SAMBAH Code File 1
Encounter Rate Analysis

Len Thomas & Louise Burt, CREEM

January 1, 2022

1 Introduction

This document (the .Rnw version) contains the R code to read in SAMBAH main survey effort (seconds of CPOD monitoring) and encounters (click positive seconds), do some exploratory summaries and export the encounter rate (click positive seconds per unit time monitoring) data needed for 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.

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 <<.

The SAMBAH main survey design consists of 304 sampling positions, in 8 countries. A rough map of positions (labelled by country number) is given in Figure 1.

2 Reading in the data

The raw data to calculate encounter rate is in the files
`detections and environment - validated and cropped - 20141013.txt`
and
`click details - validated and croppped - 20141013.txt`.
We refer to the first as the "effort file" and the second as the "click file". Below we describe how these files were read in and processed, using the R script CalculateEncounterRate_v6.r. However, we do not actually implement this process here, because it is time consuming and the raw files are very large – instead in the code and results below, we work with the output of the above R script, which is a summary of the number of click-positive seconds and number of monitoring seconds per minute per CPOD station – read in from the file n.bymonth.bymin.txt.

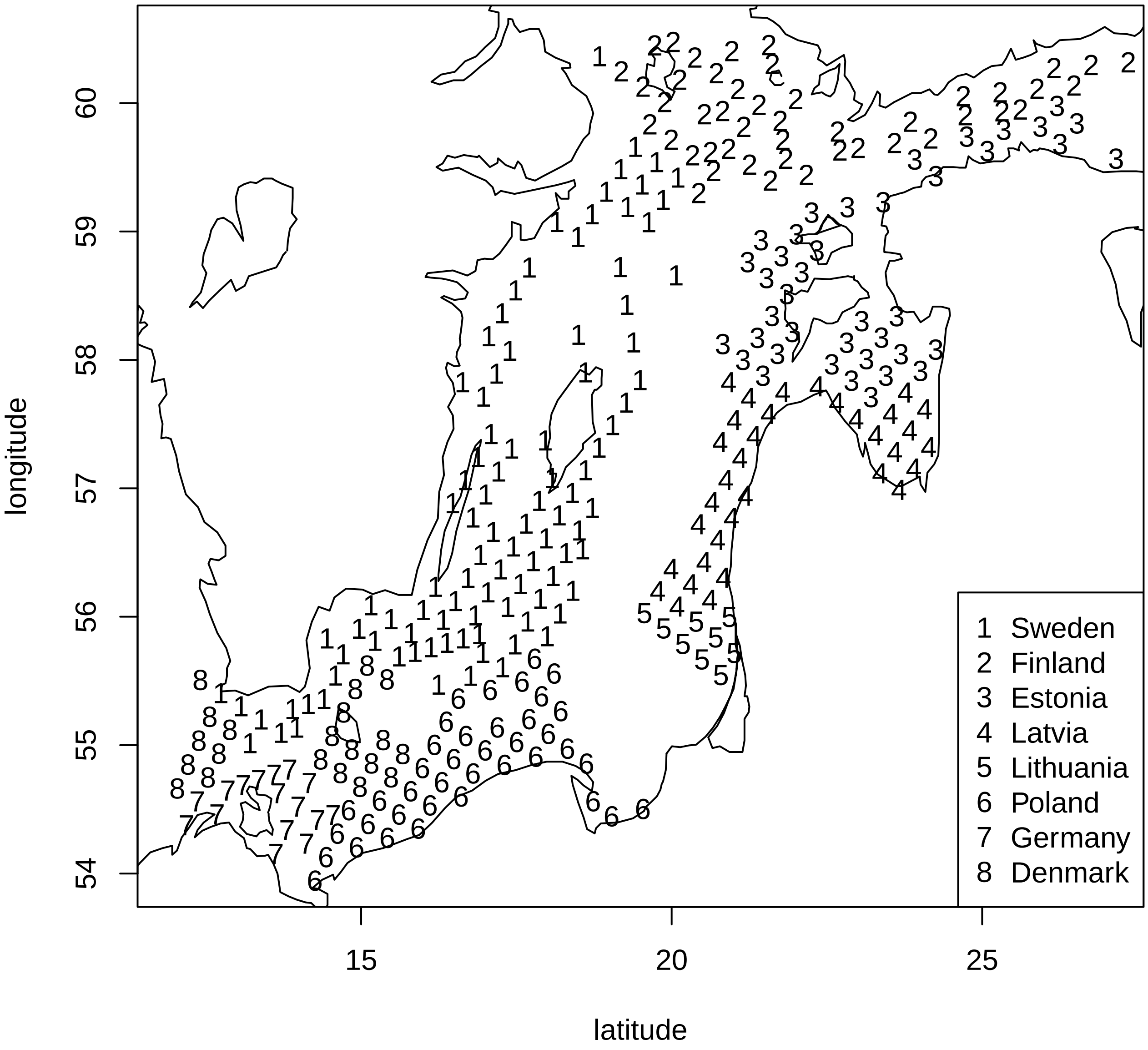

Figure 1: SAMBAH positions, labelled with their country code.

2.1 Effort file

The effort file has one line for each minute that each POD was deployed, but is truncated so that days when deployments were made are not in the file – i.e., records for a deployment always start at midnight on the day of deployment and stop one minute before midnight on the day before the deployment ended. The effort file contains 13 columns and is tab delimited. It's a large file (around 18GB). We wrote R code (the `CalculateEncounterRate` script mentioned above) to read the file.

- From the first column ("File") we extract the deployment number (it's the first 6 columns).
- From the second column ("ChunkEnd" we extract) the date and time. From this, we extract the year and month (number: 1-12), as well as the number of minutes after midnight (number: 1-1440).
- From the 6th column ("MinutesON") we extract a zero or 1. If zero, this means the POD was not on for that minute, and we skip to the next record. If 1, then from the 10th column ("%TimeLost") we extract a number between 0 and 100, representing the percentage of that minute that was "lost" – i.e., where the POD was not operating effectively - e.g., it was tipped over, etc.

For each record, we then calculate the number of seconds the POD was operating as 0 (if MinutesON is 0) or 60 times %TimeLost (if MinutesON is 1). Note that this leads to non-integer seconds. For each minute of the day within each month within each deployment, we add the number of seconds together, to make a total `effort.secs` - this is saved to the file `effort.bymonth.bymin.txt` So, we effectively aggregate each minute over days within months within deployments.

Notes:

- All times within this file are local times, with the timezone set at time of deployment. There is a GMT offset column in the master meta data file that will allow us to convert to GMT, if required.
- For stations that are at depths outside our initial criteria, we will keep them for the design-based analysis used here.

The files include data from the Russian supplement to SAMBAH (called "RUMBAH"). We excluded these data from our calculations (by excluding all data with country code of 9.)

We aggregated the results by deployment and month. Here are some summaries.

Once read in, there are 6141 deployment-months containing data, with 1356 deployments at 298 positions. Note, this means there are 6 of the original 304 positions where no data was collected. The total number of seconds of on-effort data is 1.235759e+10, which is equivalent to 391.86 years.

Table 1 and Figure 2 show the number of positions with working CPODs by month.

	1	2	3	4	5	6	7	8	9	10	11	12
2011	0	0	36	169	252	273	272	238	220	205	215	161
2012	193	186	165	135	193	226	228	228	216	216	222	209
2013	208	214	205	190	164	31	4	1	0	0	0	0

Table 1: Number of positions with working CPODs by month

From this point on, we truncate the data, so that we only work with data collected between 1st May 2011 and 30th April 2013, inclusive. This truncation deletes 6.595% of the effort data.

Figure 3 shows the spatial distribution of effort between May 2011 and April 2013.

There are strong spatial patterns in realized effort, with some isolated clusters of low effort, plus a general tendency for lower effort in Estonia, Latvia and Lithuania.

Effort can be lost for many reasons. A non-exhaustive list is: PODs not deployed; PODs failing early; PODs not serviced promptly; PODs moving from original position (e.g., being caught in trawl nets) and either lost or found later; PODs being temporarily tilted beyond tolerence (in which case any collected data is not used in those seconds); too many detections in a minute (in which case the POD stops recording for those seconds). Over two years, at each site, the total number of potential seconds of monitoring is $60 \times 60 \times 24 \times 365 = 31,536,000$. Table 2 shows give the potential and total monitoring time for each country.

Overall, there are 377.07 years of data out of a possible 608, representing 62.018% of the total possible survey effort.

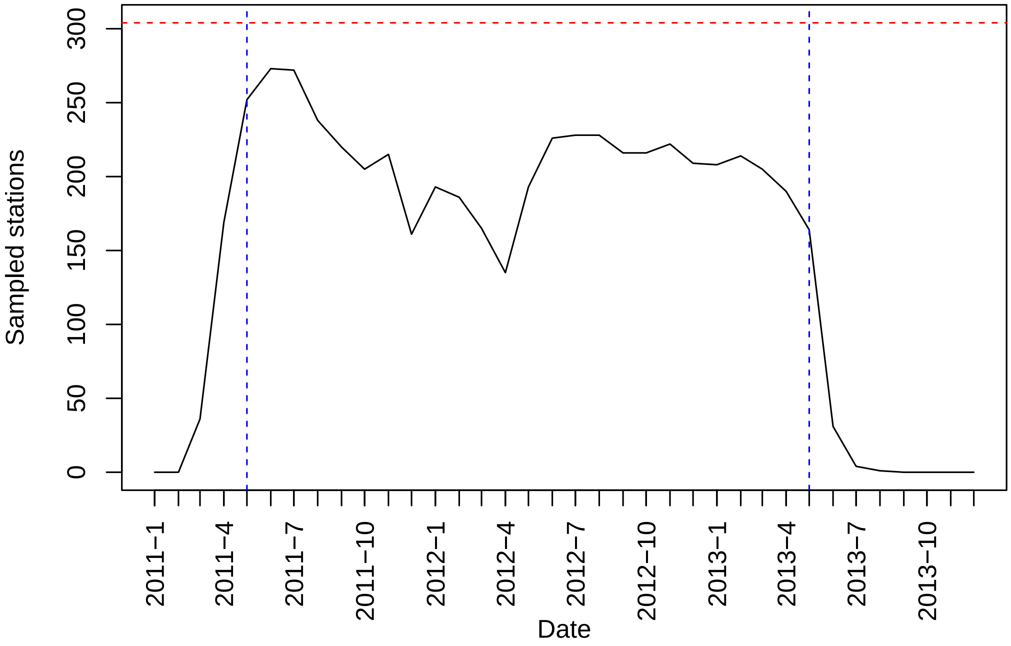

Figure 2: Number of positions with working CPODs by month. The blue vertical dashed lines mark the April/May 2011 and April/May 2013 boundaries(May 1st 2011 and April 30th 2013 are the agreed start and end points of the project) and the red horizontal dashed line showd the maximum number of positions.

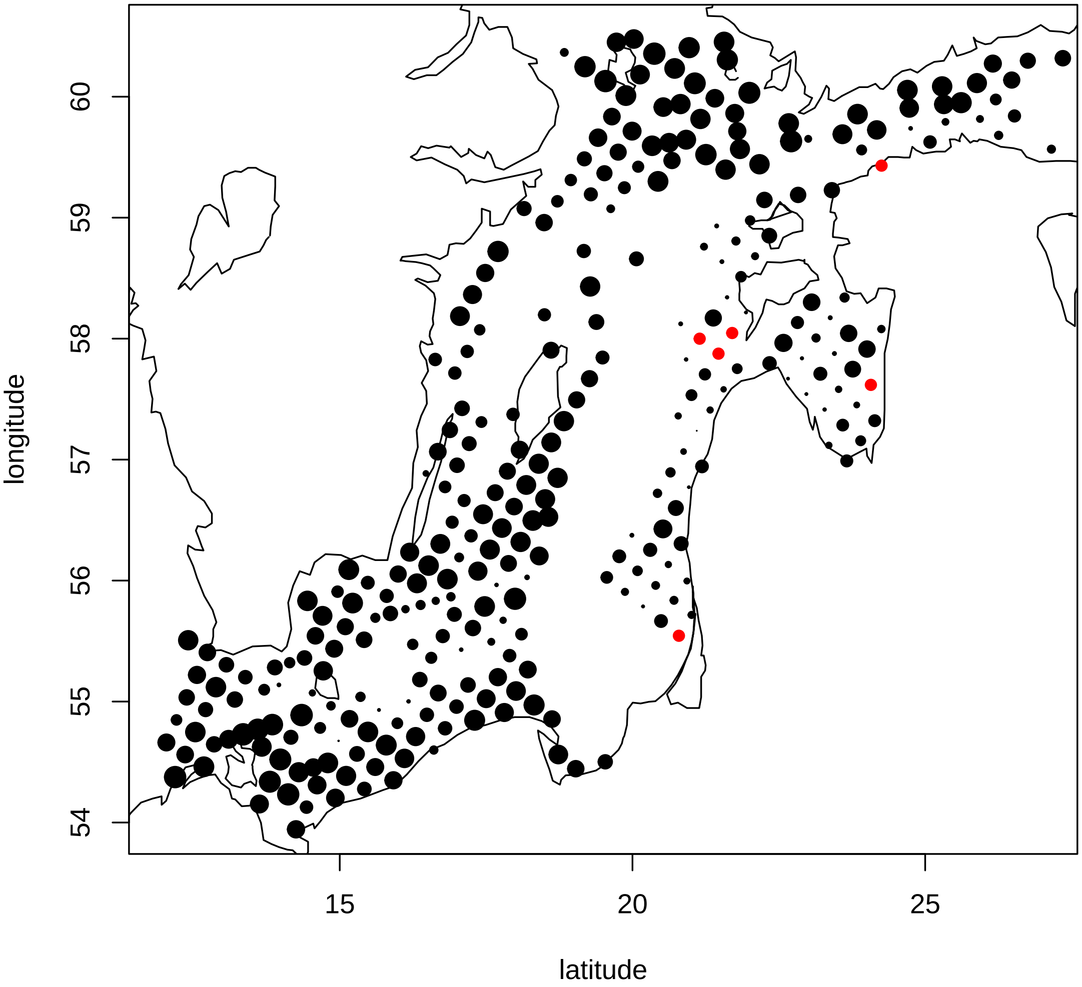

Figure 3: Spatial distribution of effort between May 2011 and April 2013. The size of each point is proportional to the number of days surveyed; red dots are positions with zero days of survey effort.

	name	positions	potential.years	years	perc.effort
1	Sweden	99	198.00	130.75	66.03
2	Finland	46	92.00	78.64	85.48
3	Estonia	40	80.00	30.20	37.74
4	Latvia	34	68.00	25.10	36.91
5	Lithuania	9	18.00	5.49	30.53
6	Poland	39	78.00	52.26	67.00
7	Germany	16	32.00	27.86	87.05
8	Denmark	21	42.00	26.77	63.75

Table 2: Number of positions by country, together with potential total monitoring time (in years) and the actual and percentage monitoring time (calculated by summing all seconds where a C-POD was operational per position).

2.2 Click file

This file contains one record for each click.

- The first column ("abbreviated file name") is the deployment number.
- From the second column ("Minute" we extract) the year, month and minute.
- From the forth column ("cycles") we extract the second the click took place in, by dividing by 1E6 and taking the quotient.

Given the above, we add up the number of click-positive seconds in each deployment-month-minute, to make the number of `click.secs` and this is saved to the file
`click.seconds.bymonth.bymin.txt`.
These have been aggregated over minutes to give deployment-months in what follows.

The data presented here have been truncated so that only data from 1st May 2011 - 30th April 2013 are presented.

Once read in, we have a total of 5.835674e+06 click positive seconds.

Interpreting patterns in the click positive seconds without correcting for effort and detectability is not particularly enlightening, but a few summary statistics are perhaps informative. Here, we give a text and graphical summary of the distribution – both show how right skewed it is. A quarter of the records are of 62 clicks or fewer, while is maximum number of clicks per month is very high, and the mean is much higher than the median (see Table 3 and Figure 4).

	location	value
1	Min.	1.00
2	1st Qu.	62.00
3	Median	344.00
4	Mean	6228.04
5	3rd Qu.	3158.00
6	Max.	179135.00

Table 3: Summary of click positive seconds per month.

2.3 Merging the effort and click files

The two datasets were merged, truncated to the period 1st May 2011 - 30th April 2013, and saved into `n.bymonth.trunc.txt`. Given this, the proportion of click-positive seconds of monitoring can be calculated. On average, over the whole dataset, given 5.835674×10^6 click positive seconds in 1.1891231×10^{10} seconds of monitoring, the proportion of click positive seconds is 0.00049075, or 0.049075%.

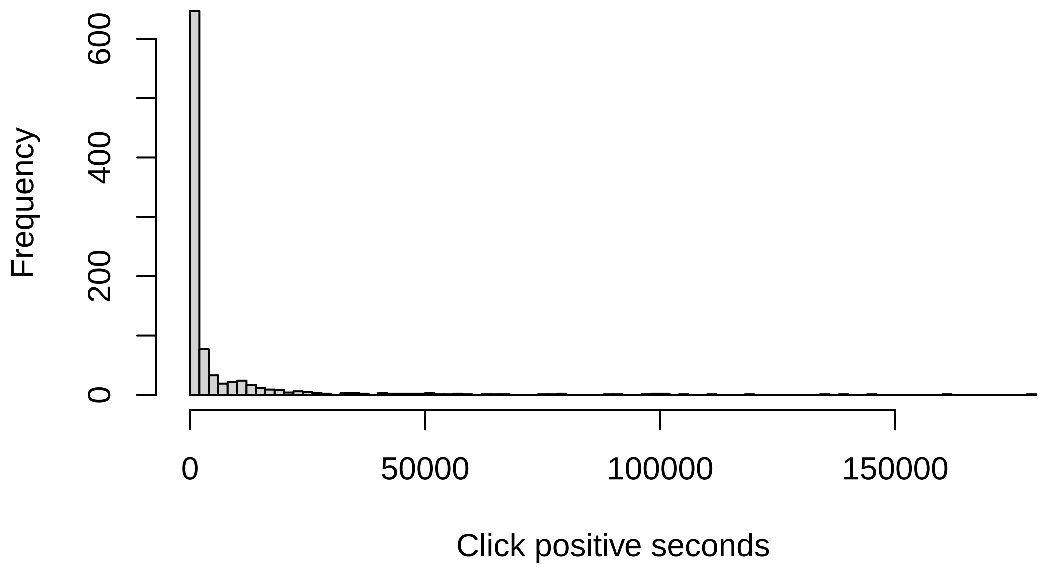

Figure 4: Histogram of click positive seconds per month.

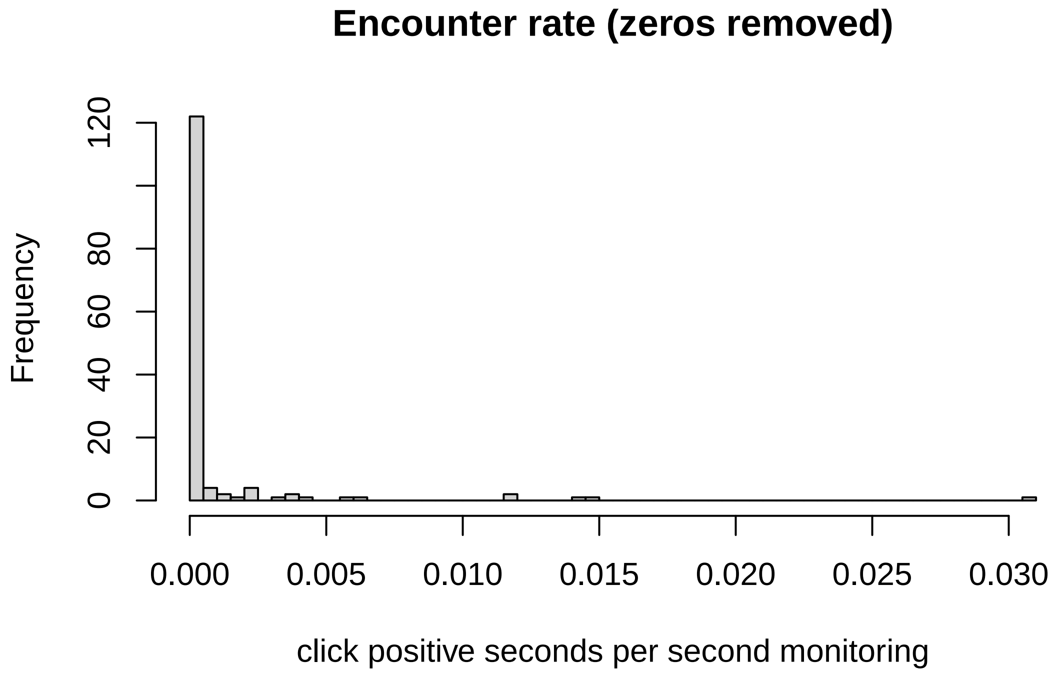

Figure 5: Histogram of encounter rate (i.e., click seconds over effort seconds) for positions with > 0 clicks.

	location	value
1	Min.	0.0000000
2	1st Qu.	0.0000017
3	Median	0.0000073
4	Mean	0.0008896
5	3rd Qu.	0.0000445
6	Max.	0.0308582

Table 4: Summary of encounter rate (click postive seconds per second of monitoring), truncated to include only positions with > 0 encounters.

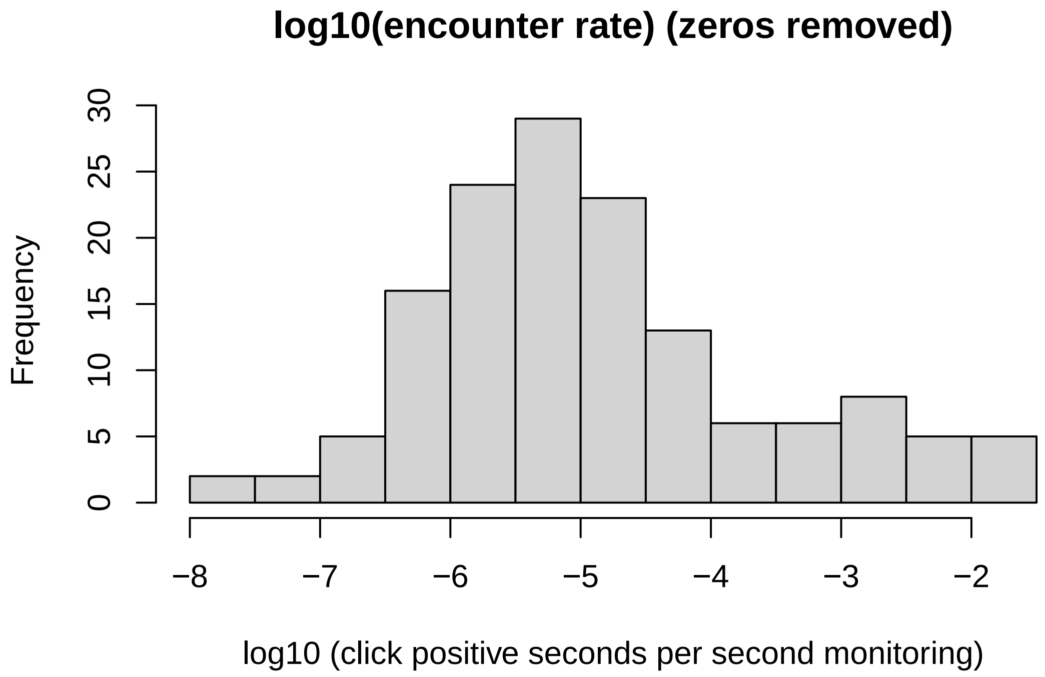

Figure 6: Histogram of log10 encounter rate (i.e., click seconds over effort seconds) for positions with > 0 clicks.

Dividing it by position, we have 6 positions with no effort, 154 that have 0 clicks counted, and 144 with at least 1 click counted. Summaries of the encounter rate where > 0 clicks counted are given in Table 4 and Figures 5 and 6.

Figures 7, 8 and 9 show how encounter rate varies over space – Figure 7 is for all data, Figure 8 is for November - April ("Winter") and 9 is for May - October ("Summer").

3 Merging in the meta file

We read in in the meta data spreadsheet. We use the file `Megametadata SAMBAH v6.csv`, which is a `csv` version of the original `xlsx` file (`csv`s are easier to read into 64 bit R). We use it for two things:

1. Check the deployment numbers and dates match what we have in the effort data file
2. Retrieve the GMT correction - we'll use the GMT correction at the start of the deployment. We need this so we can look at diurnal patterns.

Note, for this we are using the untruncated data (i.e., including all deployments, not just those in the May

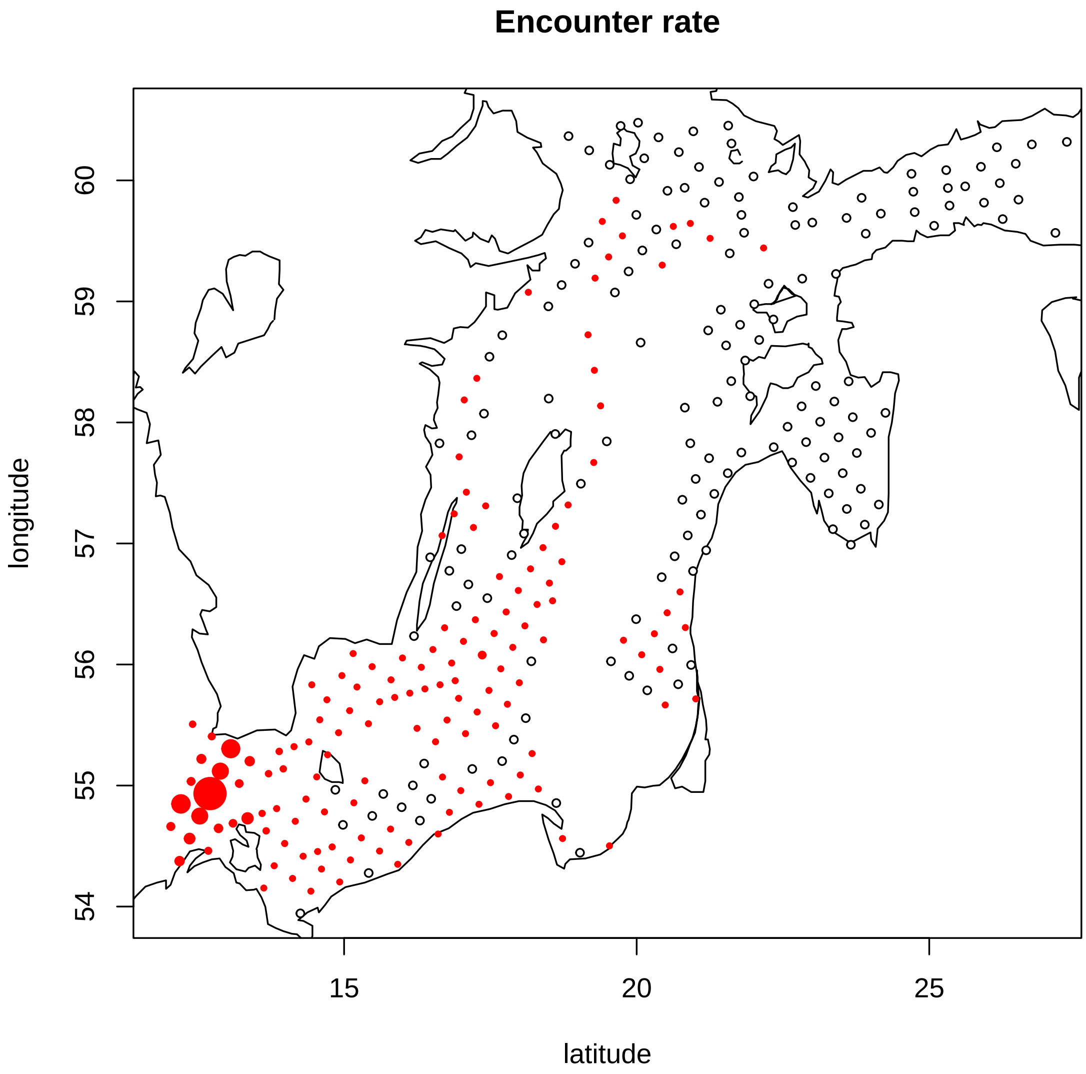

Figure 7: Map showing encounter rate (proportion of click positive seconds per second monitoring) by position. Point size is proportional to the encounter rate for the red dots (plus some offset, so that locations with almost zero encounter rate are visible); the black open circles indicate positions that were surveyed but where the encounter rate is zero.

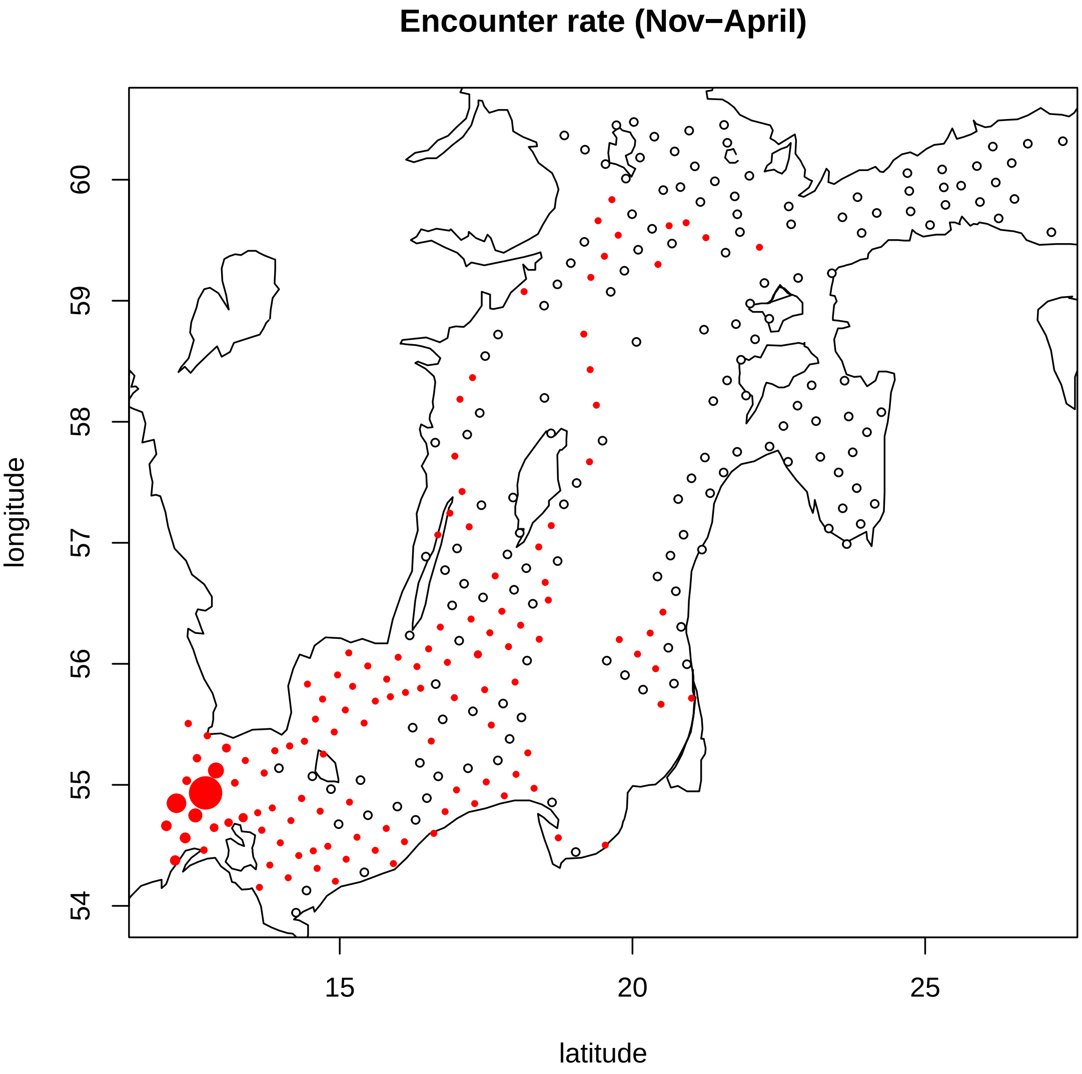

Figure 8: Map showing encounter rate by position for November - April.

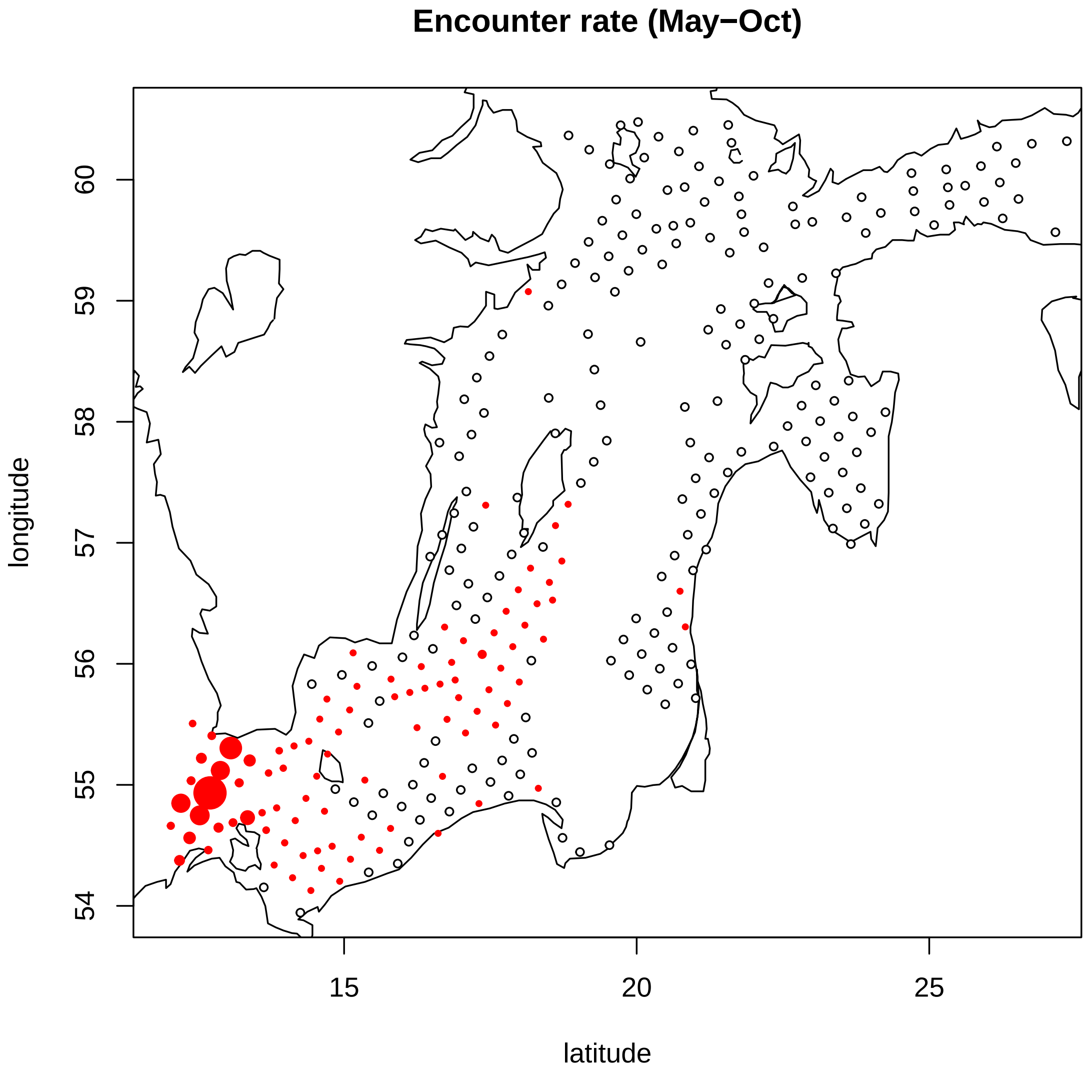

Figure 9: Map showing encounter rate by position for May - October.

2011 - April 2013 time frame). This is because we want to check all the deployments against the meta file. We go back to using truncated data later on in this report.

Regarding the first item, all we check for is missing records in the meta or effort files.

There are 0 records in the effort file that do not have any records in the meta file. (This number should be zero.)

There are 438 records in the meta file that do not have any records in the effort file. Information about these is saved into the file `NoEffort.csv`.

8 of these has a position record in the meta file but no deployments – these represent positions never surveyed. Here are the position numbers:

```
## [1] 1006 1044 1051 1067 8018 8010 8011 8012
```

The rest have deployments, but one or more of these deployments have no effort records – these are deployments that failed for one reason or another (lost C-PODS, corrupted SDs, etc.). The file `NoEffort.csv` has been checked by Daniel Wennerberg to make sure these are all correct.

4 Merging in the diel phase file

Now we read in the file of diel times, `Diel phase start times v2.csv`, and merge it with the previous data tables.

Note, we are using the date-truncated dataset for this merge.

Table 5 gives a summary of the encounter rate data by phase.

	phase	effort.secs	click.secs	er	relative.er
1	day	5461562554.20	1989148	0.0003642	1.000
2	eve	800879037.60	311268	0.0003887	1.067
3	morn	807564583.20	355208	0.0004399	1.208
4	night	4821224811.00	3180050	0.0006596	1.811

Table 5: Summary of encounter rate data by phase. First column is the total number of seconds of effort in the data, second is the total number of clicks, third is the encounter rate (i.e., number of click seconds/number of effot seconds), forth is encounter rate standardized so the smallest encounter rate has a value of 1 (just to make comparison easier).

However, any patterns seen in the above could potentially be biased, because encounter rate varies over space and time and so does the amount of daylight and night. So, it might be better to view/analyze by country/position (space) and month (time).

Here, we fit a set of models where encounter rate (per hour) is modelled as a Tweedie random variable, as a function of country, month and phase, with up to three-way interactions.

AIC for the fitted models is shown in Table 6. The AIC-best model (model 3) has a two-way interaction between country and month, plus a main effect of phase. In other words, in the most parsimonious model, encounter rate varies over large scale space and time (as we'd expect), but there is an effect of diel phase on ecounter rate that is constant over space and time. Table 7 gives the coefficients for the phase terms (day is not included as it is absorbed into the intercept term), on the log link scale (column "Estimate") and back-transformed (column "Exp(Estimate)"). Reassuringly, the results are not enormously different from those in Table 5.

	rhs	df	AIC	DeltaAIC
3	phase+country*fmonth	101	454.64	0.00
1	phase*country+fmonth	45	494.65	40.01
7	country*fmonth	98	509.85	55.21
4	phase+country+fmonth	24	539.39	84.75
2	phase*fmonth+country	57	568.63	113.99
10	country+fmonth	21	571.94	117.30
5	phase*country	34	722.25	267.62
8	phase+country	13	743.79	289.15
9	phase+fmonth	17	1176.90	722.26
6	phase*fmonth	50	1242.20	787.56

Table 6: Encounter rate model selection statistics.

	Estimate	Std. Error	t value	Pr(>|t|)	Exp(Estimate)	95 perc LCL	95 perc UCL
phaseeve	0.192	0.127	1.509	0.132	1.212	0.944	1.555
phasemorn	0.365	0.125	2.925	0.004	1.441	1.128	1.841
phasenight	0.734	0.121	6.091	0.000	2.084	1.646	2.640

Table 7: Phase coefficients from the best fitting model.