PROXIMITÉ (K10)
(21 / 11 / 2020, © Monfort, Dicostat2005, 2005-2020)

La notion de **voisinage** (définie eg dans les espaces topologiques ou les espaces métriques) est une approche mathématique de celle de **proximité**.

En **Statistique**, la notion de **proximité** se rencontre fréquemment, eg lorsque le **statisticien** procède à des comparaisons entre **unités statistiques** ie, le plus souvent, entre « **mesures** » effectuées sur ces unités.

(i) Une acception particulière est celle de « **variable proche** », ou « **variable voisine** », (en anglais : « *proxy* ») d'une variable considérée.

La **proximité entre variables** dont il s'agit s'exprime généralement, à travers leurs **observations** :

(a) soit par une « forte **corrélation** » entre les deux types de variables ;

(b) soit même par une **relation fonctionnelle**, en général non linéaire, entres ces dernières.

Cette **situation statistique** se rencontre eg lorsque la variable considérée est **inobservable**, la seconde étant **observable** (cf **variable observable**) et supposée « proche » (au sens précédent) de la première : ceci est le cas d'un modèle traité à l'aide de **variables instrumentales** (**modèle à erreurs sur les variables**, **modèle d'interdépendance**) (cf **méthode des variables instrumentales**). La corrélation dont il s'agit n'est donc pas toujours « observable », et doit être supposée.

(ii) Une notion comparable de **proximité** est aussi à la base de l'**analyse des proximités**, en **classification automatique**.

Dans les cas élémentaires, celle-ci concerne un **tableau statistique** de dimension deux (eg un **tableau de contingence**) $T \in M_{NK}$ (**R**), dont on cherche à comparer les lignes (représentant des **unités statistiques** ou des **observations**), ou les colonnes (représentant des **variables**), entre elles. Si l'**espace vectoriel** image Im T (resp Im T') est muni d'une **norme** (resp d'une **distance**) qui en fait un **espace métrique** (resp un **espace normé**), l'analyse des proximités étudie les **ressemblances** (resp **dissemblances**) à partir des distances entre lignes T_n ($n \in N_N^*$) ou entre colonnes t_k ($k \in N_K^*$) de T.

Dans le premier cas, on utilise couramment des distances tq :

$$d_1 (T_{n'} , T_{n''}) = \Sigma_{k=1}^{K} |t_{n'k} - t_{n''k}| \quad \text{(norme de } l^1\text{)},$$

(1)
$$d_2 (T_{n'} , T_{n''}) = \Sigma_{k=1}^{K} (t_{n'k} - t_{n''k})^2 \quad \text{(norme de } l^2\text{)},$$

$$d_\infty (T_{n'} , T_{n''}) = \max_{k=1}^{K} |t_{n'k} - t_{n''k}| \quad \text{(norme de } l^\infty\text{)}.$$

Lorsque les espaces vectoriels ne peuvent pas être munis de distances interprétables, on peut remplacer (au signe près) une distance d par un **coefficient de corrélation**, calculé entre les lignes ou les colonnes de T.

Lorsque les variables correspondant aux colonnes de T sont des **variables qualitatives**, on remplace souvent une distance d par un **indice de similarité** s (cf aussi **dissimilarité**, **indice de dissimilarité**), ou par un **coefficient d'association**, calculés entre les lignes ou les colonnes de T.

Souvent, T est préalablement transformé en une matrice $X \in M_{NK}$ (**R**), notamment par **codage** des variables (cf **transformation des données**).

(iii) Enfin, il existe aussi une notion de **proximité entre procédures statistiques** (cf eg **robustesse**) ou une notion de **proximité entre règles de décision** (cf eg **proximité entre estimateurs**).

(iv) Dans l'activité scientifique, la notion de proximité intervient souvent en vue de diverses comparaisons ou classifications. Ainsi (cf aussi **reconnaissance des formes**) :

(a) en archéologie (paléontologie), la découverte d'un jeu d'ossements « nouveau » conduit à en comparer ses caractéristiques avec d'autres jeux d'ossements préexistants en vue de déterminer si cette découverte est en « rupture » ou en « continuité » (évolution) pr à l'existant. Une difficulté importante réside souvent dans l'existence de **lacunes** observationnelles dans certains jeux ;

(b) en sociologie (criminologie), une trace biologique (empreinte digitale ou autre marqueur génétique) est généralement comparée à celles figurant dans diverses bases d'information en vue de déterminer un « individu proche » de celui qui est l'auteur de cette trace.