# SAMBAH Code File 4
# SAMBAH Playback Experiments Detection Function Analysis

Len Thomas & Louise Burt, CREEM

January 2, 2022

## 1 Introduction

This document (the `.Rnw` version) contains the `R` code to fit a detection function to the SAMBAH main survey area playback data. The code estimates the playback effective detection area for each station and month (EDA, $\hat{\xi}_{i,m}$), and also estimates variance, using a non-parametric bootstrap. Output files are (optionally) saved for use in the density analysis in SAMBAH Code File 6. This document is based on SAMBAH internal reports; this version has been created to accompany the paper:

Amundin et al. In press. Estimating the abundance of the critically endangered Baltic Proper harbour porpoise (Phocoena phocoena) population using passive acoustic monitoring. Ecology and Evolution.

More information about the analysis is given in the methods section of the paper.

At the end of the document, there is an exploratory analysis of the possible effect of diel phase on the playback detection function, made at the suggestion of a reviewer of the paper.

The document is a `Sweave` file – i.e., it is a mixture of `LaTeX` and `R` that is designed to be compiled into a report in pdf (or another format such as html). We have tested it using the `Knitr` package in `R version 4.1.1 (2021-08-10)`. Readers wishing to see the underlying code should view the version with the `.Rnw` suffix, and look for code chunks starting with <<.

Note that several time-consuming steps of the analysis are by default not performed if the code is re-run as is – instead results are read in from file to save time in compiling the document. Should the reader wish to re-run all the analyses, there are boolean variables at the top of the code file that can be changed to facilitate this.

| | country | name | positions | playback.stations | prop.playback.stations | playbacks | n.playbacks.per.station |
|---|---|---|---|---|---|---|---|
| 1 | 1 | Sweden | 99 | 70 | 0.71 | 70 | 1.00 |
| 2 | 2 | Finland | 46 | 25 | 0.54 | 25 | 1.00 |
| 3 | 3 | Estonia | 40 | 0 | 0.00 | 0 | |
| 4 | 4 | Latvia | 34 | 9 | 0.26 | 12 | 1.33 |
| 5 | 5 | Lithuania | 9 | 6 | 0.67 | 10 | 1.67 |
| 6 | 6 | Poland | 39 | 39 | 1.00 | 68 | 1.74 |
| 7 | 7 | Germany | 16 | 16 | 1.00 | 32 | 2.00 |
| 8 | 8 | Denmark | 21 | 16 | 0.76 | 36 | 2.25 |

Table 1: Summary of number of stations at which playbacks were performed, and number of playbacks performed at each station for which playbacks were done.

| | month | playbacks |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 2 | 8 |
| 3 | 3 | 15 |
| 4 | 4 | 27 |
| 5 | 5 | 89 |
| 6 | 6 | 25 |
| 7 | 7 | 2 |
| 8 | 8 | 48 |
| 9 | 9 | 0 |
| 10 | 10 | 17 |
| 11 | 11 | 21 |
| 12 | 12 | 1 |

Table 2: Summary of number of playbacks per month.

There were a total of 253 playback experiments, at 181 station – see Tables 1 and 2. Only Germany has a perfect record of all stations with playbacks, and 2 playbacks per station (which was our original plan). Note that more playbacks were actually performed by the SAMBAH field teams, but due to equipment damage (a damaged transponder), these additional playbacks coulnt not be used in the analyses reported here.

| | country | name | min.n.distances | median.n.distances | mean.n.distances | max.n.distances |
|---|---|---|---|---|---|---|
| 1 | 1 | Sweden | 1 | 3.5 | 3.53 | 8 |
| 2 | 2 | Finland | 2 | 4.0 | 4.16 | 7 |
| 3 | 3 | Estonia | | | | |
| 4 | 4 | Latvia | 2 | 3.0 | 2.92 | 4 |
| 5 | 5 | Lithuania | 2 | 3.0 | 2.80 | 4 |
| 6 | 6 | Poland | 3 | 4.0 | 3.99 | 4 |
| 7 | 7 | Germany | 4 | 7.0 | 6.53 | 8 |
| 8 | 8 | Denmark | 1 | 3.0 | 3.53 | 15 |

Table 3: Summary of number of distances per playback.

The original plan called for playbacks at a large number of distances, but likely for logistical reasons this was not done by many countries – see Table 3. Overall mean playback distances per playback experiment was 4.04.

| | country | name | mean.min.dist | mean.median.dist | mean.max.dist |
|---|---|---|---|---|---|
| 1 | 1 | Sweden | 68.84 | 133.44 | 190.54 |
| 2 | 2 | Finland | 51.04 | 137.45 | 226.63 |
| 3 | 3 | Estonia | | | |
| 4 | 4 | Latvia | 70.83 | 123.54 | 170.83 |
| 5 | 5 | Lithuania | 45.10 | 80.00 | 125.00 |
| 6 | 6 | Poland | 201.46 | 351.32 | 500.00 |
| 7 | 7 | Germany | 41.39 | 173.73 | 357.90 |
| 8 | 8 | Denmark | 176.16 | 211.32 | 250.36 |

Table 4: Summary of mean of min, median and max distance per playback experiment.

In addition, some countries did not have a very good distribution of distances for playbacks – see Table 4 – for example, for Poland and Denmark, the average closest (i.e., minimum) playback distance was well over 150m - a distance at which it's unlikely there'd be many positive detections at the median source level used (see Table 5). (Overall, the mean minimum distance per playback is 113.68 and the mean maximum distance is 303.44.

This means that we may not be able to model the detection function shape close to 0 distance well; it also

means that will likely not want to use the median playback source level as the reference level – rather something higher (in the event we chose something close to the maximum playback level). (We also considered discarding some of the lower source level playbacks, if they conveyed little information about the shape of the detection function, although we kept all of the data in the end.)

| | country | name | mean.min.SL | mean.median.SL | mean.max.SL |
|---|---|---|---|---|---|
| 1 | 1 | Sweden | 155.00 | 162.50 | 170.00 |
| 2 | 2 | Finland | 155.00 | 162.50 | 170.00 |
| 3 | 3 | Estonia | | | |
| 4 | 4 | Latvia | 155.00 | 162.50 | 170.00 |
| 5 | 5 | Lithuania | 155.00 | 162.50 | 170.00 |
| 6 | 6 | Poland | 154.32 | 161.82 | 169.32 |
| 7 | 7 | Germany | 138.00 | 153.00 | 168.00 |
| 8 | 8 | Denmark | 154.83 | 162.33 | 169.83 |

Table 5: Summary of mean of min, median and max of the planned (as opposed to measured, calibrated) source level per playback experiment.

The largest playback distance at which a click was detected was 200, at source level 170.

## 2 Detection function model fitting

We tried a variety of models (as described in the paper), but our final model was this:

```
mod<-gam(cbind(n.detected,n.not.detected)~te(distance,SL.plan)+geo+s(depth,k=5)+
  s(month,k=5,bs="cc")+s(SST,k=5)+s(SSsal,k=5),data=dat,family=binomial(link=logit),
  knots=list(month=c(1,12)))
```

We applied this model to the playback data and environmental covariates, but we first trimmed the top 1% of SSsal values (Sea Surface salinity), setting them to be equal to the highest 99th quantile. This is because the value was quite an outlier, and this was affecting the fit. Also, there were very few records of `geo` type 7, so we binned it with type 6.

Model output is given below[1], and the fitted smooths are shown in Figures 1 and 2.

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## cbind(n.detected, n.not.detected) ~ te(distance, SL.plan) + geo +
##     s(depth, k = 5) + s(month, k = 5, bs = "cc") + s(SST, k = 5) +
##     s(SSsal, k = 5)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.78084    0.79191   -7.30 2.88e-13 ***
## geo2         0.43557    0.02184   19.94  < 2e-16 ***
## geo3         0.99549    0.02923   34.06  < 2e-16 ***
## geo4         1.11058    0.02544   43.66  < 2e-16 ***
## geo6         1.00065    0.03836   26.08  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

[1] An interpretation of the geo levels would be in order at some point; also some model diagnostics.

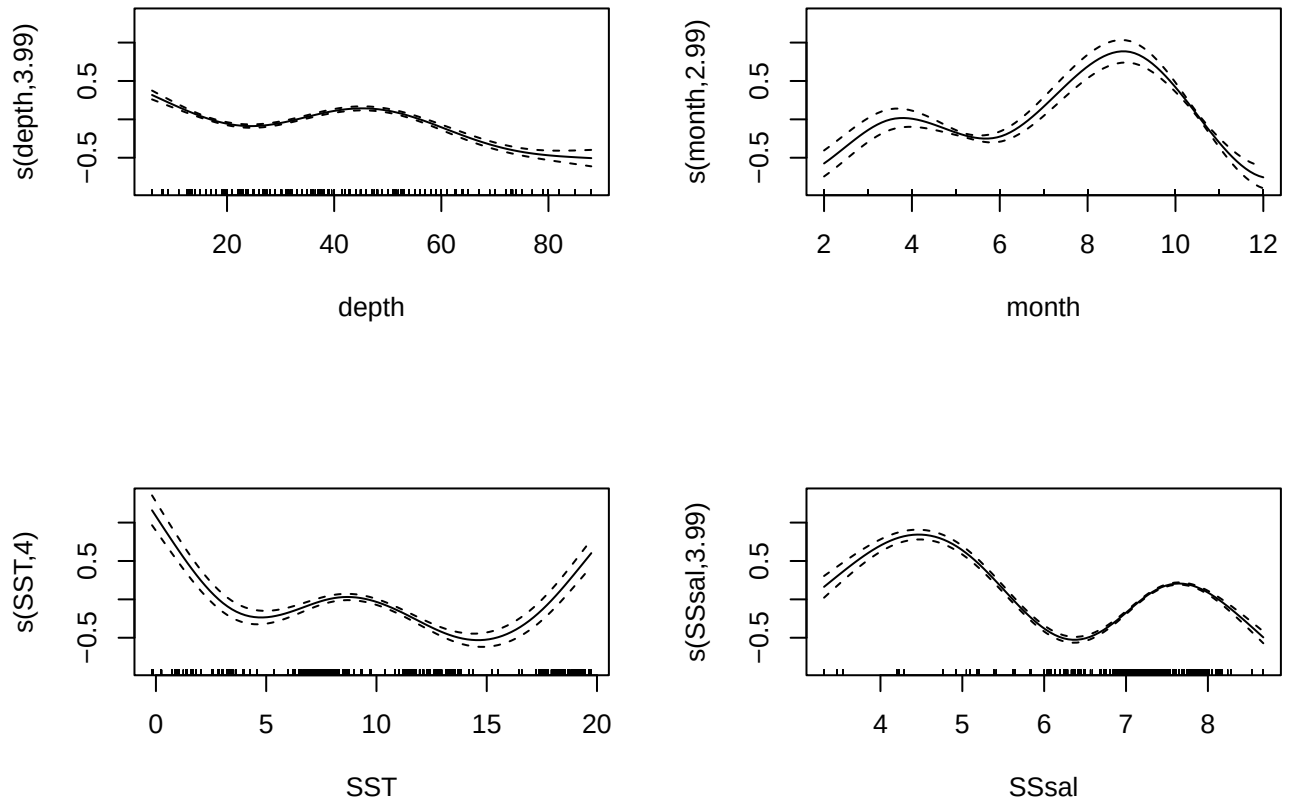


Figure 1: SAMBAH playback data fitted smooths, on the scale of the link function. The 2-d smooth of distance and SL is shown in the next figure.

```
## Approximate significance of smooth terms:
##                           edf Ref.df  Chi.sq p-value
## te(distance,SL.plan) 23.721  23.93 29656.9  <2e-16 ***
## s(depth)              3.986   4.00   432.7  <2e-16 ***
## s(month)              2.993   3.00   522.3  <2e-16 ***
## s(SST)                3.997   4.00  1575.0  <2e-16 ***
## s(SSsal)              3.992   4.00  1334.8  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.565   Deviance explained = 54.9%
## UBRE = 3.9027  Scale est. = 1         n = 14443
```

# 3 Prediction

We wish to predict effective detection area ($\nu$) for each month and location in the SAMBAH data. To do this, we need to choose a source level. As we showed in Section **??**, since many of the playbacks started rather far from the CPODs, a high source level should be used. We elected to use 168dB, which is the highest level used in the Great Belt playback experiments (recall that the outputs of this analysis, and the outputs of the Great

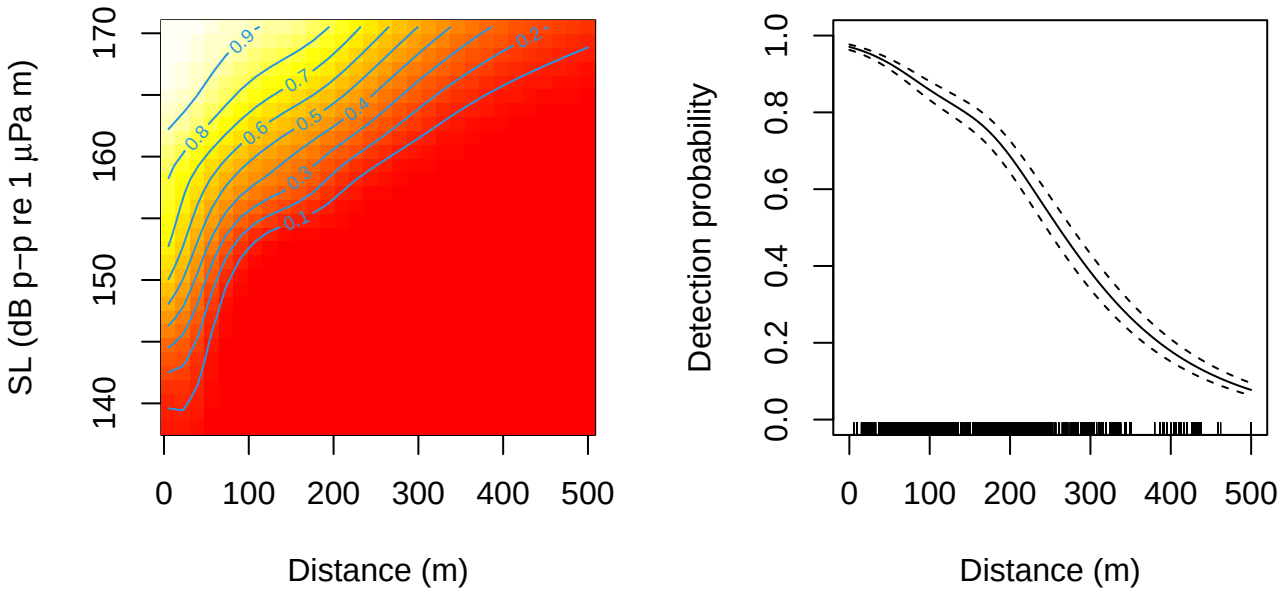


Figure 2: Two views of the SAMBAH playback data fitted 2-d smooth of distance and SL, on the response scale. The right hand plot is the marginal smooth, assuming a SL of 168 dB.

Belt playback analysis, will be used together in the density estimation calculations, so it makes some sense for the two playback analyses to use the same source level for prediction).

Before making predictions, we also constrained all covariates so that they lie within the range of values for the stations where playbacks took place.

Note that we are predicting over all SAMBAH data, not just that used in the design-based estimation, because we wish to make predictions by month and station for the model-based analysis as well.

The distribution of predicted EDAs is shown in Figure 3. The mean of these EDAs (in ha) is 21.9 and median 20.4.

We integrated out to 1000 meters. To check this is far enough, we show in Figure 4 the detection function (and density) corresponding to the record with the biggest estimated Effective Detection Area[2] – we note that the detection probability of density are both very small at 1000 meters, so we do not miss any significant detection density farther out than this distance.

# 4 Variance estimation

Variance estimation is via a non-parametric bootstrap, with the sampling unit being a playback session (i.e., a set of playbacks at a station on the same date). These are referred to as "playback" in Table 1. We performed 1000 bootstrap resamples, and for each calculated the EDA for each prediction point; these were then saved into an `.RData` file for use in the density estimation routines. Just for the record, the random seed used for resampling was 471739.

One thing to check is that the maximum distance used in the integration for the bootstrap analysis (which was 1200 meters) is far enough out so that the estimate of EDA is accurate. One way to check this is to track the estimated detection probability at this distance, which should be close to zero for all stations in all bootstrap replications. In practice, using an arbitrary value of $g(w) = 0.01$ as cause for concern, the percentage of stations

[2]For info, we also show the smallest.

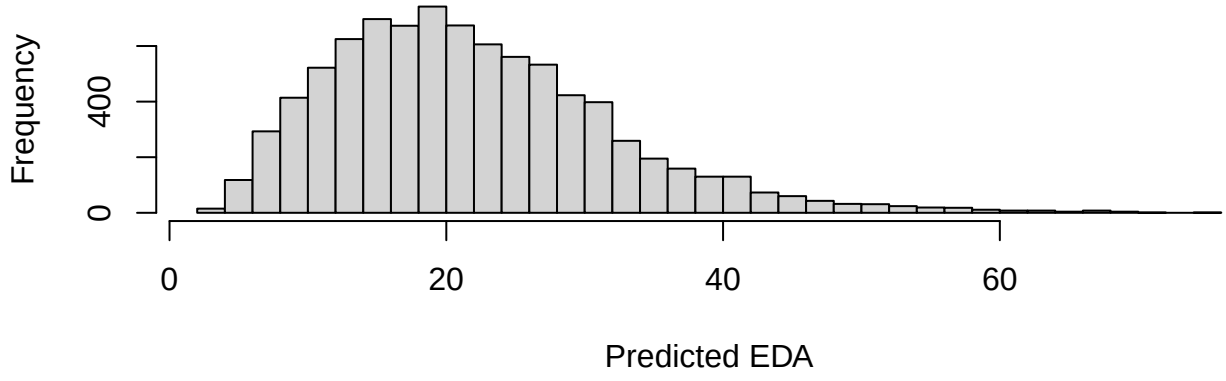


Figure 3: Histogram of predicted EDAs in hectares (i.e., 10,000 square meters).

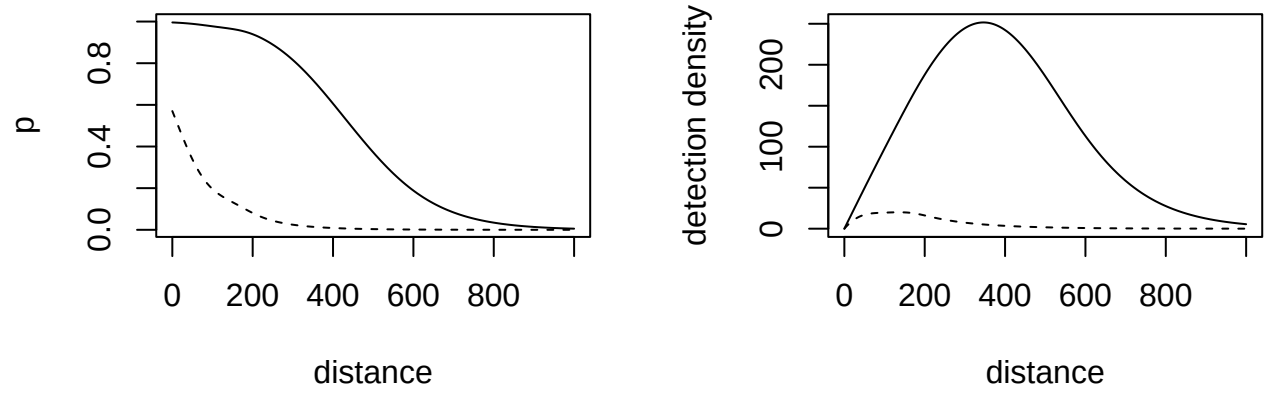


Figure 4: Detection function and distance density corresponding to the highest and lowest detectable combinations in the predicted data.

with a value of $g(w)$ greater than this across all bootstraps was 0.18%. We conclude that the value of maximum distance used was big enough.

We summarize the results as follows. We began by taking the mean of the predictions for each bootstrap replicate. The mean of these means is 222491, compared with a mean of the original data predictions of 218770. The SE is 29128 while the CV (i.e., sd(boot)/mean(original)) is 13.314%; 95% percentile confidence intervals are 175566, 291531.

The station with minimum EDA is 1097 in month 12, with EDA 34325. Standard error on this is 31320.

The station with max EDA is 3026 in month 8, with EDA 742258. Standard error on this is 212665.

# 5 Diel phase

## 5.1 Detection function modelling by diel phase

During review of our paper, we were asked to investigate the effect of diel phase on detectability. We did this by adding diel phase (as a factor variable) to our final model. The results are given below:

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## cbind(n.detected, n.not.detected) ~ te(distance, SL.plan) + geo +
##     s(depth, k = 5) + s(month, k = 5, bs = "cc") + s(SST, k = 5) +
##     s(SSsal, k = 5) + diel
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.05294    0.18350 -27.536  < 2e-16 ***
## geo2         0.37836    0.02220  17.040  < 2e-16 ***
## geo3         0.98660    0.02960  33.333  < 2e-16 ***
## geo4         1.09625    0.02566  42.727  < 2e-16 ***
## geo6         1.16355    0.03914  29.731  < 2e-16 ***
## dielDawn    -0.43488    0.03638 -11.954  < 2e-16 ***
## dielDusk    -0.21528    0.03381  -6.367 1.93e-10 ***
## dielNight    0.58992    0.01669  35.341  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                          edf Ref.df  Chi.sq p-value
## te(distance,SL.plan) 22.848  23.21 29966.1  <2e-16 ***
## s(depth)              3.989   4.00   674.7  <2e-16 ***
## s(month)              2.989   3.00   560.5  <2e-16 ***
## s(SST)                3.998   4.00  1375.4  <2e-16 ***
## s(SSsal)              3.998   4.00  1195.0  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.576   Deviance explained = 55.9%
## UBRE = 3.7881  Scale est. = 1         n = 14443
```

From this, by exponentiating the relevant coefficients, we find the odds ratios for dawn, dusk and night vs day are 0.6473, 0.8063, 1.8038. If required, a 95% CI could be obtained on these by adding and subtracting twice the stdandard error from the coefficients and then exponentiating.

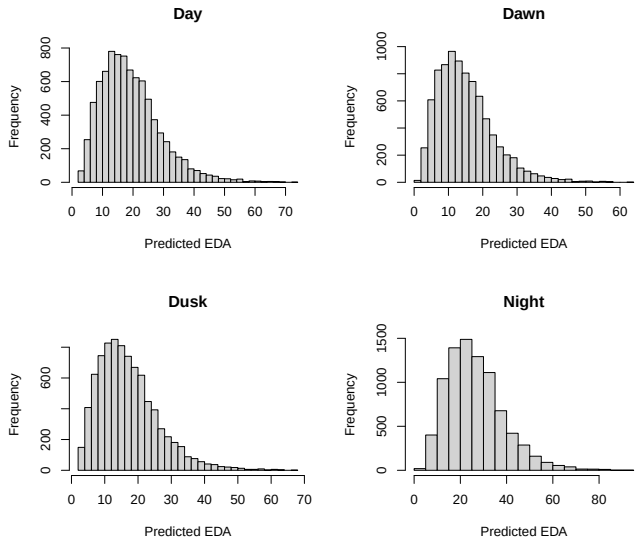


Figure 5: Histogram of predicted EDAs in hectares (i.e., 10,000 square meters) per diel phase.

## 5.2 Predictions by diel phase

To investigate the effect of introducing diel phase into the playback model, we repeat the prediction for each diel phase.

Figure 5 shows the distribution of EDA for the 4 diel phases. Mean EDA (in ha) for diel phases Day, Dawn, Dusk, Night are 19.3, 15.0, 17.1, 26.5 and medians are 17.7, 13.5, 15.5, 24.7. The ratio of mean EDA by diel phase divided by that in the Day is 1.000, 0.778, 0.885, 1.371.