Méthodes de discrétisation

Une distribution spatiale définie par un ensemble de données peut être réduite ou plutôt généralisée en effectuant un regroupement en classes; c'est ce qu'on appelle une « mise en classes » ou une « discrétisation des données ». Plusieurs méthodes de discrétisation des données sont possibles.

Pour des raisons pédagogiques, et dans le but d'établir des comparaisons entre les différentes méthodes, on utilisera la même distribution des données spatiales définies pour la variable suivante : « les densités rurales de population des 105 cantons du Kansas Centre (États-Unis d'Amérique) ». Ces données ont été originalement utilisées par George F. JENKS dans un article intitulé *« Generalization in Statistical Mapping* » publié dans les **Annals**, Association of American Geographers, Vol. 53, no 1, 1963. On présente aussi un ensemble de dix (10) figures relativement aux données brutes ainsi qu'aux méthodes de discrétisation avec la représentation cartographique correspondante (cartes choroplèthes); toutes les figures sont réunies sur les deux dernières pages.

Présentation des données brutes

Les densités rurales de population (personnes / mille carré) ainsi que le découpage spatial (unités spatiales) des 105 cantons sont présentées à la **Figure 1**.

Que peut-on constater ?

- 105 unités spatiales (cantons);
- 105 valeurs de densité rurale de population;
- étendue des données : minimum et maximum de 1,6 et 103,4 personnes / mi^2 respectivement;
- un examen des données, nous permet de remarquer que 101 cantons sur 105 ont des données variant entre 1,6 et 13,5 personnes / mi^2. Cette situation est facilement visualisée par la construction d'un graphique de dispersion des données (**Figure 2**). Il s'agit d'un outil analytique aidant le cartographe à mieux comprendre la nature des données à cartographier.

Le cartographe sera donc intéressé à utiliser une méthode discrétisation qui pourra le mieux représenter cette distribution spatiale de données.

Les méthodes de discrétisation

A) Intervalles de classe constants (valeur égale, seuils déterminés, moyenne, écarts types)

- **nombres entiers :** on choisit des nombres entiers qui conviennent en tant que limites de division; le choix étant dépendant de l'étendue des données. Par exemple, on a choisi comme nombres entiers 5, 10, 20, 50 et 100; ce qui permet d'obtenir six classes (**Figure 3**).
- **limites de classes correspondant à une division régulière selon des intervalles de valeur égale :** une fois le nombre de classes choisi, on divise l'étendue exactement ou « approximativement » en ce nombre de divisions égales. Par exemple, supposons que l'on désire six classes

  Étendue = Maximum – Minimum = 103,4 – 1,6 = 101,8
  Intervalle de classe = Étendue / Nombre de classes = 101,8 / 6 = 16,97

  La valeur de « 16,97 » sera ajoutée la valeur minimale pour donner les limites des classes, ainsi

  1,6 + 16,97 = 18,57 >> 18,6 ($1^{ère}$ classe);
  18,57 + 16,97 = 35,54 >> 35,5 ($2^{ème}$ classe);
  35,54 + 16,97 = 52,21 >> 52,5 ($3^{ème}$ classe), etc.
  Une façon plus rapide, mais approximative, consiste à choisir un nombre un peu plus grand que l'étendue (101,8) mais qui soit exactement divisible par le nombre de

classes désiré. Dans notre exemple, on emploie le nombre 102 (au lieu de 101,8) donnant ainsi un intervalle de 17 (102 / 6 = 17) et les limites des classes seront

1,6 + 17 = 18,6 (1ère classe);
18,6 + 17 = 35,6 (2ème classe);
35,6 + 17 = 52,6 (3ème classe), etc.

Ce sont ces limites de classes qui ont permis de généraliser l'information spatiale telle que présentée à la **Figure 4**.

- **limites de classes correspondant à des seuils déterminés à partir de critères:** il est souvent préférable, en climatologie, par exemple, d'utiliser comme limites de classes des seuils significatifs. Une variable définie par le taux d'accroissement de la population (augmentation de la population = taux positif; diminution de la population = taux négatif; population inchangée = taux nul) aurait avantage à être discrétisée de sorte à mettre en évidence la valeur de « 0 », c'est-à-dire un taux d'accroissement nul de la population.
- **distribution centrée sur la moyenne avec des limites de classes choisies arbitrairement :** on détermine les classes en fonction de la tendance centrale de la distribution. Dans l'exemple, ayant une moyenne de 7,58 ≅ 7,6; on fixe les limites des classes de façon à obtenir des intervalles égaux. Pour ce faire, on ajoutera et retranchera une étendue (intervalle) de 4 à partir de la moyenne 7,6 :

  < 3,6 (1ère classe);
  3,6 à **7,6** (2ème classe);
  7,6 à 11,6 (3ème classe);
  11,6 à 15,6 (4ème classe);
  > 15,6 (5ème classe)

- **distribution centrée sur la moyenne avec des limites de classes choisies en tenant compte de la dispersion des données :** on utilise l'écart type, paramètre de la dispersion des données autour de la moyenne, comme élément de discrétisation des classes. Une fois la moyenne et l'écart type déterminés, les classes sont obtenues selon une des méthodes suivantes :

  1) à partir de la valeur de moyenne, en additionnant et en soustrayant **symétriquement** une valeur d'écart type à la fois

  Moyenne – 3s à Moyenne – 2s (1ère classe);
  Moyenne – 2s à Moyenne – 1s (2ème classe);
  Moyenne – 1s à **Moyenne** (3ème classe);
  -------- << Moyenne située entre les limites de la 3ème et la 4ème classe >> --------
  Moyenne à Moyenne + 1s (4ème classe);
  Moyenne + 1s à Moyenne + 2s (5ème classe);
  Moyenne + 2s à Moyenne + 3s (6ème classe)

  Selon cette méthode, on ne fait pas bien ressortir ce qui se passe vraiment autour et près de la moyenne parce que la moyenne se trouve à la limite inférieure de la classe (**Moyenne** à Moyenne + 1s) et aussi à la limite supérieure de la classe (Moyenne – 1s à **Moyenne**) [la 3ème et la 4ème classe].

  2) à partir de la moyenne, mais en s'assurant que la moyenne soit au centre d'une classe qu'on appellera d'ailleurs **classe moyenne**. Pour ce faire, on utilise des valeurs de demis écarts types.

  Moyenne – 3,5s à Moyenne – 2,5s (1ère classe);
  Moyenne – 2,5s à Moyenne – 1,5s (2ème classe);
  Moyenne – 1,5s à Moyenne – 0,5s (3ème classe);
  Moyenne – 0,5s à Moyenne + 0,5s (4ème classe); **<< --- Classe moyenne**
  Moyenne + 0,5s à Moyenne + 0,5s (5ème classe);

Moyenne + 1,5s à Moyenne + 2,5s (6ème classe);
Moyenne + 2,5s à Moyenne + 3,5s (7ème classe)

Dans l'exemple sur les 105 divisions de cantons, on a une moyenne de 7,58 et un écart type de 10,36 (causé largement par l'effet de quelques valeurs extrêmes parmi les 105); en prenant ±1s de part et d'autre de la moyenne, on placerait ainsi 101 des 105 unités spatiales (les données) dans deux classes; ce qui offre très peu de différentiation spatiale. Autrement, avec ± 0,5s (demis écarts types : 10,36 / 2 = 5,18), on obtient

Les limites des classes :
Moyenne – 0,5s = 7,58 – 0,5(10,36) = **2,40**
Moyenne + 0,5s = 7,58 + 0,5(10,36) = **12,76**
Moyenne + 1,5s = 7,58 + 1,5(10,36) = **23,12**
Moyenne + 2,5s = 7,58 + 2,5(10,36) = **33,48**
Moyenne + 3,5s = 7,58 + 3,5(10,36) = **43,84**

Moins que 2,40 (1ère classe);
2,40 à 12,76 (2ème classe); **<< --- Classe moyenne**
12,76 à 23,12 (3ème classe);
23,12 à 33,48(4ème classe);
33,48 à 43,84 (5ème classe);
Plus que 43,84 (6ème classe)

La **Figure 5** présente la carte résultante de cette dernière mise en classes.

B) Intervalles de classe variables (intervalles mathématique et géométrique, quantiles)

Examinons quelques exemples d'intervalles mathématiques; soient les classes suivantes :

1) 0 – 4,9, 5,0 – 9,9, 10,0 – 14,9, etc.
   ➔ intervalles égaux (valeur de « 5 » entre chaque limite)

2) 0 – 4,9, 5,0 – 14,9, 15,0 – 29,9, etc.
   ➔ intervalles de 5, 10, 15, etc. (progression arithmétique)

3) 0 – 4,9, 5,0 – 14,9, 15,0 – 34,9, 35,0 – 74,9, etc.
   ➔ intervalles de 5, 10, 20, 40, etc. (progression géométrique)

- **progression arithmétique des intervalles :**

$A + X + 2X + 3X + \ldots + NX = B$ où

A = valeur minimale
B = valeur maximale
N = nombre de classes désiré

On veut généraliser l'information de la distribution de la densité rurale de population en sept (7) classes.

A = 1,6 ; B = 103,4 et N = 7

On a $A + X + 2X + 3X + \ldots + NX = B$ qui peut s'écrire comme suit

$1,6 + X + 2X + 3X + 4X + 5X + 6X + 7X = 103,4$
$1,6 + 28X = 103,4$
où $X = (103,4 – 1,6) / 28 = 101,8 / 28 = 3,64$,
ainsi, on aura les limites de classes suivantes :

1ère classe = 1,6
2ème classe = 1,6 + X = 1,6 + 3,64 = 5,24

3ème classe = 1,6 + X + 2X = 5,24 + 2(3,64) = 12,52
4ème classe = 1,6 + X + 2X + 3X = 12,52 + 3(3,64) = 23,44
5ème classe = 1,6 + X + 2X + 3X + 4X = 23,44 + 4(3,64) = 38,0
6ème classe = 1,6 + X + 2X + 3X + 4X + 5X = 38,0 + 5(3,64) = 56,2
7ème classe = 1,6 + X + 2X + 3X + 4X + 5X + 6X = 56,2 + 6(3,64) = 78,0

1ère classe : 1,6 à 5,24
2ème classe : 5,24 à 12,52
3ème classe : 12,52 à 23,44
4ème classe : 23,44 à 38,0
5ème classe : 38,0 à 56,2
6ème classe : 56,2 à 78,0
7ème classe : 78,0 à 103,4

- **progression géométrique des intervalles :**

1. Trouver la différence entre le log (Minimum) et le log (Maximum)

   Log (Maximum) – Log (Minimum) = log 103,4 – log 1,6 = 2,01452 – 0,20412 = 1,81040 et la diviser par le nombre de classes désiré, c'est-à-dire 1,81040 / 6 (nombre de classes) = 0,30173

2. En prenant le log de la valeur maximale (log 103,4 = 2,01452), lui soustraire la valeur de 0,30173 six fois (nombre de classes) :

   2,01452 – 0,03173 = 1,71279
   1,71279 – 0,03173 = 1,41106
   1,41106 – 0,03173 = 1,10933
   1,10933 – 0,03173 = 0,80760
   0,80760 – 0,03173 = 0,50587
   0,50587 – 0,03173 = 0,20414 (<<= identique au log de la valeur minimale; log 1,6 = 0,20412)

3. Les limites des classes sont déterminées en prenant l'antilog des valeurs obtenues au point 2 ci-dessus :

   antilog 1,71279 = 51,62
   antilog 1,41106 = 25,77
   antilog 1,10933 = 12,86
   antilog 0,80760 = 6,42
   antilog 0,50587 = 3,21
   antilog 0,20414 = 1,60

   Note : La progression géométrique s'applique dans cet exemple comme suit : les limites supérieures sont toujours "2,0032" (c'est-à-dire l'antilog 0,30173) fois les limites inférieures.

   Dans la **Figure 6**, on présente la représentation cartographique des classes suivantes :

   1ère classe : 1,6 à 3,21
   2ème classe : 3,21 à 6,42
   3ème classe : 6,42 à 12,86
   4ème classe : 12,86 à 25,77
   5ème classe : 25,77 à 51,62
   6ème classe : 51,62 à 103,4

- **Méthode des quantiles :**

Les **quantiles** sont des valeurs qui partagent une distribution de données “ordonnées” en des groupes plus petits (généralisation), chacun contenant le même nombre d’éléments ou presque. On pourra utiliser les quantiles en tant que base de discrétisation en cartographie. Voici quelques quantiles : la médiane qui partage une distribution en 2 groupes, les terciles, en 3 groupes, les quartiles en 4, les quintiles en 5, les sextiles en 6, les septiles en 7, les octiles en 8, les noniles en 9, les déciles en 10.

La distribution des 105 valeurs de densité rurale de population peut facilement être partagée en sept (7) groupes de 15 valeurs ou cinq (5) groupes de 21 valeurs, et ces deux mises en classes pourraient être utilisées. Par exemple, si sept (7) groupes de 15 valeurs sont considérés, les premières 15 valeurs, dont l’intervalle s’étend de 1,6 à 3,3, constitueraient la première classe; la deuxième classe s’étendrait de la 16^{ième} à la 30^{ième} valeur, soit de 3,4 à 4,8 et ainsi de suite [la limite de la classe peut être, par convention, fixée à 3,35, etc.]. Cependant, on a souvent des raisons (voir plus bas) de vouloir 4, 6 ou 8 classes et, comme on peut le constater dans notre exemple, le nombre d’observations ne se divisera pas en un nombre pair de valeurs; il faudrait alors répartir les valeurs supplémentaires parmi les classes de la façon la plus convenable. C’est ce qu’on a fait pour la carte présentée à la **Figure 7** où six (6) classes ont été choisies; compte tenu que 105 / 6 = 17 avec un reste de 3 que l’on a distribué de la façon suivante : **18**, 17, **18**, 17, **18** et 17 observations formant ainsi des sextiles. Si huit (8) classes avaient été choisies, on aurait eu 105 / 8 = 13 avec un reste de 1, donnant ainsi des classes avec 13, **14**, 13, 13, 13, 13, 13 et 13 observations.

C) <u>Méthode du graphique de dispersion</u> (discrétisation naturelle)

La méthode du graphique de dispersion permet de déterminer les classes selon des divisions « naturelles » par opposition aux autres méthodes où les limites des classes sont établies soient arbitrairement ou par un processus statistique ou mathématique. La construction du graphique de dispersion est relativement simple; en effet, on place sur l’axe des X la série de valeurs de la variable à l’étude et on localise un point pour chacune des valeurs. Une fois toutes les valeurs placées sur le graphique, on peut, **la plupart du temps**, déterminer, de façon tout à fait visuelle, des seuils pertinents en tant que limites de classes. On obtient alors un découpage empirique, c’est-à-dire non basé sur une méthode statistique ou mathématique, de l’espace selon de regroupements significatifs.

Il serait également possible d’utiliser le graphique de dispersion des données tel que construit à la **Figure 2** de notre exemple sur les densités rurales de population. Veuillez noter que ce graphique a la particularité d’ordonner les unités spatiales sur l’axe des X en présentant les valeurs de densité rurale de population sur l’axe des Y. La répartition des densités rurales de population du Kansas Centre, discrétisées selon des divisions naturelles, est présentée à la **Figure 8**.

<u>Avantages et désavantages des méthodes de discrétisation</u>

Face aux différentes méthodes de discrétisation, laquelle doit-on choisir ? Laquelle est la « meilleure » ? Il n’existe malheureusement aucune formule magique; cependant, on peut utiliser les « conseils-guides » suivants :

- Considérer toujours la possibilité d’utiliser la méthode de discrétisation naturelle. Il faut cependant s’assurer que les regroupements soient **évidents** (limites des classes); de cette façon, il n’y a pas d’ambiguïté (absence d’arbitraire).
- L’emploi des nombres entiers est à éviter si cette méthode est choisie arbitrairement pour sauver du temps et du travail. Cependant, si les données sont dispersées de façon continue sans aucune coupure marquée [situation très rare car l’information géographique est le plus souvent caractérisée par des cas particuliers et est soumise à l’effet de l’autocorrélation spatiale], on pourra utiliser des nombres entiers qui

fourniront des limites de classes facilement mémorisables et une lecture plus simple et facile de la carte.

- Emploi des quantiles : À l'opposé des méthodes des intervalles constants (valeur égale), les quantiles permettent d'obtenir des nombres égaux (ou presque) d'observations dans chaque classe; voilà un avantage à utiliser avec précaution cependant, car il est le plus souvent moins utile qu'il ne semble l'être. Par exemple, dans la représentation cartographique des 105 valeurs de densité rurale de population, regroupées en 6 classes avec les sextiles comme méthode de discrétisation (**Figure 6**), la classe 8,65 à 103,4 est « étirée » sur une grande étendue de valeurs; en effet, tout ce qui est supérieur à 8,65 doit être compris dans une classe pour combler les 17 observations requises. Il semble que la seule utilité de cette méthode de discrétisation soit de permettre l'agrégation de classes dans le but de montrer par exemple le tiers d'effectifs (d'observations) le plus élevé, le tiers au centre et le tiers le plus bas.
- L'emploi de méthodes « mathématiques » plus complexes pour déterminer les limites de classes, comme les progressions arithmétique et géométrique, possède une seule qualité, c'est-à-dire permettre de construire des classes lorsque la distribution de données a une grande étendue. Mais cette qualité peut être coûteuse dans la mesure où on ne connaît jamais d'avance le découpage spatial résultant; on pourra peut-être, après de longs calculs, constater des « lacunes cartographiques » et, on devra opter pour une autre méthode de discrétisation. Dans la **Figure 6**, presque la moitié des valeurs sont dans la classe 6,42 à 12,86, tandis que presque toute l'autre moitié inférieure à 6,42 est partagée en deux classes et les quatre autres classes comprennent seulement 6 valeurs en tout. Ce résultat est dépendant de la mathématique qu'exige la méthode; c'est donc désastreux … Même avec des étendues très grandes, une discrétisation naturelle aurait probablement produit des résultats (classes) meilleurs.
- L'emploi des valeurs d'écart type a l'avantage d'être à la fois plus réel et plus subtil. Parce que c'est une mesure de dispersion et de standardisation, l'écart type est souvent employé pour comparer les valeurs de différentes variables (caractéristiques) pour une même unité spatiale; par exemple, on peut dire que pour un secteur X, la précipitation est d'un (1) écart type au-dessus de la moyenne et que la quantité de blé produit à l'acre est de deux (2) écarts types au-dessus de la moyenne, et ainsi de suite. Attention cependant à la représentativité de la moyenne et de l'écart type; en effet, ces deux mesures sont influencées par les valeurs extrêmes d'une distribution de données lorsque celle-ci est asymétrique. C'est ce qui se produit dans la **Figure 5**; avec une moyenne de 7,58 et un écart type de 10,36, **92** des 105 valeurs sont dans la même classe (2,40 à 12,76). Comment peut-on remédier à cette lacune ? En éliminant tout simplement des valeurs extrêmes : si on élimine la valeur maximale (103,4), on obtient une moyenne de 6,65 et un écart type de 4,39; si on élimine les trois (3) valeurs les plus élevées (27,4, 34,6 et 103,4), on obtient une moyenne de 6,18 et un écart type de 2,74. On présente à la **Figure 9** une nouvelle version de la carte basée sur une mise en classes avec l'écart type tout en ayant exclu les trois (3) valeurs les plus élevées (les unités spatiales sont notées par un « X » sur la carte).

La **Figure 10** offre une façon plutôt originale de représenter les informations relatives aux densités rurales de population; en effet, à partir de données zonales (cantons = unités spatiales zonales), on attribue au centroïde de chacune des zones la valeur de densité correspondante, on trace, ensuite, des isocourbes (méthode d'interpolation des isocourbes) à partir des points centraux. On obtient des isoplèthes pour les densités rurales de population; un dégradé des trames est alors appliqué entre les isoplèthes, constituant ainsi une chorisoplèthe. La méthode de discrétisation naturelle a été utilisée (comparée avec la **Figure 8**).

Texte préparé par :
Michel Dufault
Département de géographie, UQAM
Mars 2001