach with assumptions regarding the movement process (e.g.
390 Morales et al. 2004). It is important to note that these different approaches are not exclusive, they
391 are largely complementary instead. For instance, Fauchald and Tveraa (2003)'s first passage time
392 can easily be included in a k-means cluster analysis in addition to or instead of other movement

393 characteristics. Whereas, knowledge gained by investigating patterns in movement data will aid
394 the development of models for movement processes (Patterson et al. 2008).

395 Our motivation was to demonstrate a method to empirically justify the number of
396 movement states used for movement modeling or state-dependent habitat selection studies. Apart
397 from direct observation of the animal, it is challenging to link remotely sensed location data to
398 behavior. Statistically, without a behaviorally identified response variable (i.e. foraging or
399 relocating) corresponding to each location, we must resort to grouping similar observations
400 (clustering) and then interpret those clusters relative to animal behavior. In this paper, we
401 presented an approach for the analysis of grouping patterns, without any assumptions regarding
402 the underlying processes, in multivariate movement data with k-means clustering and evaluated
403 the number of clusters using the gap statistic, which provided an empirical assessment of the
404 number of movement states within a hypothesis testing framework.

405 **Management Implications**

406 For many management applications knowledge of animal behavior in relation to its environment
407 is important. For example, if the monitored organism forages in limited habitat and this behavior
408 is identifiable from the GPS data, then identifying these presumably critical areas could have
409 consequences for management actions that may disrupt critical forage habitat. The application of
410 spatial behavioral understanding could potentially be included in resource selection functions
411 (Boyce et al. 1999, Manly et al. 2002). Similarly, it may be important to link responses in
412 behavior and not simply movement to human disturbance (Dyer et al. 2002) or other
413 environmental changes. Where direct observation is difficult or impossible, remotely sensed data
414 can provide information on behavior states; however, the indirect observation of behavior
415 requires linking behavioral states to movement and activity metrics. K-means clustering with the

416 gap-statistic can offer a defensible approach to infer these behavioral states based on directly or
417 remotely sensed data.

418 **Acknowledgments**

419 The authors would like to thank M. Lewis, A. Hurford, H. Beyer, and M. Taper for helpful
420 comments and suggestions. The senior author was financially supported by the Norwegian
421 University of Science and Technology and the Norwegian Research Council's PredClim grant to
422 B-E. Saether. The third author was supported by NSERC Discovery to M. Lewis. NSERC
423 Industrial Scholarship in partnership with Weyerhaeuser to DRV, National Science Foundation
424 (Grant # 0078130), Rocky Mountain Elk Foundation, and the Alberta Conservation Association
425 grants to EHM and JLF. The suggestions made by Scott McCorquodale and two anonymous
426 reviewers greatly improved the manuscript.

427 **LITERATURE CITED**

- 428 Adrados, C., H. Verheyden-Tixier, B. Cargnelutti, D. Pepin, and G. Janeau. 2003. GPS approach
429 to study fin-scale site used by wild red deer. *Wildlife Society Bulletin* 31:544–552.
- 430 Barraquand, F., and S. Benhamou. 2008. Animal movements in heterogeneous landscapes:
431 Identifying profitable places and homogeneous movement bouts. *Ecology* 89: 3336–3348.
- 432 Bourgoin, G., M. Garel, D. Dubray, D. Maillard, and J.-M. Gaillard. 2009. What determines
433 global positioning system fix success when monitoring free-ranging mouflon? *European*
434 *Journal of Wildlife Research* DOI 10.1007/s10344-009-0284-1
- 435 Boyce, M.S., and L.L. McDonald. 1999. Relating Populations to Habitats Using Resource
436 Selection Functions. *Trends in Ecology & Evolution* 14: 268–272.
- 437 Burnham, K.P., and D.R. Anderson. 2002. *Model Selection and Multimodel Inference: A*
438 *Practical Information-Theoretical Approach*. 2d ed. New York: Springer-Verlag.

- 439 Calinski, R. B., and J. Harabasz. 1974. A dendrite method for cluster analysis. *Communications*
440 *in Statistics* 3: 1–27.
- 441 Crist, T., D. Guertin, J. Wiens, and B. Milne. 1992. Animal movement in heterogeneous
442 landscapes - an experiment with *Eleodes* Beetles. *Functional Ecology* 6: 536–544.
- 443 Dyer, S.J., J.P. O'Neill, S.M. Wasel, and S. Boutin. 2002. Quantifying Barrier Effects of Roads
444 and Seismic Lines on Movements of Female Woodland Caribou in Northeastern Alberta.
445 *Canadian Journal of Zoology-Revue Canadienne De Zoologie* 80: 839–845.
- 446 Fauchald, P., and T. Tveraa. 2003. Using first passage time in the analysis of area restricted
447 search and habitat selection. *Ecology* 84:282–288.
- 448 Fisher, W. 1958. On grouping for maximum homogeneity. *Journal of the American Statistical*
449 *Association* 53:789–798.
- 450 Frair, J., S.E. Nielsen, E.H. Merrill, S.R. Lele, M.S. Boyce, R.H.M. Munro, G.B. Stenhouse, and
451 H.L. Beyer. 2004. Removing GPS collar bias in habitat selection studies. *Journal of Applied*
452 *Ecology* 41:201–212.
- 453 Frair, J., E.H. Merrill, D. Visscher, D. Fortin, and J. Morales. 2005. Scales of movement by elk in
454 response to heterogeneity in forage resources and predation risk. *Landscape Ecology* 20:273–
455 287.
- 456 Frair, J. L., E.H. Merrill, J.R. Allen, and M.S. Boyce. 2007. Know Thy Enemy: Experience
457 Affects Elk Translocation Success in Risky Landscapes. *Journal of Wildlife Management*
458 71:541–554.
- 459 Hurford, A. 2009. GPS Measurement Error Gives Rise to Spurious 180u Turning Angles and
460 Strong Directional Biases in Animal Movement Data. *PLoS ONE* 4: e5632.
461 doi:10.1371/journal.pone.0005632

- 462 Jerde, C., and D. Visscher. 2005. GPS measurement error influences on movement model
463 parameter estimation. *Ecological Applications* 15:806–810.
- 464 Johnson, A., B. Milne, and J. Wiens. 1992a. Diffusion in fractal landscapes: simulations and
465 experimental studies of Tenebrionid Beetle movements. *Ecology* 73:1968–1983.
- 466 Johnson, A., J. Wiens, B. Milne, and T. Crist. 1992b. Animal movements and population
467 dynamics in heterogenous landscapes. *Landscape Ecology* 7:63–75.
- 468 Johnson, C., K. Parker, D. Heard, and M. Gillingham. 2002. Movement parameters of ungulates
469 and scale-specific responses to the environment. *Journal of Animal Ecology* 71:225–235.
- 470 Johnson, R., and D. Wichern. 1998. *Applied multivariate statistical analysis*. Prentice-Hall Inc,
471 Upper Saddle River, NJ.
- 472 Jones, R. 1977. Movement patterns and egg distribution in cabbage butterflies. *Journal of Animal*
473 *Ecology* 46:195–212.
- 474 Kareiva, P.M., and N. Shigesada. 1983. Analyzing insect movement as a correlated random walk.
475 *Oecologia* 56: 234–238.
- 476 Kaufman, L., and P. Rousseauw. 1990. *An introduction to cluster analysis*. Wiley, New York.
- 477 Legendre, P., and L. Legendre. 1998. *Numerical Ecology*. Elsevier Science.
- 478 Lima, S., and P. Zollner. 1996. Towards a behavioral ecology of ecological landscapes. *Trends in*
479 *Ecology and Evolution* 11:131–135.
- 480 Lleti, A., M. Ortiz, L. Sarabia, and M. Sanchez. 2004. Selecting variables for k-means cluster
481 analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*
482 515:87–100.
- 483 Luque, S.P., and C. Guinet 2007 A maxium likelihood approach for identifying dive bouts
484 improves accuracy, precision and objectivity. *Behavior* 144:1315-1332.

- 485 MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations.
486 Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability
487 1:281–297.
- 488 Manly, B.F.J., D.L. Thomas, and L.L. McDonald. 2002. Resource selection by animals: statistical
489 design and analysis for field studies. Kluwer Academic Publishers, Boston.
- 490 Merrill, E.H. 1991. Thermal constraints on use of cover types and activity time of elk. Applied
491 Animal Behaviour Science. 29: 251–267.
- 492 Milligan, G. W., and M.C. Cooper. 1985. An examination of procedures for determining the
493 number of clusters in a data set. Psychometrika, 50, 159–179.
- 494 Morales, J., D. Haydon, J. Frair, K. Holsiner, and J. Fryxell. 2004. Extracting more out of
495 relocation data: building movement models as mixtures of random walks. Ecology 85:2436–
496 2445.
- 497 Nams, V. 2005. Using animal movement paths to measure response to spatial scale. Oecologia
498 143:179–188.
- 499 Nathan, R., W.M. Getz, E. Revilla, M. Holyoak, R. Kadmon, D. Saltz, and P.E. Smouse. 2008. A
500 movement ecology paradigm for unifying organismal movement research. Proceedings of the
501 National Academy of Sciences USA 105: 19052–19059.
- 502 Olsson, P.M.O., J.J. Cox, J.L. Larkin, D.S. Maehr, P. Widen, and M.W. Wichrowski. 2007.
503 Movement and activity patterns of translocated elk (*Cervus elaphus nelsoni*) on an active coal
504 mine in Kentucky. Wildlife Biology in Practice 3: 1–8.
- 505 Patterson, T. A., L. Thomas, C. Wilcox, O. Ovaskainen, and J. Matthiopoulos. 2008. State-space
506 models of individual animal movement. Trends in Ecology & Evolution 23: 87–94.
- 507 R Development Core Team. 2008. R: A language and environment for statistical computing. R

- 508 Foundation for Statistical Computing, Vienna, Austria.
- 509 Roitberg, B., and M. Mangel. 1997. Individuals on the landscape: behavior can mitigate
510 landscape differences among habitats. *Oikos* 80:234–240.
- 511 Steinley, D. 2006a. K-means clustering: a half-century synthesis. *The British Journal of*
512 *Mathematical and Statistical Psychology* 59:1–34.
- 513 Steinley, D. 2006b. Profiling local optima in K-means clustering: developing a diagnostic
514 technique. *Psychological Methods* 11: 178–192.
- 515 Steinley, D., and M.J Brusco. 2008. A New Variable Weighting and Selection Procedure for K-
516 Means Cluster Analysis. *Multivariate Behavioral Research* 43: 77–108.
- 517 Tibshirani, R., G. Walther, and T. Hastie. 2001. Estimating the number of clusters in a dataset via
518 the gap statistic. *Journal of the Royal Statistical Society, B* 63:411–423.
- 519 Turchin, P. 1998. *Quantitative analysis of movement*. Sinauer Associates, Inc., Sunderland, MA.
- 520 Visscher, D.R., and E.H Merrill. 2009. Temporal dynamics of forage succession for elk at two
521 scales: Implications of forest management. *Forest Ecology and Management* 257: 96-106.
- 522 Walesiak, M., and A. Dudek. 2008. clusterSim: Searching for optimal clustering procedure for a
523 data set.
- 524 Wiens, J. 1989. Spatial scaling in ecology. *Functional Ecology* 3: 385–397.
- 525 Yingqiu, L., L. Wei, and L. Yun-Chun. 2007. Network Traffic Classification Using K-means
526 Clustering. *Second International Multi-Symposiums on Computer and Computational*
527 *Sciences* 360–365.
- 528 *Associate editor: Scott McCorquodale*
- 529

530 **FIGURE CAPTIONS**

531 Figure 1

532 A represents simulated data with initial seed points. B is the simulated data after the convergence
533 of the clusters $k = 2$ and the principal components used for creating the reference data set, C is
534 the reference data created for the gap statistic.

535

536 Figure 2

537 Gap statistic for different numbers of cluster and characteristics of the variables for the selected
538 number of clusters. Panels on the left hand side (A, C, E and G) show the gap statistic with its SE
539 for 1 to 10 clusters. The selected number of clusters using a tolerance of 1 and 2 are marked with
540 respectively a black dot and a square (note, only in panel G a difference occurred between both
541 tolerances with respectively 5 and 4 clusters selected). Panels on the right hand side (B, D, F and
542 H) show for each cluster the mean and $2 * SE$ of the 4 clustering variables in our study (note,
543 that these variables were not necessarily all used for the clustering itself). The cluster analysis in
544 the upper panels (A and B) was based on all 4 variables, the second row (C and D) on the
545 trajectory measures (step length and turning angle), the third row (E and F) are the results of both
546 head movement (i.e. activity) measures, and the bottom panels (G and H) result from all four
547 measures with decreased weight of the activity measures (see main text for further explanation).

548

549 **TABLES**

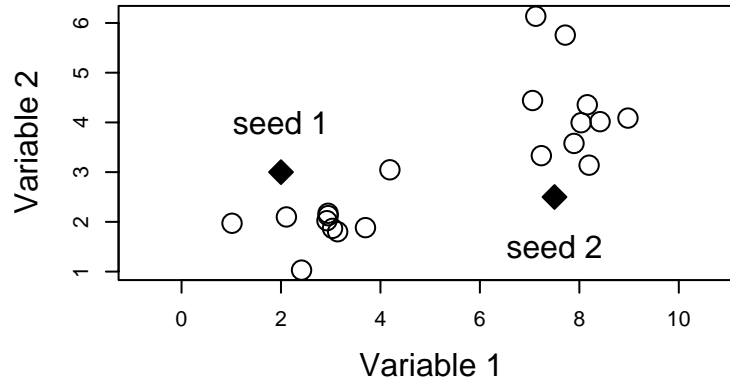
550 Table 1

551 Characteristics of the four movement states distinguished by the weighted k-means clustering on
 552 all four variables (i.e. activity measures [ACT1 and ACT2], step length [SL], and turning angle
 553 [TA]). For each state, the mean (\pm SE) of each variable, the proportion (\pm SE) during twilight and
 554 within cutblocks, and the total number of observations are shown.

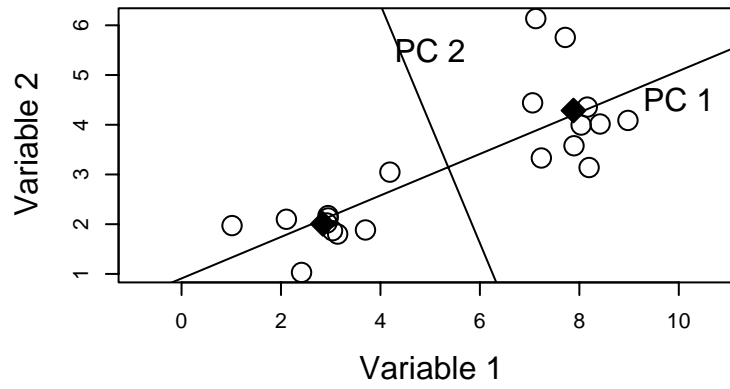
	state 1	state 2	state 3	state 4
ACT1	57.89 \pm 3.29	56.87 \pm 3.33	17.09 \pm 2.46	18.79 \pm 2.18
ACT2	71.95 \pm 3.40	58.33 \pm 3.36	2.16 \pm 0.41	1.77 \pm 0.26
SL	309.54 \pm 23.18	317.48 \pm 26.46	180.22 \pm 16.71	206.64 \pm 20.62
TA	41.23 \pm 1.51	134.16 \pm 1.77	139.19 \pm 1.67	46.56 \pm 1.74
Twilight	0.60 \pm 0.03	0.60 \pm 0.03	0.44 \pm 0.04	0.42 \pm 0.03
Cutblock	0.50 \pm 0.03	0.65 \pm 0.03	0.62 \pm 0.04	0.49 \pm 0.03
N	284	225	189	231

555

A



B



C

