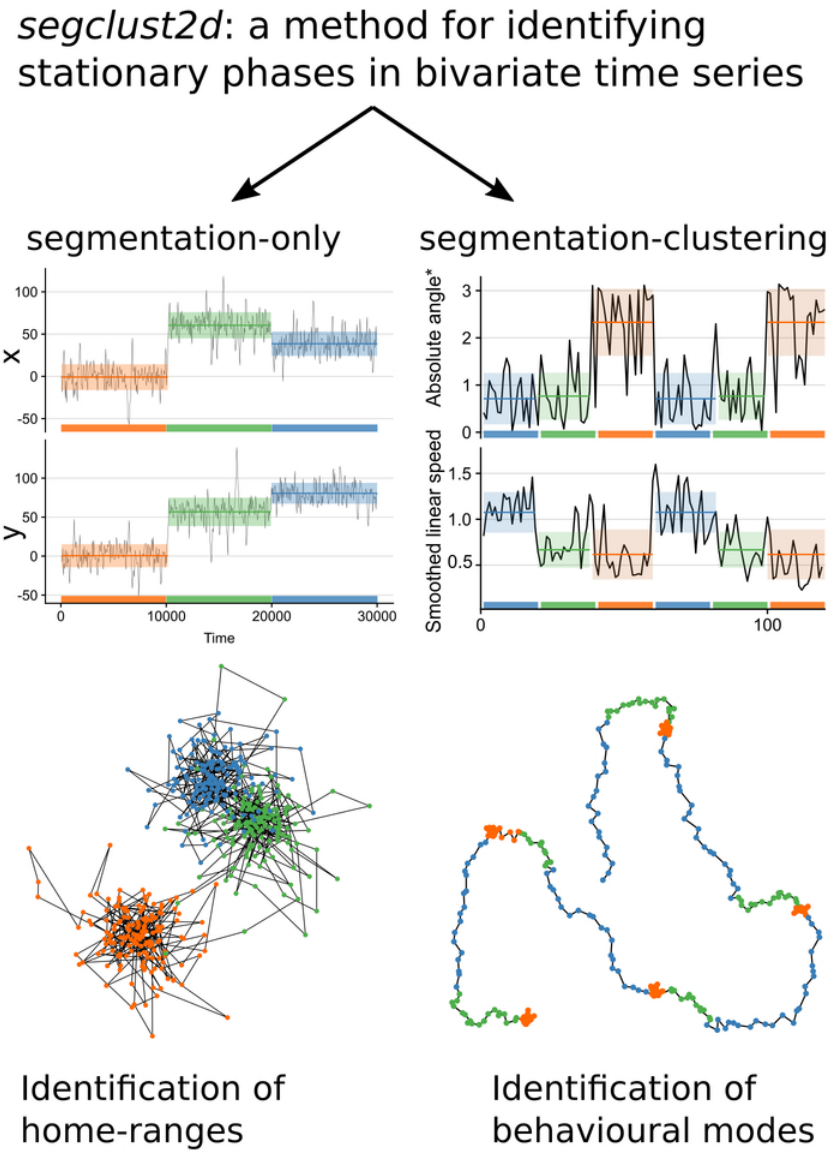
Journal of Animal Ecology

Identifying stationary phases in multivariate time-series for highlighting behavioural modes and home range settlements

Journal:	*Journal of Animal Ecology*
Manuscript ID	JAE-2018-00787.R1
Manuscript Type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Patin, Rémi; CNRS, Centre d'Ecologie Fonctionnelle et Evolutive Etienne, Marie-Pierre; Universite de Rennes 1, AgroCampusOuest Lebarbier, Emilie; AgroParisTech, Mathématiques et Informatique Appliquées Chamaillé-Jammes, Simon; CNRS, Centre d'Ecologie Fonctionnelle et Evolutive Benhamou, Simon; CEFE CNRS, Behavioural Ecology
Key-words:	movement ecology, biologging, segmentation and clustering, foraging, home range, area-concentrated searching, transit, migration



59x84mm (300 x 300 DPI)

Identifying stationary phases in multivariate time-series for highlighting behavioural modes and home range settlements

R. Patin[1], M.-P. Etienne[2], E. Lebarbier[3], S. Chamaillé-Jammes[1,4,5] & S. Benhamou[1*]

[1] Centre d'Écologie Fonctionnelle et Évolutive, CNRS et Université de Montpellier, France

[2] Institut de recherche mathématique de Rennes, Université de Rennes, AgroCampusOuest, Rennes, France

[3] Mathématiques et Informatique Appliquées, Agroparistech, Paris, France

[4] LTSER France, Zone Atelier "Hwange", Hwange National Park, Bag 62, Dete, Zimbabwe

[5] Mammal Research Institute, Department of Zoology & Entomology, University of Pretoria, Pretoria, South Africa

* Corresponding author: simon.benhamou@cefe.cnrs.fr

Authors' contributions.

RP analysed the data and contributed to the coding of the statistical model, which was developed by MPE and EL. SC provided the tracking data. SB led the project, performed computer simulations, and wrote the first draft of the manuscript, except the part describing the model which was first contributed to by MPE, EL and RP. All authors contributed significantly to the final manuscript.

Running title: Identifying stationary phases

Key-words: movement ecology, segmentation, clustering, foraging, home range, area-concentrated searching, transit, migration

Abstract

1. Recent advances in bio-logging open promising perspectives in the study of animal movements at numerous scales. It is now possible to record time-series of animal locations and ancillary data (e.g. activity level derived from on-board accelerometers) over extended areas and long durations with a high spatial and temporal resolution. Such time-series are often piecewise stationary, as the animal may alternate between different stationary phases (i.e. characterised by a specific mean and variance of some key parameter for limited periods). Identifying when these phases start and end is a critical first step to understand the dynamics of the underlying movement processes.

2. We introduce a new segmentation-clustering method we called segclust2d (available as a R package at cran.r-project.org/package=segclust2d). It can segment bi- (or more generally multi-) variate time-series and possibly cluster the various segments obtained, corresponding to different phases assumed to be stationary. This method is easy to use, as it only requires specifying a minimum segment length (to prevent over-segmentation), based on biological rather than statistical considerations.

3. This method can be applied to bivariate piecewise time-series of any nature. We focus here on two types of time-series related to animal movement, corresponding to (i) at large scale, the series of bivariate coordinates of relocations, to highlight temporary home ranges, and (ii) at smaller scale, the series two variables derived from relocations data, such as speed and turning angle, to highlight different behavioural modes such as transit, feeding and resting.

4. Using computer simulations, we show that segclust2d can rival and even outperform previous, more complex methods, which were specifically developed to highlight changes in movement modes or home range shifts (based on Hidden Markov and Ornstein-Uhlenbeck modelling), which, contrary to our method, usually require relevant initial guesses to be efficient. Furthermore we demonstrate it on actual examples involving a zebra's small scale

movements and an elephant's large scale movements, to illustrate how various movement modes and of home range shifts, respectively, can be identified.

Introduction

Landscapes are spatially and temporally variable at various scales (Levin, 1992), and animals are expected to adjust their movements to the characteristics of their local environment, so as to maximize the time spent in profitable (or safe) habitats and minimize time in adverse ones (Pyke, 1978). Recent advances in bio-logging have made it possible to acquire time-series of animal's locations, and possibly ancillary data such as activity level derived from on-board accelerometers, over extended areas and long durations with high spatial and temporal resolutions. Such locational time-series, and the other ones that can be derived from them to describe the movement behaviour (e.g. turning angle, speed), are therefore expected to be piecewise stationary, i.e. to present a specific mean and variance for limited periods corresponding to stationary phases, alternating with rapid transition phases corresponding to changes of area or behaviour (in practice, a phase can be considered as being stationary when the partial means and variances obtained for its first and second halves or for its three thirds are not markedly different). Identifying these stationary phases is a prerequisite to determine the biologically relevant scales of movement (Benhamou, 2014). It is therefore of paramount importance in two types of movement studies:

Identifying behavioural modes. Foragers are generally expected to alternate intensive (area-concentrated) searching mode, characterised by high tortuosity and low speed, and extensive searching (transit) mode, characterised by low tortuosity and high speed (see Dias et al., 2009 for contrasting examples). A number of discrete time methods have been developed to attempt to identify the different movement modes (see Gurarie et al. 2016 and Edelhoff et al. 2016 for general reviews; see also Parton & Blackwell 2017 for a continuous-time approach). As the alternation of searching modes should result in piecewise "behavioural stationarity" when considering time-series of some movement parameters, several segmentation approaches have been developed to identify behavioural modes by looking at

breakpoints (i.e. rapid transitions between stationary phases; Barraquand and Benhamou, 2008; Gurarie et al., 2009; Nams, 2014). A more sophisticated approach based on Hidden Markov Models (HMM) has gained momentum in recent years. In this approach, the joint step lengths and turning angles calculated from successive relocations are categorized among a predefined number of different modes modelled as hidden states (Morales et al., 2004; Langrock et al., 2012; McClintock et al., 2012; Beyer et al., 2013; Michelot et al., 2016). However, the convergence of HMM may require specifying relevant initial state-dependent probability distribution parameters, which can be difficult. Here, we aim at developing an alternative approach which does not require such a pre-specification. As most previous methods, it focused on speed and tortuosity but deals with such variable through a penalised likelihood criterion. Furthermore, in addition to the spatio-temporal couple of metrics classically used in HMM-based segmentation, (linear and angular speeds), we aim at testing other couples of metrics which should be more able to differentiate between the various behavioural modes by accounting for spatial and temporal effects separately (Benhamou & Bovet 1989), or which present more suitable statistical characteristics (Gurarie et al. 2009).

Identifying home range shifts. The recently emerging question of piecewise "locational stationarity" at the home range scale has been addressed in terms of movement scales (Benhamou, 2014), migration characteristics (Naidoo et al., 2012; Cagnacci et al., 2016) and of within-season shifts (Couriot et al 2018). Indeed, for an animal that exploits various temporary home ranges, the time-series of relocations coordinates can be assumed to be stationary for a relatively long time (when the animal exploited the area where it established its temporary home range), then non-stationary for a relatively short time (when the animal left its home range until it established a new one), and so on. It is worth noting that a shift in home range does not necessarily involve a shift in mean location. It may also correspond to a change in variance if the animal enlarged or shrank its home range, e.g. due to a change in season (Naidoo et al., 2012; Monsarrat et al., 2013) or in reproductive status. Various

104 methods have been proposed to detect home range shifts. The simple univariate approach
105 based on the change of the beeline distance from a starting point (Bunnefeld et al., 2011)
106 appears to be convenient in some cases but fully ignores movements leading the animal at a
107 similar distance from the starting point but in another direction. More complex approaches
108 rest on Ornstein-Uhlenbeck (OU) modelling, following Blackwell (1997). Breed et al. (2017)
109 used a Bayesian framework to estimate the number of states of an underlying multi-state OU
110 process. In this way, they could directly infer the effect of covariates on the transition rate
111 between ranges, but used a simple K-means algorithm to assign locations to the different
112 states (corresponding to the different home ranges). Gurarie et al. (2017) developed an
113 alternative approach by introducing a non-stationary state in the OU process to model the
114 shifts in home range location. However, as it requires that all home range phases and shifts
115 are explicitly modelled, this approach tends to become cumbersome when there are several
116 shifts to consider. Furthermore, it may require truly relevant initial guesses to correctly detect
117 small shifts. We therefore aim at developing an alternative approach that could be more
118 efficient than an OU-based approach to detect home range shifts and simpler to use.
119 Additionally, as detecting shifts in behavioural modes and in home ranges settlements are
120 conceptually similar, we focused on a generic approach that can be applied to both types of
121 studies.

122 Here we introduce a new method, called segclust2d, able to segment a bi- (and more
123 generally multi-) variate time-series, and to cluster similar segments (corresponding to
124 stationary phases) in a common class (corresponding to a given state) if desired. We
125 demonstrate that this method, which is easy to use, can successfully identify stationary phases
126 corresponding to temporary home ranges when based on bivariate locational time-series, as
127 well as movement modes when based on bivariate time-series of metrics such as speed and
128 tortuosity. It thus offers an efficient and user-friendly alternative to previous, more complex,
129 approaches. Furthermore, as this model applies to multivariate piecewise stationary time-

series based on any kind of metrics, it can integrate additional time-series of ancillary data (e.g. activity level derived from on-board accelerometers) for a better segmentation-clustering of movement data.

Methods

STATISTICAL MODEL AND PARAMETER ESTIMATION

General principle. Consider a multivariate time-series composed of C components (each corresponding to a univariate time-series). These components are assumed to be statistically independent, and should be normalized if they are of different nature, so as to have the same weight in subsequent procedures. The time-series is assumed to piecewise stationary, i.e. to be made of an unknown number of stationary phases. In a given phase, the values taken by any of the C components are assumed to be independent of each other. One needs a reliable statistical model to detect and locate these phases, and possibly to cluster them when they are assumed to be the expressions of a limited number of unobserved states of the underlying process (e.g. behavioural modes). Likelihood-based segmentation methods provide a suitable statistical framework to detect changes of phases but raise two main issues from a statistical and algorithmical point of view: (i) determining the optimal number of segments and (ii) for a given number of segments, finding the optimal segmentation, i.e. determining the locations of the starting/ending points of the segments (called breakpoints). The latter reduces to a well-known discrete optimization problem solved using a dynamic programming algorithm introduced by Bellman (1954; for a recent example, see Rigaill, 2015). For a time-series of n values that can potentially be cut at any point in K segments, the dynamic programming algorithm reaches the exact maximum likelihood solution with a complexity in $O(n^2K)$, drastically smaller than the complexity in $O(n^K)$ involved by a force brute algorithm when exploring the whole segmentation space. We will first introduce the models and the

estimation procedure to optimally segment a multivariate signal for a predefined number of segments K and possibly (if clustering is required) a predefined number of states M. Afterwards, we will show how the optimal number of segments and possibly of clusters can be found based on a penalized likelihood criterion. Our approach is based on Lavielle (2005)'s segmentation method of univariate signals and its extension by Picard et al. (2007) to segment and cluster DNA sequences without assuming any kind of distribution for the segment lengths, such as a geometric distribution as HMM implicitly do (Karlin & Taylor, 1975).

Optimal segmentation in K segments, with optional clustering in M states. Assume that there are K stationary phases in a multivariate time-series with total length n. A stationary phase corresponds to a segment. On each segment and for each of the C components (labelled 1 to C), the series is assumed to be a sequence of random variables sharing the exact same distribution, in particular the same mean and the same variance. As soon as one of these parameters changes, a new segment starts. Formally, the C components within a given segment $k \in [1, \ldots, K]$ starting at time $t = t_{k-1}+1$ and ending at time $t = t_k$ (with $t_0 = 0$ and $t_K = n$ by convention) are modelled as sequences of Gaussian independent random variables $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_C(t))^\intercal$ for $t = 1, \ldots, n$.

The segmentation-only model (no clustering) is written simply as

$$\mathbf{Y}\,(t \mid t \in [\,t_{k-1}+1, t_k]) \sim \mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2), \text{ with } \boldsymbol{\mu}_k = \begin{pmatrix} \mu_{k,1} \\ \vdots \\ \mu_{k,C} \end{pmatrix} \text{ and } \boldsymbol{\sigma}^2_k = \begin{pmatrix} \sigma^2_{k,1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma^2_{k,C} \end{pmatrix}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}^2_k$ are the mean vector and the variance matrix for data in segment k. As the model parameters to be estimated vary independently between segments, dynamic programming can be used to segment the multivariate signal at best in K segments. Its application is straightforward in this case, as it relies on the log-likelihood of each segment, which is simply equal to the sum of the log-likelihoods of the C components.

In the segmentation-clustering model, a state m, among M possible states, is assigned to every segment. Segments that are classified in state m are all assumed to share the same mean vector $\boldsymbol{\mu}_m$ and the same variance matrix $\boldsymbol{\sigma}^2_m$. More formally, let S_k, denotes the state of the segment k, with $k = 1, \ldots, K$. S_k is a latent random variable taking values in [1, ..., M]. It is modelled through a multinomial distribution of parameters $\pi = (\pi_m)$ with $m = 1, \ldots, M$, where π_m corresponds to the probability for a segment to belong to state m. The segmentation-clustering model can therefore be written as:

$$\mathbf{Y}(t \mid t \in [\, t_{k-1}+1, t_k], S_k = m) \sim \mathrm{N}(\boldsymbol{\mu}_m, \boldsymbol{\sigma}^2_m), \text{ with } \boldsymbol{\mu}_m = \begin{pmatrix} \mu_{m,1} \\ \vdots \\ \mu_{m,C} \end{pmatrix} \text{ and } \boldsymbol{\sigma}^2_m = \begin{pmatrix} \sigma^2_{m,1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma^2_{m,C} \end{pmatrix}$$

As the parameters (π_m, $\boldsymbol{\mu}_m$, $\boldsymbol{\sigma}^2_m$) that characterise any state m are unknown and are to be estimated, resulting in a mixture distribution where segments are linked in terms of parameters, the optimal segmentation cannot anymore be obtained using dynamic programming alone. Following Picard et al. (2007), we designed the following two-step procedure, which is iterated up to convergence.

1. Given a set of parameters (π_m, $\boldsymbol{\mu}_m$, $\boldsymbol{\sigma}^2_m$) with $m = 1, \ldots, M$, the best segmentation in K segments is obtained using dynamic programming.
2. Given this segmentation, the values of parameters are estimated using an expectation-maximization algorithm which is commonly used in latent variable modelling (Dempster et al., 1977).

By mixing dynamic programming and expectation-maximization through this iterative procedure, segmentation and clustering processes work jointly (rather than the latter after the former) leading to the optimal segmentation given K and M. An additional procedure for initialising the EM algorithm is automatically performed to solve possible convergence issues (see details in Supporting information 1). Note also that this method has to be applied

separately to data coming from different individuals because it is not designed to perform a joint estimation of parameters for disjoint time-series (i.e. concerning different individuals).

Finding the optimal numbers of segments and states. For both models (segmentation-only and segmentation-clustering), a minimum segment length L_{min} (>5 records to warrant a sufficiently reliable estimation of the parameters) has to be set not only to speed up the algorithm, but also, more fundamentally, to prevent over-segmenting, based on biological considerations. For example, setting L_{min} to a value of a few weeks when analysing locational time-series will prevent the algorithm from considering an area exploited only for a few days, corresponding to foray outside the usual home range or to stop-over during migration, as a distinct home range. Similarly, setting L_{min} to a value long enough (depending on the species) when looking for changes in behavioural modes will force the algorithm to assign a given behavioural bout to a given mode even if when it is interspersed by ephemeral events related to another behaviour (e.g. a long transit with opportunistic short feeding events on the move will be considered as a single transit phase). The likelihood is calculated for all number of segments $K < 0.75\ n/L_{min}$ (larger values of K may generate inconsistent results), and for any number of states M ($<K$) one wishes to consider if clustering has to be involved. In the latter case, the optimal values of K and M are determined as those that maximize a Bayesian Information Criterion (BIC; Schwarz, 1978)-based penalized likelihood (Supporting information 1). However, as it will be shown in the Results section, it is usually preferable to set the number of states M based on biologically relevant grounds than to let the model selection criterion determine an optimal number of states based on a statistical basis (see also Pohle et al. 2017). When no clustering is involved, the optimal number of segments K is based in agreement with Lavielle (2005) on maximizing a K-penalized likelihood curve (Supporting Information 1).

COMPUTER SIMULATIONS

We run simulations to assess the ability of our approach to detect home range shifts and changes in behavioural modes from bivariate time-series, and to compare it with that of other (OUM-based and HMM-based) methods. For each set of parameters of each type of simulation, we simulated 100 replicates. Distances are expressed in arbitrary unit length u.

Home range shifts. For simplicity, the animal was assumed to behave as a central place forager. We simulated its fine-scale movement as a central-place biased correlated random walk involving a differential klinokinetic mechanism, which results in a probability of presence decreasing exponentially with the distance D to the central place (Benhamou, 1989): at each time step, the animal turns by an angle α_i drawn from a wrapped Gaussian distribution with a null mean and standard deviation $\sigma_i = \sigma_0[1+b(D_i - D_{i-1})]$, with b=0.5 radians/u, and progresses by 1 unit length (1 u) in the new direction. The directional bias is generated by the modulation of σ_i (in the range $[0.5\sigma_0, 1.5\sigma_0]$), which leads the animal to experience a higher probability to reverse its moving direction when its moves away of ($D_i>D_{i-1}$) than towards ($D_i<D_{i-1}$) the central place. In a batch of simulations, σ_0 was set to 0.5 radians, and the central place was first set at a given location for the first 10,000 time steps (phase 1), then shifted to another location by 85 u (60 u in both X and Y) for 10,000 additional time steps (phase 2), resulting in disjoint home ranges, and then shifted to a third location by 28 u (20 u in both X and Y) for 10,000 additional time steps (phase 3), resulting in overlapping home ranges. In another batch of simulations, the central place remained at the same location for the 30,000 time steps, but σ_0 was set to 0.7 radians for time steps 10,001 to 20,000 (phase 2) and to 0.5 radians otherwise (phases 1 and 3), involving a transitory enlargement of the home range. We finally sub-sampled the data sets by keeping one location every 60. The home range phases were then defined by 166 locations each, with low serial correlation, and were thus similar to actual datasets that are commonly used in home range studies. Note that in our approach that

focuses on the contrast between the stationary phases, the actual lengths of these phases do not matter, provided they are longer than L_{min}.

Changes in behavioural modes. We simulated a random search movement as a correlated random walk where three types of activity – immobility (resting or standing), intensive (area-concentrated) searching and extensive searching (transit) – alternate, each one lasting 20 time steps, this 60-step sub-series being repeated 5 times. The step lengths L_i were drawn from a log-normal distribution with a mean equal to 0.5 u in the intensive mode or 1.0 u in the extensive mode, and with a standard deviation equal to 1/10th of the mean in both modes. Turning angles α_i were drawn from a wrapped Gaussian distribution with a null mean and a standard deviation equal to 0.4 radians in the intensive mode or 0.3 radians in the extensive mode. To mimic possible factors (e.g. GPS recording noise) that can blur the contrast between the modes, the locations obtained in this way, as well those obtained for immobility phases, were submitted to bivariate Gaussian random noise with a null mean and various standard deviations ζ. Note that in order to assess the ability of a method to segment a behavioural time-series, the precise movement rules used in simulations are not important. What really matters is the contrast between the different phases: with a high contrast, all methods should work well, whereas with a low contrast, all methods should fail, whatever the movement rules considered. In the results section, we will present the results obtained with a standard deviation of the noise ζ set to 0.2, 0.3 and 0.4, involving respectively a high, moderate and low contrast between the three modes.

METRICS

For identifying home range shifts, the two signal components considered are orthonormal Cartesian coordinates (x_i, y_i) of locations (GPS locations expressed in decimal degrees as longitude and latitude therefore require to be transformed in terms of easting and northing

through a classical projection such as UTM). For identifying behavioural modes, the two components usually considered in HMM-based approaches are the classical metrics corresponding to the step lengths L_i and the turning angles α_i, computed from locations recorded at constant time intervals Δt, and therefore acting as proxies for linear ($L_i/\Delta t$) and angular ($\alpha_i/\Delta t$) speeds, respectively. We used such metrics for comparative purpose, but we also tested some variants, assumed to improve the contrast between the different modes. We computed the linear speed as $(L_i+L_{i+1})/(2\Delta t)$. Although this basic smoothing introduces some serial correlation (r=0.5) which is not taken into account in our model, it should result in a less noisy signal. Furthermore, angular speed may show faded changes between searching modes because the intensive mode usually involves both a decrease in linear speed and an increase in path tortuosity but angular speed mechanically increases with both of them (Benhamou & Bovet, 1989; Barraquand & Benhamou, 2008). We therefore computed turning angles α_i* based on a constant step length r rather than at constant time interval. For this purpose, each location $\mathbf{X}_i = (X_i, Y_i)$ is considered the centre of a virtual circle with radius r, and the entrance and exit locations $\mathbf{P_{en}}$ and $\mathbf{P_{ex}}$ are determined through linear interpolation (Supporting information 2). The turning angle α_i* is then computed in $[-\pi, \pi]$ as the angular deviation between vectors $\mathbf{P_{en}}\rightarrow\mathbf{X}_i$ and $\mathbf{X}_i\rightarrow\mathbf{P_{ex}}$ (both with length r) rather than vectors $\mathbf{X}_{i-1}\rightarrow\mathbf{X}_i$ (with length L_i) and $\mathbf{X}_i\rightarrow\mathbf{X}_{i+1}$ (with length L_{i+1}) as done to compute α_i. When both L_i and L_{i+1} are larger than r, one gets α_i* = α_i, whereas $|\alpha_i$*$|$ tends, on average, to be larger (random search paths) or smaller (directed paths) than $|\alpha_i|$ when r is larger. In our simulations, we noticed that using a radius larger than the median of the overall step length distribution tends to improve the discrimination between the fast and slow movement modes but to worsen the discrimination between the slow movement mode and the immobility mode. We therefore chose to set r to the median of the step length distribution, and accordingly used this value in all our analyses. We also tested the two orthogonal signals provided by the

"persistence speed" $L_{i+1}\cos(\alpha_i)/\Delta t$ and "turning speed" $L_{i+1}\sin(\alpha_i)/\Delta t$ (Gurarie et al., 2009; Gloaguen et al., 2015).

PRACTICAL IMPLEMENTATION OF THE METHOD

Both segclust2d procedures (segmentation-only and segmentation-clustering) have been currently implemented for bivariate times-series (the case we considered in this paper) in an R package (available at cran.r-project.org/package=segclust2d). An integrated module makes it possible to derive the various movement variables mentioned in this paper from locations data. Because our approach requires large amounts of computer memory, it cannot deal with too long time-series (> 10000 values) on small desktop computers. It is worth noting however that, even in absence of any memory constraints, it is usually not a good practice to attempt to directly segment very long series, which encompass both very large scale phenomena (thanks to their large extent) and very small scales phenomena (thanks to their high resolution). Indeed, small-scale data are usually not relevant for analysing large-scale patterns and therefore act more as noise than as information in this context. For these reasons, some sub-sampling (thinning) is automatically performed by segclust2d when necessary. Thus, in our home range simulations, keeping a location every 60 makes it possible to drastically shorten the time-series by eliminating fine-scale movements (klinokinetic process), which are characterised by a high level of serial correlation in location and in direction. Such details are clearly not relevant for the question of home range shifts, where only the overall phase-dependent mean and/or the variance of locations matter (accordingly, our approach ignores serial correlations occurring in any stationary phase). Conversely, for fine-scale movement studies, the characteristics of the environment are liable to change (e.g. due to seasonal variations) when considering a time-series running over an extended duration, possibly leading to change in the characteristics of the behavioural classes expected. It appears therefore preferable to consider the various phases (e.g. seasons) separately rather than to

attempt to deal with the long time-series as a whole. Finally note that the time-series to be analysed are assumed to be regular (no gaps). Nevertheless, some gaps in location data do not matter when the metrics used directly correspond to northing and easting, as they will not change their mean and variance. In contrast, for derived metrics such as speed or turning angle, however, the occurrence of missing locations may result in noticeable changes in the values obtained, and is therefore likely to bias the segmentation-clustering process.

Results

IDENTIFYING HOME RANGE SHIFTS

Simulated movements. Figure 1 shows an example where the centre of the home range was shifted (by 85 u between phases 1 and 2, and by 28 u between phases 2 and 3), and an another example where the home range was enlarged during the phase 2 with respect to phases 1 and 3. The segmentation-only procedure correctly determined that the actual number of phases was equal to 3 in 98 out of the 100 replicates involving shifts in mean location (i.e. central place), and in 88 out of the 100 replicates involving shifts in variance (i.e. change in home range size). In these cases, the timing of the various shifts were also correctly determined (mean±SD = 10152±79 and 20092±188 time steps for the 85-u and 28-u shifts in mean location, respectively; mean±SD = 10035±1184 and 19942±1314 time steps for the first and second shifts in variance of the same amplitude). In contrast, the OU-based algorithm "marcher", which was specifically developed by Gurarie et al. (2017) to identify home range shifts in mean location. The "marcher" algorithm requires that the number of shifts has been specified, and is unable to detect shifts in variance. Without initial information about the possible timing of the shifts, this method relies on a simple k-means procedure to get initial guesses, and was then only able to detect the large shift in location (mean±SD = 10152±937 time steps; the small one occurred anywhere unpredictably between 15000 and 30000 time

steps, based only on 75 replicates, as the algorithm failed to provide any result for 25 replicates). However, the "marcher" algorithm proved able to correctly detect the two shifts in mean location (mean±SD = 10152±82 and 19967±239 time steps for the 85-u and 28-u shifts in mean location, respectively) when the actual home range centres and shift dates were provided as truly relevant initial guesses.

Illustrative example. We used the GPS track of an African Elephant (*Loxodonta africana*), recorded for > 2.5 years to illustrate the way the segclust2d/segmentation-only procedure can identify home ranging phases and shifts. The whole time-series of easting and northing coordinates appeared to be stationary, and thus corresponds to a large multiannual (possibly lifetime) home range. At smaller scale, it also appeared to be piecewise stationary. It can therefore be segmented to highlight temporary (possibly seasonal) smaller home ranges and the shifts between them (Fig. 2). However, some of the phases so highlighted are clearly nonstationary. In particular, segments 1 (days 1 to 46) and 5 (days 655 to 761) correspond mainly to a slow south-westwards migration between the two core areas of the multiannual home range. Segment 2 also corresponds to a nonstationary, migratory (southwards moving) phase, which went through an area used as a temporary home range during segments 4 and 6. This indicates that a same area can be used in different ways at different periods.

IDENTIFYING BEHAVIOURAL MODES

Simulated movements. An example of path with three behavioural modes (extensive searching, intensive searching and resting) is shown in Fig. 3 with the corresponding time-series in terms of turning angle α_i^* and smoothed speed $(L_i+L_{i+1})/(2\Delta t)$. In this example, the segclust2d/segmentation-clustering procedure appears able to detect the true number of modes ($M = 3$) and to attribute almost all locations to the right mode. We compared our method with a HMM-based method specifically designed to deal with movement data (Michelot et al.,

2016; McClintock & Michelot, 2018) when the true number of modes has been specified. The results obtained from 100 replicates showed that our procedure rivals with the HMM-based method although the latter was helped by initial state-dependent probability distribution parameters which were tuned to their true values for each behavioural state (Fig. 4 with medium noise level $\zeta = 0.3$). With a very low noise level ($\zeta < 0.2$), an excellent fit was obtained with all methods and metrics considered, whereas with a very high noise level ($\zeta > 0.4$) the percentage of correct state assignment became closer to the value expected for a random assignment (33%; see Supporting information 3.1 results with $\zeta = 0.2$ and $\zeta = 0.4$). It also appeared that the angular ($\alpha_i/\Delta t$) and linear ($L_i/\Delta t$) speeds are suitable metrics for detecting behavioural changes with HMM only when the noise is not too high. With a high noise level, better results were obtained with both methods when using any other of the couples of metrics considered. The best fits were obtained with turning angle α_i* or it absolute value $|\alpha_i$*| and smoothed speed. When the true number of modes is unknown, our method can also estimate this number as the most likely number of clusters, but the fraction of correct estimate is too low to consider the result as reliable (Supporting information 3.2).

Illustrative example. We used a 24-h GPS track of a plains zebra (*Equus quagga*) to illustrate the way the segclust2d/segmentation-clustering procedure can identify the occurrences of the various movement modes (Fig. 5). Although that, in this example, the most likely number of modes was estimated to be five, we present the segmentation obtained when setting this number to three, assuming that the biologically relevant modes should be resting (or any other non-moving behaviour such as standing), feeding and transiting (the other two modes detected by our procedure when using five modes were assumed to correspond to mixed behaviours).

Discussion

We showed that the method we proposed, segclust2d, which extends the methods developed by Lavielle (2005) and Picard et al. (2007) to multivariate time-series, makes it possible to reliably detect two types of changes that are of key importance when studying free-ranging animal movements: home range shifts, based on bivariate time-series of location coordinates (segmentation-only procedure), and changes in behavioural modes, based on bivariate time-series of turning angles and speed (segmentation-clustering procedure). In any case, this new method is straightforward to tune: the user has just to set the minimum segment length (L_{min}) to a biologically relevant value. Nevertheless, it proved to work at least as well as, and often better than, other recent methods specifically designed to deal with either home range shifts (Gurarie et al., 2017) or changes in behavioural modes (Michelot et al., 2016, McClintock & Michelot, 2018).

Gurarie et al. (2017) developed an OU-based method to identify home range shifts in mean location. Using computer simulations, we compared this approach, as implemented in Gurarie et al.'s "marcher" algorithm, with segclutst2d/segmentation-only. Both methods are well able to detect large shifts in mean location. However, our method is also able to detect small shifts in mean location, whereas Gurarie et al.'s one requires a priori information on the actual mean locations and the shifts dates to correctly detect them, although this is precisely in this case that such information is usually lacking (i.e., they can hardly be guesstimated from visual inspection of the data). Furthermore, contrary to Gurarie et al.'s method, which can deal only with a few number of shifts which have to be specified in advance, our method can work with any number of shifts and is able to correctly estimate this number by itself in most cases. It is also able to reveal changes in home range size. Yet, to be efficient, our method does not require any initial guess. It simply requires specifying a minimum length (L_{min}) for stationary phase to be called a temporary home range, shorter phases being assumed to correspond to

transitory exploitations of restricted areas rather than to home ranges. However, whereas our method considers migrations as simple breakpoints, Gurarie et al.'s method can estimate the duration of migrations.

The elephant we considered in our illustrative example tended to move back and forth between two main areas. This kind of space use is common in migrating birds that commute between reproductive and wintering home ranges. However, there are numerous studies showing more complex patterns, with an animal setting several distinct temporary home ranges successively (Naidoo et al., 2012; Benhamou, 2014; Cagnacci et al., 2016; Couriot et al. 2018). The segmentation of a long piecewise locational time-series in phases corresponding to temporary home ranges opens promising perspectives to understand how the occurrences and durations of home ranges are related to environmental co-variates, which is a prerequisite to infer long-term consequences for population distribution (Muller & Fagan, 2008). The elephant illustrative example also shows that, although the model underlying segclust2d looks for stationary phases, there is no guaranty that all segments obtained are really stationary. This occurs because changes between stationary phases are modelled as breakpoints but may in fact correspond to slow progressive changes.

Since the pioneering paper by Morales et al. (2004), HMM-based methods have often been considered the best way to detect changes in behavioural modes of remotely tracked animals. An alternative approach was proposed by Barraquand & Benhamou (2008). It consisted in computing the series of residence time within a virtual circle running along the path and to search for the most likely breakpoints using Lavielle (2005)'s univariate segmentation method. However, although the residence time provides a simple and reliable univariate signal easy to segment and interpret, the values obtained depend not only on the type of behaviour that is performed but also on how long it is performed, preventing the segments corresponding to the same behaviour from being easily clustered. In the present study, we show using computer simulations that the segclust2d/segmentation-clustering

451 procedure rivals (and can even outperform) a HMM-based method initialised with state-
452 dependent probability distribution parameters tuned to their true values for the different
453 behavioural modes. With actual data, as the true values of parameters are usually unknown,
454 our method, which does not require any initial guess, should have a clear advantage over the
455 HMM-based method. It is worth noting that, with both methods, the best results were usually
456 obtained with the joint use of metrics other linear and angular speeds, such as smoothed speed
457 and turning angle measured at constant step length, which were expected to improve the
458 contrast between the intensive and extensive searching modes. Interestingly, using absolute
459 rather than signed values of turning angles measured at constant step length works at best
460 with our method whereas right and left turns were balanced in any mode in our simulated
461 movements. Such metrics should be particularly useful to distinguish between intensive and
462 extensive modes when the former involves turning systematically right or left, i.e.
463 characterised by markedly either negative or positive mean turning angles, whereas the latter
464 involves balanced turning, as occurs in some species (e.g. Smith, 1974). As it results also in a
465 reliable identification when turns are balanced in both intensive and extensive searching, we
466 recommend using it systematically when using our method to distinguish between extensive
467 and intensive searching phases.

468 In the illustrative example on zebra's movements, five behavioural modes were
469 detected by the segclust2d/segmentation-clustering procedure in the time-series of smoothed
470 speed and absolute value of the turning angles measured at constant step length. Nevertheless,
471 based on behavioural observations, we chose to segment the time-series with only three
472 modes assumed to correspond to immobility, feeding and transit. Indeed, although our method
473 can estimate the number of states based on a statistical criterion, it turns out in the computer
474 simulations that this number was poorly estimated, despite the behavioural modes were
475 clearly defined in this case. With actual data, there can be some mixing between modes, for
476 instance transit and opportunistic feeding at some times, so that the estimation of the number

of relevant modes may become unreliable. Thus, we recommend using the capacity of the segclust2d/ segmentation-clustering procedure to estimate the number of states only when this number cannot be fixed a priori based on biological arguments. A similar conclusion was reached by Pohle et al. (2017) for HMM-based methods. It is also worth noting that feeding and resting can be distinguished based on movement characteristics only in animals which have to move significantly (with respect to the location recording noise) to feed. For animals which feed mainly without markedly moving, such as some browser herbivores and carcass-eating carnivores, ancillary activity data, such as those provided by on-board accelerometers, are required to distinguish these two behavioural modes. As our method can work conjointly on any number of time-series of any nature, future implementation could integrate activity (accelerometer-based) time-series for a better identification of resting vs. active phases.

The segmentation of piecewise stationary time-series, possibly complemented by the clustering of the resulting segments into functional classes, is key to understanding the dynamics of underlying processes. Based on bivariate time-series of metrics such as easting and northing (home range shift studies), or speed and turning angles (movement mode studies), segclust2d has the potential to facilitate discovery in the field of movement ecology (e.g. see Thaker et al. 2019). If necessary, this approach can potentially apply to more dimensions as well, so as to consider ancillary variables such as activity and other metrics such as distance to a nest, proxies of habitat quality, or any other variable that may be relevant when studying animal movements. As it can deal with two or more variables of any nature, our approach should be useful not only in movement ecology but also in many other fields, using appropriate metrics.

A theoretical limitation of our approach is that the metrics used are assumed to result in Gaussian and independent (for a given segment) values. The normality assumption makes computations easier but other distributions can be considered (Cleynen et al., 2013). For segmentation-only, deviation of the data from normality is not problematic, as the cost of

additional segments, based on the log-likelihood when assuming normality, can be interpreted more generally (i.e. for any distribution) as a contrast based on mean and variance. For segmentation-clustering, the way deviation from normality may affect the results remain unclear. Furthermore, dynamic programing cannot be directly applied when some dependence is considered in this framework, even if some possibilities have recently emerged (Chakar et al., 2017). However, computer simulations showed that our method is quite robust to violations of these assumptions. Thus, in home range shift study, despite northings and eastings showed a large serial correlation (even after subsampling; see Fig. 1a,b) and were not normally distributed, our method was nevertheless able to correctly detect the small shift in mean location in 98% of replicates. Our method was also able to perform an efficient segmentation-clustering in the movement mode study, despite a serial correlation (r=0.5) was mechanically introduced when considering the smoothed rather than usual speed, and the speed (drawn from a log-normal distribution) and the turns obtained for resting phases (and for any phase when turns were taken in absolute value) were not normally distributed. Some mathematical transformations may be used to attempt to normalise the data. However, using a log-transformed speed, which should a priori look like more Gaussian, did not improve the segmentation (results not showed), probably because the smoothing over two steps as well as the noise that affects the locations tend to distort the initial distribution. The choice of the metrics used should therefore first favour those that appear as being the easiest to interpret in terms of biological significance and the most able to give birth to piecewise stationary time-series (i.e. the most likely to reveal the occurrence of a breakpoint between two successive phases). The question of normality and independence certainly matters, but to a lesser extent. For home range shift studies, northings and eastings appear to be the obvious choice. For movement mode studies, the smoothed speed and the absolute value of turns at constant step length, which characterise the movement in terms of temporal and spatial component

separately, and seem able to maximize the contrast between the various modes when the recording noise is high, are certainly the two metrics to consider first.

Acknowledgments

We thank Clément Calenge for interesting discussions about segmentation issues, and two anonymous reviewers for their constructive critiques. The study was partially funded by the grant ANR-16-CE02-0001-01 of the French 'Agence Nationale de la Recherche', and the Zone Atelier program of the CNRS. Computer simulations were programmed in Pascal and ran thanks to the FreePascal compiler (www.freepascal.org).

Data accessibility

The code of the method is already publicly available as an R package (cran.r-project.org/package=segclust2d). There are only a few data used as examples, and they will be made publicly available as well.

References

Barraquand, F., & Benhamou, S. (2008). Animal movements in heterogeneous landscapes: identifying profitable places and homogeneous movement bouts. *Ecology,* 89, 3336–3348.

Bellman, R. (1954). The theory of dynamic programming. *Bull. Am. Math. Soc*., 60, 503–516.

Benhamou, S. (1989). An olfactory orientation model for mammals' movements in their home ranges. *J. Theor. Biol.*, 139, 379–388.

Benhamou, S. (2014). Of scales and stationarity in animal movements. *Ecol. Lett.,* 17, 261–272.

552 Benhamou, S., & Bovet, P. (1989). How animals use their environment: a new look at kinesis.
553 *Anim. Behav.*, 38, 375–383.
554 Beyer, H.L., Morales, J.M., Murray, D. and Fortin, M.J., 2013. The effectiveness of Bayesian
555 state-space models for estimating behavioural states from movement paths. *Meth. Ecol.*
556 *Evol.* 4, 433–441.
557 Blackwell (1997). Random diffusion models for animal movement. Ecological Modelling,
558 100(1-3), 87-102
559 Breed, G.A., Golson, E.A., & Tinker, M.T. (2017). Predicting animal home-range structure
560 and transitions using a multistate Ornstein-Uhlenbeck biased random walk. *Ecology*, 98,
561 32–47
562 Bunnefeld, N., Börger, L., van Moorter, B., Rolandsen, C.M., Dettki, H., Solberg, E.J., &
563 Ericsson, G. (2011). A model-driven approach to quantify migration patterns: individual,
564 regional and yearly differences. *J. Anim. Ecol.*, 80, 466–76.
565 Cagnacci, F., Focardi, S., Ghisla, A., van Moorter, B., Merrill, E.H., Gurarie, E., Heurich, M,
566 Mysterud, A., Linnell, J., Panzacchi, M., May, R., Nygard, T., Rolandsen, C., &
567 Hebblewhite, M. (2016). How many routes lead to migration? Comparison of methods to
568 assess and characterize migratory movements. *J. Anim. Ecol.*, 85, 54–68.
569 Chakar, S., Lebarbier, E., Lévy-Leduc, C., & Robin, S. (2017). A robust approach to multiple
570 change-point estimation in an AR(1) process. *Bernoulli*, 23, 1408-1447.
571 Cleynen, A., Dudoit, S. & Robin, S. (2013). Comparing segmentation methods for genome
572 annotation based on RNA-Seq data. *J. Agricult, Biol. & Environ. Stat.,* 19, 101-118.
573 Couriot, O., Hewison, A.J.M., Said, S., Cagnacci, F., Chamaillé-Jammes, S., et al. (2018).
574 Truly sedentary? The multi-range tactic as a response to resource heterogeneity and
575 unpredictability in a large herbivore. *Oecologia,* 187, 47–60.
576 Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete
577 data via the EM algorithm. *J. Roy. Stat. Soc. B*, 39, 1–38.

578 Dias, M.P., Granadeiro, J.P., & Palmeirim, J.M. (2009). Searching behaviour of foraging
579 waders: does feeding success influence their walking? *Anim. Behav.*, 77, 1203–1209.
580 Edelhoff, H., Signer, J. & Balkenhol, N. (2016). Path segmentation for beginners: an
581 overview of current methods for detecting changes in animal movement patterns. *Mov.*
582 *Ecol.* 4, 21.
583 Gloaguen, P., Mahévas, S., Rivot, E., Woillez, M., Guitton, J., Vermard, Y., & Etienne, M-P.
584 (2015). An autoregressive model to describe fishing vessel movement and activity.
585 *Environmetrics*, 26, 17–28.
586 Gurarie, E., Bracis, C., Delgado, M., Meckley, T.D., Kojola, I., & Wagner, C.M. (2016),
587 What is the animal doing? Tools for exploring behavioural structure in animal
588 movements. *J. Anim. Ecol.,* 85, 69–84.
589 Gurarie, E., Andrews, R.D., & Laidre, K.L. (2009). A novel method for identifying
590 behavioural changes in animal movement data. *Ecol. Lett.*, 12, 395–408.
591 Gurarie, E., Cagnacci, F., Fleming, C., Peters, W., Calabrese, J.M., Muller, T., & Fagan, W.F.
592 (2017). A framework for modelling range shifts and migrations: asking when, whither,
593 whether and will it return. *J. Anim. Ecol*., 86, 943–959.
594 Karlin, S., & Taylor, H.M. (1975). A First Course in Stochastic Processes. Academic Press,
595 New York.
596 Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D., & Morales, J.M. (2012).
597 Flexible and practical modeling of animal telemetry data: hidden Markov models and
598 extensions. *Ecology*, 93, 2336–2342
599 Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal*
600 *Processing*, 85, 1501–1510
601 Levin, S.A. (1992). The problem of pattern and scale in ecology. *Ecology*, 73, 1943–1967.

602 Michelot, T., Langrock, R., & Patterson, T.A. (2016). moveHMM: an R package for the
603 statistical modelling of animal movement data using hidden Markov models. *Meth. Ecol.*
604 *Evol*., 7, 1308–1315.
605 McClintock, B.T., King, R., Thomas, L., Matthiopoulos, J., McConnell, B.J., & Morales, J.M.
606 (2012). A general discrete-time modeling framework for animal movement using
607 multistate random walks. *Ecol. Monogr.*, 82, 335–349.
608 McClintock, B.T., & Michelot, T. (2018). momentuHMM: R package for generalized hidden
609 Markov models of animal movement. *Methods Ecol Evol*., 9, 1518–1530.
610 Monsarrat, S., Benhamou, S., Sarrazin, F., Bessa-Gomes, C., Bouten, W., & Duriez, O.
611 (2013). How predictability of feeding patches affects home range and foraging habitat
612 selection in avian social scavengers? *PLoS ONE* 8: e53077
613 Morales, J.M., Haydon, D.T., Frair, J., Holsinger, K.E., & Fryxell, J.M. (2004). Extracting
614 more out of relocation data: building movement models as mixtures of random walks.
615 *Ecology*, 85, 2436–2445.
616 Mueller, T., & Fagan, W.F. (2008). Search and navigation in dynamic environments from
617 individual behaviors to population distributions. Oikos 117, 654–664.
618 Naidoo, R., Du Preez, P., Stuart-Hill, G., Jago, M., & Wegmann, M. (2012). Home on the
619 range: Factors explaining partial migration of African buffalo in a tropical environment.
620 PLoS ONE 7: e36527.
621 Nams, V.O. (2014). Combining animal movements and behavioural data to detect behavioural
622 states. *Ecol. Lett*., 17, 1228–1237.
623 Parton, A., & Blackwell, P.G. (2017). Bayesian Inference for Multistate 'Step and Turn'
624 Animal Movement in Continuous Time. *J. Agricult. Biol. Environ. Stat.*, 22, 373–392.
625 Picard, F., Robin, S., Lebarbier, E., & Daudin, J.-J. (2007). A Segmentation/Clustering Model
626 for the Analysis of Array CGH Data. *Biometrics*, 63, 758–766.

627 Pohle, J., Langrock, R., van Beest, F.M., & Schmidt, N.M. (2017). Selecting the number of
628 states in hidden Markov models: pragmatic solutions illustrated using animal movement.
629 *J. Agricult. Biol. & Environ. Stat.*, 22, 270–293.
630 Pyke, G.H. (1978). Are animals efficient harvesters? *Anim. Behav.*, 26, 241–250.
631 Rigaill, G. (2015) A pruned dynamic programming algorithm to recover the best
632 segmentations with 1 to kmax change-points. *J. Soc. Fr. Stat.*, 156, 180–205.
633 Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist*., 6, 461–464.
634 Smith, J.N. (1974). The food searching behaviour of two European thrushes. II. The
635 adaptiveness of search patterns. *Behaviour* 49, 1–61.
636 Thaker, M., Gupte, P.R., Prins, H.H., Slotow, R., & Vanak, A.. 2019. Fine-scale tracking of
637 ambient temperature and movement reveals shuttling behavior of elephants to water.
638 *Frontiers in Ecol. Evol.* 7, 4.

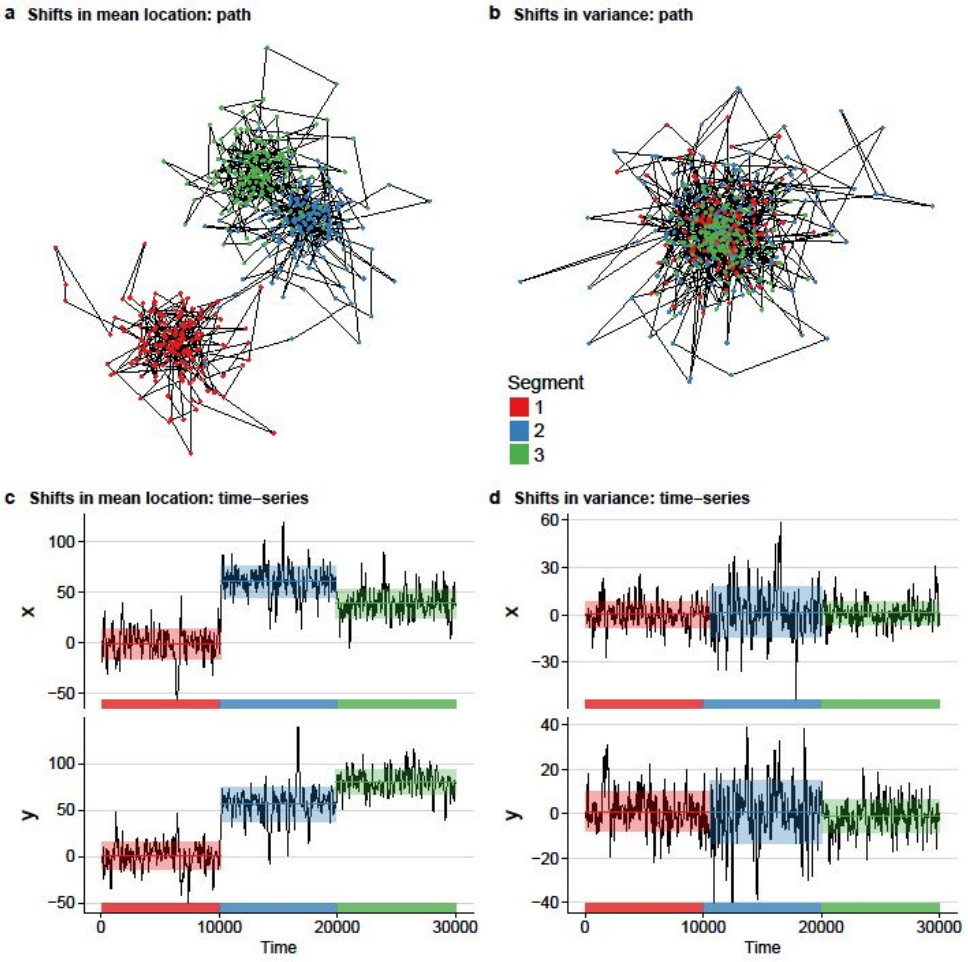


Fig. 1. Examples of application of the segclus2d/segmentation-only procedure to highlight home range phases and shifts in simulated movements. Top panels show the simulated paths (after 1/60 subsampling) corresponding to three home range phases (two shifts), either in mean location (a) or in variance (b). The corresponding time-series for both location coordinates (x, y) are presented in panel (c) and (d), respectively. The horizontal colour bars running along the time axis show the true occurrences of the three phases, whereas the coloured bands appearing over the x and y signals show their occurrences as estimated using the segclust2d/segmentation-only method with L_{min} = 45 locations (corresponding to 2700 time steps because of the 1/60 subsampling) and provide the estimated mean (plain horizontal line running in the middle of the band) ± standard deviation (band width) for each segment separately.

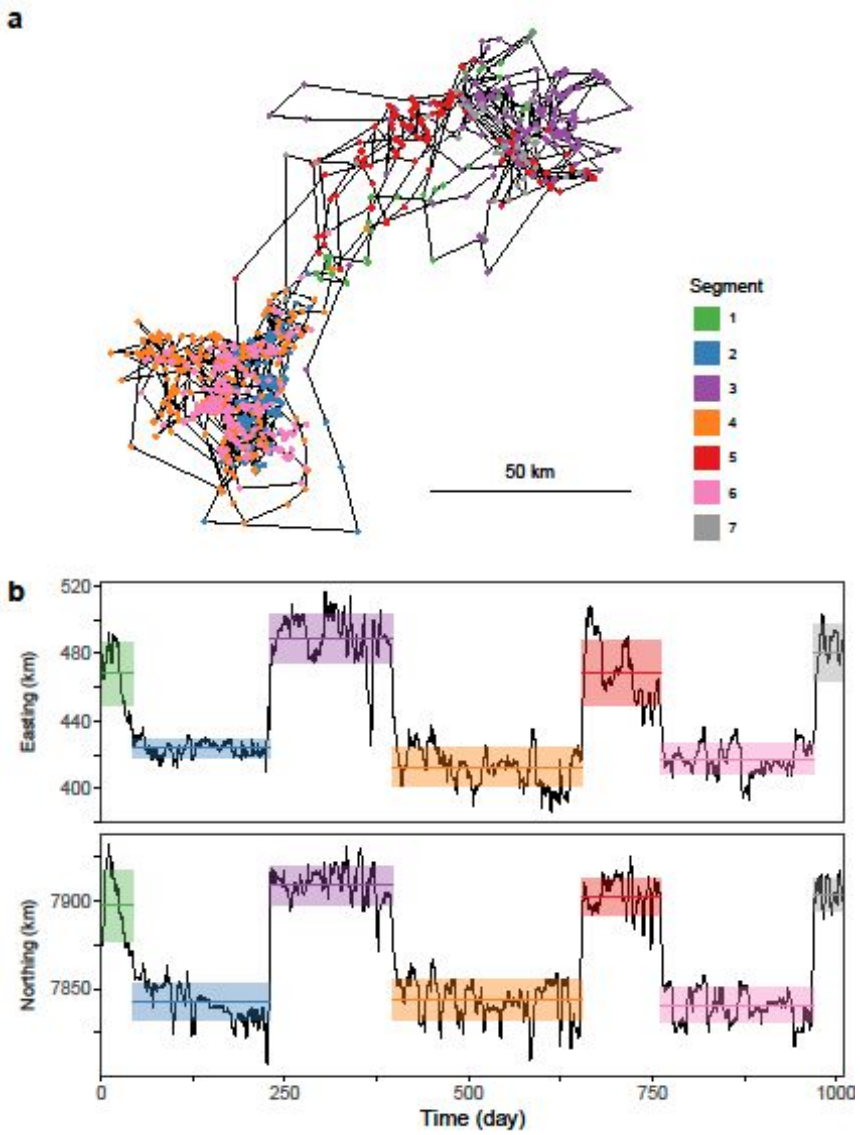


Fig. 2. Example of application of the segclust2d/segmentation-only procedure to highlight home range phases and shifts in an African elephant's movement recorded over 1000 days. (a) Rough path representation obtained by linking the locations subsampled so as to keep a single GPS location per day; (b) Corresponding time-series of locations coordinates (easting and northing). The coloured bands appearing over the time-series show the estimated mean (plain horizontal line running in the middle of the band) ± standard deviation (band width) of each of the seven segments obtained using the segmentation-only method with L_{min} = 30 days.

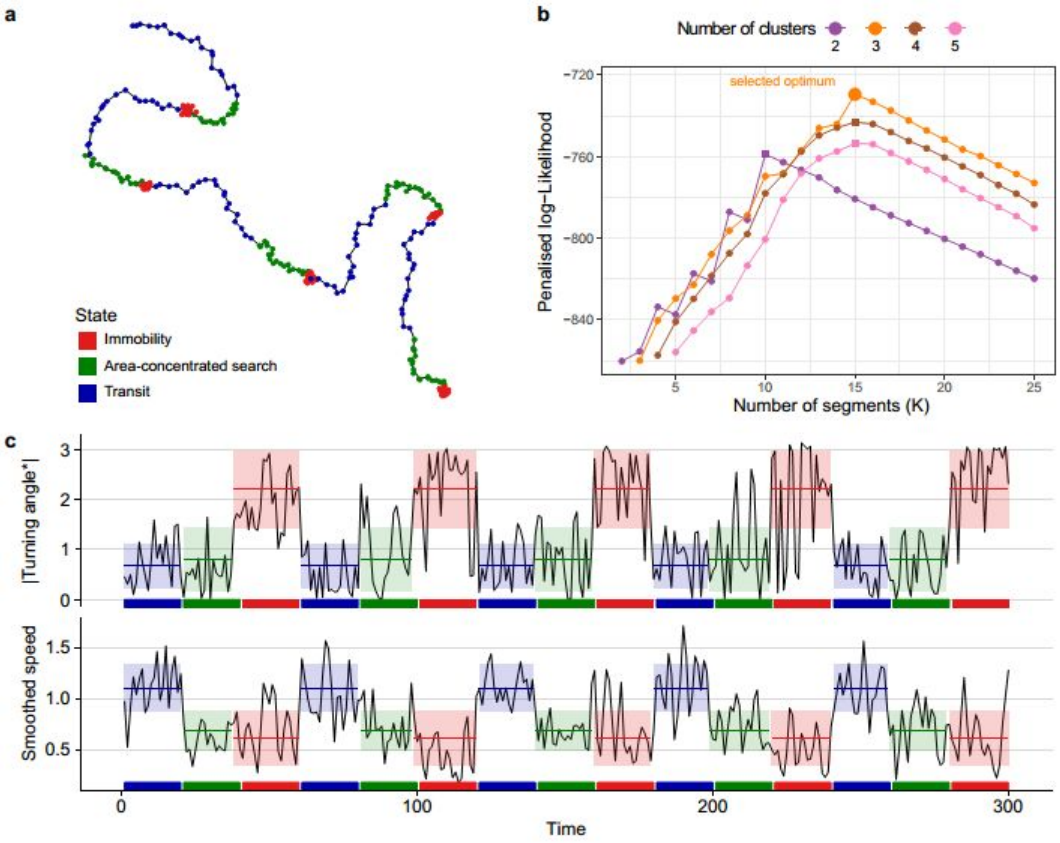
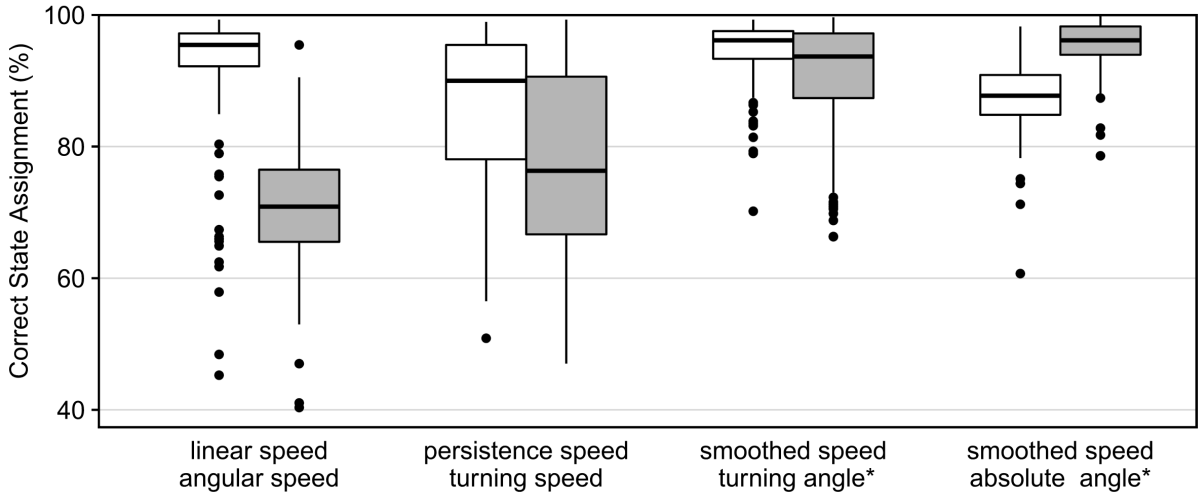
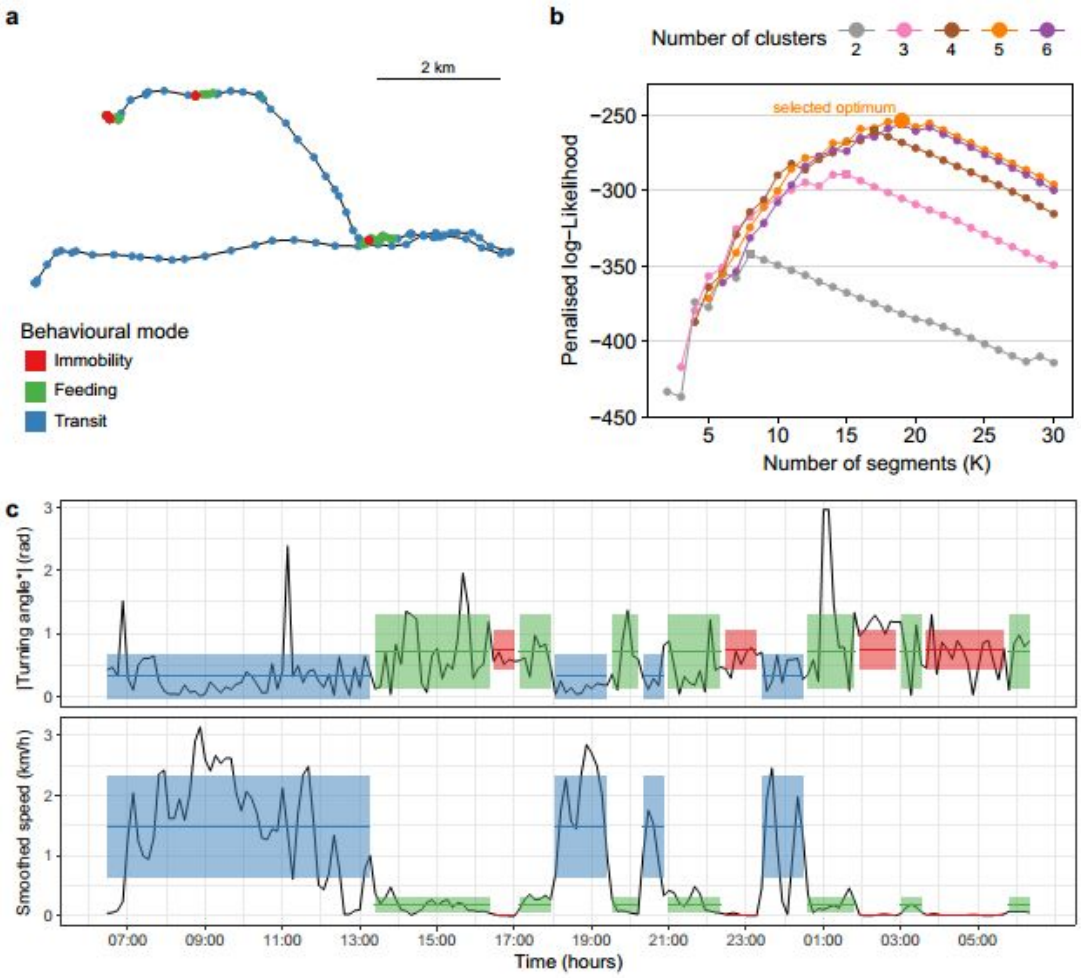


Fig. 3. Example of application of the segclust2d/segmentation-clustering procedure to highlight behavioural changes in a simulated movement. (a) Simulated path as a composite correlated random walk, incorporating an additional noise with standard deviation $\zeta = 0.3$ u that affects locations; (b) Determination using BIC-based penalised likelihood of the most likely numbers of states ($M = 3$) and segments ($K = 15$) (big orange dot), and of the most likely number of segments for the other three numbers of states considered (large squares at the top of the curves). (c) Corresponding time-series in terms of absolute turning angle computed with a constant step length, $|\alpha_i^*|$, and smoothed speed, $(L_i+L_{i+1})/(2\Delta t)$, segmented with $L_{min} = 10$ and $M = 3$; the coloured bands appearing over the two time-series show the estimated occurrences and mean (plain horizontal line running in the middle of the band) ± standard deviation (band width) for each of the three movement modes whereas the horizontal colour bars running along the time axis show the true occurrences of these modes.



673

674 **Fig. 4. Comparative performances of the segclust2d/segmentation-clustering procedure**
675 **vs. a HMM-based method for highlighting behavioural changes.** The boxplots show the
676 proportion of correct state assignments, obtained for various bivariate signals when the true
677 number of states is known ($M = 3$), as estimated from 100 replicates simulated with the same
678 parameters as to the one illustrated in Fig. 3 (noise $\zeta = 0.3$). The star (*) indicates turning
679 angles computed with a constant step length, in terms of arithmetic (α_i*) or absolute ($|\alpha_i$*|)
680 values. The white boxplots show the results obtained with HMM-based R package
681 momentuHMM (McClintock & Michelot 2017), with initial state-dependent probability
682 distribution parameters tuned to their true values for the different states (using the following
683 distributions: Gaussian for persistence and turning speeds, wrapped Cauchy for angular speed
684 and turning angle α_i*, Weibull for linear speed, smooth speed and |turning angle*|). The grey
685 boxplots shows the results obtained using the segclust2d/segmentation-clustering procedure
686 with $L_{min} = 10$.



687

688 **Fig. 5: Example of application of the segclust2d/segmentation-clustering procedure to**
689 **highlight behavioural changes in a 24-h zebra's movement.** (a) Path representation
690 obtained by linking GPS locations recorded every 8 minutes; (b) Determination using BIC-
691 based penalised likelihood of the most likely numbers of states (M = 5) and segments (K =
692 20) (big orange dot), and most likely numbers of segments for the other number of states
693 considered (large squares at the top of the curves), with L_{min} = 5 (i.e. 40 min.). (c)
694 Corresponding time-series in terms of absolute turning angle computed with a constant step
695 length, $|\alpha_i^*|$, and smoothed speed, $(L_i+L_{i+1})/(2\Delta t)$, segmented with M = 3 (leading to K = 15);
696 the coloured bands appearing over the two time-series show the estimated occurrences and
697 mean (plain horizontal line running in the middle of the band) ± standard deviation (band
698 width) for each of the three clusters considered.

Supporting Information 1: Complements about segclust2d

Initialising the Expectation-Maximization algorithm

The Expectation-Maximization (EM) algorithm used to estimate the distribution parameters in the segmentation–clustering model is known to be sensitive to initialisation, so that it may converge to local maxima of the likelihood. This behaviour has some consequences on the parameters estimates but also makes the choice of the number of segments or states complicated. The classical initialisation solution consists in running the algorithm numerous times and just choose the point with the highest value of the log-likelihood, but this strategy is too computationally demanding. To minimize the risk of reaching local maxima within an unacceptable computation time, we use the following initialisation strategy: (1) perform a pure segmentation of the signal and (2) use a hierarchical cluster algorithm, based on the log-likelihood ratio distance to assign segments to states.

Even with smart initialisation points, however, the EM algorithm may still converge to local maxima. This situation appears when looking, for a given number of states M, at the log-likelihood as a function of the number of segments K: whereas it is expected to be somewhat regular, this function can be quite noisy. To solve this issue, we use new initial points for all 'non-reliable' solution, i.e. for which an initialisation problem can be suspected. These new initial points are based on the parameter estimates of the distribution (π_m, μ_m, Σ_m) obtained for all the 'reliable' solutions, which correspond to the points that lie on the convex hull of log-likelihood curve. New initial points are provided by cutting in half a segment or merging two segments. The improvement in terms of regularity of the log-likelihood curve obtained thanks to this procedure is illustrated in figure S1-1. Although the procedure does not formally guarantee the convergence, it would be unlikely that none of the first initialisation points had converged to a 'reliable' solution that could be propagated.

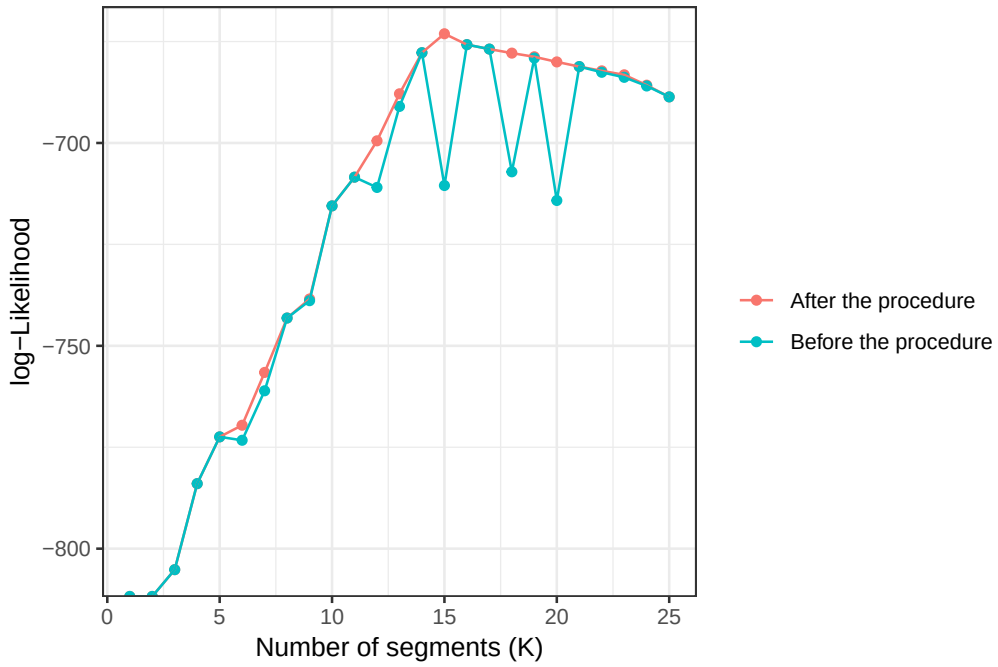


Figure S1-1: Maximum Likelihood estimates as a function of the number of segments before (in blue) and after (in red) the procedure. For instance the 'reliable' solution obtained with $K = 24$ segments was used to provide starting points for the EM algorithm for $K = 23$ and for $K = 25$, and this procedure gradually spreads over adjacent points.

Model selection

Choice of the number of segments K in the segmentation-only model

We used the adaptive model selection strategy proposed in Lavielle (2005) consisting in choosing the value of K that maximizes the following penalized log-likelihood : $\mathcal{L}_K - C\,K$ where $\mathcal{L}_K$ is the log-likelihood of the optimal segmentation in K segments and C is a unknown positive constant. The heuristic proposed by Lavielle (2005) makes it possible to shortcut the estimation of C. It consists in detecting the value of K for which the log-likelihood ceases to increase significantly. More specifically, consider the normalised log-likelihood defined as

$$\tilde{\mathcal{L}}_K = (K_{max} - 1)\frac{\mathcal{L}_{K_{max}} - \mathcal{L}_K}{\mathcal{L}_{K_{max}} - \mathcal{L}_1} + 1$$

Then, K is chosen as the value such that $\tilde{\mathcal{L}}_K$ displays the largest slope change. Namely, we take :

$$\hat{K} = argmin_K \left\{ (\tilde{\mathcal{L}}_K \tilde{\mathcal{L}}_{K+1}) - (\tilde{\mathcal{L}}_{K+1} \tilde{\mathcal{L}}_{K+2}) > S \right\}$$

where the value of threshold S is set to a predefined value (we used $S = 0.7$ as proposed in Lavielle, 2005).

As the selection relies on a predefined threshold, it is worth checking where the point corresponding to the selected number of segment lies on a plot of the log-likelihood curve (fig. S1-2). The optimal K value obtained in this way should correspond to a noticeable slope change.

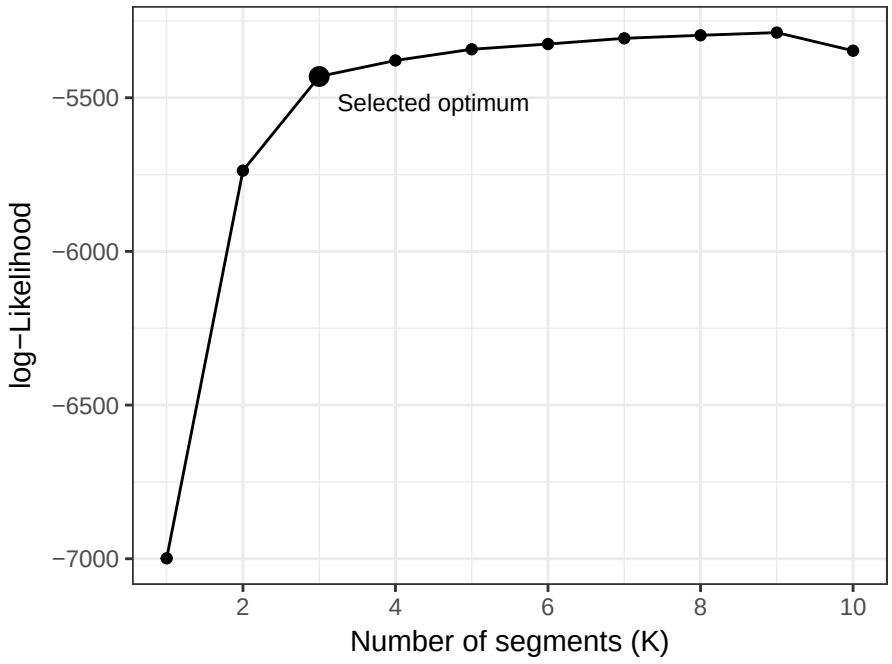


Figure S1-2: log-likelihood of a segmentation as a function of the number of segment. The optimum selected by the criterion from Lavielle (2005) should be located at a break in the increase of the curve.

Choice of the number of segments K and states M in the segmentation–clustering model

Selection of the best segmentation-clustering model (i.e. of the best couple of K and M values) is a hard task as no method has been yet proposed for this purpose. A log-likelihood is expected to increase with the number of parameters. However, as explained in Picard et al. (2007), if the log-likelihood increases with the number of clusters M, it does not always increases with the number of segments K. Indeed a phenomenon of self-penalization occurs at the 'true' number of segments when the detection of breakpoints is easy, stressing to choose K simply as the value that maximizes the log-likelihood. However when the detection of breakpoints is more difficult, choosing the maximum value would tend to overestimate K. Picard et al. (2007) suggested to add a penalty. A Bayesian Information Criterion (BIC)-based penalty appeared to be sufficient in this case, although it does not work in pure segmentation (Picard et al., 2005). As BIC is the most popular criterion to choose the optimal number of clusters in a mixture model (Frühwirth-Schnatter, 2006), we used the maximum value of the following BIC-based penalised likelihood $\mathcal{B}_{K,M}$ for the selection of both K and M parameters:

$$\mathcal{B}_{K,M} = \mathcal{L}_{K,M} - \frac{5 \times M - 1}{2} log(2n) - \frac{K-1}{2} log(2n),$$

where $\mathcal{L}_{K,M}$ stands for the log-likelihood for the optimal segmentation-clustering with K segments classified into M states. The penalization terms in the BIC criterion is half the number of parameters times the logarithm of the size of the dataset. For our model the number of parameters to be estimated is $2M$ means + $2M$ variances + $M-1$ proportions for the states, and $K-1$ breakpoints for the segments, and the size of the dataset for n bivariate values is $2n$.

Although this procedure appears to work well for choosing the optimal number of segments, it has been observed to be less reliable for choosing the optimal number of states, which tends to be overestimated. We therefore advise users to set an a priori number of states M, based on biological knowledge. We also advise to look at the plot of the BIC-penalized log-likelihood, as in figure 4b of main text, to check that the solution obtained makes sense.

References

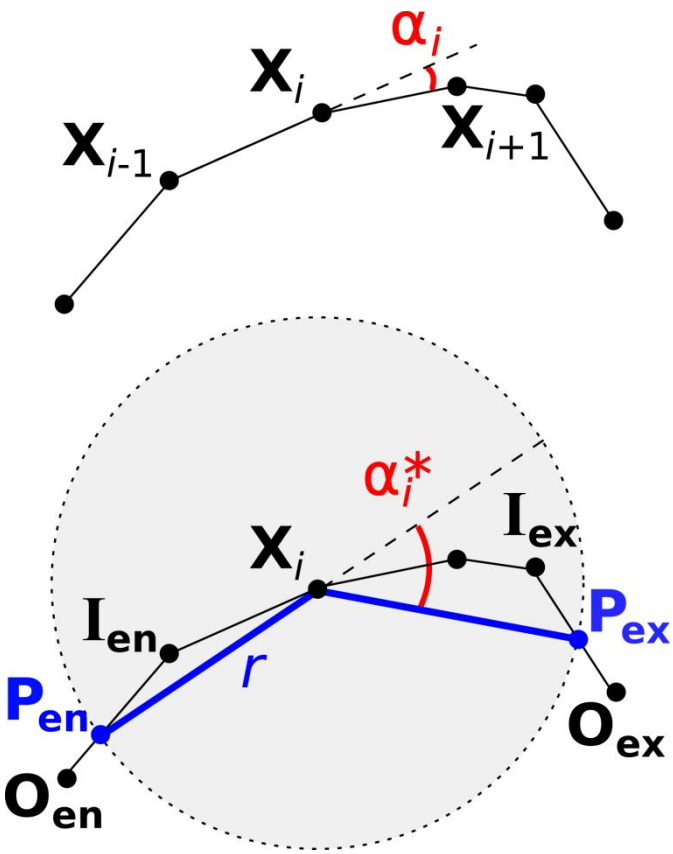
Frühwirth-Schnatter, S., 2006. Finite mixture and markov switching models. Springer Science & Business Media.

Lavielle, M., 2005. Using penalized contrasts for the change-point problem. Signal Processing 85:1501–1510. URL `http://www.sciencedirect.com/science/article/pii/S0165168405000381`.

Picard, F., S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin, 2005. A statistical approach for array CGH data analysis. BMC Bioinformatics 6:27. URL `https://doi.org/10.1186/1471-2105-6-27`.

Picard, F., S. Robin, E. Lebarbier, and J.-J. Daudin, 2007. A Segmentation/Clustering Model for the Analysis of Array CGH Data. Biometrics 63:758–766. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2006.00729.x`.

Supporting Information 2: Interpolating entrance and exit points of a circle

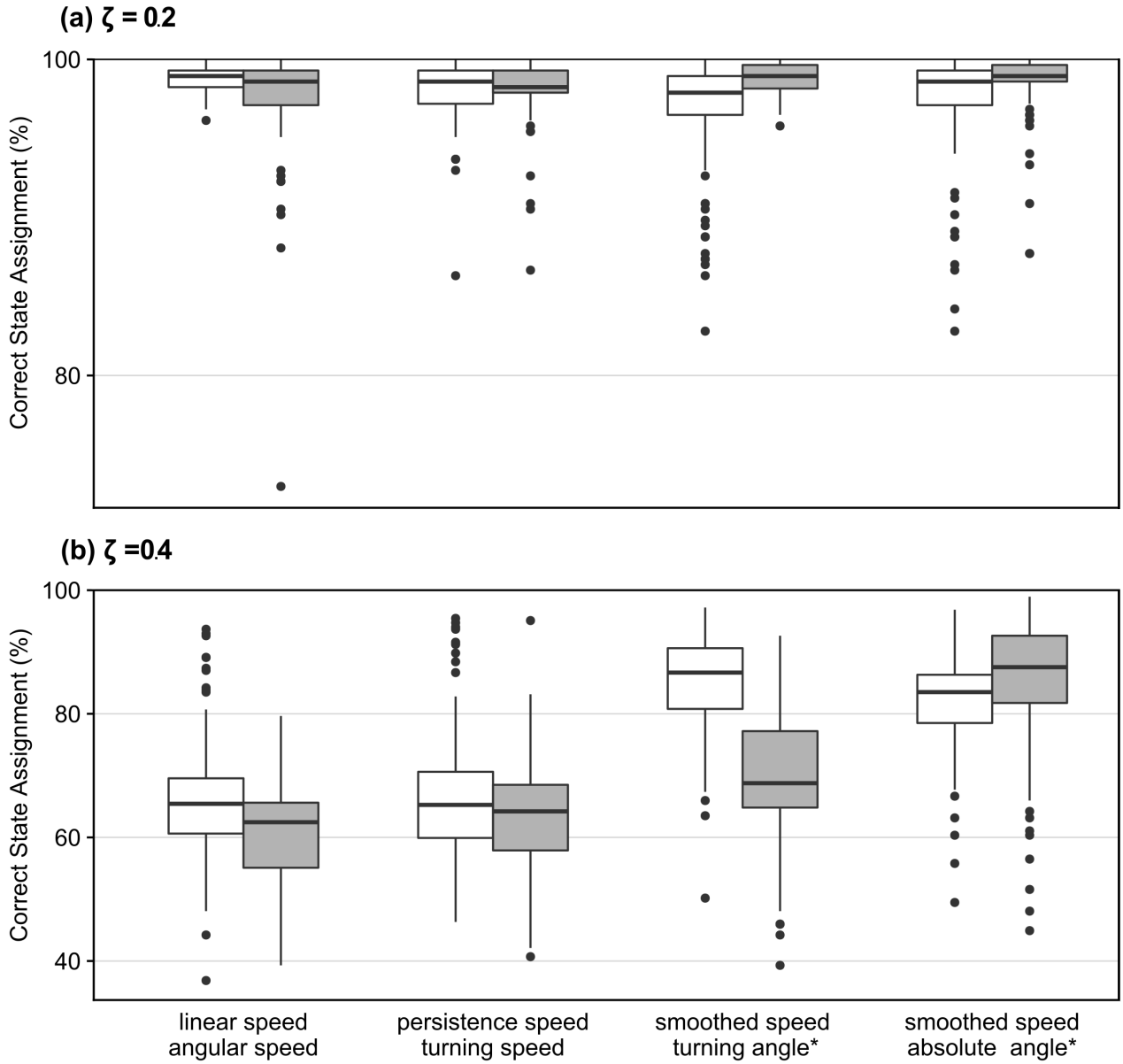


Given a series of locations $\mathbf{X}_i = (X_i, Y_i)$ recorded at constant time intervals Δt. Whereas the turning angle at constant time interval α_i (**top**) corresponds to the change in direction between vectors $\mathbf{X}_{i-1} \rightarrow \mathbf{X}_i$ (with length L_i) and $\mathbf{X}_i \rightarrow \mathbf{X}_{i+1}$ (with length L_{i+1}) and therefore acts as a proxy for angular speed ($\alpha_i/\Delta t$), the turning angle at constant length interval α_i^* (**bottom**) corresponds to the change in direction between vectors $\mathbf{P}_{en} \rightarrow \mathbf{X}_i$ and $\mathbf{X}_i \rightarrow \mathbf{P}_{ex}$, where $\mathbf{P}_{en}$ and $\mathbf{P}_{ex}$ are the last entrance and first exit locations, respectively, of a virtual circle with radius r centred on current location $\mathbf{X}_i$. Let $\mathbf{I} = (X_{in}, Y_{in})$ and $\mathbf{O} = (X_{out}, Y_{out})$ be the last inside and first outside recorded locations, respectively, of the first passage at the circle perimeter, either backwards ($\mathbf{I} = \mathbf{I}_{en}$ and $\mathbf{O} = \mathbf{O}_{en}$; $\mathbf{I}_{en} = \mathbf{X}_i$ if $L_i > r$) to determine $\mathbf{P}_{en}$, or forwards ($\mathbf{I} = \mathbf{I}_{ex}$ and $\mathbf{O} = \mathbf{O}_{ex}$; $\mathbf{I}_{ex} = \mathbf{X}_i$ if $L_{i+1} > r$) to determine $\mathbf{P}_{ex}$. The location $\mathbf{P}$ (either $\mathbf{P}_{en}$ or $\mathbf{P}_{ex}$) corresponds to the point where the vector $\mathbf{I} \rightarrow \mathbf{O}$ intersects the circle perimeter. The length of this vector is $d_{\mathbf{IO}} = (d_x^2 + d_y^2)^{0.5}$, with $d_x = X_{out} - X_{in}$ and $d_y = Y_{out} - Y_{in}$, and its orientation is θ, with $\cos(\theta) = d_x/d_{\mathbf{IO}}$ and $\sin(\theta) = d_y/d_{\mathbf{IO}}$. In a new orthonormal frame of reference (U, V) originating at $\mathbf{I}$ and with U axis running through $\mathbf{O}$, the coordinates of current location $\mathbf{X}_i$ become $U_i = (X_i - X_{in}) \cos(\theta) + (Y_i - Y_{in}) \sin(\theta)$ and $V_i = (Y_i - Y_{in}) \cos(\theta) - (X_i - X_{in}) \sin(\theta)$. By applying Pythagoras' theorem, one gets $r^2 = (d_{\mathbf{IP}} - U_i)^2 + V_i^2$, where $d_{\mathbf{IP}}$ corresponds to the distance between $\mathbf{I}$ and $\mathbf{P}$, with $d_{\mathbf{IP}} > U_i$. Entrance or exit location can therefore be linearly interpolated as $\mathbf{P} = \mathbf{I} + (\mathbf{O} - \mathbf{I})\, d_{\mathbf{IP}}/d_{\mathbf{IO}}$ (i.e. $X_P = X_{in} + \cos(\theta)\, d_{\mathbf{IP}}$ and $Y_P = Y_{in} + \sin(\theta)\, d_{\mathbf{IP}}$), with $d_{\mathbf{IP}} = U_i + (r^2 - V_i^2)^{0.5}$.

Supporting Information 3

Efficiency of segclust2d / segmentation-clustering for highlighting behavioural changes

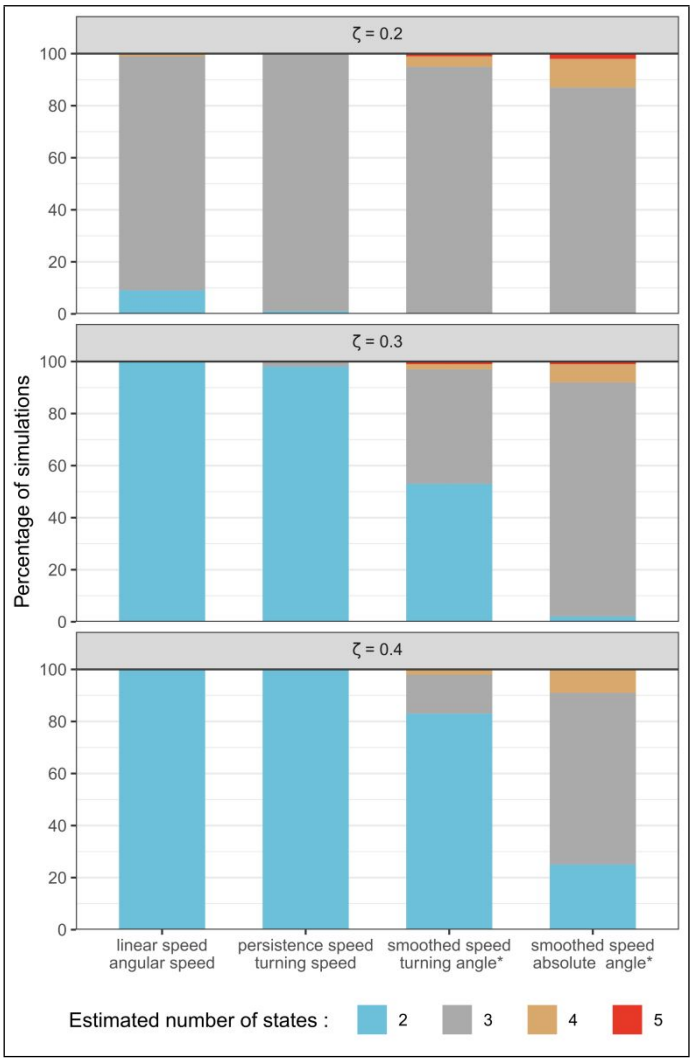
1. Comparison with HMM for low (ζ = 0.2) and high (ζ = 0.4) noise levels



The boxplots show the proportion of correct state assignments, obtained for various bivariate signals when the true number of states is known (M = 3) with noise level ζ = 0.2 u (a) or ζ = 0.4 u (b), as estimated from 100 replicates. The star (*) indicates turning angles computed with a constant step length, in terms of arithmetic (α_i*) or absolute ($|\alpha_i$*|) values. The white boxplots show the results obtained with HMM-based R package momentuHMM (McClintock & Michelot 2017), with informative initial state-dependent probability distribution parameters set to the true values of the various metrics in the different states (using the following distributions: Gaussian for persistence

and turning speeds, wrapped Cauchy for angular speed and turning angle α_i*, Weibull for linear speed, smooth speed and |turning angle*|). The grey boxplots shows the results obtained using the segclust2d/segmentation-clustering procedure with L_{min} = 10.

2. Estimation of the number of states



The various bars show the proportions of simulations resulting in a predicted number of states (i.e. behavioural modes) equal to 2, 3, 4, or 5, for the three noise levels considered (ζ = 0.2 u, ζ = 0.3 u and ζ = 0.4 u) and the four types of couples of metrics considered. The true number of states is 3. The star (*) indicates turning angles computed with a constant step length, in terms of arithmetic (α_i*) or absolute ($|\alpha_i|$*) values. The couple of metrics leading to best segmentation when the true number of states is known – absolute turning angle computed with a constant step length and smoothed speed – also leads to the best estimation of the number of states, but this latter estimation is not fully satisfactory, and should be worse with actual data because of possible mixing of movement behaviours.