## Types de base de données avec Stata

Le tableau suivant présente, de manière un peu abstraite mais compacte, les différents types de bases de données rencontrées en économétrie (coupe, chronique, panel, etc.). Il ne faut pas confondre le type de bases de données et le type de variable (quantitative, qualitative, etc.). Par exemple, une série temporelle peut mesurer la trajectoire d'une variable quantitative (en $t$, $t+1$, …) telle que le taux de chômage. Mais elle peut aussi mesurer la trajectoire d'une variable qualitative telle que la décision de chercher du travail.

Supposons $1,2,\dots,K$ variables aléatoires (v.a.). Notons les valeurs de la v.a. $k$ par $y_{itk}$. L'indice $i \in \{1,\dots,N\}$ caractérise l'individu (ménage ou entreprise ou commune, pays, etc.), $t \in \{1,\dots,T\}$ la période (seconde, jour, mois, année, etc.). Il y a $NT$ observations pour chaucne des $K$ v.a. Pour **Stata**, une ligne = une observation, quel que soit le nombre de variables, et une colonne = une variable. On a donc $KNT$ valeurs dans la base. Caractérisons les différents types de base selon les valeurs de $K$, $N$ et $T$.

**Tableau 1.** Différents types de bases de données (noms en anglais)

| | $N=1$ | | $N>1$ | |
|---|---|---|---|---|
| | $T=1$ | $T>1$ | $T=1$ | $T>1$ |
| $K=1$ univariate data | scalar variable | time series | cross section | panel ≠ repeated cross section |
| $K>1$ multivariate data | $K$- vector of random variables | multivariate time series | multivariate cross section | multivariate panel |

Dans le cas :

- $K=N=1$, $T>1$, on parle aussi d'« univariate time series ». Par exemple, la population française sur la période 1901-2023 (voir par là, sur le site de l'INED, première colonne).
- $K>1$, $N>1$, $T=1$, on parle de « multivariate cross section ». Par exemple, les dépenses de recherche des pays de l'UE-27 en 2020 (« rd_e_gerdfund » sur Eurostat, sélectionner cette page).
- $N>1$, $T>1$, on parle de « panel data », mais on trouve aussi « longitudinal panel data » ; si $K>1$, il est correct de parler de « multivariate longitudinal panel data », mais personne n'a envie d'être aussi précis.

**Q** : Quelle est la différence entre « longitudinal data » et « pseudo-panel » ?