USGS
science for a changing world

DS – Software for Analyzing Data Collected Using Double Sampling

Open-File Report 2011–1269

U.S. Department of the Interior
U.S. Geological Survey

DS – Software for Analyzing Data Collected Using Double Sampling

By Jonathan Bart and Dana Hartley

Open-File Report 2011–1269

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
KEN SALAZAR, Secretary

U.S. Geological Survey
Marcia K. McNutt, Director

U.S. Geological Survey, Reston, Virginia: 2011

For more information on the USGS—the Federal source for science about the Earth,
its natural and living resources, natural hazards, and the environment,
visit http://www.usgs.gov or call 1–888–ASK–USGS.

For an overview of USGS information products, including maps, imagery, and publications,
visit http://www.usgs.gov/pubprod

To order this and other USGS information products, visit http://store.usgs.gov

Suggested citation:
Bart, Jonathan, and Hartley, Dana, 2011, DS – Software for analyzing data collected using double sampling: U.S. Geological Survey Open-File Report 2011-1269, 24 p.

Contents

Figures

Tables

This page left intentionally blank

DS – Software for Analyzing Data Collected Using Double Sampling

By Jonathan Bart and Dana Hartley

Introduction

DS analyzes count data to estimate density or relative density and population size when appropriate. The software is available at *http://iwcbm.dev4.fsr.com/IWCBM/default.asp?PageID=126*. The software was designed to analyze data collected using double sampling, but it also can be used to analyze index data. *DS* is not currently configured to apply distance methods or methods based on capture-recapture theory.

Double sampling for the purpose of this report means surveying a sample of locations with a rapid method of unknown accuracy and surveying a subset of these locations using a more intensive method assumed to yield unbiased estimates. "Detection ratios" are calculated as the ratio of results from rapid surveys on intensive plots to the number actually present as determined from the intensive surveys. The detection ratios are used to adjust results from the rapid surveys. The formula for density is (results from rapid survey)/(estimated detection ratio from intensive surveys). Population sizes are estimated as (density)(area). Double sampling is well-established in the survey sampling literature—see Cochran (1977) for the basic theory, Smith (1995) for applications of double sampling in waterfowl surveys, Bart and Earnst (2002, 2005) for discussions of its use in wildlife studies, and Bart and others (in press) for a detailed account of how the method was used to survey shorebirds across the arctic region of North America.

Indices are surveys that do not involve complete counts of well-defined plots or recording information to estimate detection rates (Thompson and others, 1998). In most cases, such data should not be used to estimate density or population size but, under some circumstances, may be used to compare two densities or estimate how density changes through time or across space (Williams and others, 2005). The Breeding Bird Survey (Sauer and others, 2008) provides a good example of an index survey. Surveyors record all birds detected but do not record any information, such as distance or whether each bird is recorded in subperiods, that could be used to estimate detection rates. Nonetheless, the data are widely used to estimate temporal trends and spatial patterns in abundance (Sauer and others, 2008).

DS produces estimates of density (or relative density for indices) by species and stratum. Strata are usually defined using region and habitat but other variables may be used, and the entire study area may be classified as a single stratum. Population size in each stratum and for the entire study area also is estimated for each species. For indices, the estimated totals generally are only useful if (a) plots are surveyed so that densities can be calculated and extrapolated to the entire study area and (b) if the detection rates are close to 1.0. All estimates are accompanied by standard errors (SE) and coefficients of variation (CV, that is, SE/estimate).

Preparing the Input Files for *DS*

Defining the Sampling Plan

The first step in preparing the input files for *DS* is to define the sampling plan. Seeking advice from someone with expertise in survey sampling is recommended but we cover the basics here.

DS is usually used to analyze data from surveys that employ stratification with strata defined by regions and habitats. In the rest of this guide we assume this to be the case but we note here:

1. *DS* will run with only one region, only one habitat, or only one of each (that is, no stratification). Instructions for filling out the required tables in these cases are provided below.
2. Although the words "region" and "habitat" are used in various files and column headings, strata may actually be defined in other ways (for example, by season and time of day). This point will be clearer after the input files and output tables have been described.

DS permits three sampling plans within strata: simple random (or systematic) sampling, cluster sampling, and what we call "clusters with zones," a phrase defined below. In simple random or systematic sampling, the sample locations are well-distributed across the stratum, without distinct clumping, whereas in cluster sampling they occur in well-defined groups (fig. 1). If a formal sampling plan was not used (or at least not described), you can decide whether cluster sampling was used by plotting the locations surveyed. If they occur in distinct groups as in figure 1 (right), the groups are clusters. Simple random and systematic samples are analyzed using the same formulas so in *DS*, we refer to them both just as "simple random sampling."

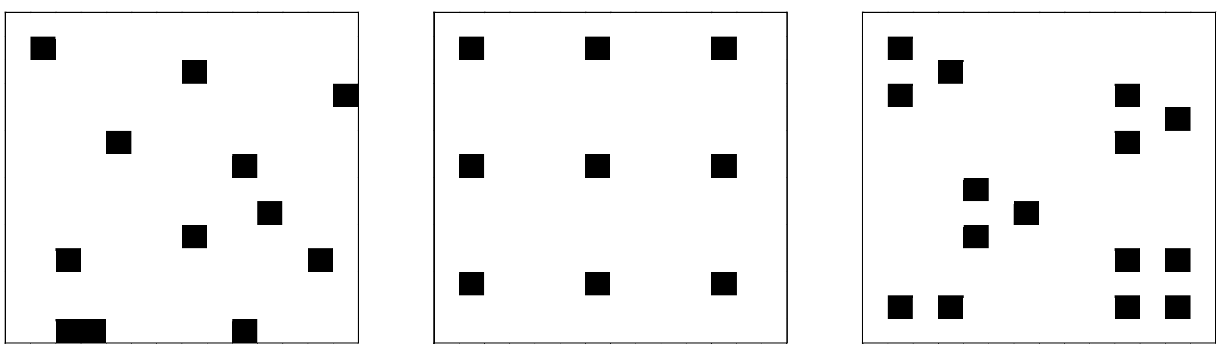

Figure 1. Examples of simple random sampling (left), systematic sampling (center), and cluster sampling (right).

The concept of zones is complex, and we recommend that you seek assistance if you think your design may include zones. The concept was introduced in the arctic (Bart, 2010). In that survey (Sample Data Set 3), the study area was subdivided into plots and plots were assigned to strata using region and habitat. Helicopters were used to transport surveyors each one or two of whom surveyed one plot. Plots surveyed at the same time needed to be fairly close together to reduce travel costs. The unusual feature of the design, from a survey sampling perspective, was that the plots at a given location were often in different habitats and thus in different strata. Usually, in stratified sampling, each cluster is within a single stratum. It turns out that having clusters contain plots in more than one stratum needs to be acknowledged in the analysis. Zones were thus defined as the set of plots that might be selected around a single helicopter location. The same issue would arise with data from the Breeding Bird Survey (BBS) if locations for 3-minute counts were assigned to different habitats and habitats were used in defining strata. Locations within clusters (routes) would thus be in different strata. For additional explanation of zones and how to analyze data when they occur, see Bart and others (in press). *DS* carries out the calculations described there.

Constructing the Input Tables

DS requires 6–11 input files depending on the design used (table 1). Most of the tables are self-explanatory but see notes below table 1 regarding the *IntKey.csv* file. Table 1d lists the tables needed if the design included intensive plots. *DS* comes with three sample data sets that may clarify how to format the input files.

It may be seen by studying table 1 that *DS* uses 14 variables as column headings. The definitions of these variables are the same in all input tables and are described below. In the definitions, the data type is listed immediately after each term in parentheses. "Numeric" means that any number may be used. "Alphanumeric" means that most standard characters may be used in the column. If *DS* encounters a character that is not allowed, it changes the character to the "_" symbol and continues processing.

Table 1. Files needed for the various designs that *DS* analyzes.

a. Sampling Plan 1: No clusters or zones

File	Columns
Habitats.csv	Habitat number, Habitat name
Regions.csv	Region number, Region name
RapCts.csv	Region number, Plot, Date, Species, Number recorded
RapHdr.csv	Region number, Habitat number, Plot, Date, Area
Species.csv	Species
Strata.csv	Region number, Habitat number, N plots, Area

b. Sampling plan 2: Clusters but not zones

File	Columns
ClAreas.csv	Region number, Habitat number, Cluster, Area
Habitats.csv	Habitat number, Habitat name
Regions.csv	Region number, Region name
RapCts.csv	Region number, Cluster, Plot, Date, Species, Number
RapHdr.csv	Region number, Cluster, Plot, Date, Area
Species.csv	Species
Strata.csv	Region number, Habitat number, N clusters, Area

c. Sampling Plan 3: Clusters and zones

File	Columns
ClAreas.csv	Region number, Habitat number, Zone, Area
Habitats.csv	Habitat number, Habitat name
Regions.csv	Region number, Region name
RapCts.csv	Region number, Plot, Date, Species, Number
RapHdr.csv	Region number, Habitat number, Zone, Plot, Date, Area
Species.csv	Species
Strata.csv	Region number, Habitat number, N clusters, Area

d. Tables for intensive surveys

File	Columns
IntAct.csv	Region number, Cluster, (Plot?), Species, Birds present
IntCts.csv	Region number, (Cluster?), Plot, Date (?), Species, Number
IndHdr.csv	Region number, Cluster, Plot, Date
IntKey.csv	Species, (*species1*), (*species2*), etc.[a]

DS uses a "headers and counts" format for the input data. This means that information about a survey, such as date and time, is kept in a "headers" table while the numbers recorded are kept in a "counts" table. For rapid surveys, the headers file is *RapHdr.csv*. The counts file is *RapCts.csv*. *DS* uses the combination *Region number-Habitat number-Plot-Date* to link records in the *Headers* and *Counts* tables. Thus the combination *Region number-Habitat number-Plot-Date*, sometimes referred to as the "Survey ID", must be unique for each survey. The definitions for column headings below provide additional discussion of how to insure that Survey IDs are unique.

Area (numeric) – area covered by a plot, cluster, zone, or stratum. Units must be the same in all tables. For example, it would not make sense to report plot sizes in hectares but stratum sizes in square kilometer. Densities in the *DS* output tables are in whatever units are used for Area.

Birds present (numeric) – number of birds (or other objects) determined by intensive surveys to be present on the intensive plots. In some projects, surveyors record partial territories so the entries in *IntAct.csv* may not be integers (for example, 1.5).

Birds recorded (numeric) – number of birds (or other objects) recorded on rapid surveys. In some projects, the entries are means from more than one count so entries may not be integers (for example, 3.5).

Cluster (alphanumeric) – name or numeric designation for a cluster. Cluster names must be unique within strata. Thus, you could number clusters sequentially starting with 1 in each stratum but you could not have clusters 1, 2, 3, ... in each zone. If the data were originally designated using the same numbers in different zones (within a stratum), then unique names must be created. For example, you could concatenate the Zone and Cluster ID and use it in the Cluster column. *Important*: Note that if clusters were not defined, then you do not include a *Cluster* column in *RapHdr.csv* or R*apHdr.csv* (table 1a) but the files for intensive plots (table 1d) always include a *Cluster* column. We did this to keep the number of different potential tables manageable. *If intensive plots were not selected in clusters, then enter the plot number in the Cluster column.*

Date (alphanumeric) – usually the date including day, month, and year. If replicate counts were made on rapid plots, then this field identifies them so the actual date (or any other unique entry for the survey) must be entered. If only one rapid survey was conducted per plot, then this field may be used to help the user in proofing the data or any value (for example, "1") may be entered in this column.

Habitat name (alphanumeric) – identifier for the habitat used in the output table. Can equal the entries in the *Habitat number* column or can be other, more descriptive, terms (for example, Wet, Moist, Dry for habitat numbers 1, 2, and 3).

Habitat number (alphanumeric) – names or numbers of the habitats used in the headers and counts tables. The reason for having both the *Habitat number* and *Habitat name* columns is that the former (for example, 1, 2, 3) are often used in the headers and counts tables but users may wish more descriptive terms (for example, Wet, Moist, Dry) in the output tables. *DS* uses the *Habitat nam*e, not the *Habitat number*, column in labeling output.

N clusters (numeric) – number of clusters in a stratum (always a whole number). This is the *total* number in the entire stratum, not the number surveyed.

N plots (numeric) – number of plots in a stratum (always a whole number). This is the *total* number in the entire stratum, not the number surveyed.

Plot (alphanumeric) – the plot identifier; part of the Survey ID (*Region number-Habitat number-Plot-Date*). Any values may be used to identify plots but we have found that short, numeric values are usually best.

Region name (alphanumeric) – identifier for the region used in the output table. Can equal the entries in *Region number* or can be other, more descriptive, terms (for example, North, South for region numbers 1 and 2).

Region number (alphanumeric) - names or numbers of the regions used in the headers and counts tables. The reason for having both the *Region number* and *Region name* columns is that the former (for example, , 1, 2) are often used in the headers and counts tables but users may wish more descriptive terms (for example, North, South) in the output tables. *DS* uses the *Region name*, not the *Region number*, column in labeling output.

Species (alphanumeric) – names, abbreviations, or numbers used to identify species (or whatever is being counted).

Zone (alphanumeric) – names or numbers used to identify zones.

Several important points may now be made about the input files.

1. Table names. You can use the default file names, defined above, or define new names for any or all input files. If the default names are used, then you need only tell *DS* the folder where they occur. Otherwise, you must tell *DS* which files to use (see "*Running DS*" below). File names may be any name valid for comma-delimited files.
3. Workbook. We recommend that you prepare an Excel or similar spreadsheet workbook with all the input tables. You can add additional columns (for example, surveyor name) in these tables that *DS* does not use, and having them all in one workbook makes them easy to review. The sample datasets follow this practice.
4. *ReadMe.doc* file. We recommend that you prepare a short *ReadMe.doc* file which briefly describes the project, and especially the sampling plan, and ideally refers to other documents that describe the project in more detail.
5. Format. All files must be in comma-delimited format (.csv) which is easily produced by Microsoft® Excel and many other programs. No empty fields are permitted.
6. Header rows. The first row in each table contains the column headings. This record helps you study the entries in the table. *DS* discards this record so it may contain any entries though we recommend entering the text as in the examples above and in the sample data sets. *Note, however, that because DS discards the header record, the columns must be in the order specified.*
7. Causes of errors. Spelling errors or inconsistencies usually cause errors in the output. For example, you cannot use "wetlands" in *RapHdr.csv* but "wetland" in *RapCts.csv*. A good way to avoid errors is to use the *Filter* command in Microsoft® Excel to examine the unique values for each column in each table.

8. Units. *DS* assumes that units are the same in all tables, and it produces output using these units. Inconsistency in units will cause errors. For example, do not enter plot area in ha but stratum sizes in square kilometers.
9. Stratum definition. *DS* assumes that strata will be defined by "region" and "habitat" and it produces output tables organized to facilitate comparisons across regions and habitats (see "*Output produced by DS*"). However, the values entered under region or habitat can be any values, they do not have to refer to actual regions or habitats. For example, weather conditions (for example. clear, cloudy) could be entered in the *Region* column and time of day (for example, morning evening) could be entered in the *Habitat* column. *DS* would then produce separate estimates for each weather-time of day combination. As noted above, you could label the columns *Weather* and *Time* rather than *Region* and *Habitat* because the header row is discarded by *DS*.
10. Single region or habitat. The files *Regions.csv* and *Habitats.csv* and the columns for region and habitat numbers and names are required. If no regions were defined, enter "All" in any column labeled *Region number* or *Region name* and enter two records in *Regions.csv*, the header row (*Region number, Region name*) and "All, All". If no habitats were defined, enter "All" in any column labeled *Habitat number* or *Habitat name* and enter two records in *Habitats.csv*, the header row (*Habitat number, Habitat name*) and "All, All". (You may use any entry, it does not have to be "All", and you may use one entry for *Region number*, *Region name*, and *Regions.csv* and a different entry for *Habitat number*, *Habitat name*, and *Habitats.csv*.)
11. Indicated pairs. Many people using double sampling record "indicated pairs" rather than birds. Thus, a nest, pair, or single male each contributes 1 to the number recorded. *DS does not convert these entries to birds.* If surveyors recorded indicated pairs but you want densities and population sizes in birds, then multiply each count by 2 (assuming you believe the sex ratio is 1:1).
12. Birds present. *DS* expects count data to be the numbers of birds in a plot, not densities. It calculates densities itself using plot area. Although ways can be devised to use densities as counts, we recommend using counts.
13. Means per survey. If more than one rapid survey was carried out on a given plot, *DS* computes the simple mean of the survey results. As noted above, in such cases the survey date (or some other value distinguishing surveys) must be entered in the *Date* column.

The IntKey.csv table

This file is only present if intensive plots were surveyed (and are being used in the analysis). The file is a table with species names as both rows and columns. During the analysis, *DS* reads one row at a time and completes the analysis for the species named in column 1 of this row (referred to as the focal species). One step in the analysis is to calculate the estimated detection ratio (birds recorded on the rapid surveys of intensive plots)/(birds present on the intensive plots as determined by the intensive surveys). With abundant species, this calculation may be based solely on the focal species. With less common species, too few data are available for a meaningful estimate based solely on data (from the intensive plots) for the focal species. For such species, *DS* needs to know which species to use in calculating the detection ratio for the focal species. It obtains this information from *IntKey.csv*. A "1" in the focal species' row means "use data for the species at the top of this column"; a "0" means "do not use data for this species."

An example may help clarify how to construct the *IntKey.csv* file. Table 2 shows the numbers of four species recorded on intensive plots. Opinions vary on how much data should be available in estimating detection rates. Let us assume that the analyst wishes to have a sample size of at least 100 (that is, 50 indicated pairs in many studies) to estimate each detection ratio. Fewer than 100 were recorded of the first three species. The analyst believes that the two plovers may have similar detection rates. By combining them, the analyst obtains a sample of 124. The analyst suspects that the detection rate for dunlin is more similar to semipalmated sandpipers than to the plovers, however, data for semipalmated sandpipers are sufficient so data from other species do not have to be used. Table 3 reflects these choices.

Table 2. Sample data from intensive plots for four species.

Species	Number recorded on intensive plots
American golden-plover	46
Black-bellied plover	78
Dunlin	62
Semipalmated sandpiper	374

Table 3. Sample *IntKey.csv* files for the data in table 2.

Species	American golden-plover	Black-bellied plover	Dunlin	Semipalmated sandpiper
American golden-plover	1	1	0	0
Black-bellied plover	1	1	0	0
Dunlin	0	0	1	1
Semipalmated sandpiper	0	0	0	1

IntKey.csv should be constructed using sample sizes, knowledge of the species, and estimated standard errors (SEs) of the detection rates. Adopting a formal model-selection procedure based on AICs or analysis of variance is often desirable especially if numerous analyses are being conducted. Our experience has been that calculating detection rates based on fewer than 100 individuals is seldom worthwhile.

Validations Carried Out by *DS*

DS carries out extensive checks on the input files to identify errors or inconsistencies that might cause unnoticed errors in the output (table 4). Understanding the validation procedure will help you construct input tables that *DS* will accept. The checks vary between designs in fairly obvious ways. For example, if intensive plots were not surveyed then none of the checks on *IntHdr.csv*, *IntCts.csv*, or *IntAct.csv* are made, and if clusters are not used then none of the checks involving clusters are made.

Table 4. Validations made by *DS* prior to starting the analysis.

File	Validation	No clusters or zones	Clusters but no zones	Clusters and zones
All	The file is not empty	x	x	x
	All entries are valid characters	x	x	x
Regions.csv	Values in the *Region number* and *Region name* columns are unique	x	x	x
Habitats.csv	Values in the *Habitat number* and *Habitat name* columns are unique	x	x	x
Species.csv	Values in the *Species* column are unique	x	x	x
RpdHdrs.csv	The *Region number* column contains all values in the *Region number* column of *Regions.csv*	x	x	x
	Each *Region number-Habitat number* combination occurs in *ClAreas.csv*		x	x
	Each *Region number-Habitat number-Zone* combination occurs in *ClAreas.csv*			x
RpdCts.csv	Each value in the *Plot* column occurs in *RpdHdrs.csv*	x	x	x
IntCts.csv	>1 record occurs in each *Region*	x		
	>1 record occurs in each *Region-Cluster* combination		x	x
IntAct.csv	All entries in the *Plot* column occur in the *Plot* column of *IntHdrs.csv*	x	x	x
	>1 record occurs in each *Region*	x		
	Each Survey ID occurs in *IntHdrs.csv*	x		
	>1 record occurs in each *Region-Cluster* combination		x	x
	Each *Region-Cluster-Plot* combination occurs in *IntHdrs.csv*		x	x
StraData.csv	Each *Region-Habitat* combination is unique	x	x	x
ClAreas.csv	Each value in the *Region* column occurs in the *Region* column of *Regions.csv*	x	x	
	Each value in the *Habitat* column occurs in the *Habitat* column of *Habitats.csv*	x	x	
	Each *Region-Habitat* combination is unique	x		
	Each *Region-Habitat-Cluster* combination is unique		x	
	Each *Region-Habitat-Zone-Cluster* combination is unique			x

Running *DS*

Once *DS* has been installed and the input files have been created or modified, you are ready to run *DS*. When you start *DS* (usually by double clicking on the *DS* icon), it produces a welcome screen with a list of projects (fig. 2).

Welcome Screen for *DS*

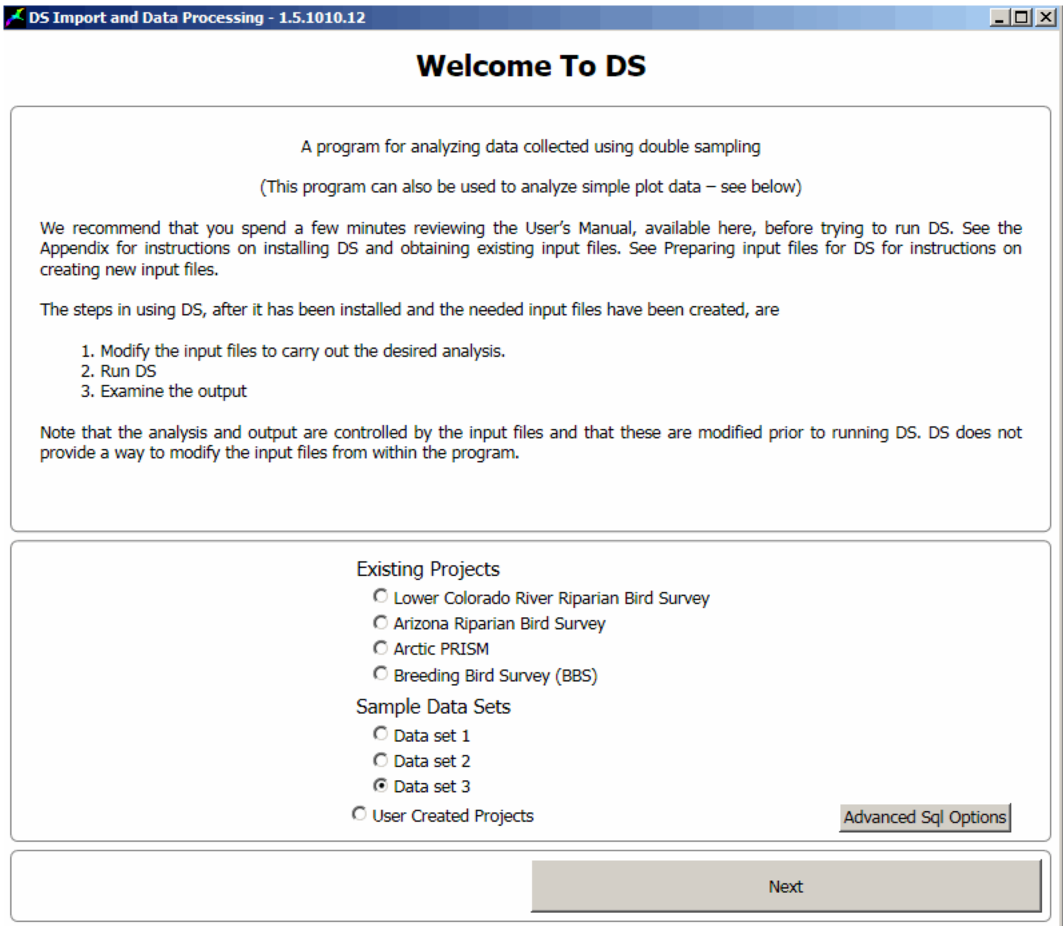

Figure 2. Welcome screen for *DS*.

Existing projects refers to projects for which the input files have been constructed by personnel associated with the project. *DS* has built-in instructions concerning what files, and columns in each file, are needed for these projects. Each project also has a page briefly describing the project and providing links to more information about the project. The input files for some of these are shipped with *DS*. The input files for others must be requested from the project managers.

Sample data sets are small datasets shipped with *DS* intended to illustrate input and output. Descriptions of the three datasets may be obtained by clicking on their names.

User-created projects are projects for which the needed input files do not ship with *DS* and *DS* does not contain any built-in instructions for how to analyze the data. Someone associated with the project must prepare the input files and document how *DS* should access them in a *ReadMe.doc* file (see section, “Preparing Input Files for DS”).

Choose one of the projects. If you click on an existing project or a sample dataset, a page will display telling you a little about the project and giving you links for more information. When you are ready to continue, click *Next*. If you click on *User-defined project* a screen will display asking you to define the sampling plan (and giving you three choices) and to state whether intensive plots were surveyed. This information should be available in a *ReadMe.doc* file stored with the input files. Answering these questions tells *DS* what input files it needs and what columns each of them must contain. If you have SQL Server on your system and would like to use it instead of the default, click the *Advanced Sql Opt*ions Button. Once you have provided the needed information, click *Next*.

After you click *Next*, a page will display (fig. 3) on which you identify the input and output files and provide other information needed by *DS* to carry out your analysis. If you have used the default file names (see section, “Preparing Input Files for DS”), then click the top button under “*Choose Files*” and navigate to the correct folder. *DS* then looks in this folder for the default file names. If it does not find them, it displays an error message at the right hand side of the screen listing the files it did not find. If it finds all required files, no message is displayed.

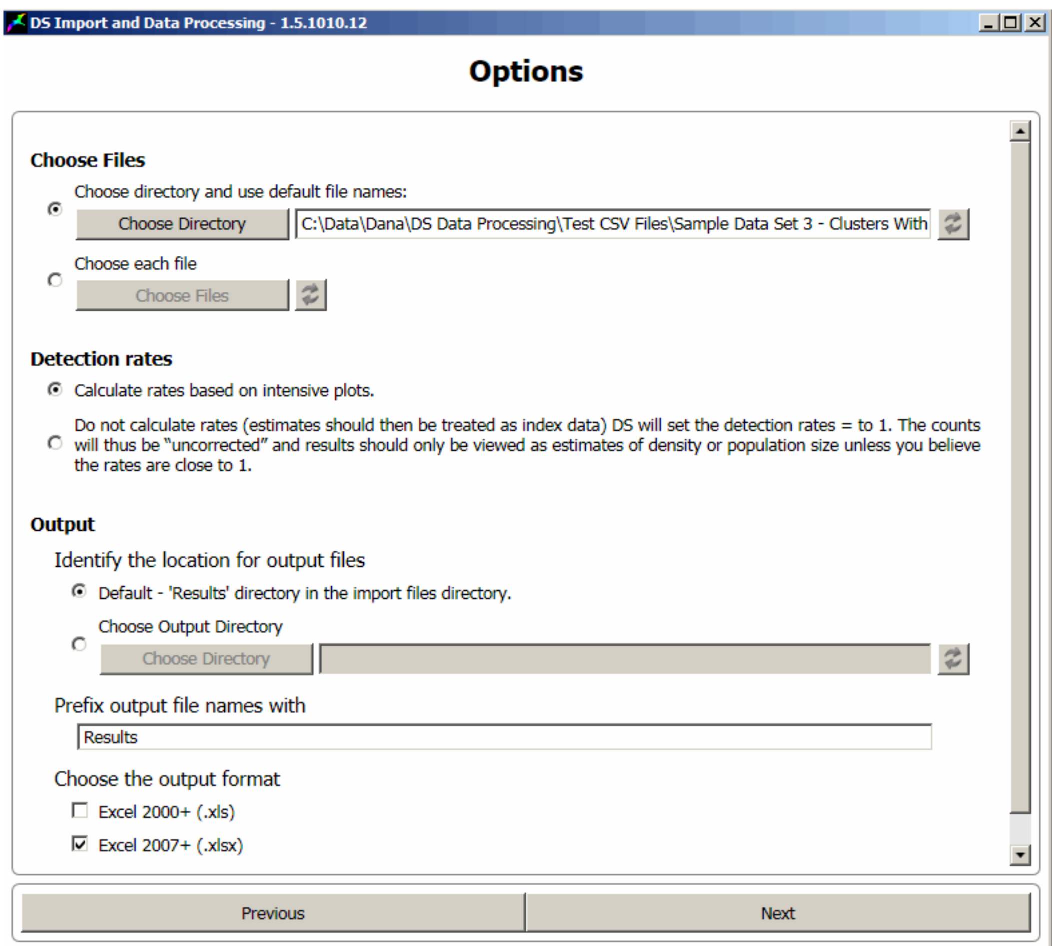

Figure 3. The *Running DS* screen.

If you did not use the default file names, then click on the second button under "*Choose Files.*" *DS* then opens a window listing the needed files on the left and asking you to identify each of them (fig. 4). Initially, all files have a red circle on the right, indicating that the file has not been identified. As you identify the files, these symbols disappear. When you have identified all the files, click "*OK*" and the previous screen will appear (if you have not identified all needed input files a red circle with an "!" appears to the right of the "*Choose Files*" button).

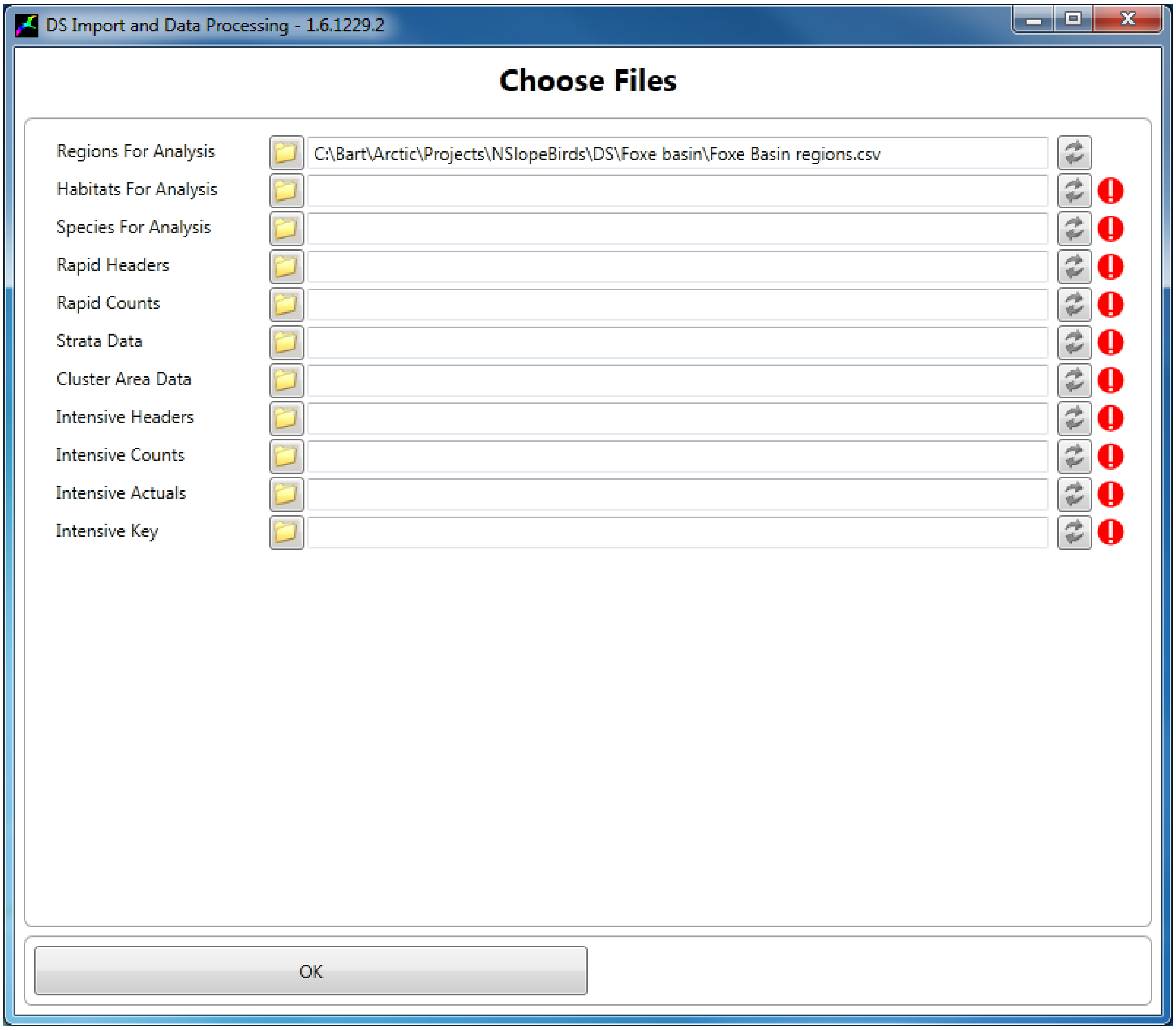

Figure 4. *DS* screen used to select files with non-default names. The file listing region names has been selected.

The “Options”screen also asks whether you want detection rates calculated and gives you some options for how the output will be delivered. When you have answered the questions on this screen, click *Next* and *DS* will begin the analysis. It displays numerous messages documenting progress. After operations cease, these messages can be inspected to determine if any errors occurred. At the end of a successful analysis, *DS* writes the results to the output directory you have selected. Each file name begins with a date and time stamp and a prefix if you have specified one.

Output

The output depends on the project design. The most extensive output occurs when more than one species, region, and habitat are included in the *Species.csv*, *Regions.csv*, and *Habitats.csv* files and the design includes intensive plots. We refer to this as the “full output” and describe it below, noting which tables or parts of tables are not included when only one species, region, or habitat is defined or if intensive plots are not included. We use the phrase “estimates of density and population sizes” below, but, as noted above, if detection rates were not estimated, the results should usually be treated as index data, that is “relative” not absolute densities and population sizes.

The full output includes six tables either in an Microsoft® Excel workbook or written as separate comma-delimited files. Two tables contain the estimated densities and population sizes by habitat and by both region and habitat. One contains the detection ratios. Three provide sample size information. Each file is described below.

Estimates_1 – estimated densities by habitat (table 5). For each habitat, and for all habitats combined, *DS* prints the estimated density, its estimated SE and CV, and the number of birds recorded on rapid surveys. The estimated population sizes, for the entire study area, along with their SEs, CVs, and lower and upper bounds for 95% confidence intervals are also reported. This table does not give region-specific estimates (see *Estimates_2.csv*).

Table 5. Example of Estimates_1 (slightly reformatted).

	Wetlands					All habitats						95% CI	
Species	Density	SE	CV	N birds		Density	SE	N birds	Pop size	SE	CV	LCB	UCB
AMGP	1.01	0.7	0.68	86		2.29	0.7	159	282,249	84,902	0.3	448,657	115,841
BARG	0.46	0.3	0.72	43		0.66	0.4	66	81,208	54,336	0.7	187,706	0
BASA	0.49	0.2	0.37	4	data for two habitats omitted	0.19	0.1	7	22,797	15,650	0.7	53,470	0
BBPL	4.03	0.5	0.12	146		2.13	0.5	189	262,733	66,564	0.3	393,199	132,267
...													
All	2.35	1	0.43	532		1.34	0.3	735	1,035,478	110,659	0.1	1,252,370	818,586

Estimates_2 – identical to *Estimates_1.csv* except that estimates are given for each region (table 6). The 2nd and 3rd columns are *Region number* and *Region name*. All other columns are the same as in *Estimates_1*.

Table 6. Example of Estimates 2 (slightly re-formatted)

Species	Regon number	Region name	Wetlands Density	SE	CV	N birds	
AMGP	13.1	Arctic NWR	2.44	0.78	0.32	11	Other columns identical to Estimates_1
AMGP	13.2	Central NS	1.05	1.1	1.05	1	
AMGP	13.3	Foothills east	0.00	0.00	-	0	
AMGP	14.1	NPRA NE	2.86	0.68	0.24	62	
AMGP	14.2	NPRA coast	2.49	0.82	0.33	12	
AMGP	14.3	Foothills west	0.00	0.00	-	0	
AMGP	14.4	Colville River	0.05	0.03	0.60	1	
All	All	All	10.32	31.2	3.02	235	

Detection Ratios – estimated densities from the rapid surveys and detection ratios from the intensive plots with SEs, CVs, and numbers of birds recorded (table 7). Comparing the CVs is often useful for considering whether to change the allocation of effort between rapid plots and intensive plots. This table is omitted if there were no intensive plots. In the example below, *IntKey.csv* was filled with 1s so all species were used in computing estimates for each species (see *Using DS*).

Table 7. Example of Detection Ratios (slightly re-formatted).

	Data from rapid plots				Data from intensive plots			
Species	Density	SE	CV	N birds	Detection ratio	SE	CV	N birds
AMGP	1.85	0.5	0.3	159	0.81	0.1	0.1	2928
BARG	0.53	0.4	0.7	66	0.81	0.1	0.1	2928
BASA	0.15	0.1	0.7	7	0.81	0.1	0.1	2928
...	1.72	0.4	0.2	189	0.81	0.1	0.1	2928
All	0.28	0.1	0.4	25	0.81	0.1	0.1	2928

Sample sizes_1 – for each region, the number of plots surveyed, the area they covered, and the area covered by all plots in the stratum (table 8). Also, the same information for all habitats in the region (right hand 3 columns).

Table 8. Example of Sample Sizes 1 (slightly re-formatted).

		Wetlands			Uplands			Region		
Region ID	**Region name**	**N plots**	**Surveyed area**	**Stratum area**	**N plots**	**Surveyed area**	**Stratum area**	**N plots**	**Surveyed area**	**Stratum area**
13.1	Arctic NWR	85	9.9	1207	70	3.7	1353	179	23.3	4793
13.2	Central NS	15	2	2398	26	0	266	42	6	10826
13.3	Foothills east	2	0.3	68	8	2.2	1266	24	3.6	2727
...										
All	All	330	98.3	15034	230	234.2	165237	695	375.7	209514

Sample sizes_2 - numbers reported on rapid surveys by species and region (table 9). The final column gives the sum for all species. The final row gives the sum for all regions.

Table 9. Example of Sample sizes 2 (slightly re-formatted).

Region Number	**Region Name**	**AMGP**	**BARG**	**BASA**	**BBPL**	**BBSA**	**BLTU**		**WRSA**	**All**
13.1	Arctic NWR	29		4	0	8			0	428
13.2	Central NS	8	1	1	5	3		several species omitted		190
13.3	Foothills east	1	0		1				0	16
...										
All	All	652	153	23	330	65	237	________	23	20434

Sample sizes 3 - mean numbers recorded on rapid surveys, and actual numbers present, on intensive plots (table 10). This table is omitted if there were no intensive surveys. It has one record for each cluster of intensive camps (which is the same as each plot if intensive plots are not selected in clusters – see *Preparing the input files for DS*). *DS* calculates the mean number recorded per rapid survey on each plot. The sum of these numbers is reported for each cluster as is the sum of the numbers actually present on each plot in the cluster. Statistical analyses are carried out on these totals, rather than on plot-specific results, to avoid pseudoreplication. The final line gives sums for all regions, species, and clusters.

Table 10. Example of Sample Sizes 3 (slightly re-formatted).

Region ID	Region name	Species	Cluster	Estimate	Actual
13.1	Arctic NWR	AMGP	1	1	2
13.1	Arctic NWR	AMGP	2	1	0
14.1	NPRA NE	AMGP	1	2.8	2
14.1	NPRA NE	AMGP	2	0.1	0
14.1	NPRA NE	AMGP	3	3	5
14.1	NPRA NE	BARG	1	2	2
...					
All	All	All	All	43	52

Estimation

This section is adapted from Bart and others (in press). It presents estimators for density and population size, and their SEs, for the most general case (>1 region, >1 habitat, clusters, and zones) and then discusses how the formulas simplify if restrictions (for example, 1 habitat, no zones) are added. For a given species, the estimate of population density is calculated as

$$d = \frac{\hat{\bar{X}}}{\hat{R}} \tag{1}$$

where $\hat{\bar{X}}$ is an estimate of the density of birds that would have been recorded if an indefinitely large sample of rapid surveys had been conducted and $\hat{R}$ is an estimate of the detection ratio (birds recorded/birds present) on the rapid surveys. $\hat{\bar{X}}$ is obtained from the rapid surveys; $\hat{R}$ is obtained from the intensive surveys. From the standard equation for the estimated variance of a ratio of independent random variables (Cochran, 1977, chapter 6),

$$\hat{V}(d) = d^2\left(\frac{\hat{V}(\hat{\bar{X}})}{\hat{\bar{X}}^2} + \frac{\hat{V}(\hat{R})}{\hat{R}^2}\right) \tag{2}$$

The estimated population size is

$$\hat{Y} = Ad \tag{3}$$

where A is the size of the study area. The variance of $\hat{Y}$ is estimated as

$$\hat{V}(\hat{Y}) = A^2\hat{V}(d). \tag{4}$$

Estimators for the terms in expressions (1) – (4) are derived below. They depend on the sampling plan used to select plots within strata. We start with the most general case, clusters that may contain plots in different habitats (and thus in different strata) with unequal sized plots, and then present the simpler cases.

Cluster Sampling within Strata

To derive estimators for the terms in expression (2), let
$\bar{z}_{uhi}$ = mean number of birds recorded per plot in the i^{th} cluster of habitat *h* plots in region *u*
$\bar{b}_{uhi}$ = mean area covered per surveyed plots in the i^{th} cluster of habitat *h* plots in region *u*
a_{uhi} = area covered by all plots in the i^{th} cluster of habitat *h* plots in region *u*
n_{uh} = number of clusters of habitat *h* plots surveyed in region *u*
N_{uh} = number of clusters of habitat *h* plots in region *u*

$\hat{\hat{X}}$ is estimated using the "combined approach" (Cochran, 1977) for ratios with stratification

$$\hat{\hat{X}} = \frac{\hat{Z}}{a} = \frac{\sum_u^U \sum_h^H \frac{N_{uh}}{n_{uh}} \sum_i^{n_{uh}} a_{uhi}\left(\bar{z}_{uhi} / \bar{b}_{uhi}\right)}{\sum_u^U \sum_h^H \frac{N_{uh}}{n_{uh}} \sum_i^{n_{uh}} a_{uhi}} = \frac{\sum_u^U \sum_h^H \frac{N_{uh}}{n_{uh}} \sum_i^{n_{uh}} \hat{Z}_{uhi}}{\sum_u^U \sum_h^H \frac{N_{uh}}{n_{uh}} \sum_i^{n_{uh}} a_{uhi}} \tag{5}$$

The quantity $\hat{Z}_{uhi} = a_{uhi}(\bar{z}_{uhi} / \bar{b}_{uhi})$ is the estimated number of birds that would be recorded if all the plots were surveyed in the i^{th} cluster of habitat *h* plots in region *u*. The numerator in (5) is the mean of the $\hat{Z}_{uhi}$ times the number, N_{uh}, of clusters in the stratum. The numerator may thus be viewed as an estimate of the number of birds that would be recorded on rapid surveys if all plots in all regions were surveyed. The denominator is an estimate of the total area, based on the surveyed plots. The ratio is thus an estimate of density (uncorrected for the detection rate). The rationale for this estimator may be explained as follows. If the mean area of the surveyed zones is larger than the mean area of all zones, then the numerator will tend to exceed the true number of birds present, but the denominator will also tend to be greater than the true area so the ratio will tend to be closer to the actual density.
It will be convenient in deriving the variance estimator to express $\hat{\hat{X}}$ as

$$\hat{\hat{X}} = \frac{\hat{Z}}{a} = \frac{\sum_u^U \sum_h^H N_{uh}\hat{\bar{Z}}_{uh}}{\sum_u^U \sum_h^H N_{uh}\bar{a}_{uh}} \tag{6}$$

where $\hat{\bar{Z}}_{uh}$ and $\bar{a}_{uh}$ are the means of the $\hat{Z}_{uhi}$ and a_{uhi}, respectively, in stratum *u-h* (*i.e.*, region *u* and habitat *h*). $V(\hat{\hat{X}})$ was estimated using the standard formula for the estimated variance of a ratio of correlated random variables (Cochran 1977, Chapter 6),

$$\hat{V}(\hat{\hat{X}}) = \left(\frac{\hat{Z}}{a}\right)^2 \left(\frac{\hat{V}\left(\hat{Z}\right)}{\hat{Z}^2} + \frac{\hat{V}\left(a\right)}{a^2} - \frac{2C\hat{o}v(\hat{Z},a)}{\hat{Z}a}\right) \tag{7}$$

As noted above, sampling in different strata was not independent. To acknowledge this dependence, let the subscripts g and h indicate habitat, n_{ugh} = the number of zones in region u in which at least one habitat g plot and one habitat h plot were surveyed, and $\hat{\bar{Z}}_{ug|b}$ = the mean of the $\hat{Z}_{ugi}$ among the n_{ugh} plots. The "|b" (for "both") notation means the sum is restricted to zones in which both habitats g and h were surveyed. Let $\hat{\bar{Z}}_{uh|b}$, $\bar{a}_{ug|b}$ and $\bar{a}_{uh|b}$ be defined in a similar manner. With this notation, and using standard survey sampling methods, it may be shown that

$$\hat{V}(\hat{Z}) = \sum_{u}^{U}\sum_{g}^{H}\sum_{h}^{H} N_{ug}N_{uh}\frac{n_{ugh}}{n_{ug}n_{uh}}\sum_{i}^{n_{ugh}}(\hat{Z}_{ugi} - \hat{\bar{Z}}_{ug|b})(\hat{Z}_{uhi} - \hat{\bar{Z}}_{uh|b})/(n_{ugh} - 1), \tag{8}$$

$$\hat{V}(a) = \sum_{u}^{U}\sum_{g}^{H}\sum_{h}^{H} N_{ug}N_{uh}\frac{n_{ugh}}{n_{ug}n_{uh}}\sum_{i}^{n_{ugh}}(a_{ugi} - \bar{a}_{ug|b})(a_{uhi} - \bar{a}_{uh|b})/(n_{ugh} - 1), \tag{9}$$

$$C\hat{o}v(\hat{Z},a) = \sum_{u}^{U}\sum_{g}^{H}\sum_{h}^{H} N_{ug}N_{uh}\frac{n_{ugh}}{n_{ug}n_{uh}}\sum_{i}^{n_{ugh}}(\hat{Z}_{ugi} - \hat{\bar{Z}}_{ug|b})(a_{uhi} - \bar{a}_{uh|b})/(n_{ugh} - 1). \tag{10}$$

When n_{ugh} = 1, the corresponding term in expressions (8) – (10) cannot be evaluated even though the variance or covariance the term represents does exist. Models could be used to estimate the missing terms (*e.g.* using the mean of the terms for which n_{ugh} >1) but we have not investigated this approach and therefore omitted terms in expressions (8) – (10) when n_{ugh} was 1.

The detection ratio, R, is estimated as

$$\hat{R} = \frac{\bar{x}}{\bar{y}} \tag{11}$$

where $\bar{x}$ is the mean number of birds recorded on rapid surveys of the intensive plots and $\bar{y}$ is the mean number of birds determined to be present on these plots through intensive surveys.

Let m = the number of clusters, $\bar{x}_i$ = the mean number recorded per rapid survey at cluster i, and $\bar{y}_i$ = the mean number actually present at all plots in cluster i, then $\bar{x} = \sum \bar{x}_i / m$ and $\bar{y} = \sum \bar{y}_i / m$. The $\bar{x}_i$ were calculated as the simple means of the means/plot because sometimes plots in a cluster were not all surveyed the same number of times by rapid surveyors. Clusters were widely distributed across the study area so they were treated as a simple random sample. Under this assumption, the variance of $\hat{R}$ may be estimated as

$$\hat{V}(\hat{R}) = \hat{R}^2 \left(\frac{\hat{V}(\bar{x})}{\bar{x}^2} + \frac{\hat{V}(\bar{y})}{\bar{y}^2} - \frac{2C\hat{o}v(\bar{x},\bar{y})}{\bar{x}\bar{y}}\right) \tag{12}$$

where

$$\hat{V}(\bar{x}) = \frac{1}{m}s^2(\bar{x}_i), \quad \hat{V}(\bar{y}) = \frac{1}{m}s^2(\bar{y}_i), \quad C\hat{o}v(\bar{x},\bar{y}) = \frac{1}{m}cov(\bar{x}_i,\bar{y}_i)$$

and s^2 and $c\hat{o}v$ indicate the sample variance and covariance respectively.
If plots are the same size, then a is a constant so expression (5) becomes

$$\hat{\hat{X}} = \frac{\hat{Z}}{a} = \frac{1}{a}\sum_u^U\sum_h^H\frac{N_{uh}}{n_{uh}}\sum_i^{n_{uh}}a_{uhi}\left(\bar{z}_{uhi}/\bar{b}_{uhi}\right) = \frac{1}{a}\sum_u^U\sum_h^H\frac{N_{uh}}{n_{uh}}\sum_i^{n_{uh}}\hat{Z}_{uhi} \tag{13}$$

Expressions (9) and (10) equal 0 so expression (7) becomes

$$\hat{V}(\hat{\hat{X}}) = \frac{1}{a^2}\sum_u^U\sum_g^H\sum_h^H N_{ug}N_{uh}\frac{n_{ugh}}{n_{ug}n_{uh}}\sum_i^{n_{ugh}}(\hat{Z}_{ugi} - \hat{\hat{Z}}_{ug|b})(\hat{Z}_{uhi} - \hat{\hat{Z}}_{uh|b})/(n_{ugh}-1), \tag{14}$$

In summary, with cluster sampling in strata, the point estimate of density is given by expression (1) with $\hat{\hat{X}}$ given by expression (5) and $\hat{R}$ given by expression (11). With plots that vary in size, the $V(\hat{\hat{X}})$ is given by expression (7) with the equations for the components given by expressions (6) and (8) - (10). If plots are all the same size then $\hat{\hat{X}}$ is given by expression (13) and $V(\hat{\hat{X}})$ is given by expression (14). In either case, the $V(\hat{R})$ is given by expression (12).

Estimates for a single habitat or region are easily derived from the expressions above. Within a single stratum (denoted using the subscripts u and h) we have simple random sampling. From (6), $\hat{\hat{X}}_{uh} = \hat{\hat{Z}}_{uh} / \bar{a}_{uh}$. The estimated variance has the same structure as expression (7) but $\hat{V}(\hat{\hat{Z}}_{uh}) = s^2(\hat{Z}_{uhi})/n_{uh}$, $\hat{V}(\bar{a}_{uh}) = s^2(a_{uhi})/n_{uh}$. For the mean within one habitat, across >1 region, expressions (6) – (10) apply except that g and h are equal and constant so expressions (8) – (10) simplify:

$$\hat{V}(\hat{Z}) = \sum_u^U N_{uh}^2 s^2(\hat{Z}_{uhi})/n_{uh} \tag{15}$$

$$\hat{V}(a) = \sum_u^U N_{uh}^2 s^2(a_{uhi})/n_{uh} \tag{16}$$

$$C\hat{o}v(\hat{Z},a) = \sum_u^U N_{uh}^2 cov(\hat{Z}_{uhi},a_{uhi})/n_{uh} \tag{17}$$

The estimate for >1 habitat within a single region is provided directly by expressions (6) – (10) with the sums in u having a single value. For all three cases, the equations for the detection rate and its variance do not change except that the expressions may have fewer terms.

Clusters but No Zones

"Clusters but no zones" means that all plots in each cluster are in the same habitat and thus the same stratum. This is the normal approach in survey sampling. The only change in the estimation equations is that in expression (8) – (10) g and h are always equal so that the expressions become

$$\hat{V}(\hat{Z}) = \sum_u^U\sum_h^H N_{uh}^2 s^2(\hat{Z}_{uhi})/n_{uh} \tag{18}$$

$$\hat{V}(a) = \sum_{u}^{U} \sum_{h}^{H} N_{uh}^{2} s^{2}(a_{uhi}) / n_{uh} \tag{19}$$

$$C\hat{o}v(\hat{Z}, a) = \sum_{u}^{U} \sum_{h}^{H} N_{uh}^{2} cov(\hat{Z}, a_{uhi}) / n_{uh} \tag{20}$$

For estimates within a single stratum, the double sums and N_{uh} in expression (6) drop out leaving

$$\hat{\bar{X}}_{uh} = \frac{\hat{\bar{Z}}_{uh}}{\bar{a}_{uh}} \tag{21}$$

and the terms in expression (7) become

$$\hat{V}(\hat{Z}_{uh}) = s^{2}(\bar{z}_{uhi}) / n_{uh} \tag{22}$$

$$\hat{V}(a_{uh}) = s^{2}(\bar{a}_{uhi}) / n_{uh} \tag{23}$$

$$C\hat{o}v(\hat{Z}_{uh}, a_{uh}) = cov(\bar{z}_{uhi}, \bar{a}_{uhi}) / n_{uh} \tag{24}$$

Expressions (11) and (12) for the detection ratio remain the same.

Simple Random Sampling within Strata

If clusters are not defined, then we do not use the a_{uhi}. Expression (5) becomes

$$\hat{\hat{X}} = \frac{\hat{Z}}{a} = \frac{\sum_{u}^{U}\sum_{h}^{H}\frac{N_{uh}}{n_{uh}}\sum_{i}^{n_{uh}} z_{uhi}}{\sum_{u}^{U}\sum_{h}^{H}\frac{N_{uh}}{n_{uh}}\sum_{i}^{n_{uh}} b_{uhi}} \tag{25}$$

where z_{uhi} and b_{uhi} are the mean number of birds recorded and the plot size for the i^{th} plot in habitat h and region u. Expression (7) remains the same but the terms are calculated as follows

$$\hat{V}(\hat{Z}) = \sum_{u}^{U}\sum_{h}^{H} N_{uh}^2 s^2(z_{uhi}) / n_{uh} \tag{26}$$

$$\hat{V}(a) = \sum_{u}^{U}\sum_{h}^{H} N_{uh}^2 s^2(b_{uhi}) / n_{uh} \tag{27}$$

$$C\hat{o}v(\hat{Z}, a) = \sum_{u}^{U}\sum_{h}^{H} N_{uh}^2 cov(z_{uhi}, b_{uhi}) / n_{uh} \tag{28}$$

Expressions (11) and (12) for the detection ratio remain the same except that $\overline{x}_i$ and $\overline{y}_i$ are no longer means and may be replaced by x_i and y_i the estimated and actual numbers present on the i^{th} plot.
For estimates within a single stratum, the double sums disappear and the N_{uh} may be cancelled yielding

$$\hat{\hat{X}}_{uh} = \frac{\hat{Z}}{a} = \frac{\frac{1}{n_{uh}}\sum_{i}^{n_{uh}} z_{uhi}}{\frac{1}{n_{uh}}\sum_{i}^{n_{uh}} b_{uhi}} \tag{29}$$

and in expression (7)

$$\hat{V}(\hat{Z}_{uh}) = s^2(z_{uhi}) / n_{uh} \tag{30}$$

$$\hat{V}(a_{uh}) = s^2(a_{uhi}) / n_{uh} \tag{31}$$

$$C\hat{o}v(\hat{Z}, a) = cov(z_{uhi}, a_{uhi}) / n_{uh} \tag{32}$$

References Cited

Bart, J., and Earnst, S.L., 2002, Double sampling to estimate density and population trends in birds: Auk, v. 119, p. 36-45.

Bart, J., and Earnst, S.L., 2005, Methods for shorebird surveys in the arctic, *in* Ralph, C.J., and Rich, T.R., eds., Bird Conservation Implementation and Integration in the Americas: Proceedings of the Third International Partners in Flight Conference, March 20–24, 2002: Asilomar, California, Volume 2. General Technical Report PSW-GTR-191: Albany, CA, Pacific Southwest Research Station, Forest Service, U.S. Department of Agriculture, 643 p.

Bart, J., Johnston, V.J., Smith, P,A., Manning, A., Rausch, J., and Brown, S., in press, Chapter 2. Methods, *in* Bart, J., and Johnston, V.J., Arctic shorebirds in North America: results of a ten-year monitoring program: Studies in Avian Biology.

Cochran, W.G., 1977, Sampling techniques: New York, NY, Wiley.

Sauer, J.R., Hines, J.E., and Fallon, J., 2008, The North American Breeding Bird Survey, Results and Analysis 1966 – 2007, Version 5.15.2008: Laurel, MD, accessed September 12, 2011, at http://www.pwrc.usgs.gov/.

Smith, G., 1995, A critical review of the aerial and ground surveys of breeding waterfowl in North America, Technical report, USFWS, accessed September 12, 2011, at http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA322667&Location=U2&doc=GetTRDoc.pdf.

Thompson, W.L., White, G.C., and Gowan, C., 1998, Monitoring vertebrate populations: New York, NY, Academic Press.

Williams, B.K., Nichols, J.D., and Conroy, M.J., 2005, Analysis and management of animal populations: New York, NY, Academic Press.

Appendix A. Installing *DS*.

Installing *DS* is a two step process if you do not already have SQL Server running on your computer.

1. Install SQL Server Express

   DS using SQL Server to process most of its data. If you do not have SQL Server on your computer or network, then the easiest way to run *DS* is to install the free program SQL Server Express onto your computer. The SQL Server Express installer was either shipped to you with the *DS* installer or you can download it directly from here:

   http://go.microsoft.com/fwlink/?linkid=65212

   Run the installer and accept all of the default options.

14. Install *DS*

   There are two way to install – Manual or Automatic.

   Automatic

   If you are using the automatic installer, you will have received a .exe or .msi install file. Run the install file and follow the onscreen options. This will install *DS* program files and add a shortcut link to your Start Menu and/or Desktop

   Manual

   DS also can be distributed as a zip file. Simply unzip the file and place *DS* directly anywhere on your computer. Then open the folder and click the *DS*_ImportAndProcess.exe file to begin. Note that this installation method will not add a shortcut link to your Start Menu or Desktop. This method is particularly useful if you have to install the program on a computer where you do not have permissions to run an installation program.

NOTE: If you already have a version of SQL Server Express on your computer but still cannot get *DS* to run correctly, try uninstalling SQL Server Express and reinstalling it using the link above. That will reset it back to all default parameters which should allow *DS* to access the program.

Contact Information

If you are having trouble with this program, please feel free to contact Leah Dunn at (208) 426-2696, *ldboise@gmail.com*, or Jon Bart at (208) 426-5216), *jon_bart@usgs.gov*.

This page left intentionally blank

Publishing support provided by the U.S. Geological Survey
Publishing Network, Tacoma Publishing Service Center

For more information concerning the research in this report, contact the
Director, Forest and Rangeland Ecosystem Science Center
U.S. Geological Survey
777 NW 9th Street, Suite 400
Corvallis, Oregon, 97330
http://fresc.usgs.gov/

Bart and Hartley—***DS*** **– Software for Analyzing Data Collected Using Double Sampling**—Open-File Report 2011–1269