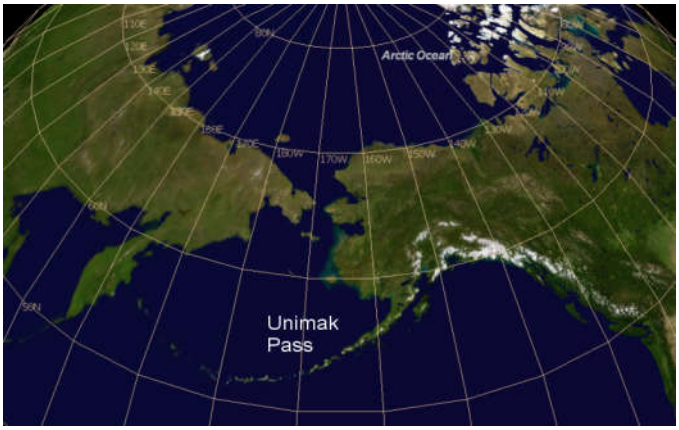
Introduction to Distance Sampling

Covariates in the detection function

This exercise consists of four datasets of increasing difficulty. The first will show you the rudiments of conducting an analysis, while the remaining analyses take you deeper into the heart of understanding multiple covariates. Feel free to complete as many as you like!

1 A whale of a dataset

Imagine you are a research biologist collecting distance sampling data during December on gray whales as they migrated through the Aleutian chain near Unimak Pass en route to their wintering grounds off Baja California (some luckier, more senior researcher, got the job of data collection on their wintering grounds). These data will now be the focus of your attention for this exercise examining the potential utility of covariates in explaining variation in animal detectability.



Detections were of individuals (not groups), and you chose to record not only distance, but also time of observation (at this latitude at this time of year, the crew was restricted to making observations between 1000 and 1500 during the day). However, because of the low sun angles during much of this time, there was some reason to believe that time of day might play a role in whale detectability. [In what manner might you wish to incorporate this covariate?]

Under extreme weather conditions, observer motion sickness can influence the performance of the observers. An additional covariate, "motion sickness tablet effective dosage at time of observation (MSTDO)" was recorded each time a whale was detected.

The data are available for your inspection in the archived Distance project **CovarWhaleSim.zip**. Notice the extreme precision with which the perpendicular distances were measured (how do you suppose this could happen on a rolling ship in the Bering Sea?).

Describe your candidate model set (what models did you construct) and your rationale for the final estimates you provide. You may also comment upon the use of time of observation as a measure of glare from oblique sun angles.

If you have been successful in performing the analysis of this dataset (which can now be revealed to have been simulated), you can continue to sharpen your skills in using covariates in your analysis of distance sampling data by exploring two other data sets, that are considerably more elaborate.

2 Golf tees Data

2.1 The data and distance project

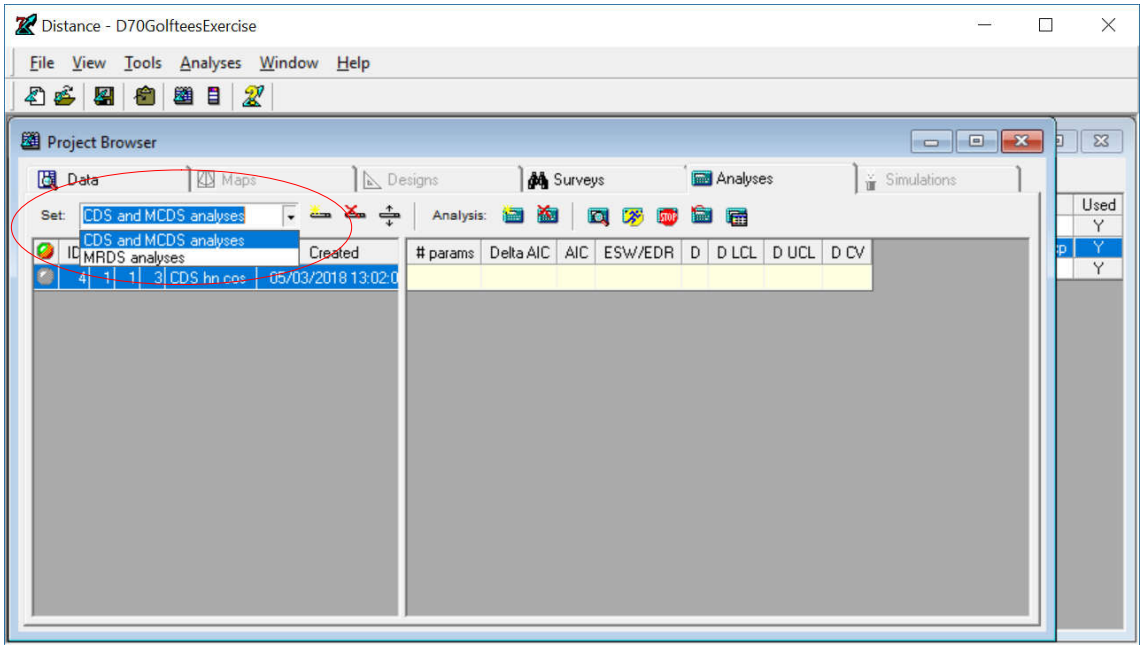
The data come from a survey of clusters of golf tees in grass, conducted by 3rd and 4th year statistics students at the University of St Andrews. It was conducted as a double platform survey; double-platform methods are covered in the Intermediate/Advanced

Workshops – we will just just use data for one observer (or platform) for the purposes of this exercise.

Assume that all the data were collected on one 210 metre long transect line, and that this comprises the study area. There were 250 clusters of tees in the study area and 760 individual tees in total.

The population was independently surveyed by two observer teams, of which we will use the data recorded by observer 1. The following data were recorded for each detected group: perpendicular distance, cluster size, observer (team 1 or 2), "sex" (males are yellow=1, females green=0 and golf tees occur in single-sex clusters), and "exposure". Exposure was a subjective judgment of whether the cluster was substantially obscured by grass (exposure=0) or not (exposure=1). The lengths of grass varied along the transect line, and the grass was also slightly more yellow along one part of the line compared with the rest.

The data are stored in the archived distance project **GolfteesExercise**. Open the project. Notice that there are already some data filters and model definitions set up. If you look in the **Analysis** tab and then at the **Set:** drop down list, you will see there are two sets of analyses already set up – named "CDS and MCDS analyses" and "MRDS analyses" (see below).



We will be working in the CDS and MCDS analyses set, using only the data recorded by observer 1.

2.2 CDS analysis of the golf tee data

A CDS analysis has been created already. Have a look at the **Data Filter** for this analysis. Notice that it uses data selection (the **Data Selection** tab) to select only observations from Observer 1 and with Detected=1 (i.e., observations detected by observer 1). (If you're curious, feel free to have a look in the **Data** tab of the **Project Browser** to see how double-platform data are set up.) Check the **Model Definition** for this analysis – it should specify a standard half-normal key function with automated selection of cosine adjustments. Run this analysis.

Look at the results (in the **Analysis** details, **Results** tab). Don't worry about the warning – this is because there is only one transect and so the encounter rate variance is estimated assuming that the observations are from a Poisson distribution so that

$\hat{V}(n) = n$ rather than from inter-transect variation. Make a note of the estimated abundance and associated coefficient of variation (CV). Also have a look at the percentage of variance that was due to the detection function.

2.3 MCDS analysis of the golf tee data

Create a new analysis and a new model definition. This time choose the MCDS analysis engine.
Check that under the **Detection function** tab, the selected key function is half normal and under the Adjustment terms button we have manual selection of zero adjustment terms. MCDS analyses are much harder for the analysis engine to fit than single covariate ones (and a different algorithm is used). In general, it is better to avoid automated selection of adjustment terms and use manual selection instead. Start with zero adjustments terms, and gradually build up 1, 2 etc. checking AIC or one of the other criteria to see if this gives a better fit. It is also a good idea to tick the option in the **Misc.** tab to 'Report results for each iteration of detection function fitting routine' (it is ticked by default for the MCDS engine) – this will help you to diagnose any problems that may occur during fitting.

There were 3 additional covariates recorded along with perpendicular distance; cluster size, sex and exposure. Obviously, sex and exposure are factor variables. Sometimes cluster size can be treated as both a factor variable or as a continuous variable: if there are only a few cluster sizes then it can be treated as a factor; however, if cluster size ranged over a large number of values it would have to be treated as a continuous variable. In this data, cluster sizes ranged from 1 to 8 and it is debatable as to whether you would want to treat it as a factor variable as there are very few large clusters detected. When including cluster size don't forget to tick the cluster size box on the **Covariates** tab – this tells Distance that this covariate is the cluster size covariate. When cluster size is included as a covariate, Distance uses a 'Horvitz-Thompson-like' estimator of abundance (this will have been covered in lectures). In this case, Distance changes a number of options in the **Estimate** and **Cluster size** tabs. In **Estimate**, it changes the 'Sample definition' option and doesn't allow stratification and in **Cluster size** it removes all the options.

Select each of these terms in turn and also in combination on the **Covariates** tab. After running a model, look at the results. The presentation of results is like that in CDS analyses, with a **Log** tab where any warnings or error messages are written, and the **Results** tab which contains details of the analysis. Make a note of the AIC value and look at the detection function plots – notice the difference in the detection function plots when the covariate is specified as a factor variable or a continuous variable.

Once you have decided on the best model, make a note of the estimated abundance, associated CV and percentage of variance accounted for by the detection function. How has this changed?

3 Dolphin Sightings Data

In this exercise there are several potential covariates and no 'right' answers!

3.1 Reviewing the data

In this example we have a sample of eastern tropical Pacific (ETP) offshore spotted dolphin sightings data, collected by observers placed on board tuna vessels (the data were kindly made available to us by the Inter-American Tropical Tuna Commission – IATTC). In the ETP, schools of yellow fin tuna commonly associate with schools of certain species of dolphins, and so vessels fishing for tuna often search for dolphins in the hopes of also locating tuna. For each school detected by the tuna vessels, the observer records the species, sighting angle and distance (later converted to perpendicular distance and truncated at 5 nautical miles), school size, and a number of covariates associated with each detected school. Many of these covariates potentially affect the detection function, as they reflect how the search was being carried out.

A variety of search methods are used to find the dolphins, but currently the most commonly used are 20x binoculars from the crow's nest, 20x binoculars from another location on the vessel, from a helicopter, or through "bird radar" (high power radars which are able to detect seabirds flying above the dolphin schools). In the example dataset these are coded as 0, 2, 3, and 5, respectively. Some of these methods may have a wider range of search than the others, and so it is possible that the effective strip width varies according to the method being used.

For each sighting the initial cue type is recorded. This may be birds flying above the school, splashes on the water, floating objects such as logs, or some other unspecified cue. In the example they have been coded as 1, 2, 4 and 3, respectively.

Another covariate that potentially affects the detection function is sea state, as measured by Beaufort. In rougher conditions (i.e. higher Beaufort levels), visibility and/or detectability may be reduced. For this example, Beaufort levels are grouped into two categories, the first including Beaufort values ranging from 0 to 2 (coded as 1) and the second containing values from 3 to 5 (coded as 2).

The sample data encompasses sightings made over a three month period: June, July and August (months 6, 7 and 8, respectively).

Begin by extracting and opening the project from the archive **Dolphin.zip**. Once it is open, you will see the **Project Browser**, from which you can have a look at the data (**Data** tab).

3.2 Analysis of Dolphin Sightings data

Start by running a set of conventional distance analyses. Are there any problems in the data and if so how might you mitigate them? (Hint – check out the q-q plot, and also try dividing the data into a large number of intervals in the Model Definition | Detection Function | Diagnostics.)

As there are a number of potential covariates to be used in this example, try fitting models with different covariates and combinations of the covariates. All of the covariates in this example are factor covariates except cluster size.

Keep in mind that this is a large dataset (> 1000 observations), and hence estimation may take a while, particularly if you are allowing up to 5 adjustment terms to be fitted. It will be generally more efficient to start fitting models without any adjustment terms, and then adding one at a time if appropriate. Consider also whether to standardize by w or by σ (or try both!).

You will likely end up with quite a few models, so think about how you are going to name and organize them in the Project Browser (for analyses) and Analysis Components window (for model definitions).

4 Passerine data from Marques et al. (2007)

The data from the Auk paper by Marques et al. (2007) is also available. It is zipped as the project **fTAMAUK07.zip**. See if you can produce results comparable to those presented in the manuscript.

Reference: Marques, T.A., L. Thomas, S.G. Fancy and S.T. Buckland. 2007. Improving estimates of bird density using multiple covariate distance sampling. The Auk 127: 1229-1243.