Comment ?

Discrétiser des données continues en données discrètes dans le cadre de la réalisation d'une carte thématique quantitative.

Fiche outil inspirée par l'article « *Cartographie numérique : précis de discrétisation pour les nuls (blog.m0le.net) et de l'article « discrétisation » (www.hypergeo.eu).*

L'information géographique peut revêtir plusieurs formes. Parmi ces dernières, la représentation cartographique présente l'avantage de la lecture instantanée d'une image, une lecture qui doit aller du général au particulier. L'avantage que constitue cette lecture ne peut exister qu'au prix d'un investissement préalable fait lors de la conception et de la réalisation de la carte. Cet investissement concerne aussi bien la sélection pertinente de l'information que le traitement de cette information. Le traitement préalable de l'information dépend du type de caractère statistique que l'on veut cartographier. En fonction de ce caractère les choix des méthodes de discrétisation sont plus ou moins nombreux et les résultats cartographiques peuvent donner des images très variées. (Source : hypergeo)

*La **discrétisation** est l'opération qui permet de découper en classes une série de variables qualitatives ou de variables quantitatives. Cette opération simplifie l'information en regroupant les objets géographiques présentant les mêmes caractéristiques en classes distinctes.*

En discrétisant, on découpe sa carte **en un certain nombre de "classes"** dans lesquelles sont rangées des valeurs colorées avec une teinte unique.
Sauf qu'il y a **différentes manières de discrétiser** une carte, et qu'**aucune d'entre elles n'est parfaite**. Bien les connaître permet en revanche de se faire **une rapide idée de celle qui est la plus judicieuse** à appliquer.

Une discrétisation est satisfaisante lorsqu'elle permet la création de **classes homogènes et distinctes entre elles** : les objets géographiques d'une même classe doivent se ressembler plus entre eux qu'ils ne ressemblent aux objets des autres classes.

Etape 1 : Choisir le nombre de classes.

Le nombre de classes optimum à réaliser dans une partition est toujours fonction du nombre d'individus observés (unités spatiales). Il existe un indice permettant de connaître le nombre de classes idéales pour une distribution, il faut le considérer uniquement comme une aide indicative.

Il s'agit de l'indice de Huntsberger :

$$N(cl) = 1 + 3{,}3 \log_{10}(N)$$

N = nombre d'observations
N(cl) = nombre de classes

Etape 2 : Choisir la méthode de choix des seuils (discrétisation).

Méthode 1 : Discrétisation en classes d'amplitude égale

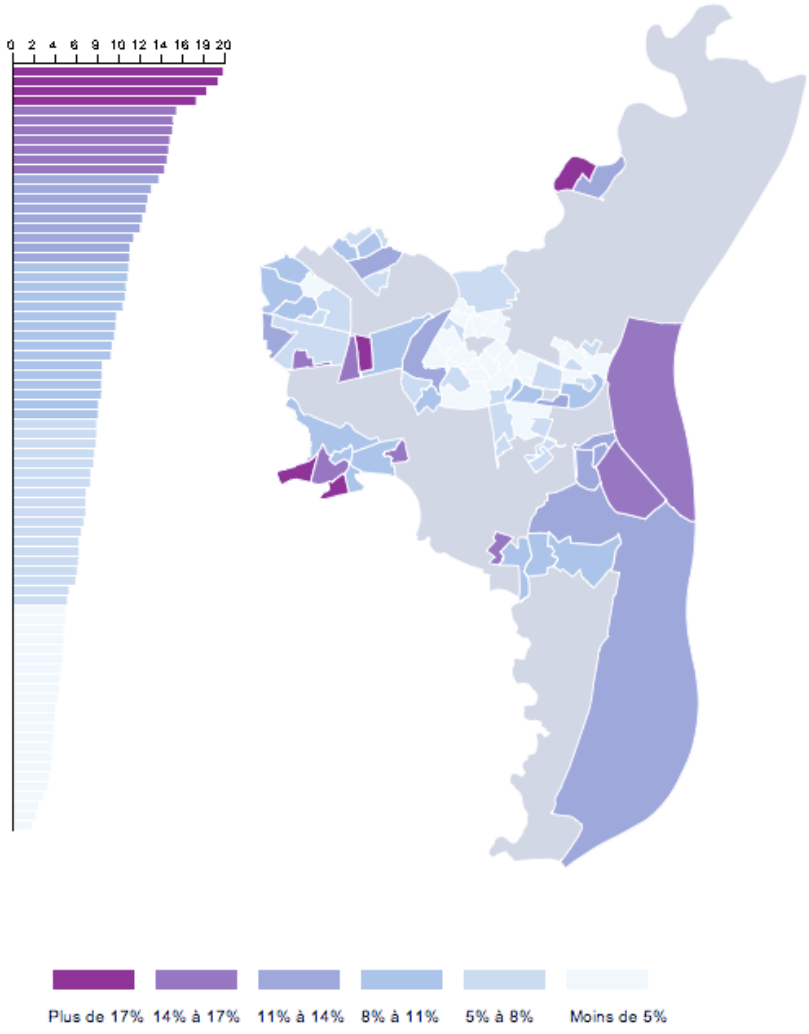
Une des questions que l'on se pose le plus souvent quand on traite d'importants volumes de données est : **que faire des valeurs extrêmes ?**

On peut soit **les considérer comme négligeables, voire parasitaires** et ne pas les valoriser, ou **bien articuler toute une histoire autour d'elles**.

La discrétisation en classes d'amplitude égale, l'une des plus répandues, permet précisément **de valoriser les valeurs extrêmes**.

Pour obtenir l'amplitude type, **on divise simplement l'étendue de l'échantillon** (maximum-minimum) **par le nombre de classes**.

Exemple : si la valeur maximale est de 19.88 et la minimale de 1.83 notre amplitude sera donc de (20-2)/6=3 pour une légende de 6 classes. La première classe ira donc de 2 à 5, la deuxième de 5 à 8, et ainsi de suite jusqu'à la dernière, de 17 à 20.



Avantages :

Ultra simple à réaliser très courante, donc **très facilement interprétable par tout le monde.**

Inconvénients :

Ce modèle est **"idéal" pour des distributions uniformes** et dans une moindre mesure symétriques, **cas plutôt rares**. Dans notre cas, la distribution est franchement dissymétrique, et ça se ressent dans **les effectifs très déséquilibrés** : la classe la plus élevée contient à peine quatre zones, tandis que la moins élevée en contient une grosse vingtaine.

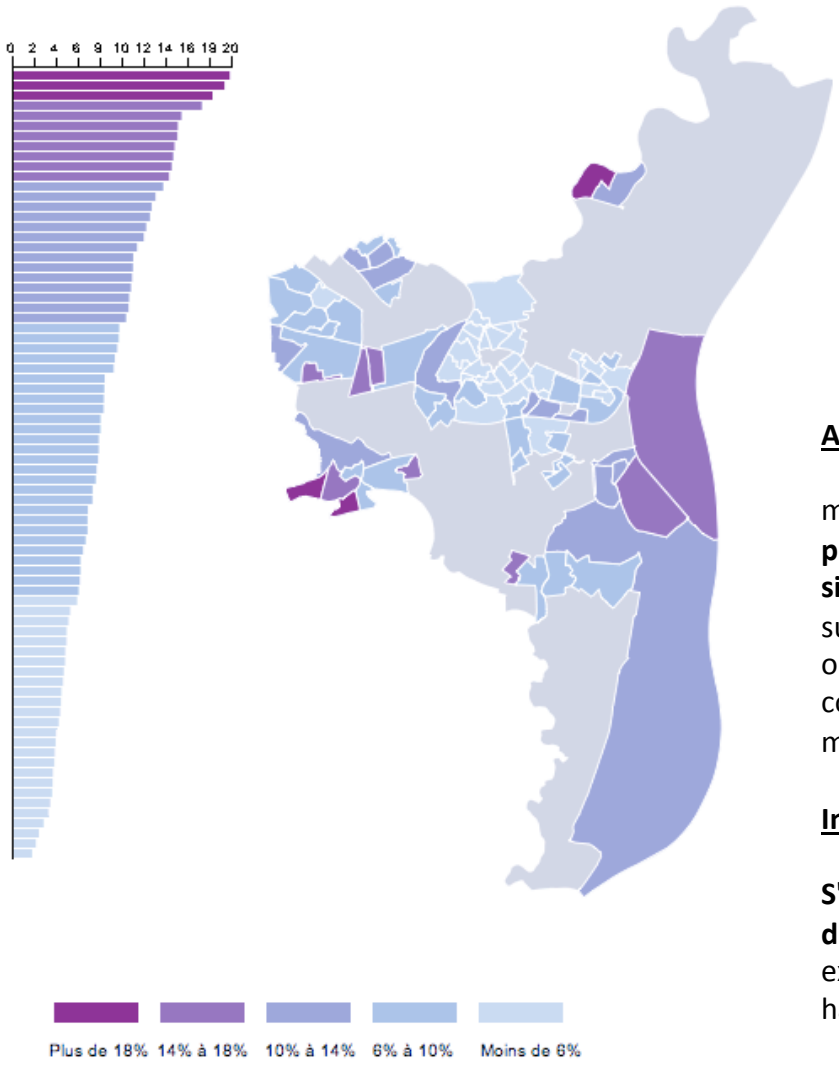
Méthode 2 : Discrétisation avec moyenne et écart-type

La **moyenne** et l'**écart-type** sont deux mesures ultra courantes en **statistique descriptive**, mais qui ont le défaut d'être **très sensibles aux valeurs extrêmes**. Malgré tout, dans le cas où on compare plusieurs cartes, la moyenne et l'écart type peuvent entièrement se justifier.

Par exemple, remarquer qu'une même zone est systématiquement deux écarts types au-dessus de la moyenne peut amener un éclairage intéressant quand on analyse des données.

Avec cette méthode, la moyenne finira borne de classe si leur nombre est pair, et centre d'une classe si leur nombre est impair.

Exemple : Si la moyenne et l'écart type sont de 8 et de 4, une fois arrondis. On aura donc une classe de 8 à 12 (un écart type au-dessus de la moyenne), une de 8 à 4 (un écart type en-dessous de la moyenne), puis une de 12 à 16, etc...



Calculez l'écart type :
Il représente une mesure de dispersion de données par rapport à la moyenne.

Écart-type = $\sqrt{[(\Sigma((\text{valeur-moy})^2))/(N)]}$.

Avantages :

méthode particulièrement **taillée pour les comparaisons** elle est assez **simple et intuitive à reproduire** : il suffit d'utiliser deux options d'Excel ou d'appliquer la formule et de construire ses classes à partir de la moyenne

Inconvénients :

S'adapte mal aux distributions dissymétriques, dont les valeurs extrêmes tirent la moyenne vers le haut ou vers le bas.

Méthode 3 : Discrétisation par progression géométrique

La méthode de l'écart-type s'applique mal à une distribution de données dissymétrique. On peut du coup se demander : **comment faire pour traiter correctement une distribution dissymétrique ?**
Une des discrétisations très utilisées propose de traiter les classes de la distribution cartographiée comme les membres d'une suite géométrique.
Ce choix par de l'observation que les données produites dans la nature et dans nos sociétés ont tendance à mieux coller à un modèle multiplicatif qu'additif, par exemple les distribution de type exponentielle lors de la propagation d'une épidémie.

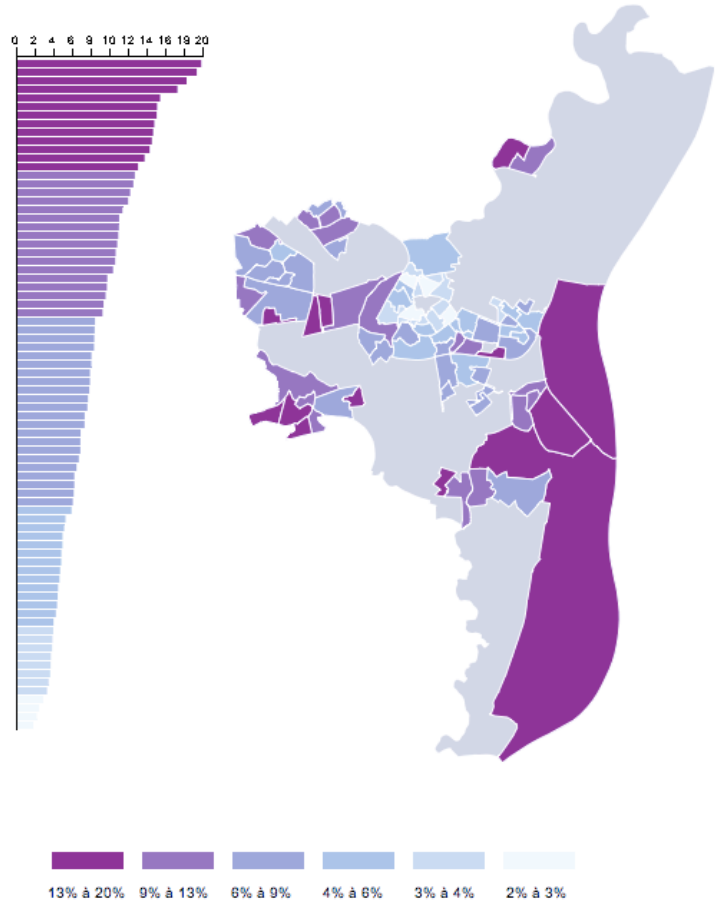
Pour revenir à la suite géométrique, sa **raison** s'obtient grâce à un logarithme (en base 10) : comme on part d'une distribution "log-normale", la conversion en logarithme nous donne une distribution normale (autrement dit : ce que l'on recherche).

Formule de la RAISON :
Il représente une mesure de dispersion de données par rapport à la moyenne.

raison = 10^((log(max)-log(min)/ nombre de classes)

<u>Exemple</u> : Avec un maximum de 19.88, un minimum de 1.83 et six classes à construire, on obtiendrait :
log raison =(log(19.88)-log(1.83))/6 = 0.17
donc raison = 10^0.17 =1.48

*On peut construire les classes comme suit : d'abord 1.83, puis 2.71 (1.83*1.48), puis 4.01 (2.71*1.48), etc. jusqu'à la valeur maximale.*



<u>Avantages :</u>

cette méthode **s'adapte très bien aux distributions dissymétriques**, qui sont parmi celles que l'on rencontre le plus souvent

<u>Inconvénients :</u>

pas très conseillée pour les comparaisons et **pas la plus simple** à mettre en œuvre.

Méthode 4 : Discrétisation par la méthode des plus grands écarts naturels

Cette expression est un peu ambigüe, puisque **son application n'est pas franchement "naturelle"**. La méthode des **plus grands écarts** revient à **décréter que les plus fortes discontinuités observées dans la distribution constituent des paliers marquants**, qu'il convient donc de séparer en classes. Cette méthode aisée à comprendre n'est par contre pas la plus simple et rapide à mettre en œuvre.

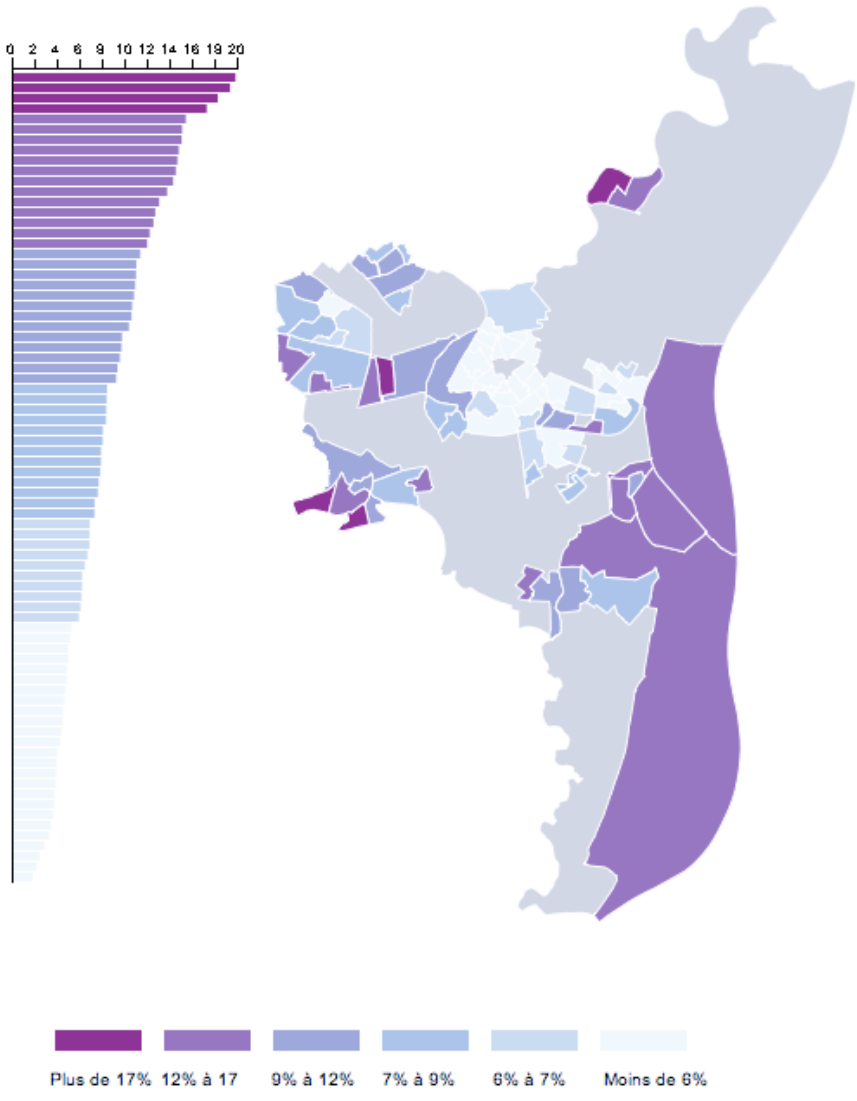
Etape 1 : Classer les données dans l'ordre croissant
Etape 2 : Calculer chaque écart entre chaque couple de données
Etape 3 : Repérer les écarts les plus importants correspondant au nombre de seuils
Etape 4 : borner chaque seuil en faisant la moyenne du couple de données du seuil.

Exemple :

2
5 *5-2 = 2*
9 *9-5 = 4*
16 *16-9=7*
24 *24-16=8*

Pour une légende avec 3 classes, il faut 2 seuils. Les 2 plus grands écarts sont sélectionnés comme seuils.
On calcule la valeur des seuils : (16+9)/2 = 12,5 et (24+16)/2=20

La 1ère classe est donc de 0 à 12,5, la deuxième de 12,5 à 20 et la troisième de plus de 20.



Avantages :

cette méthode est particulièrement conseillée quand on observe une **distribution dont les zones ont leur dispersion caractéristique**. Si on considère par exemple le taux de natalité dans le monde, il y aura un seuil de rupture entre les pays européens et africains, et établir des classes entre les deux se justifiera aisément

Inconvénients :

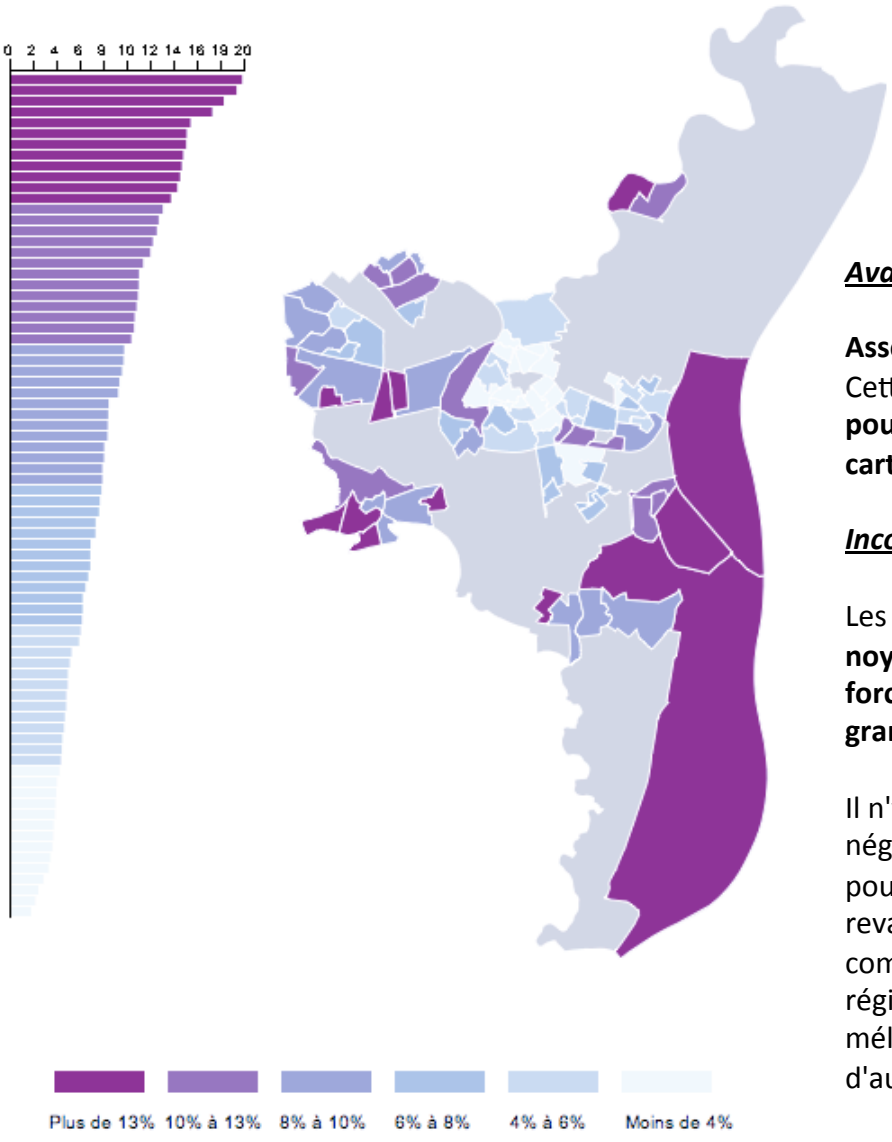
méthode **très aléatoire** : à partir des mêmes données, différents choix de seuils peuvent apparaître. Elle est très sensible aux valeurs extrêmes.

Longue et fastidieuse à traiter.

pas idéale quand il s'agit de comparer

Méthode 5 : Discrétisation par quantiles

La discrétisation par quantiles est très simple à comprendre : elle attribue **le même nombre d'unités géographiques aux classes de la carte**.



Avantages :

Assez simple à réaliser.
Cette discrétisation est **bien profilée pour dresser des cartes comparatives**

Inconvénients :

Les valeurs extrêmes se retrouvent **noyées avec des zones n'ayant pas forcément le même ordre de grandeur**.

Il n'y a donc que lorsqu'elles sont négligeables comme ici que l'on pourra se servir des quantiles. En revanche, dans un cas comme des comparaisons de populations régionales, il serait peu judicieux de mélanger l'Île-de-France avec d'autres régions...

Bibliographie :

Cette fiche s'inspire librement d'un article du site Hypergéo et
du blog blog.m0le.net, lui même inspiré d'un chapitre de «La représentation des données géographiques, statistique et cartographie » de Michèle Béguin et Denise Pumain.