Contact : fabien.guerreiro@educagri.fr. 

*Les Incontournables*

# Classification et discrétisation

## 1. Pourquoi classer ?

Cela induit une certaine perte d'information, mais la classification permet une meilleure lecture des données qualitatives ou quantitatives (classement d'interprétation d'images satellites, de données statistiques, d'altitude, de catégories …).
Pour réaliser une classification, il faut choisir le nombre de classes et les bornes de chaque classe. Il faut toujours être en mesure de justifier les classes. De la même façon, le choix de ne pas faire de classes se justifie. La classification (ou l'absence de classification) conditionne l'interprétation de la carte finale !

## 2. Quand doit-on discrétiser les valeurs numériques ?

En général, lorsque l'on a des valeurs quantitatives absolues (variables brutes absolues), on les représente, à l'aide de cercles proportionnels aux valeurs (population…).
La discrétisation intervient lorsqu'on a des valeurs quantitatives relatives, dites variables continues (PIB/Hab, densité…) ; c'est le passage de ces variables continues qui peuvent prendre toutes les valeurs à des variables discrètes : Les classes !
Par exemple, un pourcentage aux valeurs de 0 à 100 prend 4 ou 5 valeurs après classification.
La discrétisation est donc le regroupement de valeurs statistiques en classes.

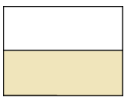
## 3. Discrétisation

**Règles :**

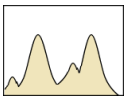
- Les classes doivent contenir toutes les valeurs, être contigües et significativement différentes.
- Les valeurs ne peuvent être que dans une seule classe et considérées semblables au sein d'une même classe.

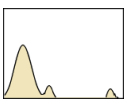
**Choix du nombre de classes** : L'idéal pour une carte lisible est de trois ou quatre classes, avec un maximum de sept classes à ne jamais dépasser. (Pour n individus, on ne dépasse jamais k classes ; k = partie entière $[1+3{,}3\log_{10}(n)]$).

**Choix de la méthode** : Il n'existe malheureusement pas de méthode miracle, mais il existe des orientations, ou conseils pour orienter au mieux les choix. **Il est donc indispensable d'observer l'histogramme de répartition des valeurs avant de fixer les classes**. De la même façon, observez la moyenne et l'écart-type pour avoir une idée de la dispersion des données.
Nous retiendrons trois types essentiels de distribution des données, et les tendances de classifications qui leurs sont associées :

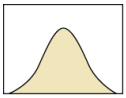


- **Uniforme** : On retrouve ces histogrammes très rarement ; la répartition est associée à la **méthode de seuils d'amplitude égale**. On utilise l'amplitude égale pour mettre en évidence une classe par rapport à une autre (pourcentages).





- **Dissymétrique** : C'est la distribution la plus fréquemment rencontrée ; la **méthode des effectifs égaux** permet de bien représenter ces données, mais la représentation par les **seuils naturels** (ou Jenks) peut être indiquée.
  La dissymétrie particulière où les données sont regroupées sur les valeurs faibles peut amener la **méthode de progression géométrique** qui apporte un regard intéressant sur les données, puisqu'elle minimise l'écart avec les valeurs fortes, faiblement représentées.



- **Symétrique** (courbe de Gauss) : Les données étant centrées autour de la moyenne, la **méthode de la moyenne et de l'écart-type** est souvent la plus appropriée.