

Preparing your data to use in the zoolog package



In many applications, the most time consuming step is normally preparing the data. R and RStudio are very robust tools, but they also need information in a dataset to be unambiguous in order to work properly. Zooarchaeological datasets 'in the wild' normally require some taming before they can be manipulated in R. 'Cleaning' data into an organized and standardized dataset can be done in R, but we normally find it easier to work in Excel or Access, and then save the dataset as a CSV file for importing into R.

This PDF has instructions on how to tame your zooarchaeological dataset for use with the zoolog package in R.

In summary, your dataset should have:

- Column names conforming with R naming conventions
- 1 column with taxon/species names
- 1 column with skeletal element names
- Columns for each measurement. Measurement names must exactly match the von den Driesch abbreviation. Measurements each need their own column (the only exception is for astragalus GLL, which can be included in GL or GLL column).
- Measurements should only be numbers – delete any marks like *- ()
- Remove all semi-colons ';' from the dataset. Remove also all commas ',' if your language uses periods '.' as decimal points.
- Save as a CSV file

**Data sets
in tutorials**



**Data sets in
the wild**



Remember to sign up to a personal 2-hour session on Wednesday 20th January if you want individual help getting your data into zoolog:

https://doodle.com/poll/qgr9kgbn2k978qrf?utm_source=poll&utm_medium=link

1. General organization

First open your file in Excel / similar software. For the data to work with zoolog, we need:

- 1 column containing the taxon/species names
- 1 column containing the skeletal element names
- 1 column for each measurements (Bd, Dd, DD, BatF, etc.). The name of each measurement column must match the von den Driesch abbreviation. the only exception is for astragalus GLI, which can be included in GL or GLI column.

Here is some example data. There is a species column (TAX) and an element column (EL). It is OK if the names of these column are different in your dataset.

However, when recording the measurements I put scapula GLP in the GL column to save time (here in red)

Biometry ID	Site name	Area	TAX	EL	GL	Bp	DD	GLI	GLm	SLC	BT	Individual	Indi dup	Precision	Comments
247	Marzabotto	V; 3	B.T.	AS				59.4*							Check, typo?
248	Casale di Rivalta		B.T.	SC	62.2					46.9					
250	Rubiera	Ca del Cristo	B.T.	SC	59.3					44.8					
251	Case Vandelli		B.T.	SC						(47.4)					
252	Bologna	Via Ca Selvatica	B.T.	SC						46**					
253	Marzabotto	US 8	B.T.	SC						57.7		A			
254	Marzabotto	US 8	B.T.	HU							60.0	A			

GLP now needs its own column, so I will add it to the table and put the GLP measurements in it. **Do this for any other measurements that you have combined.**

Biometry ID	Site name	Area	TAX	EL	GL	Bp	DD	GLI	GLm	GLP	SLC	BT	Individual	Indi dup	Precision	Comments
247	Marzabotto	V; 3	B.T.	AS				59.4*								Check, typo?
248	Casale di Rivalta		B.T.	SC						62.2	46.9					
250	Rubiera	Ca del Cristo	B.T.	SC						59.3	44.8					
251	Case Vandelli		B.T.	SC							(47.4)					
252	Bologna	Via Ca Selvatica	B.T.	SC							46**					
253	Marzabotto	US 8	B.T.	SC							57.7		A			
254	Marzabotto	US 8	B.T.	HU								60.0	A			



Check that your column names match the von den Driesch abbreviations. You can see I have a 'DD' column. I will want to check that this really is DD (depth of the diaphysis) and not Dd (depth of the distal epiphysis).

I'm going to leave in all my extra columns with Site name, Comments, and any information on period etc. because I will want to see this information after I calculate the log ratios in R. You will have a LOT of columns. This is normal!

2. Column names

Column names should be compatible with R naming conventions:

- Avoid names with special symbols: ?, \$, *, +, #, (,), -, /, }, {, |, >, < etc. However underscore '_' and period '.' can be used.
- Column names must be unique. Duplicated names are not allowed.
- R is case sensitive. This means that Name is different from Name or NAME
- The R functions read.table and read.csv will automatically replace blank spaces in column names (header) by periods ('.'). Thus, you may want avoid names with blank spaces in headers. Example: names like "Site_name" or "Site.name" will not be changed, but "Site name" will be automatically converted to "Site.name".

Looking at the table below, we can see there are 3 columns with spaces in the name.

Biometry ID	Site name	Area	TAX	EL	GL	Bp	DD	GLI	GLm	GLP	SLC	BT	Individual	Indi dup	Precision	Comments
247	Marzabotto	V; 3	B.T.	AS				59.4*								Check, typo?

I can change these to better fit the conventions by adding an underscore.

Biometry_ID	Site_name	Area	TAX	EL	GL	Bp	DD	GLI	GLm	GLP	SLC	BT	Individual	Indi_dup	Precision	Comments
247	Marzabotto	V; 3	B.T.	AS				59.4*								Check, typo?

3. Taxon and skeletal element names

Now we need to check that zoolog will understand the names we use for different species and body parts. The zoolog pack contains a Thesaurus that will recognize common variations and abbreviations of different species and body parts.

Check that your species and elements names are in the lists below.

It is OK if the capitalization, punctuation, accented characters, and spacing are different, but the main characters need to be the same.

For example: 'Bos', 'bos', and 'BOS' are all equivalent. Similarly, 'B.T.', 'BT', and 'B. T.' are equivalent, and 'Phalanx 2 ant./post.' is equivalent to 'Phalanx2antpost'.

If you are using one of these names for different taxon or element, you will need to change that name to something else. For example: If I were using 'SS' for wild boar and not domestic pig, I would need to change my code for wild boar from 'SS' to something else like 'SSwild'. Otherwise zoolog will think it is a domestic pig!

If you need to add categories to Thesaurus, you can use the WriteThesaurus function in zoolog. Please let us know about the changes at svalenzuela@imf.csic.es so we can include them in the next versions of the package. Your name will be added to the list of contributors as you will help to make the package more robust and comprehensive.

Taxa

bos taurus	ovis aries	capra hircus	ovis capra	sus domesticus	cervus elaphus
bota	ovar	cahi	oc	sudo	ceel
bos	ovis	capra	caprine	pig	cervus
cattle	sheep	goat	ovis/capra	sus	red deer
BT	OA	CH	s/g	SS	CE
bovino	oveja	cabra	sheep/goat	cerdo	ciervo
Grands Bovides Boeuf	Ovicaprines Mouton	pecora	Sh/G	Suides Porc	Cervides Cerf
vaca	ov	cabra	ovicaprino	S	CEE
	ovella	CAH	O/C		cervol
	OVA		Ovicaprines Ovis-Capra		ciervo
			ovicapri		
			O		

Elements

scapula	humerus	radius	metacarpus	metacarpus III	metacarpus IV	pelvis
SC	HU	RA	MC	metacarpal III	metacarpal IV	PE
scapula	hum	rad	metac	MC III	MC IV	coxal
scapola	Humerus	radi	MC1	metac. III	metac. IV	cox
esc	humer	radio	MC2	MC3	MC4	innominate
scap.	omero	Membre-anterieur radius	metacarp			
escap.	Membre-anterieur humerus		metacarpo			
			Metacarpe Canon			
			metacarpal			

femur	tibia	calcaneum	astragalus	metatarsus	metatarsus III	metatarsus IV
FE	TI	CAL	AS	MT	metatarsal III	metatarsal IV
fem	TIB	calca	talus	metat	MT III	MT IV
femore	Membre-posterieur tibia	Calcani	AST	MT1	metat. III	metat. IV
Membre-posterieur Femur		calcan	astragalo	MT2	MT3	MT4
		calcagno	astrag	metatars		
		Tarse calcaneum	Tarse Talus	metatarso		
		Calcaneus		Metatarse Canon		
				metatarsal		

anterior first phalanx	first phalanx	posterior first phalanx	anterior second phalanx	second phalanx	posterior second phalanx
Phalange 1 ant	Phalange 1	Phalange 1 post	Phalange 2 ant	Phalange 2	Phalange 2 post
Phalanx 1 ant.	Phalanx 1	Phalanx 1 post.	Phalanx 2 ant.	Phalanx 2	Phalanx 2 post.
P1 ant	P1	P1 post	P2 ant	P2	P2 post
falange 1 ant	falange 1	falange 1 post	falange 2 ant	falange 2	falange 2 post
PH1A	PH1	PH1P	PH2A	PH2	PH2P
phal 1 ant	phal 1	phal 1 post	phal 2 ant	phal 2	phal 2 post
Phalange I anterior	Phalange I	Phalange I post	Phalange II anterior	Phalange II	Phalange II posterior
fal1 ant	fal1	fal1 post	fal2 ant	fal 2	fal 2 post
ant 1fal	1fal	post 1fal	ant 2fal	2fal	post 2fal
ant fal 1	fal 1	post fal 1	ant fal 2	fal 2	post fal 2
1 fal ant	1 fal	1 fal post	2 fal ant	2 fal	2 fal post

4. Check measurements are all numbers and only numbers

In order to function, all the measurements must be stored as numbers, not as text.

- Check that all measurements are a single number.
- Remove any other characters from the cells – for example, they cannot contain -*() or spaces

You can see in my example data below, I have some cells with * and () that I used to indicate estimated measurements.

Biometry_ID	Site_name	Area	TAX	EL	GL	Bp	DD	GLI	GLm	GLP	SLC	BT	Individual	Indi_dup	Precision	Comments
247	Marzabotto	V; 3	Bos	AS				59.4*								Check, typo?
248	Casale di Rivalta		Bos	SC						62.2	46.9					
250	Rubiera	Ca del Cristo	Bos	SC						59.3	44.8					
251	Case Vandelli		Bos	SC							(47.4)					
252	Bologna	Via Ca Selvatica	Bos	SC							46**					
253	Marzabotto	US 8	Bos	SC							57.7		A			
254	Marzabotto	US 8	Bos	HU								60.0	A			

Now you have two options:

- If I only want to use the most precise measurements, I can delete these rows.
- If I want to keep the measurements and also the information that they are estimated, I can put this information into another column. Here I have marked that a measurement is estimated in the 'Precision' column. I also have one measurement from an old report that did not take very precise measurements. I have also put this in the Precision column. Now only numbers are stored in the Measurement columns.

Biometry_ID	Site_name	Area	TAX	EL	GL	Bp	DD	GLI	GLm	GLP	SLC	BT	Individual	Indi_dup	Precision	Comments
247	Marzabotto	V; 3	Bos	AS				59.4							Estimated	Check, typo?
248	Casale di Rivalta		Bos	SC						62.2	46.9					
250	Rubiera	Ca del Cristo	Bos	SC						59.3	44.8					
251	Case Vandelli		Bos	SC							47.4				Estimated	
252	Bologna	Via Ca Selvatica	Bos	SC							46				Rounded to 0.5	
253	Marzabotto	US 8	Bos	SC							57.7		A			
254	Marzabotto	US 8	Bos	HU								60.0	A			

5. Separators

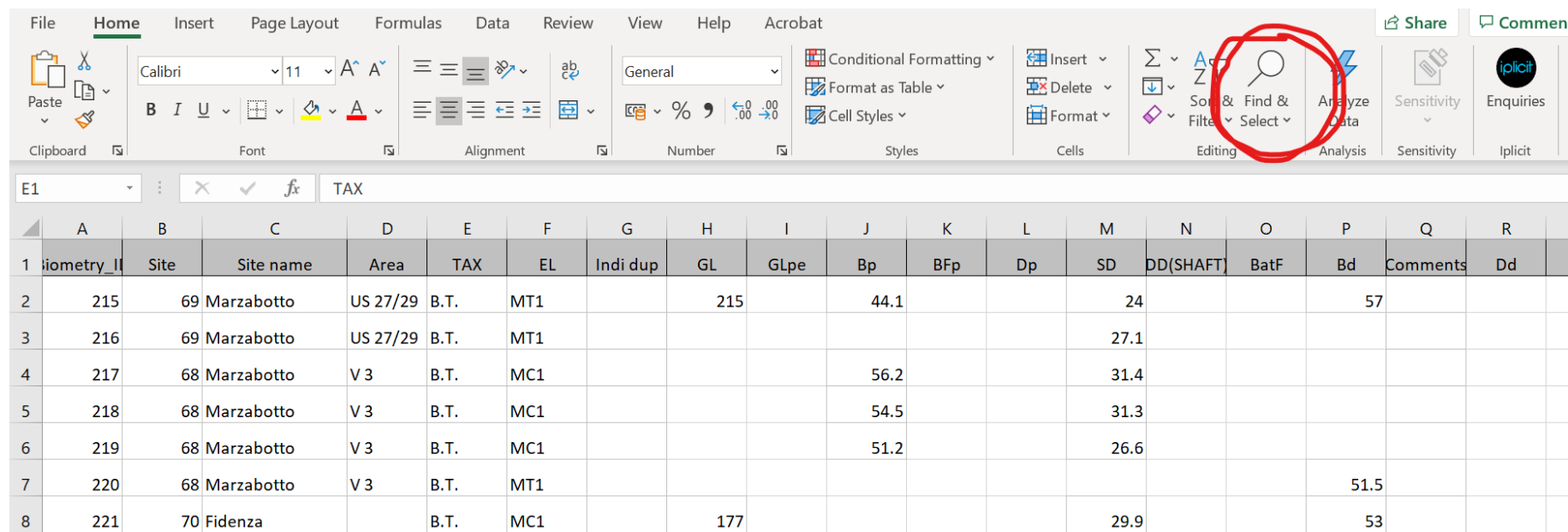
Remove all semi-colons ';' from your dataset. Remove also all commas ',' if your language uses periods '.' as decimal points.

You can use the 'Replace' function to change them to '.' or '_' or just take them out.

In my dataset, I've got a few cells that have ';' and ',' in them.

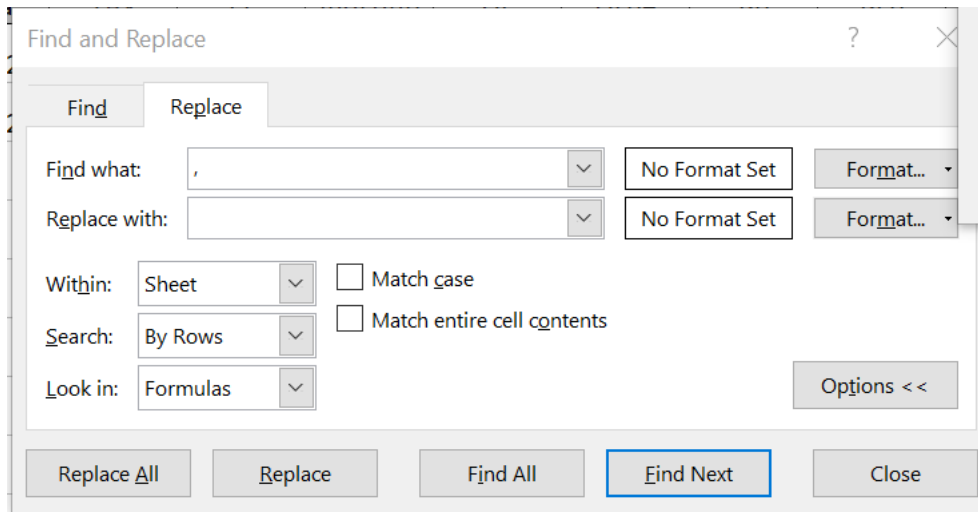
Biometry_ID	Site_name	Area	TAX	EL	GL	Bp	DD	GLI	GLm	GLP	SLC	BT	Individual	Indi_dup	Precision	Comments
247	Marzabotto	V; 3	Bos	AS				59.4							Estimated	Check, typo?
248	Casale di Rivalta		Bos	SC						62.2	46.9					
250	Rubiera	Ca del Cristo	Bos	SC						59.3	44.8					
251	Case Vandelli		Bos	SC							47.4				Estimated	
252	Bologna	Via Ca Selvatica	Bos	SC							46				Rounded to 0.5	
253	Marzabotto	US 8	Bos	SC							57.7		A			
254	Marzabotto	US 8	Bos	HU								60.0	A			

I'm just going to take them out using the Replace function (found under Find & Select in the Home tab). Uncheck the box 'match entire cell contents'.



The screenshot shows the Microsoft Excel interface with the 'Home' tab selected. The 'Find & Select' dropdown menu is open, and the 'Find & Select' option is circled in red. Below the ribbon, a portion of an Excel spreadsheet is visible, showing columns A through R and rows 1 through 8. The spreadsheet contains data for biometry sites, including Marzabotto and Fidenza.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Biometry_ID	Site	Site name	Area	TAX	EL	Indi dup	GL	GLpe	Bp	BFp	Dp	SD	DD(SHAFT)	BatF	Bd	Comments	Dd
2	215	69	Marzabotto	US 27/29	B.T.	MT1		215		44.1			24			57		
3	216	69	Marzabotto	US 27/29	B.T.	MT1							27.1					
4	217	68	Marzabotto	V 3	B.T.	MC1				56.2			31.4					
5	218	68	Marzabotto	V 3	B.T.	MC1				54.5			31.3					
6	219	68	Marzabotto	V 3	B.T.	MC1				51.2			26.6					
7	220	68	Marzabotto	V 3	B.T.	MT1										51.5		
8	221	70	Fidenza		B.T.	MC1		177					29.9			53		



We need to remove these marks because we will be saving our dataset as a CSV file. This is a text file type that allows information to be saved in tables. In the CSV text file, the cells of the table are separated by commas ‘,’ (or by semi-colons ‘;’ if your language uses commas as decimal points).

For example, in a CSV file this text would produce this table:

Marzabotto, V3, Bos, AS, 26.1, Check this

Marzabotto	V3	Bos	AS	26.1	Check this
------------	----	-----	----	------	------------

Or if my version of Excel uses ‘;’ for the decimal:

Marzabotto; V3; Bos; AS; 26,1; Check

Marzabotto	V3	Bos	AS	26,1	Check this
------------	----	-----	----	------	------------

However, if I have commas in my text, the CSV will mess up the data by making a new cell every time there is a comma. We don’t want this!

Marzabotto, V,3, Bos, AS, 26.1, Check, this

Marzabotto	V	3	Bos	AS	26.1	Check	this
------------	---	---	-----	----	------	-------	------

6. Saving

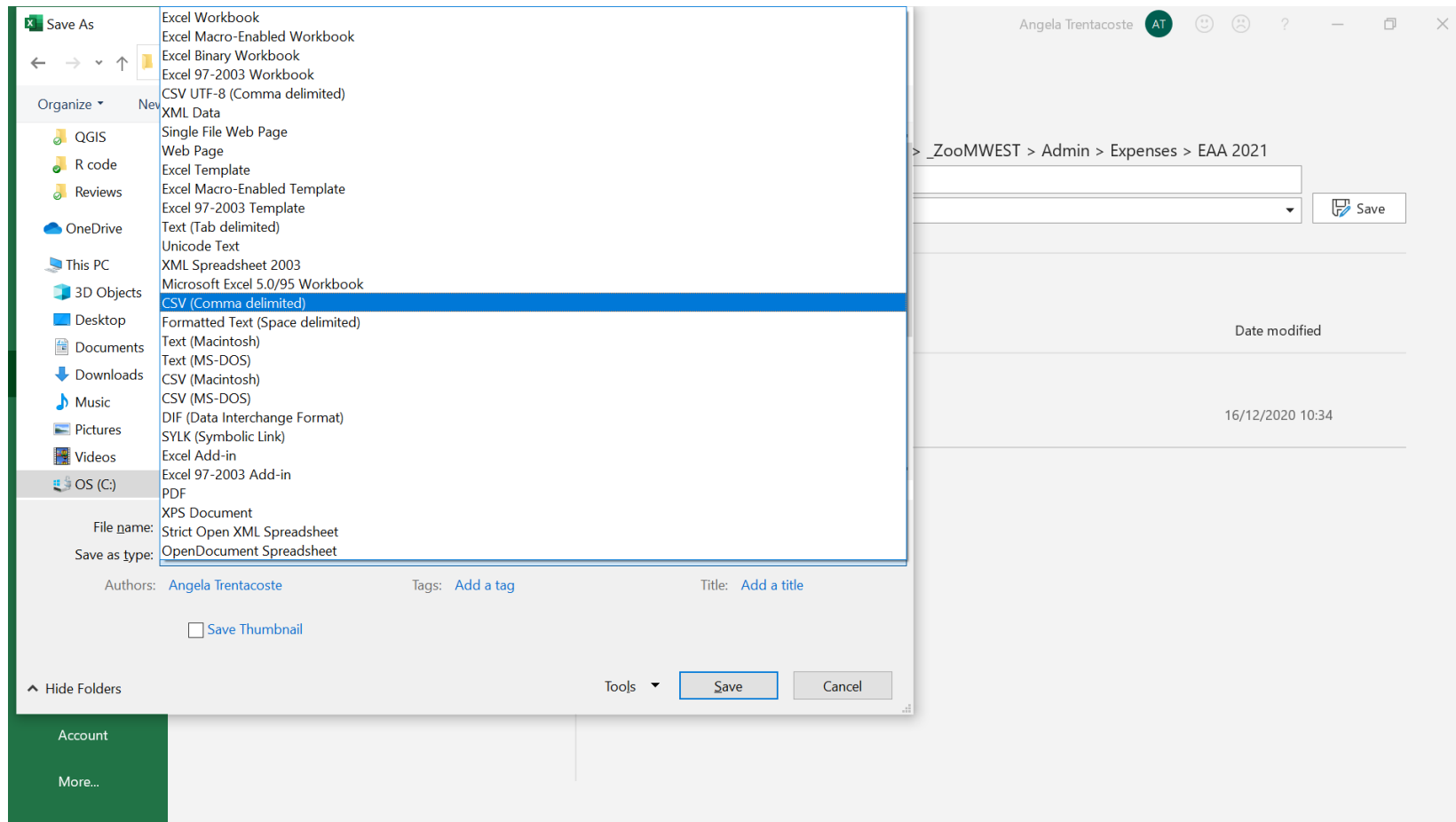
Now we are ready to save the file in CSV format. First save this new version as a normal Excel file.

Now choose File >> Save As

For 'Save as type' choose CSV (comma delimited) from the list.

If you used accents or special characters in your database, you will need to set an encoding (e.g. "UTF-8") for these to be correctly displayed.

Done! Your data should now be ready for R and zoolog 😊



Bonus: What do I do with bones from the same individual?

When using log ratios we normally want to exclude most bones from skeletons or partial skeletons, so we don't count the same individual many times.

You can see in my data, I have a scapula and humerus from the same skeleton: partial skeleton 'A' in the 'Individual' column.

Biometry_ID	Site_name	Area	TAX	EL	GL	Bp	DD	GLI	GLm	GLP	SLC	BT	Individual	Indi_dup	Precision	Comments
247	Marzabotto	V; 3	Bos	AS				59.4							Estimated	Check, typo?
248	Casale di Rivalta		Bos	SC						62.2	46.9					
250	Rubiera	Ca del Cristo	Bos	SC						59.3	44.8					
251	Case Vandelli		Bos	SC							47.4				Estimated	
252	Bologna	Via Ca Selvatica	Bos	SC							46				Rounded to 0.5	
253	Marzabotto	US 8	Bos	SC							57.7		A			
254	Marzabotto	US 8	Bos	HU								60.0	A			

I will want to include one measurement from this partial skeleton in my log ratios analysis, but not all of them. So I have another column where I indicate which measurements I want to exclude: 'Indi_dup'. I can use this column to mark which rows duplicate that same individual and are therefore redundant in the analysis. I want to keep the humerus measurement in the analysis, so I will mark scapula as redundant in the Indi_dup column.

Biometry_ID	Site_name	Area	TAX	EL	GL	Bp	DD	GLI	GLm	GLP	SLC	BT	Individual	Indi_dup	Precision	Comments
247	Marzabotto	V; 3	Bos	AS				59.4							Estimated	Check, typo?
248	Casale di Rivalta		Bos	SC						62.2	46.9					
250	Rubiera	Ca del Cristo	Bos	SC						59.3	44.8					
251	Case Vandelli		Bos	SC							47.4				Estimated	
252	Bologna	Via Ca Selvatica	Bos	SC							46				Rounded to 0.5	
253	Marzabotto	US 8	Bos	SC							57.7		A	Yes		
254	Marzabotto	US 8	Bos	HU								60.0	A			

After calculating my log ratios in zoolog, I can easily remove rows with 'Yes' in the duplicate individual 'Indi_dup' column if I don't want to count the same skeleton more than once.

Save your file as csv! If you used accents or special characters in your database, you will need to set an encoding (e.g. "UTF-8") for these to be correctly displayed.