

Missing data estimation (Clyde study)

Transformation options for the Clyde environmental matrix, **Clyde environment**, are discussed in more detail in the following (PCA) section, but the tool to carry out separate transforms on sets of variables, **Pre-treatment>Transform(individual)**, rather than transforming the whole array, **Pre-treatment>Transform(overall)**, was met in Section 4, applied to the environmental data from the Ekofisk oil-field study. Here, all heavy metals and organics (10 of the 11 variables) will benefit from log transformation, to reduce their right-skewness and so bring these continuous variables closer to normality across the sites (in so far as that can be judged from only 12 samples!). Thus, highlight all variables except Water Depth (*Dep*) and take **Pre-treatment>Transform(individual)>(Expression: $\log(V+0.1)$)** & (☒ Rename variables), renaming the result **Clyde log abiotic**. Give the variables in this sheet shorter names (e.g. **InCu**, **InMn** etc) with **Edit>Labels>Variables**.

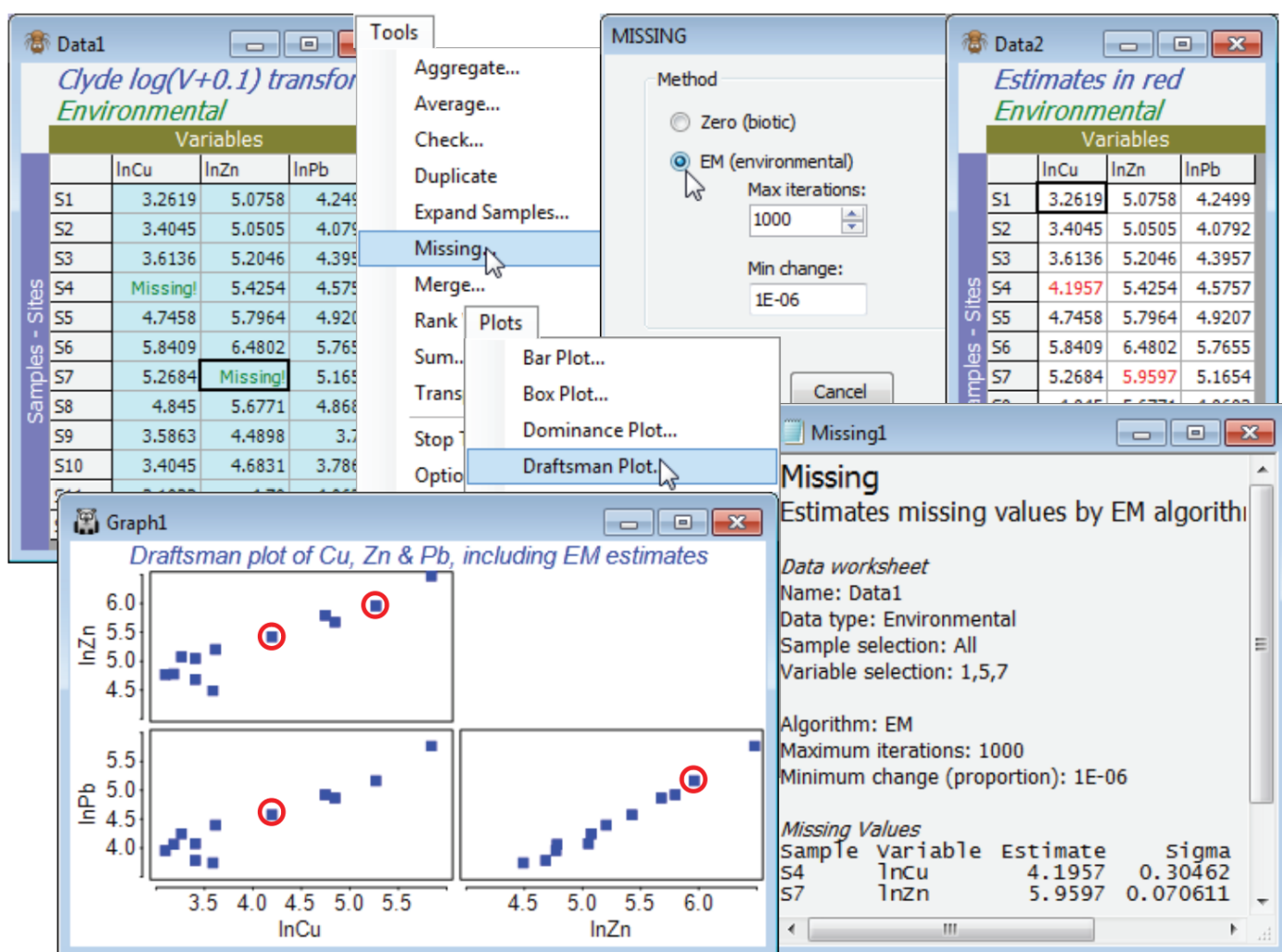
The screenshot shows three windows from a data analysis software. The 'Clyde environment' window displays a table of variables (Cu, Mn, Co, Ni, Zn, Cd, Pb, Cr, Dep, %C, %N) for 12 samples (S1-S12). The 'TRANSFORM' window shows the expression 'log(V+0.1)' and the 'Labels' window shows the variable names being renamed (InCu, InMn, etc.).

Samples - Sites	Cu	Mn	Co	Ni	Zn	Cd	Pb	Cr	Dep	%C	%N
S1	26	2470	14	34	160	0	70	53	144	3	0.53
S2	30	1170	15	32	156	0.2					
S3	37	394	12	38	182	0.2					
S4	74	349	12	41	227	0.5					
S5	115	317	10	37	329	2.2	1				
S6	344										
S7	194										
S8	127										
S9	36										
S10	30										
S11	24										
S12	22										

Samples - Sites	InCu	InMn	InCo	InNi	InZn	InCd	InPb	InCr	Dep	In%C	In%N
S1	3.2619	7.812	2.6462	3.5293	5.0758	-2.3026	4.2499	3.9722	144	1.131	-0.4620
S2	3.4045	7.0648	2.7147	3.4689	5.0505	-1.204	4.0792	2.7147	152	1.131	-0.5798
S3	3.6136	5.9766	2.4932	3.6402	5.2046	-1.204	4.3957	4.3451	140	1.098	-0.7765
S4	4.3054	5.8554	2.4932	3.716	5.4254	-0.51083	4.5757	4.7283	106	1.335	-0.5798
S5	4.7458	5.7592	2.3125	3.6136	5.7964	0.83291	4.9207	5.1767	112	1.740	-0.2357
S6	5.8409	5.3986	2.3125	3.6136	6.4802	1.7579	5.7655	5.7497	82	2.424	0.157
S7	5.2684	5.5495	2.4069	3.5293	6.0523	1.335	5.1654	5.4254	74	1.974	-0.1984
S8	4.845	5.5057	2.3125	3.4995	5.8471	0.83291	4.8683	5.2046	70	1.931	-0.3856
S9	3.5863	5.2684	1.8083	2.7788	4.4898	-0.69315	3.74	4.0448	64	0.693	-0.9416
S10	3.4045	5.7872	2.4069	3.2619	4.6831	-1.6094	3.7865	3.9532	80	1.193	-0.7339
S11	3.1822	6.0847	2.4932	3.5293	4.78	-1.6094	4.0622	3.5863	83	0.788	-0.7985
S12	3.0956	6.686	2.4932	3.4995	4.7715	-2.3026	3.9532	3.9338	83	0.875	-0.5978

Take a copy with **Tools>Duplicate** and from this remove a couple of cells at random – perhaps (S4, InCu) and hit the delete key, then (S7, InPb) and delete again. Both cells will now be displayed as **Missing!**. **Analyse>Draftsman Plot** on this transformed data shows that normality assumptions are probably now acceptable (see the following section) but the above *DpP* criterion for the whole matrix fails badly ($n = 12$, $p = 11$, $m = 2$, so $DpP = 1.7$) and we should not trust the outcome even if **Tools>Missing** converges (it does not, here). The correlation matrix output with the draftsman plot does, however, show some very high correlations between e.g. Cu, Pb and Zn,

which gives a better basis for prediction than the whole matrix. So, select just these three variables (highlight them then **Select>Highlighted**), and **Tools>Missing** produces credible missing data estimates of 4.18 (S4, ln Cu) and 5.26 (S7, ln Pb), compared with the original 4.31 and 5.17. Note that the ratio $DpP = 3.3$, which is still some way from respectability, but clearly is capable (sometimes at least) of producing useful results. The results window shows that the imprecision (under the assumption that the value is missing at random, of course) is lower for the estimated (S7, Pb) reading than the (S4, Cu) value, though both are rather well determined. The standard deviation of the estimate for (S7, Pb) is about 0.07 and for (S4, Cu) about 0.30, so that rough confidence intervals are (3.6, 4.8) and (5.8, 6.1) respectively. The reason for this difference in precision is clear from the draftsman plot for these three Cu, Zn, Pb variables, on which the respective points are manually circled (the plot window was copied and pasted to Powerpoint with Ctrl-C and Ctrl-V). The linear relationship between Pb and one of the other variables (Pb) is seen to be extremely tight, whereas Cu is not so highly correlated with either Zn or Pb, so there is inevitably greater uncertainty in the interpolation – it is a consequence of the multivariate normality condition that these relationships are estimated as straight lines. The estimates now need to be individually copied (click in the cell and Ctrl-C) and pasted back into the full matrix (Ctrl-V at the cursor). Of course the process is more automatic in less borderline cases, with larger n , when the full matrix can be input to **Tools>Missing**.



Revision #5

Created 24 September 2024 01:47:13 by Arden

Updated 11 February 2025 22:13:00 by Abby Miller