# Overview of the **overlap** package

Mike Meredith and Martin Ridout

May 3, 2018

## 1   Introduction

Camera traps – cameras linked to detectors so that they fire when an animal is present – are a major source of information on the abundance and habitat preferences of rare or shy forest animals. Modern cameras record the time of the photo, and the use of this to investigate diel [1] activity patterns was immediately recognised (Griffiths and van Schaik, 1993).

Initially this resulted in broad classification of taxa as diurnal, nocturnal, crepuscular, or cathemeral (van Schaik and Griffiths, 1996). More recently, researchers have compared activity patterns among species to see how overlapping patterns may relate to competition or predation (Linkie and Ridout, 2011; Carver *et al.*, 2011; Ramesh *et al.*, 2012; Carter *et al.*, 2012; Kamler *et al.*, 2012; Ross *et al.*, 2013; Azevedo *et al.*, 2018).

Ridout and Linkie (2009) presented methods to fit kernel density functions to times of observations of animals and to estimate the coefficient of overlapping, a quantitative measure ranging from 0 (no overlap) to 1 (identical activity patterns). The code they used forms the basis of the **overlap** package.

Although motivated by the analysis of camera trap data, **overlap** could be applied to data from other sources such as data loggers, provided data collection is carried out around the clock. Nor is it limited to diel cycles: tidal cycles or seasonal cycles, such as plant flowering or fruiting or animal breeding seasons could also be investigated.

## 2   Kernel density curves

### 2.1   Example data set

To demonstrate the use of the software we will use camera-trapping data from Kerinci-Seblat National Park in Sumatra, Indonesia (Ridout and Linkie, 2009).

```
> library(overlap)
> data(kerinci)
> head(kerinci)

  Zone   Sps  Time
1    1 tiger 0.175
2    1 tiger 0.787
3    1 tiger 0.247
4    1 tiger 0.591
5    1 tiger 0.500
6    1 tiger 0.564

> table(kerinci$Zone)
```

---

[1] We use "diel" for 24-hour cycles, and reserve "diurnal" to mean "not nocturnal".

1

```
   1    2    3    4
 104  425  280  289

> summary(kerinci$Sps)

   boar clouded  golden macaque muntjac  sambar   tapir   tiger
     28      86     104     273     200      25     181     201

> range(kerinci$Time)

[1] 0.003 0.990
```

The data provide time-of-capture data from 4 Zones within the Park for 8 species: wild pig ("boar"), clouded leopard, golden cat, pig-tailed macaque, common muntjac, sambar deer, tapir, and tiger.

The unit of time is the day, so values range from 0 to 1. Package overlap works entirely in *radians*: fitting density curves uses trigonometric functions (sin, cos, tan), so this speeds up bootstraps and simulations. The conversion is straightforward:

```
> timeRad <- kerinci$Time * 2 * pi
```

## 2.2   Fitting kernel density

We will extract the data for tigers in Zone 2 (which has the most observations) and plot a kernel density curve:

```
> tig2 <- timeRad[kerinci$Zone == 2 & kerinci$Sps == 'tiger']
> densityPlot(tig2, rug=TRUE)
```
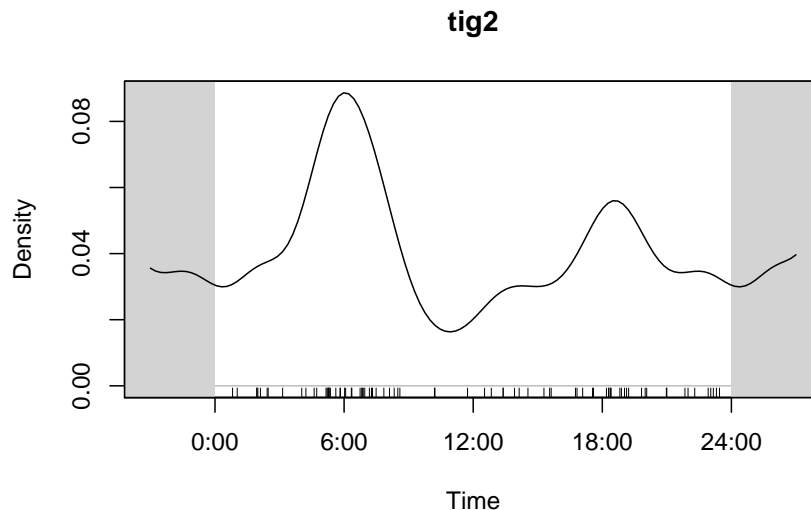


Figure 1: *Fitted kernel density curve for tigers in Zone 3, using default smoothing parameters.*

Figure 1 shows the activity pattern from 21:00 to 03:00, a reminder that the density is *circular*. Unlike the usual density plot that uses a Gaussian kernel, we use a von Mises kernel, corresponding to a circular distribution.

The actual data are shown at the foot of Figure 1 as a 'rug'.

Density estimation involves smoothing the information in the data, and the degree of smoothing is controlled by the argument `adjust` to the `densityPlot` function. Increasing `adjust` above the default value of 1 gives a flatter curve, reducing it gives a more 'spiky' curve, as shown in Figure 2. The choice of `adjust` affects the estimate of overlap, as we discuss below.
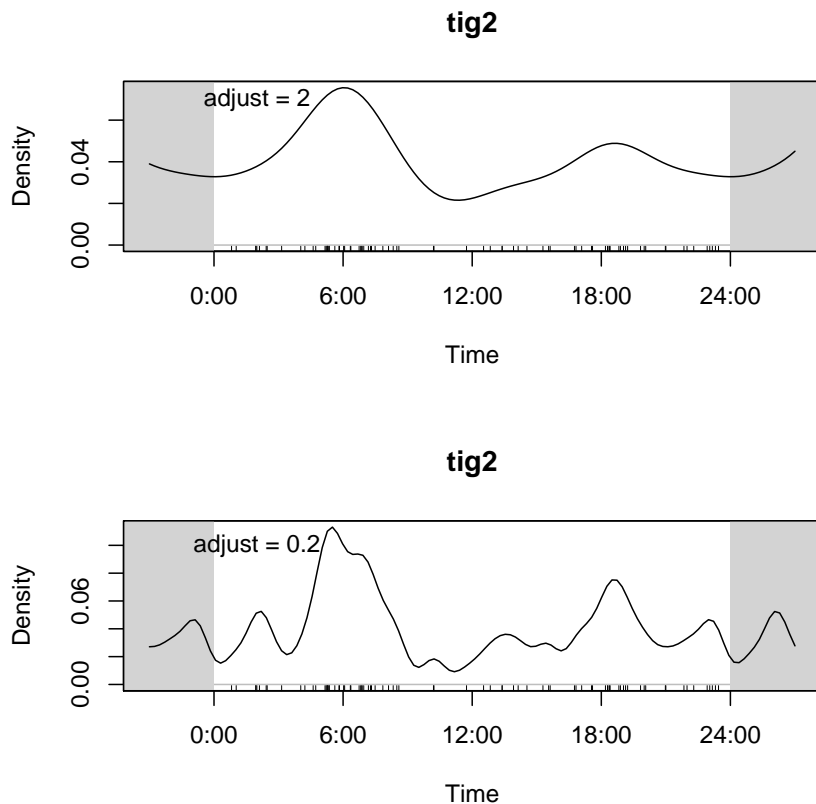
**tig2**



**tig2**



Figure 2: *Kernel density curves fitted with different smoothing adjustments.*

# 3 Quantifying overlap

Various measures of overlap have been put forward: see Ridout and Linkie (2009) for a review. We use the *coefficient of overlapping* proposed by Weitzman (1970).

## 3.1 Coefficient of overlapping

As shown in Figure 3, the coefficient of overlapping, $\Delta$, is the area lying under *both* of the density curves. (Remember that the area under a density curve is, by definition, one.) Mathematically, if the two density curves are $f(x)$ and $g(x)$, this is:

$$\Delta(f, g) = \int \min\{f(x), g(x)\} \mathrm{d}x \tag{1}$$

This works if we know the true density distributions, $f(x)$ and $g(x)$; but we usually only have samples and need to estimate $\Delta$ from these.

## 3.2 Estimators

Five general nonparametric estimators of the coefficient of overlapping were proposed by Schmid and Schmidt (2006). For circular distributions, the first two are equivalent and the third is unworkable (Ridout and Linkie, 2009). We retain $\hat{\Delta}_1$, $\hat{\Delta}_4$ and $\hat{\Delta}_5$.

The first, $\hat{\Delta}_1$, matches the definition in equation (1), but in practice it is estimated numerically, taking a large number of values, $t_1, t_2, ..., t_T$, equally spaced between 0 and $2\pi$ ($t_i = 2\pi i/T$) and summing:

$$\hat{\Delta}_1 = \frac{1}{T} \sum_{i=1}^{T} \min\{\hat{f}(t_i), \hat{g}(t_i)\} \tag{2}$$

For $\hat{\Delta}_4$ and $\hat{\Delta}_5$, we compare the densities at the observed values, $x_1, ..., x_n$ for one species and $y_1, ..., y_m$ for the other:

$$\hat{\Delta}_4 = \frac{1}{2} \left( \frac{1}{n} \sum_{i=1}^{n} \min\left\{1, \frac{\hat{g}(x_i)}{\hat{f}(x_i)}\right\} + \frac{1}{m} \sum_{i=1}^{m} \min\left\{1, \frac{\hat{f}(y_i)}{\hat{g}(y_i)}\right\} \right) \tag{3}$$

$$\hat{\Delta}_5 = \frac{1}{n} \sum_{i=1}^{n} I\left\{\hat{f}(x_i) < \hat{g}(x_i)\right\} + \frac{1}{m} \sum_{i=1}^{m} I\left\{\hat{g}(y_i) \leq \hat{f}(y_i)\right\} \tag{4}$$

where $I(.)$ is 1 if the condition in the parenthesis is true, 0 otherwise.

The terms $\hat{f}(.)$ and $\hat{g}(.)$ refer to the fitted kernel density functions, and as such they are affected by the choice of the smoothing constant, `adjust`. On the basis of simulations, Ridout and Linkie (2009) recommend using `adjust = 0.8` to estimate $\hat{\Delta}_1$, `adjust = 1` for $\hat{\Delta}_4$, and `adjust = 4` for $\hat{\Delta}_5$. (Note that `adjust` in the `overlap` functions corresponds to $1/c$ in Ridout and Linkie (2009)). These are the default values used in `overlap` functions.

## 3.3 Choice of estimator

Ridout and Linkie (2009) carried out simulations with a variety of scenarios where the true overlap was known, and compared the resulting estimates with the truth, calculating the root mean squared error (RMSE) for each estimator. The present authors have carried out further simulations in the same manner.

We found that the best estimator depended on the size of the *smaller* of the two samples: When the smaller sample was less than 50, $\hat{\Delta}_1$ performed best, while $\hat{\Delta}_4$ was better when it was greater than 75.

In no case was $\hat{\Delta}_5$ found to be useful. It is unstable, in that small, incremental changes in the data produce discontinuous changes in the estimate, and it can give estimates greater than one.

## 3.4 Examples

We will see how this works with the `kerinci` data set. We will extract the data for tigers and macaques for Zone 2, calculate the overlap with all three estimators, and plot the curves:

```
> tig2 <- timeRad[kerinci$Zone == 2 & kerinci$Sps == 'tiger']
> mac2 <- timeRad[kerinci$Zone == 2 & kerinci$Sps == 'macaque']
> min(length(tig2), length(mac2))

[1] 83

> tigmac2est <- overlapEst(tig2, mac2, type="Dhat4")
> tigmac2est
```

```
      Dhat4
0.4205464

> overlapPlot(tig2, mac2, main="Zone 2")
> legend('topright', c("Tigers", "Macaques"), lty=c(1,2), col=c(1,4), bty='n')
```
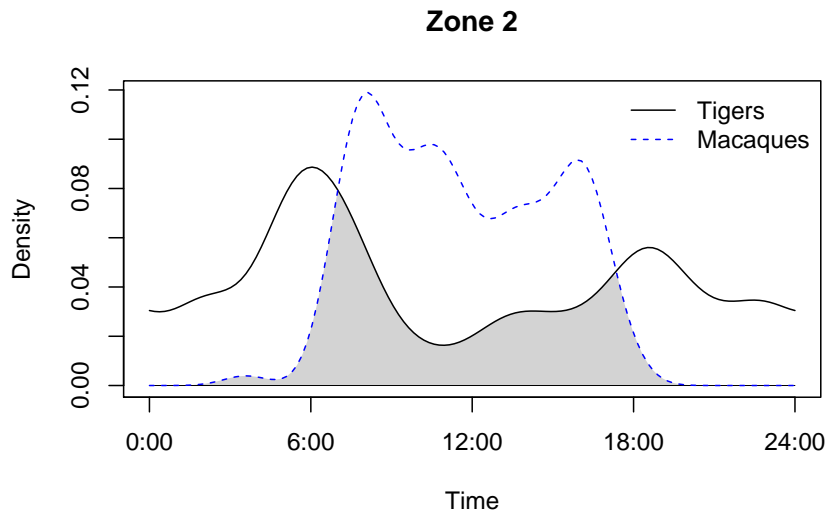
**Zone 2**



Figure 3: *Activity curves for tigers and macaques in Zone 2. The coefficient of overlapping equals the area below both curves, shaded grey in this diagram.*

Both of these samples have more than 75 observations, so we chose to use the $\hat{\Delta}_4$ estimate, `Dhat4` in the R code, giving an estimate of overlap of 0.42.

# 4  Confidence intervals

To estimate confidence intervals we need to know the sampling distribution which our coefficient of overlapping is drawn from, ie, the distribution we would get if we had a very large number of independent samples from nature. The best way to investigate this is to use a bootstrap.

## 4.1  The bootstrap

The usual bootstrap method treats the existing sample as representative of the population, and generates a large number of new samples by randomly resampling observations with replacement from the original sample. For the case of estimating activity patterns, this may not work very well: suppose our original sample for a nocturnal species has observations ranging from 20:58 to 03:14; resampling will never yield an observation outside that range, while a fresh sample from nature may do so.

An alternative is a smoothed bootstrap. We begin by fitting a kernel density to the original data then draw random simulated observations from this distribution. Faced with original values between 20:58 and 03:14, most simulated observations would fall in the same range, but a few will fall outside.

In the `overlap` package, we generate bootstrap samples with `resample`, which has a `smooth` argument; if `smooth = TRUE` (the default), smoothed bootstrap samples are generated. For this

example, we will generate just 1000 smoothed bootstrap samples for tigers and macaques in Zone 2; for a real analysis 10,000 bootstrap samples would be better:

```
> tig2boot <- resample(tig2, 1000)
> dim(tig2boot)

[1]   83 1000

> mac2boot <- resample(mac2, 1000)
> dim(mac2boot)

[1]  125 1000
```

This produces matrices with a column for each bootstrap sample. The bootstrap sample size is the same as the original sample size.

Now we pass these two matrices to `bootEst` to generate estimates of overlap from each pair of samples. We select $\hat{\Delta}_4$:

```
> tigmac2 <- bootEst(tig2boot, mac2boot, type="Dhat4")  # takes a few seconds
> ( BSmean <- mean(tigmac2) )

[1] 0.4746636
```

Note that the bootstrap mean, $\overline{BS}$, differs from $\hat{\Delta}$: 0.47 versus 0.42. The difference, $\overline{BS} - \hat{\Delta}$, is the *bootstrap bias*, and we need to take this into account when calculating the confidence interval.

If the bootstrap bias were a good estimate of the original sampling bias, a better estimator of $\Delta$ would be $\tilde{\Delta} = 2\hat{\Delta} - \overline{BS}$. Our simulations show that $\tilde{\Delta}$ results in higher RMSE than the original $\hat{\Delta}$, so we do not recommend applying this correction.

## 4.2   Extracting the CI

One way to estimate the confidence interval is simply to look at the appropriate percentiles of the set of bootstrap estimates (interpolating between values if necessary): for a 95% confidence interval these would be the 2.5% and 97.5% percentiles. This is `perc` in the output from `overlap`'s `bootCI` function.

We noted at the end of Section 4.1 that, on average, the bootstrap values differ from the estimate: this is the bootstrap bias. The raw percentiles produced by `perc` need to be adjusted to account for this bias. The appropriate confidence interval is `perc` $-(\overline{BS} - \hat{\Delta})$; this is `basic0` in the `bootCI` output.

An alternative approach is to use the standard deviation of the bootstrap results, $(s_{BS})$, as an estimate of the spread of the sampling distribution, and then calculate the confidence interval as $\hat{\Delta} \pm z_{\alpha/2} s_{BS}$. Using $z_{0.025} = 1.96$ gives the usual 95% confidence interval. This is `norm0` in the `bootCI` output. This procedure assumes that the sampling distribution is normal. If that's the case, `norm0` will be close to `basic0`, but if the distribution is skewed – as it will be if $\hat{\Delta}$ is close to 0 or 1 – `basic0` is the better estimator.

For the tiger-macaque data from Zone 2 we have the following estimates of a 95% confidence interval:

```
> bootCI(tigmac2est, tigmac2)

           lower      upper
norm   0.2652980 0.4675602
norm0  0.3194153 0.5216774
basic  0.2650490 0.4718550
basic0 0.3151205 0.5219265
perc   0.3692378 0.5760438
```

`bootCI` produces two further estimators: `basic` and `norm`. These are analogous to `basic0` and `norm0` but are intended for use with the bias-corrected estimator, $\tilde{\Delta}$. They match the `basic` and `norm` confidence intervals produced by `boot.ci` in package `boot`.

The coefficient of overlapping takes values in the interval [0,1]. All the confidence interval estimators except `perc` involve additive corrections which might result in values outside of this range. This can be avoided by carrying out the corrections on a logistic scale and back-transforming. This is done by `bootCIlogit`:

```
> bootCIlogit(tigmac2est, tigmac2)

          lower     upper
norm   0.2790988 0.4680405
norm0  0.3249769 0.5224676
basic  0.2793647 0.4736311
basic0 0.3200740 0.5221380
perc   0.3692378 0.5760438
```

In this example, the CIs are well away from 0 or 1, so the difference is small (and `perc` is exactly the same as there's no correction anyway).

## 4.3  Choice of CI method

If a series of X% confidence intervals are calculated from independent samples from a population, we would expect X% of them to include the true value. When running simulations we know the true value and can check the actual proportion of confidence intervals which contain the true value: this is the *coverage* of the estimator. Ideally the coverage should equal the nominal confidence interval, ie, 95% coverage for a 95% confidence interval.

We ran a large number of simulations with different true distributions and sample sizes (see Ridout and Linkie (2009) for details). For each scenario, we ran both smoothed and unsmoothed bootstraps, extracted all nine 95% confidence intervals, and checked the coverage for each.

Each estimator gave a range of coverages. We looked for a method which gave median coverage closest to the nominal 95% and all or most values above 90%. This was satisfied by the `basic0` estimator with smoothed bootstraps.

With small samples (smaller sample < 75) and $\Delta > 0.8$, coverage sometimes fell below 90%, but none of the other options fared better.

# 5  Summary of recommendations

- Use the $\hat{\Delta}_4$ estimator (`Dhat4`) if the smaller sample has more than 75 observations. Otherwise, use the $\hat{\Delta}_1$ estimator (`Dhat1`).

- Use a smoothed bootstrap and do at least 1000 resamples, preferably 10,000.

- Use the `basic0` output from `bootCI` as your confidence interval; be aware that this confidence interval will be too narrow if you have a small sample and $\Delta$ is close to 1.

# 6  Caveats

## 6.1  Pooling data

Pooled data give higher estimates of overlap than the original, unpooled data (Ridout and Linkie, 2009). Suppose we find a species of bat that emerges immediately after sunset and a hawk which goes to roost just before sunset: their activity patterns do not overlap and

presumably the hawk will not be feeding on the bats. But the time of sunset changes; data from December only or from June only show no overlap, but the pooled data do, and this apparent overlap is an artefact of pooling.

This is a clear-cut example. In general, differences in activity patterns across sites or time periods will be smaller, but any heterogeneity will inflate the overlap estimates from pooled data. Care is needed when comparing coefficients of overlap among study areas or periods of varying extent or degree of heterogeneity.

One way to mitigate these differences is to map "clock time" to "sun time" (Nouvellet *et al.*, 2012). The new function `sunTime` allows this to be done, see its help page. Azevedo *et al.* (2018) used this approach for their study of puma.

## 6.2 What "activity" is observed?

Camera traps set along animal trails – as they often are – record instances of animals moving along trails. The resulting "activity pattern" refers to walking on trails, and overlap indicates the extent to which two species are walking on trails at the same period of the day. A browsing herbivore and the carnivore stalking it are probably both "inactive" by this definition.

In view of this, conclusions about species interactions need to be drawn with care. In a study in Lao PDR, Kamler *et al.* (2012) found that dhole and pig were active during the day and deer at night. This might suggest that dhole feed on pig rather than deer. But examination of dhole faeces showed that dhole consumed mainly deer and very little pig.

# 7  References

Azevedo FC, Lemos FG, Freitas-Junior MC, Rocha DG, Azevedo FCC (2018). "Puma activity patterns and temporal overlap with prey in a human-modified landscape at Southeastern Brazil." *Journal of Zoology*, **0**(0), 0.

Carter NH, Shrestha BK, Karki JB, Pradhan NMB, Liu J (2012). "Coexistence between wildlife and humans at fine spatial scales." *Proceedings of the National Academy of Sciences*, **109**(38), 15360–15365.

Carver BD, Kennedy ML, Houston AE, Franklin SB (2011). "Assessment of temporal partitioning in foraging patterns of syntopic Virginia opossums and raccoons." *Journal of Mammalogy*, **92**(1), 134–139.

Griffiths M, van Schaik CP (1993). "Camera-trapping: a new tool for the study of elusive rain forest animals." *Tropical Biodiversity*, **1**, 131–135.

Kamler JF, Johnson A, Vongkhamheng C, Bousa A (2012). "The diet, prey selection, and activity of dholes (*Cuon alpinus*) in northern Laos." *Journal of Mammalogy*, **93**(3), 627–633.

Linkie M, Ridout MS (2011). "Assessing tiger-prey interactions in Sumatran rainforests." *Journal of Zoology*, **284**(3), 224–229.

Nouvellet P, Rasmussen GSA, Macdonald DW, Courchamp F (2012). "Noisy clocks and silent sunrises: measurement methods of daily activity pattern." *Journal of Zoology*, **286**(3), 179–184.

Ramesh T, Kalle R, Sankar K, Qureshi Q (2012). "Spatio-temporal partitioning among large carnivores in relation to major prey species in Western Ghats." *Journal of Zoology*, **287**(4), 269–275.

Ridout MS, Linkie M (2009). "Estimating overlap of daily activity patterns from camera trap data." *Journal of Agricultural, Biological, and Environmental Statistics*, **14**(3), 322–337.

Ross J, Hearn AJ, Johnson PJ, Macdonald DW (2013). "Activity patterns and temporal avoidance by prey in response to Sunda clouded leopard predation risk." *Journal of Zoology*, **290**(2), 96–106.

Schmid F, Schmidt A (2006). "Nonparametric estimation of the coefficient of overlapping — theory and empirical application." *Computational Statistics and Data Analysis*, **50**, 1583–1596.

van Schaik CP, Griffiths M (1996). "Activity periods of Indonesian rain forest mammals." *Biotropica*, **28**(1), 105–112.

Weitzman MS (1970). "Measure of the Overlap of Income Distribution of White and Negro Families in the United States." *Technical report 22*, U.S. Department of Commerce, Bureau of the Census, Washington, DC.