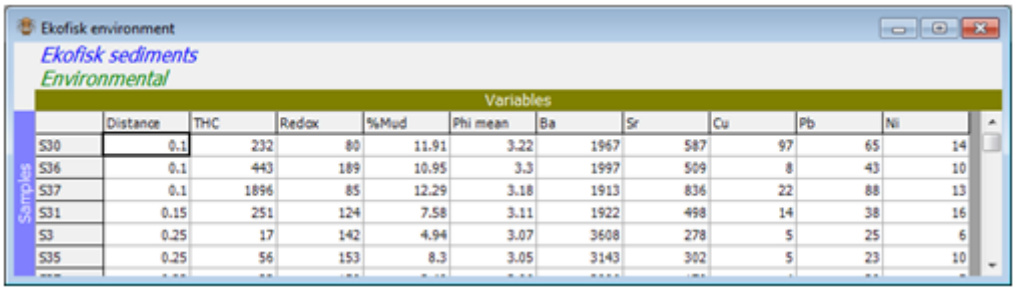
Missing or zero values?

The final option is whether a blank cell in the Excel sheet should be interpreted as a Missing value or a Zero. Typically, it will be Zero for species variables and Missing for environmental or other data. The distinction is important for subsequent analysis: most species-by-samples matrices have large numbers of species that are not present in many samples – they are indicated by zeros, and this information is properly catered for by an appropriate choice of similarity coefficient. If an environmental variable is not detected at a sample site then that should also be recorded as a zero, or as the lower detection limit (or perhaps half that limit). If a specific variable is not measured at a site, through random loss of a sample, then that is properly a Missing value. Inputting a blank cell from Excel, with the (Blank=•Missing value) option, or editing it to a blank after it has been read into PRIMER, will display a Missing! entry.

There are then three possible approaches. For environmental type data which might be transformable to approximate multivariate normality, and for which there are relatively few missing cells, a good option may be to attempt statistical estimation of the (randomly) missing values using the **Tools>Missing** routine. This uses the EM routine to give maximum likelihood estimates of the missing cells by exploiting the correlations among variables (see Section 12), thus completing the matrix. However, in many cases these normality assumptions are not viable, or there are simply too many parameters to estimate. Thus, secondly (and new to v7), PRIMER now automatically takes the simpler approach of calculating resemblance measures after removing, separately for each pair of samples, all variables which have a missing value for either sample. All resemblance measures are then automatically adjusted for the crude bias which results from such *pairwise eliminated data* input to totalled measures, such as Euclidean and Manhattan distance (without this adjustment some pairs of samples would be given greater distance simply because they are summed over more variables), see Section 5. Of course, a third possibility is simply to select a subset of samples and variables for which there are no missing values, e.g. by **Select>Variables** >(•No missing values).

It is important to appreciate that random loss of a whole sample (for all variables), e.g. loss of a replicate community sample from a balanced sampling design, is not thought of as producing missing values. If all species (or variables) are lost for that sample, it is simply omitted, and the design becomes a slightly unbalanced one, which is perfectly well catered for in most of the PRIMER (or PERMANOVA+) routines, e.g. in the ANOSIM or PERMANOVA hypothesis tests.



	Distance	THC	Redox	%Mud	Phi mean	Ba	Sr	Cu	Pb	Ni
S30	0.1	232	80	11.91	3.22	1967	587	97	65	14
S36	0.1	443	189	10.95	3.3	1997	509	8	43	10
S37	0.1	1896	85	12.29	3.18	1913	836	22	88	13
S31	0.15	251	124	7.58	3.11	1922	498	14	38	16
S3	0.25	17	142	4.94	3.07	3608	278	5	25	6
S35	0.25	56	153	8.3	3.05	3143	302	5	23	10

Save the workspace in the C:\Examples v7\Ekofisk directory with **File>Save Workspace As**>(File name: Ekofisk ws.pwk), for later use, and **File>Close Workspace** to clear the workspace. Further

files will now be opened from C:\Examples v7\Tasmania meiofauna, to demonstrate text file input.

Revision #5
Created 15 May 2024 21:51:50 by Arden
Updated 8 January 2025 01:41:12 by Abby Miller