View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Horizon / Pleins textes

A multi-level dissimilarity index between soil profiles

R. VAN DEN DRIESSCHE, A.-M. GARCIA GOMEZ, A.-M. AUBRY
ORSTOM Soil Data Base, 70 route d'Aulnay, F 93140 Bondy (France)

Available data from each selected profile is introduced in coded form by a retrieval program in a 10×39 data matrix wherein rows correspond to 10 depths (1, 5, 13, 25, 41, 61, 85, 113, 145, 181 cm) and columns to 23 ordinal field variables + 16 interval laboratory variables. Missing data is allowed under the -1 code.

A dissimilarity index $0 \leqslant d'_{ij} \leqslant 1$ is computed with v variables as follows, between the 10 successive depths of each profile:

$$d'_{ij} = \left[\frac{1}{v}\sum_{k=1}^{v}\left(\frac{x_{ik}-x_{jk}}{G_k}\right)^2\right]^{1/2}$$

where constant G_k is the largest value variable k has ever taken.

The 9 indices $(d'_{1,2}, \ldots, d'_{9,10})$ are then introduced with the original data in every row of a $m \times \{9 + (10 \times 39)\}$ data matrix, where m stands for the number of profiles. The algorithm is used again, and yields a $\binom{m}{2}$ dissimilarity matrix between the profiles. Program DISSIM allows for $2 \leqslant m \leqslant 250$ and $1 \leqslant v \leqslant 400$. Note $x_{ik} \neq -1$ and $x_{jk} \neq -1$. The proposed index is not dependent on the names given to the soil horizons. Furthermore, independency between the horizon descriptions is not assumed and, for the first time, within-profile dissimilarities are included in the between-profiles dissimilarities. Tape output (matrix) and lineprinter output ($m-1$ dissimilarities ordered m times) are provided.

Computer is a Univac 1108, 192 K, under exec 8. Terminal is an Ordoprocesseur 300 card reader + lineprinter.

Fortran program DISSIM, 47 K, developed 1975, is available without charge upon individual application. Total CPU time is 7 minutes for 100 profiles, retrieval included.

Un programme de sélection booléenne sert à placer les données de terrain et de laboratoire qui sont disponibles dans un tableau à 10 lignes et 39 colonnes. Les lignes correspondent à 10 profondeurs croissantes (1, 5, 13, 25, 41, 61, 85, 113, 145, 181 cm), les intitulés de colonne correspondent à un choix de 23 variables qualitatives ordonnées et de 16 variables quantitatives. Le zéro des échelles qualitatives et quantitatives est codé 0, alors que les manquants sont symbolisés par le code -1.

Un indice de dissemblance variant entre 0 et 1 est calculé comme suit avec v variables, entre les 10 profondeurs successives de chaque profil pris individuellement :

$$d'_{ij} = \left[\frac{1}{v}\sum_{k=1}^{v}\left(\frac{x_{ik}-x_{jk}}{G_k}\right)^2\right]^{1/2}$$

algorithme dans lequel G_k est une constante égale à la donnée maximale connue à ce jour pour la variable k.

On entre ensuite les 9 indices $(d'_{1,2}, \ldots, d'_{9,10})$ avec les données d'origine sur chaque ligne d'un tableau de données à m lignes (m profils) et $\{9 + (10 \times 39)\}$ colonnes. Le même algorithme donne, cette fois, une matrice de $\binom{m}{2}$ indices de dissemblance entre les profils. La capacité maximale du programme DISSIM est de 250 profils et 400 variables par profil. Rappelons que $x_{ik} \neq -1$ et $x_{jk} \neq -1$. La nomenclature des horizons choisie par le pédologue n'a aucune influence sur l'indice que nous proposons. De plus, l'hypothèse d'une indépendance entre les descriptions des horizons d'un même profil n'a pas à être faite. Enfin, la nouveauté réside dans l'adjonction de dissemblances intra-profil au tableau de données brutes servant au calcul des dissemblances interprofils.

L'ordinateur auquel est raccordé le terminal Ordoprocesseur modèle 300 (lecteur de cartes perforées et imprimante) est un Univac modèle 1108 qui a 192 K mots de mémoire et qui opère sous exec 8.

Le programme DISSIM est écrit en Fortran ; il a été validé en 1975 et occupe 47 K. Le lecteur intéressé peut l'obtenir de la Banque de Données Pédologiques. Il est à noter que le temps cumulé de calcul pour un fichier de 100 profils est de 7 minutes CPU, sélection du fichier comprise.