Comment ?

Construire un indicateur statistique composite ?

Un indicateur est une donnée chiffrée dont l'intervalle varie généralement de 0 à 1 ou de 0 à 100%. Il permet de refléter une réalité particulière de notre monde.

Le degrés Celsius est un indicateur de la température, la moyenne pondérée en % d'un bulletin est un indicateur de réussite scolaire, tout comme la croissance du PIB est un indicateur de la santé économique d'un pays.

Les indicateurs composites sont des indicateurs composés de plusieurs données chiffrées différentes mais réduit à une seule donnée chiffrée.

Pour exemple, L'IDH est un indicateur composite sensé représenter le Développement Humain par une valeur numérique.
La particularité de cet indicateur est de prendre en compte plusieurs facteurs différents, à savoir, un indicateur reflétant la santé (espérance de vie), un indicateur reflétant l'éducation, lui-même composé de deux statistiques différentes, le taux d'alphabétisation et le taux de scolarisation et enfin un indicateur économique.

La construction d'un indicateur composite est à priori simple puisqu'il s'agit de réaliser une moyenne des données chiffrées des différents facteurs constitutifs du futur indicateur.

Indicateur composite = Σ Indicateur / n

Dans le cas de l'IDH, celui-ci s'obtient en calculant la moyenne de l'indicateur de l'espérance de vie I(ev), de l'indicateur de l'éducation I(ed) et l'indicateur économique I(eco).

$$\text{IDH} = \frac{\text{I(ev)} + \text{I(ed)} + \text{I(eco)}}{3}$$

avec $\text{I(ed)} = \frac{2 \times \text{T\% alphabetisation} + \text{T\% scolarisation}}{3}$

Nous voyons ainsi que l'indicateur de l'éduction n'est autre que la moyenne pondérée du T% d'alphabétisation et du T% de scolarisation en donnant 2 x plus d'importance au T% d'alphabétisation qu'au T% de scolarisation.

La plupart des séries statistiques utilisées sont exprimées dans des unités différentes des intervalles (pourcentage ou 0 à 1) de l'indicateur composite. Cette réalité constitue un inconvénient majeur puisque la moyenne est très sensible aux grandes valeurs. La conséquence est alors que l'indicateur composite sera plus le reflet des indicateurs à valeurs élevées que faibles.

Il est alors absolument nécessaire de transformer toutes les séries de données afin que celles-ci soient bornées entre 0 et 1 ou 0 et 100%. Cette opération porte le nom de NORMALISATION.

Dans le cas de l'IDH, deux séries sont normalisées puisque l'indicateur de l'éducation lui est déjà exprimé en %.

Les données à normaliser sont donc l'espérance de vie et le PIB.

Il existe plusieurs méthodes de normalisation. Nous en retiendrons deux en particulier. La première permet de normaliser des données dont la **distribution statistique est linéaire**.

$$\textbf{Vnormalisée} = \frac{\textbf{Vvariable} - \textbf{Vmin}}{\textbf{(Vmax} - \textbf{Vmin)}}$$

Dans le cas de l'IDH, on considère que l'espérance de vie dans le monde varie de 25 ans à 85 ans.
La normalisation de la série donne ainsi
pour la variable = 25 ans : 25-25/(85-25) = 0/60 = 0. Dans ce cas, 25 ans est bien égal à 0.
Pour la variable = 85 ans : (85-25)/(85-25) = 60/60 = 1. Dans ce cas, 85 ans est bien égal à 1.

Si la variable = 55 ans alors (55-25)/(85-25)=30/60 = 0,5.

L'espérance est ainsi bien bornée entre 0 et 1. Pour travailler en % comme l'indicateur de l'éducation, il faut multiplier le résultat par 100.

Par contre, certaines séries statistiques présentent des distributions non linéaires. Il s'agit souvent de distribution exponentielle. Leur normalisation passe donc pas l'application d'un logarithme en base 10.

$$\textbf{Vnormalisée} = \frac{\textbf{LOG}_{10}\,\textbf{Vvariable} - \textbf{LOG}_{10}\,\textbf{Vmin}}{\textbf{LOG}_{10}\textbf{Vmax-LOG}_{10}\textbf{Vmin}}$$

Dans le cas de l'IDH, on considère que le PIB/ha dans le monde varie de 100$ à 40000$.
La normalisation de la série donne ainsi
pour la variable = 100$: $(LOG_{10}100 - LOG_{10}100)/(LOG_{10}40000-LOG_{10}100) = (2-2)/(4,67-2) = 0$.
Dans ce cas, 100$ est bien égal à 0.
Pour la variable = 40000 : $(LOG_{10}40000 - LOG_{10}100)/(LOG_{10}40000-LOG_{10}100) = (4,67-2)/(4,67-2)=1$.
Dans ce cas, 40000$ est bien égal à 1.

Maintenant vos données normalisées, il vous est possible de construire n'importe quel indicateur composite complexe.

Petit conseil : pour des longues séries de données, n'hésitez pas à utiliser un tableur informatique !