A new selection criterion for statistical home range estimation*

A. Baíllo[a] and J. E. Chacón[b]

[a] Departamento de Matemáticas, Universidad Autónoma de Madrid (Spain)

[b] Departamento de Matemáticas, Universidad de Extremadura (Spain)

July 16, 2020

Abstract

The home range of an animal describes the geographic area where this individual spends most of the time while doing its usual activities. From a statistical viewpoint the problem of home range estimation can be considered as a set estimation one. In the ecological literature there are a variety of home range estimators. We address the open question of choosing the "best" home range from a collection of them constructed on the same sample. We introduce the penalized overestimation ratio, a numerical index to rank the estimated home ranges. The key idea is to balance the excess area covered by the estimator (with respect to the sample) and a shape descriptor measuring the over-adjustment of the home range to the data. To our knowledge, apart from computing the home range area, our ranking procedure is the first one both applicable to real data and to any type of home range estimator. Further, optimization of the selection index provides a way to select the tuning parameters of nonparametric home ranges. For illustration purposes, we apply our selection proposal to a dataset of a Mongolian wolf and we carry out a simulation study.

Keywords: nonparametric; penalization; set estimation; utilization distribution

1 Introduction

There has long existed an interest in identifying different characteristics (degrees of use, geographical limits, environmental descriptors...) of space use by whole animal species

*We acknowledge that we gave an oral presentation on the contents of this manuscript in the 13th International Conference on Computational and Financial Econometrics (CFE 2019) and 12th International Conference of the ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on Computational and Methodological Statistics (CMStatistics 2019). The title and abstract of the talk can be found in the Programme and Abstracts book available at `https://air.unimi.it/retrieve/handle/2434/691965/1353585/BoACFECMStatistics2019.pdf`

or individuals. Ecologists, for instance, are interested in estimating space use maps from animal tracking data for conservation planning. Delineation of the home range boundary is a popular and simple way of describing the space use of a monitored individual. Burt (1943) is credited with the first formalization of the idea of home range, as "that area traversed by the individual in its normal activities of food gathering, mating and caring for young".

Estimating a home range is based on a sample of locations from the individual of interest. Some decades ago the observations, quite spaced apart in time, were considered independent. Several home range estimators (see Section 2), derived from set estimation techniques and still in wide use today, were designed under this independence assumption. From the 1990s tracking data are a collection of two- or three- dimensional coordinates obtained from a GPS receiver attached to the animal. Due to technological developments, the GPS signals have an increasing time resolution. Analyzing or incorporating the effect of autocorrelation of the resulting observations into the home range estimator is currently a matter of great interest (see Kie 2010, Fleming *et al.* 2015, Cholaquidis *et al.* 2016, Noonan *et al.* 2018).

There are diverse statistical proposals in the literature to estimate the home range of a specific animal. For instance, individual space use can be described by the utilization distribution, the probability distribution of the animal locations over a period of time (van Winkle 1975). Many procedures aiming to obtain the home range usually define it as a level set, with high probability content (95% in general), of the utilization density, f. In this sense, home range estimation techniques are mostly based on set estimation approaches, either through direct geometrical procedures to estimate a set or through density estimation of the utilization distribution and posterior obtention of its level set. See Cuevas and Fraiman (2010) for a survey on general set estimation, Saavedra-Nieves *et al.* (2014) or Chen *et al.* (2017) for level set estimation.

The purpose of this work is to tackle the problem of, given such a wealth of possible home range estimators, choosing the "best" one. This is an important question that, to our knowledge, still remains open. Here we propose to construct a numerical index that balances overestimation and overfitting of the home range with respect to the observed locations. An advantage of our proposal is that it can be computed for any type of home range estimator, while previously existing selection procedures only work for specific types of home ranges (for example, those defined as the level set of a utilization density). The procedure is illustrated through both a simulation study and an analysis of a real data set.

The real sample of locations is the data set with ID 14291019 from Movebank (`www.movebank.org`), an animal movement database. It contains the locations of Zimzik, a Mongolian male wolf, tracked with GPS technology from March 2004 to September 2005 (Kaczensky *et al.* 2006, Kaczensky *et al.* 2008). Signal transmission took place between one and three times a day, at irregularly-spaced times. The trajectory, with 1455 observed locations (Figure 1), suggests a home range with interesting mathematical features (more than one connected component, non-convex components, ...). The map comprises part of

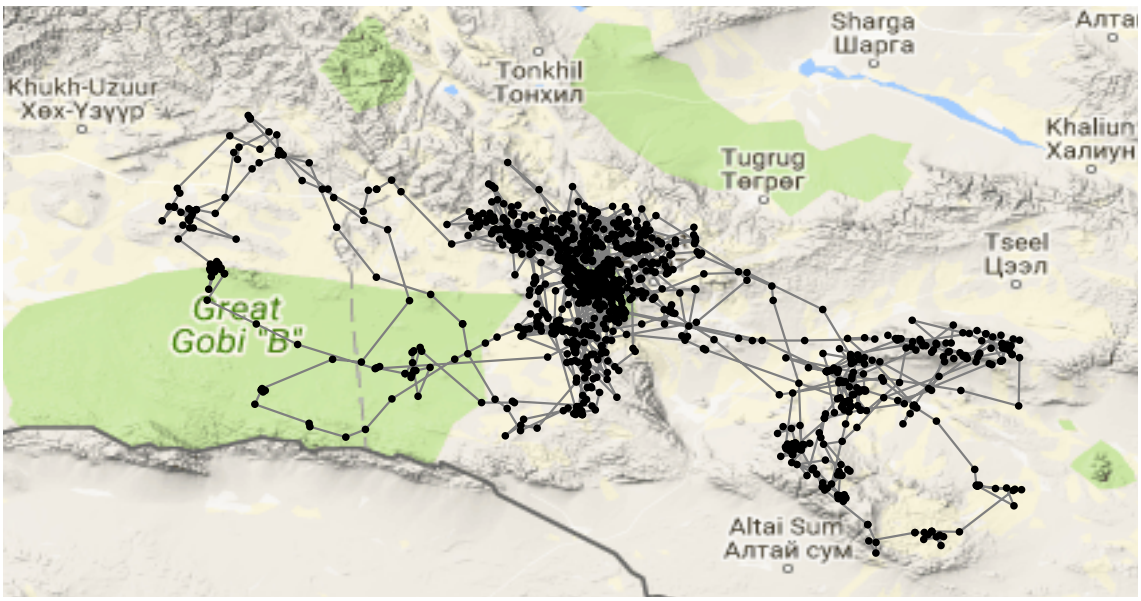


Figure 1: Locations and trajectories of the Mongolian wolf Zimzik on Google Maps.

Mongolia and China (the border is in solid grey). Figure 1 shows that Zimzik preferred mountainous terrains (slope $> 5°$), which hinder wolf hunting, over the flat steppe. Thus, in this case elevation contours contribute to shape of the home range.

In Section 2 we review, comment and illustrate some well-known home ranges. In Section 3 we summarize the existing home range selection procedures and introduce our own proposal, a numerical index based on penalization, which is then implemented on the home ranges constructed in Section 2. We also discuss several practical aspects of our selection index and how these considerations help to identify the most satisfactory home range among the available ones. In Section 4 we carry out simulations to compare the true home range of the model with that selected by our penalization procedure.

2 Some statistical home range estimators

In this section we introduce some home range estimators, popular in the ecological literature, to be compared later using our proposed selection technique (for a more complete survey of the existing home range proposals see Baíllo and Chacón 2020). We denote the animal locations by $\mathbf{x}_i$, $i = 1, \ldots, n$, and assume they are realizations of a random vector $\mathbf{X}$ in $\mathbb{R}^2$.

Minimum convex polygon (MCP) or convex hull

The convex hull of a sample of points is the minimum convex set enclosing them all. A variant of the MCP home range estimator is obtained by removing a proportion α (usually α=5%) of the points farthest from the data centroid. Obviously, the convexity restriction has serious drawbacks such as home range overestimation, as illustrated in Figure 2a, show-

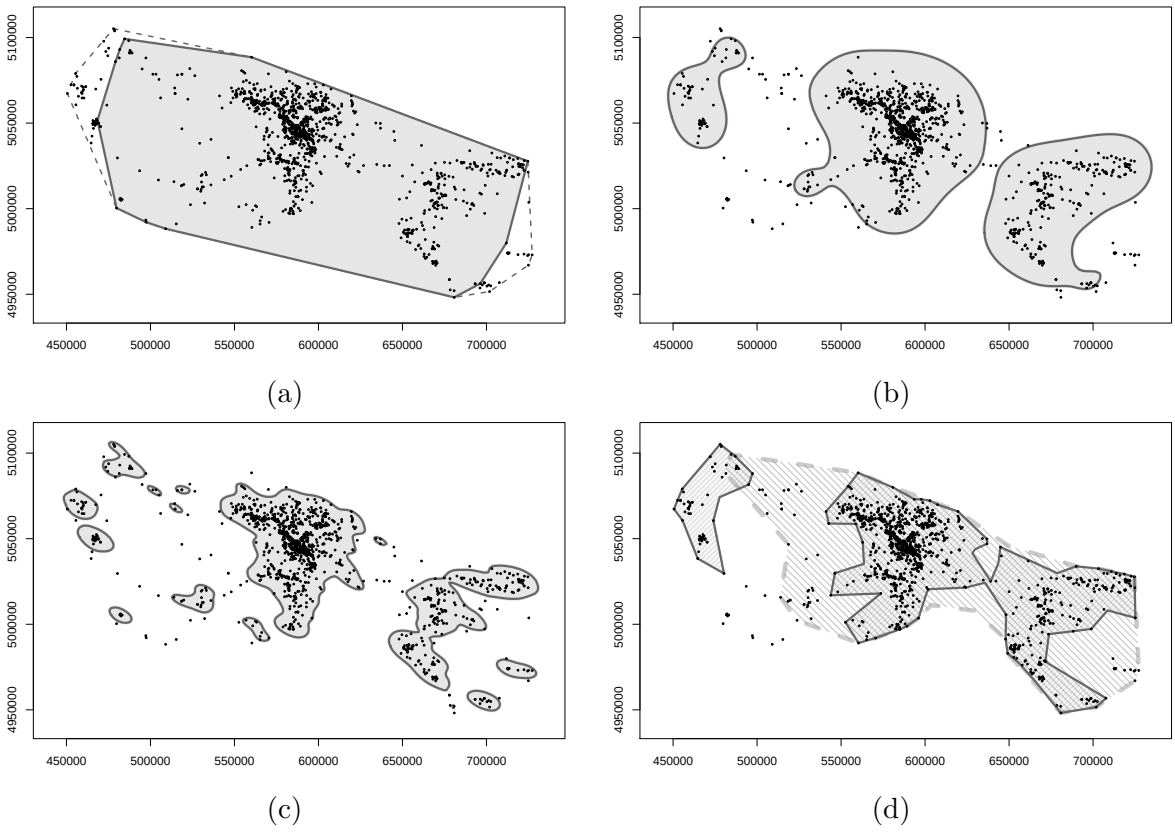


Figure 2: Home range estimates for Zimzik locations: (a) MCP containing 95% of sample points (in grey) and MCP of the whole sample (dashed line); (b) KDE with one-dimensional ad hoc bandwidth h; (c) KDE with unconstrained plug-in bandwidth matrix $\mathbf{H}$; (d) LoCoH isopleths with probability content 95% and $k = 35$ neighbours (darker continuous perimeter) and $k = 90$ (lighter dashed perimeter).

ing the MCP of Zimzik locations, computed using the R package `adehabitatHR` (Calenge 2006). This home range does not adapt to the mountainous territory to which Zimzik usually circunscribed its movements.

Kernel density estimation (KDE)

The home range of an animal is frequently defined as the level set $\{f \geq c\}$ of the utilization density f attaining a 95% probability content, that is, $0.95 = \int_{\{f \geq c\}} f$. Worton (1989) introduced kernel density estimators as a nonparametric approach to estimate f in home range procedures. The general expression of a kernel density estimator is

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i), \tag{1}$$

where $K : \mathbb{R}^2 \to [0, \infty)$ is the kernel function (a probability density in $\mathbb{R}^2$), $\mathbf{H} = (h_{ij})$ is a symmetric, positive-definite 2×2 bandwidth matrix, and the scaling notation $K_{\mathbf{H}}(\mathbf{x}) :=$

$|\mathbf{H}|^{-1/2}K(\mathbf{H}^{-1/2}\mathbf{x})$ has been used, with $\mathbf{H}^{-1/2}$ standing for the inverse of the matrix square root of $\mathbf{H}$ (see the recent monograph Chacón and Duong 2018). The kernel home range estimator is the level set $\{\hat{f} \geq \hat{c}\}$ of the kernel density estimator $\hat{f}$ in (1), with $\hat{c}$ chosen so that $0.95 = \int_{\{\hat{f}\geq\hat{c}\}} \hat{f}$.

It is well known that the choice of the kernel K has little effect on the accuracy of the estimator $\hat{f}$, compared to the effect of the bandwidth $\mathbf{H}$. Worton (1989) chose a constrained bandwidth matrix $h^2\mathbf{I}$ (with $\mathbf{I}$ the identity matrix) depending on a single smoothing parameter $h > 0$, proposed to be selected either via the "ad hoc" method (optimal for the Gaussian distribution) or via least-squares cross-validation.

The KDE home range for the wolf data is displayed in Figure 2b for the scalar ad hoc bandwidth $h = 12269.09$ (computed with `adehabitatHR`) and in Figure 2c for the unconstrained plug-in bandwidth matrix

$$\mathbf{H} = \begin{pmatrix} 23727441 & -9074807 \\ -9074807 & 10663700 \end{pmatrix} \tag{2}$$

of Chacón and Duong (2010), obtained with the package `ks` (Duong 2018). Bauder *et al.* (2015) observed that the use of this unconstrained plug-in bandwidth matrix outperforms the single smoothing parameter approach in some cases. Moreover, these home ranges based on estimating the utilization density clearly reveal that just trimming points furthest away from the centroid to obtain a region containing 95% of the data might not be a good idea, since those points may not necessarily correspond to locations within low density zones.

Local convex hull (LoCoH) or k-nearest neighbor convex hulls

This is a localized version of the MCP. For a fixed integer $k > 0$, Getz and Wilmers (2004) construct the convex hull of each sample point $\mathbf{x}_i$ and its $k-1$ nearest neighbors (NN) with respect to the Euclidean metric (although other metrics could be employed). Then these hulls are ordered according to their area, from smallest to largest. The LoCoH home range is the isopleth that results of progressively taking the union of the hulls from the smallest upwards, until a specific percentage (e.g., 95%) of sample points is included. In Figure 2d we display the LoCoH home range for Zimzik data with $k = 35$ and $k = 90$ neighbours, obtained with the R package `tlocoh` (Lyons *et al.* 2018). Clearly, this procedure is far more flexible than the convex hull, although it retains the simplicity of the latter. It is also evident that the choice of the parameter k affects the topological properties (connected components, holes, ...) of the resulting LoCoH home range. In this work we chose $k = 35$ due to the reasonable graphical similitude of the resulting LoCoH to the original tracking data. We also consider the value of $k = 90$ for illustration purposes.

3 Methodology: Ranking the home ranges

A key problem after computing several home ranges based on the same tracking data is how to select the most adequate one. In most case studies the ecologist chooses the soundest result according to previous information on the animal or its species. The possibility of making an automatised choice, based on objective mathematical measures, among a collection of home range estimates is still an open question.

As an application of set estimation, the performance of a home range estimator could be evaluated through the Hausdorff distance or the distance in measure between the estimator and the population counterpart (see Cuevas and Fraiman 2010). These are the usual measures to evaluate the quality of a set estimator and, as such, they would be the theoretically ideal criterion to optimise; however, they are unfeasible in practice since they depend on the unknown true HR.

3.1 Existing selection procedures

There have been some attempts to compare home ranges according to different criteria. Next we briefly mention the proposals valid for real data (methods only working for simulated samples are not considered). One possibility, especially if locations are treated as independent data, is to separate the sample into two subsamples of locations: a training sample to construct the home range estimator and a test sample to check its predictive accuracy (Kranstauber *et al.* 2012, Tarjan and Tinker 2016). Fleming *et al.* (2015) compare the areas of the home range estimates obtained with the complete sample and with the first half of the data. Long and Nelson (2015) compute the areas of different home range estimates. Kenward *et al.* (2001) study the relationship between the logarithm of the home range area and, for instance, environmental factors (e.g., food availability) known to influence the animal behaviour. Apart from the area, Steiniger and Hunter (2013) use shape complexity as given, for example, by holes and patches and the presence or not of corridors in the home range. Cumming and Cornélis (2012) and Walter *et al.* (2015) compare home range estimates using the area-under-the-curve corresponding to receiver operating characteristic (ROC) curves. However, a ROC curve can only be computed in the context of binary supervised classification, where we have a training sample, all of whose observations are correctly classified into their corresponding population. This is certainly not the case in home range estimation, where the sampled locations are not classified.

3.2 The penalized overestimation ratio

We propose a new way of choosing the "best fitting" home range among a collection of them, or at least a criterion for discarding the least satisfactory ones. The idea is to minimize the *penalized overestimation ratio*, a measure of the excess extension incurred by the home

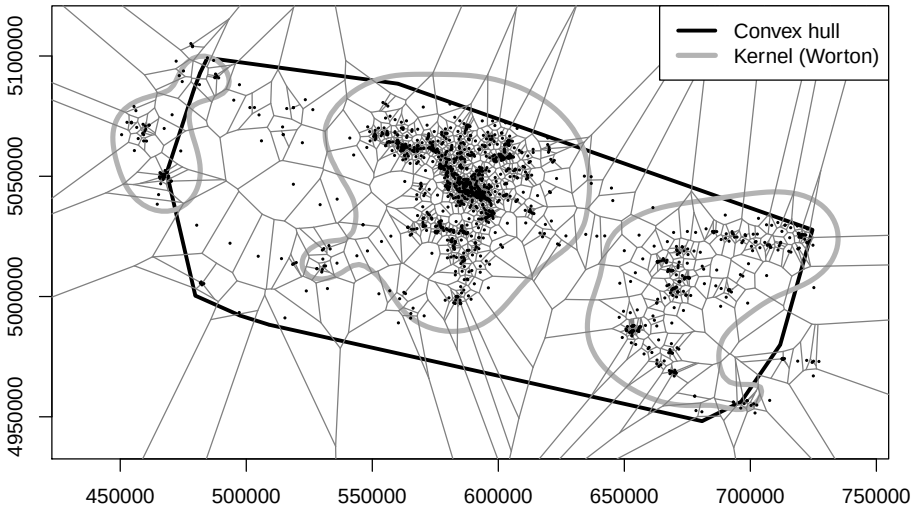


Figure 3: Voronoi tessellation of Zimzik relocations and two superimposed home ranges.

range as compared with the original locations, penalized by a measure of sample overfitting. One advantage of the procedure is that it works for any type of home range, regardless of the way it was constructed.

To quantify the excess area of a certain home range with respect to the observed sample, we propose to intersect the given home range with the Voronoi tessellation of the sample (see Figure 3). Observe that zones contained in the set estimator but never or seldom visited by the animal usually correspond to large (intersected) Voronoi cells. Then we sort the resulting (intersected) Voronoi cells according to their area (see Figure 4). We denote by $S_{(i)}$ the area of the i-th largest cell after intersection with the home range. We take the area of the largest cells as a proxy for the measure of the home range "hollowness". Specifically, we have considered either $S_{(1)}$, the area of the largest Voronoi cell, or $\sum_{i=1}^{10} S_{(i)}/10$, the mean area of the ten largest cells. Other more general approaches could be investigated, such as taking a fixed proportion of cells or a weighted average $\sum_{i=1}^{n} w_i S_{(i)}$, with $0 \leq w_i \leq 1$, $w_i \geq w_{i+1}$ for all i and $\sum_{i=1}^{n} w_i = 1$.

An advantage of this proposal with Voronoi cells is that it is instantly computed in R for any usual home range estimator. Another advantage is that the Euclidean distance in $\mathbb{R}^2$, used to construct the Voronoi diagrams in this work, can be extended to many other metrics, such as Euclidean distance in $\mathbb{R}^3$ or geographical distance if the altitude of the locations is taken into account, or a distance incorporating the time lapse between pairs of observations or extra information on locations that can affect the presence of the animal (such as human population density).

We computed $S_{(1)}$ and $\sum_{i=1}^{10} S_{(i)}/10$ for the home ranges of Figure 2: the convex hull (containing 95% of the sample), the LoCoH with $k = 35$ and $k = 90$, the kernel estimator

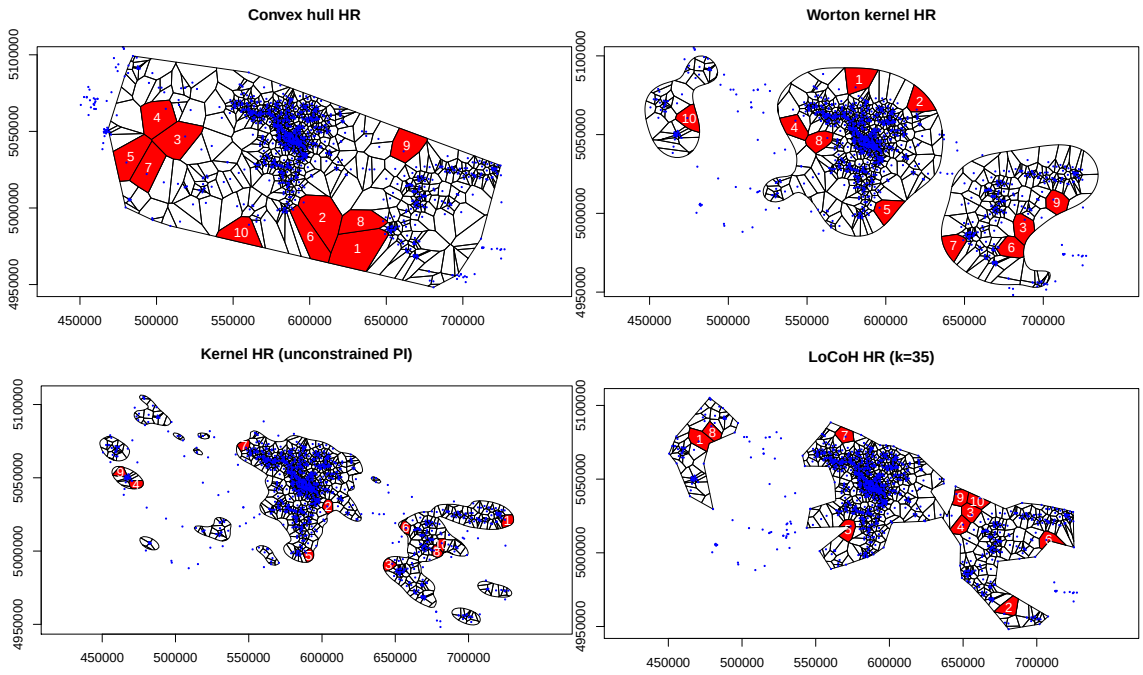


Figure 4: Intersection of Voronoi tessellation with some home ranges and highlight of the ten largest (intersected) cells sorted and numbered by area.

with the "ad hoc" smoothing parameter $h = 12269.09$ and the kernel estimator with the unconstrained plug-in bandwidth matrix (2). We also computed the area B of the smallest bounding box containing all these home ranges. In Table 1 we display the ratios $R^{(1)} = 100\, S_{(1)}/B$ and $R^{(10)} = 100 \sum_{i=1}^{10} S_{(i)}/10B$, respectively, to quantify the overestimation degree of the home range. As expected, in both cases the convex hull is the home range exhibiting the largest overestimation ratio R, followed by the group formed by the LoCoH with $k = 90$ and the kernel estimator with the "ad hoc" bandwidth h. The kernel estimator with the plug-in bandwidth always attains the lowest overestimation ratio, followed by the LoCoH with $k = 35$.

Home range	$R^{(1)}$	$R^{(10)}$	C	$1/C$
Convex hull	1.49	0.92	0.7090	1.4104
LoCoH ($k = 35$)	0.31	0.22	0.1510	6.6215
LoCoH ($k = 90$)	0.70	0.49	0.4738	2.1104
Kernel (Worton)	0.51	0.37	0.2548	3.9251
Kernel (plug-in)	0.11	0.09	0.0723	13.8232

Table 1: Ratios and circularities for some home ranges based on Zimzik data.

To measure how much a home range estimator overfits the tracking data, we compute

the circularity ratio $C = 4\pi\,\text{area}/\text{perimeter}^2$, a shape descriptor taking values in (0,1] (see González and Woods 2008, Ch. 11). In particular, C takes a value of 1 for a circular shape and is close to 0 for a shape full of spikes (such as an undersmoothed home range with deep intrusions). The two rightmost columns of Table 1 display the circularity and its inverse for the home ranges considered before. The results are not surprising: the two estimators more adapted to the sample (namely, LoCoH with $k = 35$ and the kernel estimator with plug-in bandwidth) are the ones with lowest circularity ratio (< 0.17).

	LoCoH		Kernel		LoCoH	Kernel
MCP	$k = 35$	$k = 90$	Worton	plug-in	$\min R_p$	$\min R_p$
1.583	**0.762**	0.843	<u>0.775</u>	1.064	0.728	0.681
1.015	0.678	**0.637**	<u>0.639</u>	1.049	0.614	0.621

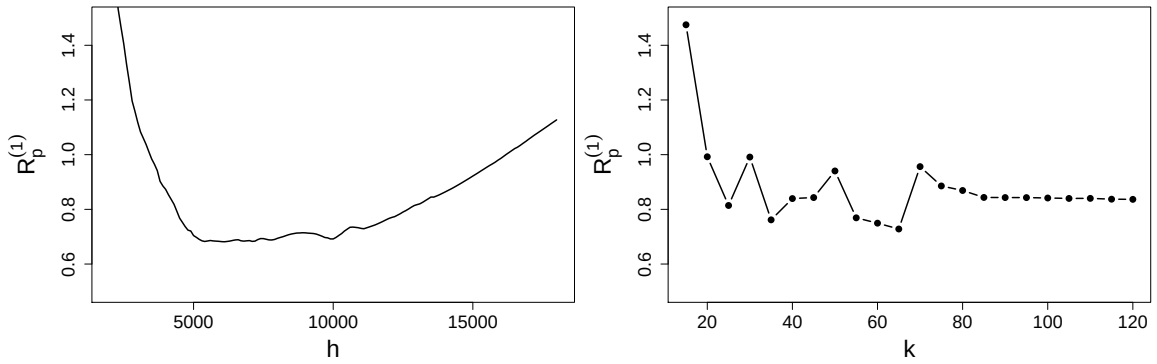
Table 2: Values of $R_p^{(1)}$ (upper line) and $R_p^{(10)}$ (lower line) for Zimzik data, with $\lambda = 0.0687$. For the five leftmost home ranges, the minimum R_p values appear in bold and the second lowest value is underlined.

The circularity C and the overestimation ratio $R = R^{(1)}$ or $R = R^{(10)}$ are combined into the *penalized overestimation ratio* $R_p := R + \lambda/C$, where $\lambda > 0$ is a penalization parameter regulating the trade-off between the home range size and its goodness of fit to the locations. A natural question is how to choose the tuning parameter λ. If the home range S is a rectangle (coinciding with the smallest bounding box containing it) and the sample of animal locations is uniformly distributed in S, then $\text{area}(S_{(i)}) \simeq B/n$, for all $i = 1, \ldots, n$, and $R_p \simeq 100/n + \lambda/C$. For the convex hull or a very circular home range, we have values of C relatively close to 1 and, thus, $R_p \simeq 100/n + \lambda$. On the basis of this reference scenario, not to overbalance any of the two terms in R_p, we suggest setting $\lambda = 100/n$.

For Zimzik data, we reproduce in columns 1–5 of Table 2 the values of $R_p^{(1)} = 100\, S_{(1)}/B + \lambda/C$ and $R_p^{(10)} = 100 \sum_{i=1}^{10} S_{(i)}/(10B) + \lambda/C$, with $\lambda = 100/n = 0.0687$, for all the home ranges considered in Table 1. Both R_p criteria lead to similar conclusions: first, the convex hull does not seem an adequate home range estimator for this data set, since its R_p scores are substantially higher than for the best of the alternatives; second, the LoCoH appears to yield the best results for this sample, with the subjective choice of $k = 35$ being preferred regarding $R_p^{(1)}$ and $k = 90$ being a better option if the $R_p^{(10)}$ criterion is employed; and third, the kernel estimator with the ad hoc bandwidth comes second in the ranking, not far from the LoCoH.

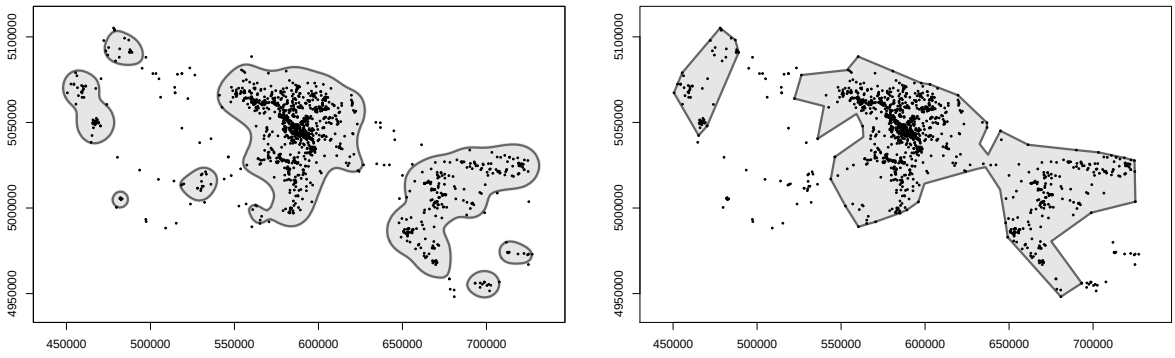
3.3 Tuning the parameters

The penalized overestimation ratio R_p can also be used as a tool for choosing the parameter of a family of home range estimators, by selecting the parameter value that minimises R_p.



(a) (b)

Figure 5: $R_p^{(1)}$ as a function of (a) the bandwidth h for the kernel home range and (b) the number of neighbours k for the LoCoH.



(a) (b)

Figure 6: (a) Kernel home range using the automatic data-driven bandwidth $h_{R1} = 6039$; (b) LoCoH using automatic data-driven value of $k_{R1} = 65$.

As an illustration, Figure 5 shows $R_p^{(1)}$ (a) as a function of the bandwidth h for the kernel home range and (b) as a function of the number of neighbours k for the LoCoH, for $\lambda = 0.0687$. The bandwidth minimising $R_p^{(1)}$ for the kernel home range is $h_{R1} = 6039$, and the resulting $R_p^{(1)}$ value is 0.681. This confirms that the ad-hoc bandwidth $h = 12269.09$ severely oversmooths, as suggested by Figure 2(b). The corresponding home range with $h = h_{R1}$, shown in Figure 6(a), clearly displays a compromise by presenting a small overestimation area while, at the same time, avoiding sharp inlets into the location data shape. It is also possible to minimise the $R_p^{(1)}$ criterion over the class of positive-definite matrices $\mathbf{H}$, leading to a slightly smaller $R_p^{(1)}$ value of 0.678, but the final kernel home range is very similar to the previous one, so its plot is omitted. Analogously, the bandwidth minimising the $R_p^{(10)}$ criterion is $h_{R10} = 7755$, with an $R_p^{(10)}$ value of 0.621.

Similarly, the penalized overestimation ratio provides a non-subjective number of neigh-

bours k for the LoCoH. The minimal $R_p^{(1)}$ value of 0.728 is attained with $k_{R1} = 65$ (see the corresponding LoCoH in Figure 6(b)). The minimal $R_p^{(10)}$ of 0.614 is reached at $k_{R10} = 65$.

To facilitate the comparison with the other studied methods, the R_p values for these optimal home ranges are included in the two rightmost columns of Table 2. As expected, in terms of the R_p criteria they improve over the previous home ranges.

4 Application: A simulation study

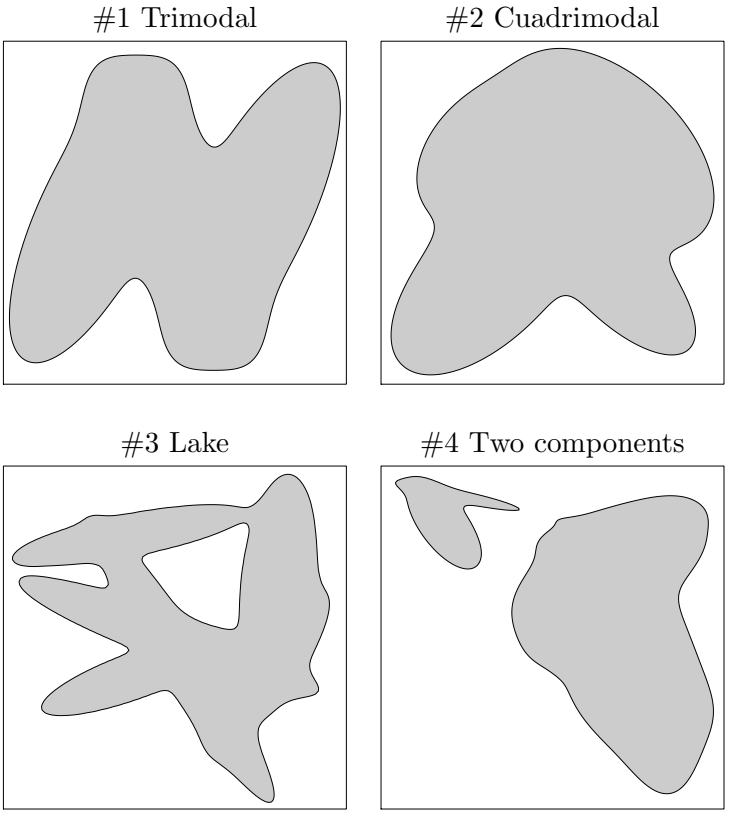
As a second illustration of the performance of our home range selection criterion, we report the results of a simulation study under different models. For a fixed known model with utilization density f, the true home range is the level set $\{f \geq c\}$ with 95% probability content. Assuming locations in the sample are independent, we use four models of normal mixtures giving rise to home ranges with different geometrical and topological features (see Table 3 and Figure 7). Models #1 (Trimodal) and #2 (Quadrimodal) are models (J) and (L) in Wand and Jones (1993). Models #3 (Lake) and #4 (Two connected components) were generated following the second general model in Seaman and Powell (1996). We have also generated locations from a fifth model with autocorrelation, namely, an Ornstein-Uhlenbeck model, which has been extensively used in the context of home range analysis (see, e.g., Dunn and Gipson 1977, Hooten *et al.* 2017). This process is defined as the solution of the bivariate stochastic differential equation

$$d\mathbf{X}(t) = \boldsymbol{\theta}(\boldsymbol{\mu} - \mathbf{X}(t))dt + \boldsymbol{\sigma} d\mathbf{W}(t), \tag{3}$$

where $\boldsymbol{\theta}$ is an 2×2 invertible real matrix, $\boldsymbol{\mu}$ is an bidimensional real vector, $\boldsymbol{\sigma}$ is a 2×2 positive definite real matrix and $\mathbf{W}$ is a 2-dimensional standard Wiener process (see, e.g., Vatiwutipong and Phewchean 2019). In our work we have chosen $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\sigma} = \mathbf{I}$, the identity matrix, $\boldsymbol{\theta} = 0.5\mathbf{I}$ and $\mathbf{X}(0)$, the initial value, a random vector following a standard bivariate normal distribution. Under this specific choice of the model parameters there exists a stationary solution to (3): $\mathbf{X}(t)$ follows a bivariate standard normal distribution for all t. Thus, this is the utilization distribution, from which the population home range (the circle centred in the origin and with radius $(\log 0.05^{-2})^{1/2}$) can be explicitly derived.

For each model we performed 500 Monte Carlo runs. In each of them we generated $n = 1000$ observations and used them to construct the home ranges derived from the convex hull and the kernel density estimator with the ad-hoc and plug-in bandwidths and the LoCoH with $k = 10$ and $k = 100$ neighbours. Other sample sizes were also tested, yielding similar results; thus, we only show the details for $n = 1000$ to save space. In each simulation run, we also took $\lambda = 100/n = 0.1$ and computed the penalized overestimation ratio $R_p^{(1)}$ for all these home ranges. We determined the bandwidth h_{R1} and the number of neighbours k_{R1} minimizing $R_p^{(1)}$ for the KDE and the LoCoH home ranges, respectively. This makes a total of seven home range estimates (denoted generically by HR). For each

Figure 7: True home ranges for the four normal mixture simulation models.



Density	$w_1 N(\boldsymbol{\mu}_1^T; \text{vech}\, \boldsymbol{\Sigma}_1) + \ldots + w_m N(\boldsymbol{\mu}_m^T; \text{vech}\, \boldsymbol{\Sigma}_m)$
#1 Trimodal	$\frac{1}{3}N\left((-\frac{6}{5}, 0)\,; \frac{1}{25}\left(9, \frac{63}{10}, 9\right)\right) + \frac{1}{3}N\left(\left(\frac{6}{5}, 0\right)\,; \frac{1}{25}\left(9, \frac{63}{10}, 9\right)\right)$ $+\frac{1}{3}N\left((0,0); \frac{1}{25}\left(9, -\frac{63}{10}, 9\right)\right)$
#2 Quadrimodal	$\frac{1}{8}N\left((-1,1); \frac{4}{9}\left(1, \frac{2}{5}, 1\right)\right) + \frac{3}{8}N\left((-1,-1); 4\left(\frac{1}{9}, \frac{1}{15}, \frac{1}{9}\right)\right)$ $+\frac{1}{8}N\left((1,-1); \frac{4}{9}\left(1, -\frac{7}{10}, 1\right)\right) + \frac{3}{8}N\left((1,1); \frac{2}{9}(2,-1,2)\right)$
#3 Lake	$0.19N((5.72, 2.27); (14.62, -21.26, 34.51)$
	$+0.15N((0.97, 13.77); (19.76, 3.62, 1.34))$
	$+0.14N((1.48, 1.09); (13.05, 4.44, 2.38))$
	$+0.22N((4.41, 4.63); (33.63, -12.02, 5.31))$
	$+0.24N((14.32, 9.43); (2.19, 1.62, 17.49)$
	$+0.06N((13.70, 1.54); (14.81, 16.51, 24.73)$
#4 Two connected components	$0.15N((13.16, 2.94); (4.01, -2.00, 12.26))$
	$+0.18N((11.55, 9.06); (8.92, 5.41, 7.85))$
	$+0.03N((10.90, 12.32); (13.39, 2.03, 1.65))$
	$+0.08N((10.18, 3.12); (0.75, -0.12, 0.25))$
	$+0.05N((0.18, 12.97); (1.34, -1.11, 2.17))$
	$+0.06N((13.44, 0.70); (13.53, -2.42, 5.71))$
	$+0.01N((1.41, 14.84); (13.29, -2.92, 0.81))$
	$+0.09N((9.51, 8.66); (4.65, 1.34, 2.44))$
	$+0.09N((12.66, 8.81); (3.02, 1.56, 1.39))$
	$+0.26N((14.17, 1.68); (1.47, -1.23, 8.95))$

Table 3: Parameters for the four normal mixture simulation models.

of them we recorded the 500 Monte Carlo values of $R_p^{(1)}$, $R_p^{(10)}$ and the area (denoted by $|\quad|$) of the symmetric difference between the true and the estimated home ranges, $|\{f \geq c\} \Delta\, \text{HR}| = |\{f \geq c\} \cap \text{HR}^c| + |\{f \geq c\}^c \cap \text{HR}|$, as a measure of the error attained in the estimation of the real home range. One of the aims of this simulation study is to check to which extent the penalized overestimation ratio identifies the home range with the lowest error.

The results for the first four models (independent observations) are summarized in Figure 8, where each row corresponds to a different model. Each graphic in Figure 8 displays the boxplots for the seven home ranges previously described. The variables represented in the boxplots are, from left to right column of the figure, $|\{f \geq c\} \Delta\, \text{HR}|$, $R_p^{(1)}$ and $R_p^{(10)}$.

We observe that, for the four parametric models considered, the penalized overestimation ratio $R_p^{(1)}$, a strictly nonparametric index that only uses the largest Voronoi cell, reflects well the behaviour of the error attained by the estimated home range as measured by $|\{f \geq c\} \Delta\, \text{HR}|$. Home ranges attaining large errors (such as the convex hull) also attaing

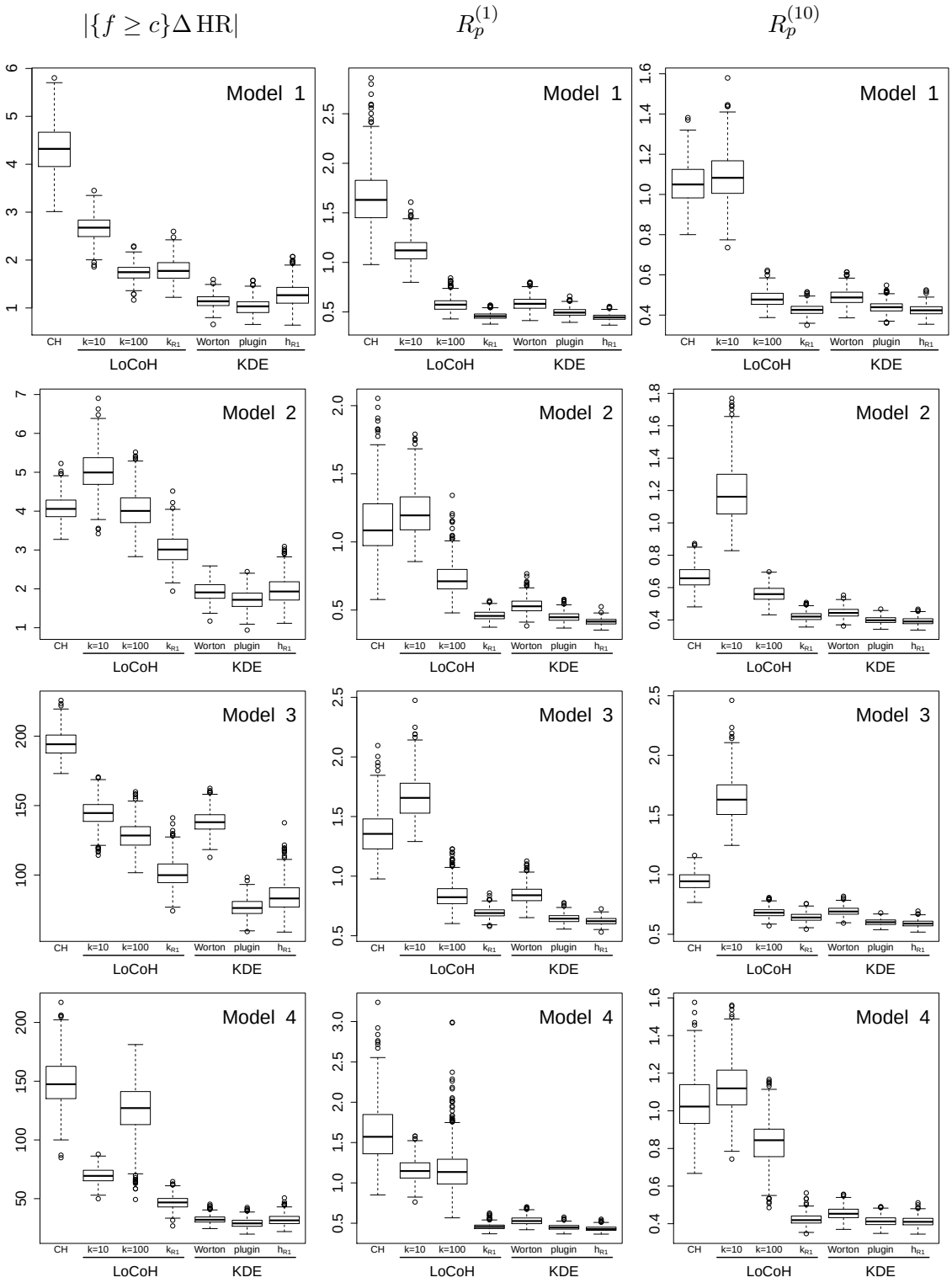


Figure 8: Simulation results for the four normal mixture models (independent locations).

the largest values of $R_p^{(1)}$, while those with small $|\{f \geq c\}\Delta\,\mathrm{HR}|$ (such as the KDE with plug-in bandwidth or with the h_{R1}) correspond to the lowest values of $R_p^{(1)}$. The overestimation ratio $R_p^{(10)}$, that averages over the 10 largest Voronoi cells, does not exhibit such a clear relationship with $|\{f \geq c\}\Delta\,\mathrm{HR}|$. Thus, we recommend the use of $R_p^{(1)}$ over $R_p^{(10)}$.

It is interesting to note that, for all four models, the KDE with the plug-in bandwidth or with h_{R1} leads to low errors in the estimation of the true home range, so, as a general recommendation to practitioners, these two home ranges would be a safe choice if observation times are not taken into account in the estimation. Observe also that the LoCoH with $k = k_{R1}$ always attains the lowest value of the error $|\{f \geq c\}\Delta\,\mathrm{HR}|$ among the three LoCoH estimators. Thus, our proposed $R_p^{(1)}$ provides also a fully automatic way of choosing a good value of the tuning parameter k in LoCoH. Since $R_p^{(1)}$ appears to be directly related to the performance of the home range estimator, it is fully nonparametric, and its computation time in R is negligible, it follows that the criterion $R_p^{(1)}$ provides a useful tool in practice to compare several home range estimates (of any kind) and to make an appropriate choice amongst them.

The results for the Ornstein-Uhlenbeck model are summarized in Figure 9. To interpret these results we have to take into account that the population home range is a level set of the multivariate standard normal, i.e., a circle. This makes the convex hull of the 5% trimmed sample perform extremely well in this case, although the error $|\{f \geq c\}\Delta\,\mathrm{HR}|$ still exhibits greater variability than in the Worton and plug-in kernel home ranges. The convexity of the population home range also has the consequence that the higher the number k of neighbours in the LoCoH the better it performs: the LoCoH with $k = 10$ is full of holes, while the LoCoH with $k = 100$ is generally very similar to the convex hull, although not necessarily convex. However, the optimal k in terms of the penalized overestimation ratio is generally low: it has a median of 40 and a third quartile of 205. This is justified by the convexity of the level set and the high concentration of points near the origin, as a consequence of which the largest (external) Voronoi cells are systematically eliminated in the construction of the LoCoH, and the penalized overestimation ratio is very stable for those k's producing LoCoH's without holes. The kernel home ranges with the Worton and plug-in bandwidths are systematically coincident in all the 500 samples and their error is the lowest among all home ranges considered. In this case, the bandwidth minimizing the penalized overestimation ratio is not able to perform as well as the other two, although we suspect that this problem is not due to the autocorrelation of the sample, but to the fact that the Worton and plug-in bandwidths are both optimal for the population target of this model (a level set of a normal distribution).

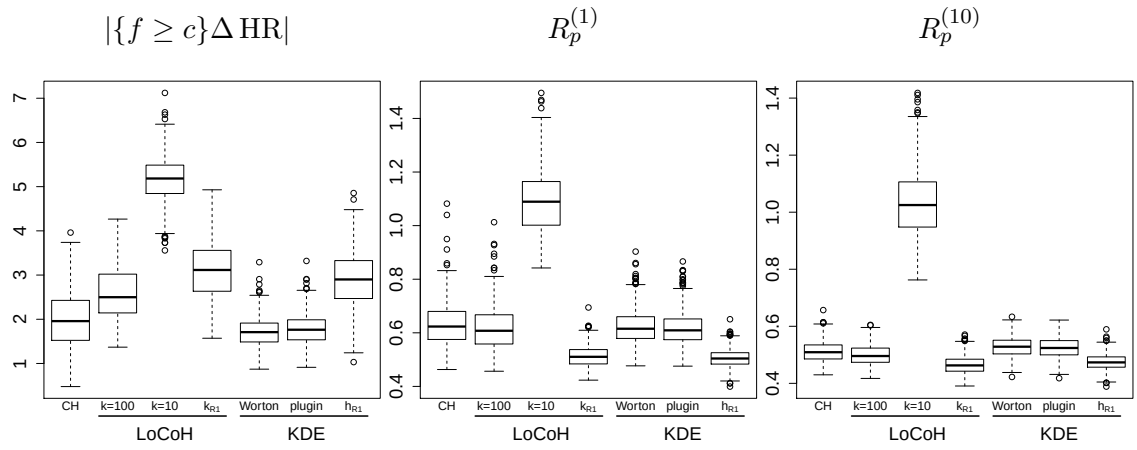


Figure 9: Simulation results for the Ornstein-Uhlenbeck model.

5 Conclusion and Discussion

Many different methods are nowadays available to estimate the home range of an animal. The problem can be formally cast as the estimation of the essential support of the utilization distribution, so it can be directly addressed via existing set estimation techniques. However, only few and rather limited proposals can be found for the task of comparing the output of different home range techniques. Here, a new and effective procedure is introduced to select the optimal home range among a collection of them based on the same set of locations. The selection is performed via an index measuring the excess extension of the home range with respect to the tracking data, penalized by the shape circularity to prevent over-fitting to the sample.

It must be stressed that this novel approach is applicable to any kind of home range estimator, no matter its nature and construction and, hence, it appears most useful to compare the performance of very diverse competing methodologies. The results of our analysis of a real data set and several simulation scenarios suggest that the new penalized criterion is also a valuable tool for selecting the tuning parameters involved in the existing home range estimation processes.

In any case, even if we provide a first working solution to this relatively unexplored selection issue, some open problems still persist. While a practical choice of the penalization parameter λ is offered above, perhaps more elaborated options might lead to even better performance. Moreover, our setting is closer to the classical, original context in which data locations are considered independent. Extending this setup for autocorrelated data constitutes an interesting avenue of future research.

Acknowledgements

Research by the authors was supported by the Spanish MEyC grants MTM2013-44045-P and MTM2016-78751-P. The research on the Mongolian wolves was conducted by Kaczensky *et al.* (2006, 2008) within the framework of the Przewalskii's horse reintroduction project of the International Takhi Group (ITG), in cooperation with the Mongolian Ministry of Nature and Environment, the National University in Ulaanbaatar, Mongolia and the Great Gobi B Strictly Protected Area Administration.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Baíllo, A. and Chacón, J.E. (2020). Statistical outline of animal home ranges, an application of set estimation. Under review for *Handbook of Statistics, Vol. 44: Data Science: Theory and Applications*. Elsevier.

Bauder, J.M., Breininger, D.R., Bolt, M.R., Legare, M.L., Jenkins, C.L. and McGarigal, K. (2015). The role of the bandwidth matrix in influencing kernel home range estimates for snakes using VHF telemetry data. *Wildlife Research*, 42, 437–453.

Burt, W.H. (1943). Territoriality and home range concepts as applied to mammals. *Journal of Mammalogy*, 24, 346–352.

Calenge, C. (2006). The package "adehabitat" for the R software: A tool for the analysis of space and habitat use by animals. *Ecological Modelling*, 197, 516–519.

Chacón, J.E. and Duong, T. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, 19, 375–398.

Chacón, J.E. and Duong, T. (2018). *Multivariate Kernel Smoothing and Its Applications*. Chapman & Hall.

Chen, Y.-C., Genovese, C.R. and Wasserman, L. (2017). Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, 112, 1684–1696.

Cholaquidis, A., Fraiman, R., Lugosi, G. and Pateiro-López, B. (2016). Set estimation from reflected Brownian motion. *Journal of the Royal Statistical Society: Series B*. 78(5), 1057–1078.

Cuevas, A. and Fraiman, R. (2010). Set estimation. In *New Perspectives on Stochastic Geometry*, W.S. Kendall and I. Molchanov, eds., pp. 374–397. Oxford University Press.

Cumming, G.S. and Cornélis, D. (2012). Quantitative comparison and selection of home range metrics for telemetry data. *Diversity and Distributions*, 18, 1057–1065.

Dunn, J.E. and Gipson, P.S. (1977). Analysis of radio telemetry data in studies of home range. *Biometrics*, 33, 85–101.

Duong, T. (2018). ks: Kernel Smoothing. R package version 1.11.0. https://CRAN.R-project.org/package=ks

Fleming, C.H., Fagan, W.F., Mueller, T., Olson, K.A., Leimgruber, P. and Calabrese, J.M. (2015). Rigorous home range estimation with movement data: a new autocorrelated kernel density estimator. *Ecology*, 96, 1182–1188.

Getz, W.M. and Wilmers, C.C. (2004). A local nearest-neighbor convex-hull construction of home ranges and utilization distributions. *Ecography*, 27, 489–505.

González, R.C. and Woods, R.E. (2008). *Digital Image Processing*. Third edition. Pearson International Edition.

Hooten, M.B., Johnson, D.S., McClintock, B.T. and Morales, J.M. (2017). Animal Movement: Statistical Models for Telemetry Data. CRC Press.

Kaczensky, P., Ganbaatar, O., Enksaikhaan, N. and Walzer, C. (2006). Wolves in Great Gobi B SPA GPS tracking study 2003-2005 dataset. Movebank Data Repository.

Kaczensky, P., Enkhsaikhan, N., Ganbaatar, O. and Walzer, C. (2008). The Great Gobi B Strictly Protected Area in Mongolia - refuge or sink for wolves Canis lupus in the Gobi. *Wildlife Biology*, 14, 444–456.

Kenward, R.E., Clarke, R.T., Hodder, K.H., and Walls, S.S. (2001). Density and linkage estimators of home range: nearest-neighbor clustering defines multinuclear cores. *Ecology*, 82, 1905–1920.

Kie, J.G., Matthiopoulos, J., Fieberg, J., Powell, R.A., Cagnacci, F., Mitchell, M.S., Gaillard, J.-M. and Moorcroft, P.R. (2010). The home-range concept: are traditional estimators still relevant with modern telemetry technology?. *Philosophical Transactions of the Royal Society B*, 365, 2221–2231.

Kranstauber, B., Kays, R., LaPoint, S.D., Wikelski, M. and Safi, K. (2012). A dynamic Brownian bridge movement model to estimate utilization distributions for heterogeneous animal movement. *Journal of Animal Ecology*, 81, 738–746.

Long, J. and Nelson, T. (2015). Home range and habitat analysis using dynamic time geography. *The Journal of Wildlife Management*, 79, 481–490.

Lyons, A., Getz, W. and the R Development Core Team (2018). T-LoCoH: Time Local Convex Hull Homerange and Time Use Analysis. R package version 1.40.05.

Noonan, M.J., Tucker, M.A., Fleming, C.H., Alberts, S.C., Ali, A.H., Altmann, J., Antunes, P.C., Belant, J.L., Berens, D., Beyer, D., Blaum, N., Böhning-Gaese, K., LauryCullen, Jr. , de Paula, R.C., Dekker, J., Farwig, N., Fichtel, C., Fischer, C., Ford, A., Goheen, J.R., Janssen, R., Jeltsch, F., Kappeler, P., Koch, F., LaPoint, S., Markham, A.C., Medici, E.P., Morato, R.G., Nathan, R., Oliveira-Santos, L.G.R., Patterson, B.D., Paviolo, A., Ramalho, E.E., Roesner, S., Selva, N., Sergiel, A., Silva, M.X., Spiegel, O.,

Ullmann, W., Zieba, F., Zwijacz-Kozica, T., Fagan, W.F., Mueller, T., Calabrese, J.M. (2019). A comprehensive analysis of autocorrelation and bias in home range estimation. *Ecological Monographs*, 89 (2), e01344.

Saavedra-Nieves, P. González-Manteiga, W. and Rodríguez-Casal, A. (2014). Level set estimation. In *Topics in Nonparametric Statistics*, M.G. Akritas *et al.*, eds., pp. 299–307. Springer Science+Business Media, New York.

Seaman, D.E. and Powell, R.A. (1996). An evaluation of the accuracy of kernel density estimators for home range analysis. *Ecology*, 77 (7), 2075–2085.

Steiniger, S. and Hunter, A.J.S. (2013). A scaled line-based density estimator for the retrieval of utilization distributions and home ranges from GPS movement tracks. *Ecological Informatics*, 13, 1–8.

Tarjan, L.M. and Tinker, M.T. (2016). Permissible home range estimation (PHRE) in restricted habitats: a new algorithm and an evaluation for sea otters. *PLoS One*, 11(3), e0150547.

Vatiwutipong, P. and Phewchean, N. (2019). Alternative way to derive the distribution of the multivariate Ornstein–Uhlenbeck process. *Advances in Difference Equations*: 276.

Walter, W.D., Onorato, D.P. and Fischer, J.W. (2015). Is there a single best estimator? Selection of home range estimators using area-under-the-curve. *Movement Ecology*, 3:10.

Wand, M.P. and Jones, M.C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, 88, 520–528.

van Winkle, W. (1975). Comparison of several probabilistic home-range models. *The Journal of Wildlife Management*, 39, 118–123.

Worton, B.J. (1989). Kernel methods for estimating the utilization distribution in home-range studies. *Ecology*, 70, 164–168.