Appendix B: Dissimilarity formulae

Dissimilarity in ordinal variables (NI/RRD)

The neighbor interchange (NI) approach defines the similarity between two observations i and j for variable k (s_{ijk}) as:

$$s_{ijk} = \begin{cases} 1 & \text{if } r_{ik} = r_{jk} \\ 1 - \frac{|r_{ik} - r_{jk}| - \frac{T_{ik}-1}{2} - \frac{T_{jk}-1}{2}}{(\max r_k - \min r_k - \frac{T_{(k,\max)}-1}{2} - \frac{T_{(k,\min)}-1}{2}} \in [0,1] & \text{otherwise} \end{cases} \quad (1)$$

Where r_{ik} is the rank score of observation i for variable k; T_{ik} is the number of observations which have the same rank score for variable k as object i (including i); $T_{(k,max)}$ and $T_{(k,min)}$ are the number of objects which have, respectively, the maximum and minimum rank for variable k. As explained by Podani (1999), the numerator is the minimum number of steps (interchanges of neighbouring values in the ordering) between value x_{ik} and x_{jk}. However, Podani offers a second metricised approach, the relative rank difference (RRD), which is a simplification of the NI formula, but still respects the ordinal nature of variables:

$$s_{ijk} = 1 - \frac{|r_{ik} - r_{jk}|}{\max r_k - \min r_k} \in [0,1] \quad (2)$$

Extended Gower distance

According to Pavoine et al. (2009)'s extended Gower distance, the global distance between two observations i and j (D_{ij}) is the square root of the average squared distances between n observations for n variables; where variables weights (w_{ijk}) and the distance function itself (d_{ijk}) vary depending on variables' nature:

$$D_{ij} = \sqrt{\frac{\sum_{k=1}^{n} d_{ijk}^2 \delta_{ijk} w_k}{\sum_{k=1}^{n} \delta_{ijk} w_k}} \in [0,1] \quad (3)$$

Following the philosophy of Gower's general coefficient of similarity (1971), from which this distance derives, δ_{ijk} is equal to 1, unless the value of the kth variable is missing for one or both observations i and j, in which case it is equal to 0 (i.e., invalidates the comparison). Concerning ordinal variables, we took advantage of this feature to consider specific values in certain variables to add nothing to this distance-i.e., they do not differentiate from any other value. For instance, in most variables regarding voids and inclusions, "none" values count as the minimum category of the ordinal gradient representing frequency, assuming that "none" is more similar to "few" than to "dominant". Exceptionally, when "none" equals the complete absence of a measured trait (e.g., in inclusions distribution and grain form), it is considered a missing value. Where the frequency of missing values causes distances to violate the metric axioms of Euclidean space, Lingoes transformation is applied (Lingoes 1971; after Pavoine et al. 2009).

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **27**: 857–871.

Lingoes, J. C. (1971). Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika* **36**: 195–203.

Pavoine, S., Vallet, J., Dufour, A.-B., Gachet, S. and Daniel, H. (2009). On the challenge of treating various types of variables: application for improving the measurement of functional diversity. *Oikos* **118**: 391–402.

Podani, J. (1999). Extending Gower's General Coefficient of Similarity to Ordinal Characters on JSTOR. *Taxon* **48**: 331–340.