Data formatting for Russian BOSS seal surveys

May 17, 2013

This document describes data requirements for formatting Russian BOSS data into a suitable configuration for abundance estimation. We provide several optional formats that differ in the level of post-processing required. Table 1 summarizes data needs as belonging to one of five groups: Effort data, Hotspot-photo data, species misidentification data, disturbance data, and detection probability data.

Table 1: Data needs for BOSS survey analysis. Each component is described in further detail in susequent sections.

Data type	Option 1	Option 2
Effort data	Effort table	"on effort" tracks shapefile with metadata
Hotspot-photo data	Hotspot-photo grid cell data	Hotspot-photo lat-long data
Species misidentification data	Gold standard experiment	Double observer experiment
Disturbance data	Forward facing camera trials	NA
Detection probability	Distances from centerline	Independent search trials

1 Effort data

Effort data ultimately needs to link the amount of area surveyed with the underlying grid cells that are surveyed and the dates/times that surveys are conducted. This can be accomplished in two ways: (1) the Russian analysis team can calculate the effective area survey in each grid cell by date, or (2) the Russian analysis team could provide a shapefile for flight tracks, together with metadata describing the altitude, dates, and times when each flight segment was flown. If data are provided at the grid cell level, one possible format is given in Table 2. Each data type can be provided in an Excel spreadsheet or R data frame.

Table 2: Structure of effort data if provided on a grid cell basis. Example data are also provided in desired format, where "Grid_cell" designates the BOSS grid cell (provided previously), "Date_time" gives the average time while surveying was being conducted in that cell (GMT), "altitude" gives the average altitude while that cell was being surveyed, and "area_surveyed" gives the area on the ground that was covered by the infrared (thermal) swath while sampling in that cell. Multiple entries per cell should be recorded if a given cell is surveyed more than once. The rows do not have to be in any partiular order

Record	Grid_cell	Date_time	altitude(m)	area_surveyed (km^2)
1	1	2012-04-22 13:01:04	250	18.2
2	1	2012-05-01 08:57:12	288	22.4
3	236	2012-04-28 09:34:45	299	4.5
4	17	2012-05-13 15:12:54	276	30.1
5	236	2012-04-25 11:23:53	260	12.5

An alternative approach is to provide the U.S. analysis team shapefiles and enough metadata that we can construct such an effort table. For instance, we have been using tools in the R `sp` (and related) packages to overlay flight lines on the analysis grid. In this case, we would need additional metadata describing the date and time each track began and ended, the speed and altitude flown, as well as the average thermal swath width.

2 Hotspot-photo data

The next major data component we need is a data table giving all detected hotspots (not just ones where seals were detected). These would ideally be cross referenced by effort record (see above), but if hotspots had a date-time stamp and grid cell associated with them, we should be able to cross reference effort and hotspot records. There are really two options for hotspot data: (1) Russian analysts could determine which grid cell each hotspot belongs to, or (2) Russian analysts could provide latitude or longitude for each hotspot. An example of (1) is provided in Table 3.

Table 3: Structure of hotspots data if records are provided by grid cell or effort record number. `Cell` indicates the particular BOSS grid where the hotspot record was obtained (or, preferably, the 'record' number from the Effort table). `Date_time` gives date & time in GMT. `Photo` indicates whether there was (photo=1) or was not (photo=0) a digital photograph available for the hotspot. `Hotspot_type` describes whether the hotspot was a seal ("seal"), another species ("other"), or an anomaly such as a melt pool ("anomaly") [note that `hotspot_type`=NA for hotspots without accompanying photographs]. `Species` gives the particular species of phocid seal ("bearded," "ribbon," "ringed," "spotted"). The field can also be entered as "unknown" if the individual looking at photographs can tell the observation is a phocid seal but cannot determine the species; for records where `Hotspot_type` is not a seal, `Species` should be set to NA. Note also that we require one species per hotspot; if there are multiple species per hotspot, we suggest assigning the hotspot the species that is most prevalent (or in the case of a tie, making a random selection). The field `Num_seals` gives the number of seals associated with a given hotspot. Finally, `Conf` is a field giving observer confidence in their species determination ("certain" (100% positive), "likely" (51-99% confident) or "guess" (50% confident)).

Hotspot$_ID$	Cell	Date_time	Photo	Hotspot_type	Species	Num_seals	Conf
1	1	2012-04-22 12:55:34	1	seal	ribbon	2	certain
2	1	2012-04-22 12:56:20	1	seal	ribbon	1	likely
3	1	2012-04-22 12:57:54	0	NA	NA	NA	NA
4	1	2012-04-22 12:59:12	1	anomaly	NA	NA	NA
5	1	2012-04-22 13:10:23	1	other	NA	NA	NA
6	236	2012-04-28 09:30:41	1	anomaly	NA	NA	NA
7	236	2012-04-28 09:34:20	1	seal	spotted	1	certain
8	236	2012-04-28 09:37:45	1	seal	spotted	1	likely
9	236	2012-04-28 09:38:01	1	seal	bearded	1	guess
10	17	2012-05-13 15:10:15	1	seal	unknown	1	NA
11	17	2012-05-13 15:11:55	1	seal	ringed	1	likely
12	17	2012-05-13 15:11:58	0	NA	NA	NA	NA

Note that the certainty level provided is to match what we have done in the U.S. Depending on what is done to quantify misclassification (see below), it may be possible to omit this field. If it is easier to provide data in lat-long instead of by grid cell, the format would be identical except `Cell` would be replaced by `Latitude` and `Longtitude` fields.

3 Species misidentification data

At our March 2013 BOSS meeting in Seattle, we discussed possible approaches for calculating species classification probabilities. We realize the Russian analysis team may try to estimate these differently than the U.S. analysis team, given that the Russian team has access to data from hand-held high resolution cameras. If the Russian team can conclusively ascertain species from high resolution cameras, these may be used as a test for how species is classified in photographs from fixed cameras. The data would likely take one of several forms, depending upon whether photographs are routinely scored for certainty (Table 5) or not (Table 4). In both cases, we desire summary counts from records for which hand-held photographs can be matched with fixed camera photographs.

Table 4: Structure of misclasification data if high resolution photos from hand-held cameras can provide conclusive species identifications and for the case where hotspots with seals are not assigned a species uncertainty level; in this case $C^{[i|j]}$ provides a count of animals that were assigned observation type i (from examination of fixed camera photographs) when the species was truly j (as determined from hand held cameras).

	True Species			
Obs species	Bearded (1)	Ribbon (2)	Ringed (3)	Spotted (4)
Bearded (1)	$C^{[1|1]}$	$C^{[1|2]}$	$C^{[1|3]}$	$C^{[1|4]}$
Ribbon (2)	$C^{[2|1]}$	$C^{[2|2]}$	$C^{[2|3]}$	$C^{[2|4]}$
Ringed (3)	$C^{[3|1]}$	$C^{[3|2]}$	$C^{[3|3]}$	$C^{[3|4]}$
Spotted (4)	$C^{[4|1]}$	$C^{[4|2]}$	$C^{[4|3]}$	$C^{[4|4]}$
Unknown (5)	$C^{[5|1]}$	$C^{[5|2]}$	$C^{[5|3]}$	$C^{[5|4]}$

Table 5: Structure of misclassification data if high resolution photos from hand-held cameras can provide conclusive species identifications and when all hotspots with seals are assigned a species uncertainty value; in this case $C^{[i|j]}$ provides a count of animals that were assigned observation type i (from fixed cameras) when the species was truly j (as determined from hand-held cameras).

			True Species			
Obs code	Obs species	Confidence	Bearded (1)	Ribbon (2)	Ringed (3)	Spotted (4)
1	Bearded	Certain	$C^{[1|1]}$	$C^{[1|2]}$	$C^{[1|3]}$	$C^{[1|4]}$
2	Bearded	Likely	$C^{[2|1]}$	$C^{[2|2]}$	$C^{[2|3]}$	$C^{[2|4]}$
3	Bearded	Guess	$C^{[3|1]}$	$C^{[3|2]}$	$C^{[3|3]}$	$C^{[3|4]}$
4	Ribbon	Certain	$C^{[4|1]}$	$C^{[4|2]}$	$C^{[4|3]}$	$C^{[4|4]}$
5	Ribbon	Likely	$C^{[5|1]}$	$C^{[5|2]}$	$C^{[5|3]}$	$C^{[5|4]}$
6	Ribbon	Guess	$C^{[6|1]}$	$C^{[6|2]}$	$C^{[6|3]}$	$C^{[6|4]}$
7	Ringed	Certain	$C^{[7|1]}$	$C^{[7|2]}$	$C^{[7|3]}$	$C^{[7|4]}$
8	Ringed	Likely	$C^{[8|1]}$	$C^{[8|2]}$	$C^{[8|3]}$	$C^{[8|4]}$
9	Ringed	Guess	$C^{[9|1]}$	$C^{[9|2]}$	$C^{[9|3]}$	$C^{[9|4]}$
10	Spotted	Certain	$C^{[10|1]}$	$C^{[10|2]}$	$C^{[10|3]}$	$C^{[10|4]}$
11	Spotted	Likely	$C^{[11|1]}$	$C^{[11|2]}$	$C^{[11|3]}$	$C^{[11|4]}$
12	Spotted	Guess	$C^{[12|1]}$	$C^{[12|2]}$	$C^{[12|3]}$	$C^{[12|4]}$
13	Unknown	NA	$C^{[13|1]}$	$C^{[13|2]}$	$C^{[13|3]}$	$C^{[13|4]}$

If it is found that the high resolution photographs are not sufficient to enable 100% species detection certainty, it may be desirable to conduct a misclassification experiment in the same manner as the U.S. team. In particular, the U.S. team used double observer data with certainty levels to estimate misclassification rates (note that in this case a certainty level also needs to be provided for each seal detection, as suggested in Table 3). In this case a random sample of photographs could be evaluated by two observers. Each observer would independently categorize each hotspot as to species and certainty level.

4 Disturbance probability data

This is another component of detection which is likely to be assessed differently for U.S. and Russian analysis teams. During 2013 surveys, the U.S. team conducted trials by looking forward through bubble windows and recording whether seals that were first detected at or greater than a prespecified distance in front of the airplane (approximately 300m) flush into the water prior to the point at which they would have been detectable by infrared cameras, or evaded detection by moving along the ice and outside of the path of thermal cameras.

Following our March 2013 BOSS meeting in Seattle, our understanding was that the Russian team will be using forward facing video to make the same assessment. If species is identifiable from forward facing video, the data we would need could take several forms. Ideally, each trial would also be associated with an altitude so that we could correct abundance differently depending on altitude (this is especially important since the Russian 2012 and 2013 surveys are conducted at different altitudes). If altitude is available as a continuous variable, data could be provided as in Table 6.

Table 6: Structure of disturbance data if altitude is available as a continuous variable. Note that the "Flush" column is only recorded as a yes if a seal becomes unavailable to thermal detection by the Malakhit swath as a result (i.e., animals who exhibit disturbance behavior but are still detected by thermal video receive a "No"). Each row represents a single animal)

Obs species	Altitude (m)	Flush (Yes-No)
Bearded	252	No
Ribbon	255	No
Spotted	256	Yes
Spotted	255	No
Ringed	275	No
Bearded	300	No
⋮	⋮	⋮

An alternate approach would be to bin data according to altitude class (for example, one altitude class representing typical 2012 data, and one typical for 2013 data). These data could be summarized as in Table 7. We do not provide actual cutoffs for altitude bins; we imagine the Russian analysis team may have a better idea of what bins make the most biological sense given their experience. Ideally, every observation would be classified by species (it is perhaps less important to correct for species misclassification in this case).

Table 7: Structure of disturbance data if altitude is binned into different classes (here showing two altitude bins). As with Table 6, the "Num.flushed" column only represents seals that became unavailable to thermal detection by the Malakhit swath as a result of flushing.

Species	Altitude.bin	Num.trials	Num.flushed
Bearded	1	42	15
Bearded	2	125	10
Ribbon	1	137	23
Ribbon	2	387	35
Ringed	1	142	15
Ringed	2	225	7
Spotted	1	230	87
Spotted	2	691	131

5 Detection probability data

A final piece of data needed to produce abundance estimates is information to characterize detection probability, or the probability of thermal imagery detecting a seal given that it is present on ice. For example, our belief is that seals can be missed by thermal imagery if they have recently emerged from water, or if their thermal signature is not sufficiently different from surrounding media. The U.S. team has used independent searches of automated digital photographs (our photographs are taken at a constant rate and are not dependent on thermal detections) to try to estimate a detection probability. From this experiment, 66 out of 70 seals were identified by our thermal detection method.

We are unclear to what extent this issue occurs for the Russian thermal detection method, but it would be good to have data from a similar experiment. We would also like to make adjustments for distortion at the edges of the Russian thermal swath. This was another issue that came up at the March 2013 BOSS workshop in Seattle, and our impression was that the Russian team was going to decrease the swath angle to minimize distortion. However, even if this eliminates the problem for 2013 surveys, we would still need to make adjustments for 2012 surveys.

To adjust for distortion, one approach would be to provide data on the distance from the centerline where hotspots are detected. For instance, data could be provided as in Table 8.

Table 8: Detection distance data to correct abundance estimates for imperfect detection/distortion at the edge of the thermal swath. Here, "Distance" is standardized as the proportion of maximum detectable distance. Each row gives proportional distance and altitude for each detected hotspot.

Hotspot	Altitude (m)	Distance
1	250	0.71
2	255	0.24
3	256	0.03
⋮	⋮	⋮

We should then be able to fit a model to these data and correct for imperfect detction on the edge. It would be helpful to also have data to inform detection probability near the centerline. One option would be to use data from the forward facing video as trials for thermal detection; these could be summarized in the same format as the disturbance data.