10. Traitement statistique de l'information : La discrétisation

Exprimer (carto-)graphiquement une information implique souvent sa simplification. Ainsi en découpant en classes une série de données, la discrétisation réduit une variable quantitative en variable ordonnée.

> *«Perdre le moins possible d'information, du point de vue de la statistique, reviendrait à cartographier toutes les valeurs, sans discrétisation préalable. Or, nous savons que les règles de la perception visuelle contredisent cette exigence, la lisibilité de l'image cartographique étant d'autant plus grande que le nombre de signes distincts, donc de classes, est petit. Il est donc indispensable de procéder à une «réduction» de l'information statistique.» (BEGUIN & PUMAIN, 1994)*

La réduction statistique, appelée discrétisation, doit conserver au mieux l'information tout en la simplifiant. Ces objectifs, a priori contradictoires, peuvent être conciliés en respectant certaines règles :

- Indiquer les descripteurs clé d'une distribution. L'expression graphique fait surtout ressortir les inégalités, les variations dans l'espace. Pour leur interprétation correcte, il convient d'indiquer certains des descripteurs suivants: moyenne, mode, minimum, maximum et écart-type.
- Respecter la forme de la distribution. La plupart des séries de données ressemblent à un type de distribution: uniforme, symétrique (normale), dissymétrique (lognormale ou non). La méthode de discrétisation tiendra compte du «modèle» de distribution que la série épouse.
- Appliquer le principe de ressemblance / dissemblance. Toute discrétisation s'efforce de minimiser la variance intraclasse tout en maximisant la variance interclasse.

A. Présentation de méthodes

La discrétisation soulève deux questions:

- Combien de classes créer?
- Où placer les bornes des classes?

Le nombre de classes

Il doit évidemment être en rapport avec le nombre d'individus de la série de valeurs: sa longueur. Ainsi pour des valeurs relatives aux 22 régions métropolitaines, inutile de discrétiser en 2 classes (appauvrissement excessif) ou en 10 classes (risque de classes vides).

Outre la longueur de la série, les règles sémiologiques imposent certaines limites. L'expression graphique - et notamment cartographique - ne supporte qu'un nombre de classes limité. Si on veut que le lecteur puisse aisément distinguer les classes (perception sélective), leur nombre se limite à 5 - 7 (selon le nombre d'individus). Si, au contraire, le but de la carte est de montrer une progression plus ou moins continue, un gradient, le nombre de classes peut être nettement supérieur. Tel est le cas des cartes de diffusion.

Si on veut mathématiquement définir le nombre de classes en fonction de la longueur de la série, on peut se référer à l'une des deux méthodes les plus usuelles:

- Méthode de Huntsbeerger: le nombre optimal de classes est

$$\boxed{k = 1 + 3.3 \log_{10} n}$$

où k = nombre de classes et n = longueur de la série, son effectif total

- Méthode de Brooks-Carruthers: le nombre optimal de classes est

$$k \leq 5 \log_{10} n$$

La discrétisation

Seront présentées ici les méthodes de discrétisation les plus utilisées: mathématiques, statistiques et graphiques.

- **Discrétisation par équidistance**: appelée aussi **discrétisation en classes d'amplitude égale**. L'écart entre les classes est constant. Il suffit de diviser la dynamique de la série (maximum – minimum) par le nombre de classes souhaité.
  Avantages et inconvénients: Facile à réaliser, mais la qualité de la discrétisation est tributaire de la distribution des valeurs de la série. Ainsi cette méthode risque de produire des classes vides si la distribution est asymétrique ou comporte des fortes discontinuités pour une étendue donnée (par exemple la densité de la population par région où l'Ile de France se détache nettement du reste du territoire français (cf. l'exemple développé ci- dessous).

- **Discrétisation par progression arithmétique**: contrairement à la discrétisation par équidistance, l'amplitude des classes augmente selon une progression arithmétique à raison de R. R étant la dynamique (étendue) de la série divisée par l'addition des classes. Exemple pour 5 classes:

$$\boxed{R = (max – min) / (1 + 2 + 3 + 4 + 5)}$$

  L'amplitude des classes: classe n°1: min – min+R, classe n° 2: min+R – min+2R, etc.
  Avantages et inconvénients: La méthode est conçue pour les distributions asymétriques avec beaucoup de valeurs faibles et peu de valeurs fortes (cf. l'exemple développé ci-dessous). Si, au contraire, la distribution s'apparente à la loi normale, la méthode n'est pas adaptée.

- **Discrétisation par équifréquence**: appelée également **discrétisation selon les quantiles** ou **par équipopulation**. Toutes les classes ont, dans la mesure du possible, le même nombre d'individus. La réalisation repose sur un tri par ordre croissant des individus de la série et le calcul du nombre idéal d'individus par classe (nombre d'individus / nombre de classes). On affecte alors les individus de la série selon leur rang aux classes successives. Les bornes des classes peuvent être définies par les valeurs extrêmes de chaque classe. Une autre méthode, qui ne change en rien l'effectif des classes, consiste à calculer les bornes par interpolation simple entre les deux valeurs successives appartenant à deux classes successives: (valeur maximale de la classe n + valeur minimale de la classe n+1)/2. Pour le confort de lecture, on peut les arrondir.
  Avantages et inconvénients: La discrétisation selon les quantiles diminue le poids des valeurs extrêmes. Elle les nivelle en les regroupant avec des valeurs plus proches de la moyenne pour équilibrer les classes. De ce point de vue son effet est comparable à un filtre passe-bas (lissage) en traitement d'image. Il s'agit, en cartographie, de la méthode qui affecte à chaque classe le même nombre d'unités spatiales. Par conséquent, la carte est visuellement la plus riche, la plus équilibrée mais pas forcément la plus juste car elle masque les fortes discontinuités (par exemple l'Ile de France sera regroupée avec des régions nettement moins peuplées, cf. l'exemple ci-dessous). Plus généralement, la méthode peut regrouper des individus assez éloignés et, par conséquent, ne pas respecter les seuils dans la distribution.

- **Discrétisation selon les moyennes emboîtées** : la moyenne arithmétique divise la série en deux classes primaires. Leurs moyennes respectives les divisent à leur tour en deux et de suite. Le nombre de classes est toujours un multiple de 2 ce qui peut s'avérer contraignant.
  Avantages et inconvénients: La méthode est basée sur la moyenne considérée comme centre de gravité de la distribution ainsi que sur les moyennes des sous-ensembles. L'inconvénient majeur est la rigidité de la méthode en nombre de classes: multiple de 2. La moyenne peut également découper des ensembles de valeurs proches (non respect des seuils naturels de la distribution).

- **Discrétisation standardisée**: elle se base sur une distribution supposée proche de la loi normale. On calcule la moyenne arithmétique (point de référence) et l'écart-type en tant qu'indicateur de l'amplitude des classes. Si le nombre de classes est pair, la moyenne sépare les deux classes centrales. Si le nombre de classes est impair, la moyenne est au centre de la classe centrale. L'ampleur des classes est généralement d'un écart-type.
  Avantages et inconvénients: Cette méthode est d'autant plus adaptée que la distribution des valeurs de la série est proche de la distribution normale. Elle est également utilisée pour la comparaison (séries de cartes). Si la distribution des valeurs n'est pas normale mais pourrait la devenir en tenant compte d'un facteur de progression (lognormal ou autre), il convient de l'appliquer. Dans le cas contraire, il est possible (attention toutes les distributions ne s'y prêtent pas) de la normaliser: la transformation (cf., entre autres, LAHOUSSE & PIEDANNA, 1998).

- **Discrétisation selon la méthode de Jenks** : elle se base sur le principe de ressemblance / dissemblance en calculant la distance paramétrique entre toutes les valeurs de la série. La méthode minimise la variance intraclasse et maximise la variance interclasse. Il s'agit d'un algorithme itératif formant, dans un premier temps, autant de couples qu'il y a de combinaisons de valeurs (triées par ordre croissant) pour un nombre de classes donné. On calcule alors la variance intraclasse et interclasse.

Avantages et inconvénients: Il s'agit d'une méthode respectant l'allure de la série mais elle ne permet pas de les rendre comparables entre elles. Cette méthode est certainement la plus adaptée des méthodes statistiques. Sa mise en œuvre nécessite l'emploi d'un logiciel statistique.

- **Discrétisation manuelle**: appelée aussi méthode des seuils observés. Contrairement aux autres méthodes, il s'agit d'une méthode manuelle où l'opérateur découpe la série visualisée graphiquement (histogramme de fréquence) selon les discontinuités (seuils). Le nombre de classes résultant est un compromis entre l'allure de la série (seuils, discontinuités) et le nombre initialement projeté.
  Avantages et inconvénients: Le méthode convient d'autant plus que la distribution est «typée» (série asymétrique, plurimodale). De ce point de vue il s'agit de la méthode opposée à la discrétisation par équifréquence notamment préconisée en cas de répartition uniforme des valeurs. La discrétisation manuelle s'appuyant sur des seuils observés est certainement la méthode de bon sens dans la mesure où elle produit des cartes équilibrées graphiquement tout en épousant la distribution de la série. Elle a l'avantage d'obliger l'opérateur à analyser graphiquement la série de valeurs. Un tel examen graphique (histogramme de fréquence) devrait précéder toute discrétisation (aveugle).
  La significativité des discontinuités observées dépend néanmoins de la longueur de la série. Ainsi, le constat d'un seuil dans une population de 500 individus est autrement plus significatif que l'observation de la même discontinuité dans une population nettement plus réduite. D'autre part, l'observation même dépend étroitement du choix de la finesse graphique de l'échelle de l'histogramme de fréquence. De ce point de vue, la méthode paraît bien subjective.
  Bien que manuelle, des possibilités d'automatisation existent: une fois triés les individus par ordre croissant de leur valeur, on calcule l'écart entre deux rangs successifs. Dès que l'écart dépasse un seuil préalablement établi ou calculé de manière circonstancielle (moyennes mobiles), l'algorithme place la borne d'une classe.

B. Exemple: la densité de population par région

Mise à part la méthode de Jenks, pour laquelle il faut disposer d'un programme statistique, nous illustrons les méthodes de discrétisation par un exemple simple: la densité de population (habitants / km²) par région (données du recensement 1990, France métropolitaine).

Région	hab/km²
CORSE	28.77
LIMOUSIN	42.67
AUVERGNE	50.81
BOURGOGNE	50.97
CHAMPAGNE-ARDENNE	52.65
MIDI-PYRENEES	53.62
CENTRE	60.56
POITOU-CHARENTES	61.82
AQUITAINE	67.69
FRANCHE-COMTE	67.73
LANGUEDOC-ROUSSILLON	77.30
BASSE-NORMANDIE	79.14
PICARDIE	93.35
PAYS DE LA LOIRE	95.31
LORRAINE	97.95
BRETAGNE	102.75
RHONE-ALPES	122.48
PROVENCE-ALPES-COTE D'AZUR	135.66
HAUTE-NORMANDIE	141.09
ALSACE	196.20
NORD-PAS-DE-CALAIS	319.49
ILE-DE-FRANCE	887.59

Descripteurs de la série et nombre de classes souhaitées (5):

Indicateurs		
Nombre d'individus	n	22
Maximum	max	887.59
Minimum	min	28.77
Moyenne arithmétique	x	131.16
Ecart-type	σ	180.41

Nombre de classes souhaité	k	5

Discrétisation par équidistance		
Amplitude a=(max-min)/k=171.76		valeur
Borne 0	min	28.77
Borne 1	min + a	200.53
Borne 2	min +2a	372.30
Borne 3	min +3a	544.06
Borne 4	min +4a	715.82
Borne 5	max	887.59

Discrétisation par progression arithmétique		
Amplitude a = n° de classe * R où R = (max-min)/(1+2+3 …		valeur
Borne 0	min	28.77
Borne 1	min +R	86.02
Borne 2	borne2 +2R	200.54
Borne 3	borne3 +3R	372.30
Borne 4	borne 4 + 4R	601.32
Borne 5	max = borne5+5R	887.59

Discrétisation par équifréquence			
Amplitude a =n/k = 4.4			valeur (interpolé)
Borne 0	1	1.00	28.77
Borne 1	a	4.4 -> 4	51.81
Borne 2	2a	8.8 –>9	67.71
Borne 3	3a	13.20 -> 13	94.53
Borne 4	4a	17.60 -> 18	138.36
Borne 5	n	22.00	887.59

Nombre d'individus par classe: classe n°1=4, classe n°2=5, classe n°3=4, classe n°4=5, classe n°5=4

Discrétisation par moyennes emboîtées		
(4classes)		valeur
moyenne 1ère classe primaire	x2a	70.92
moyenne 2ème classe primaire	x2b	336.00
Borne 0	min	28.77
Borne 1	x2a	70.92
Borne 2	x	131.16
Borne 3	x2b	336.00
Borne 4	max	887.59

Discrétisation standardisée		
		valeur
Borne 0	x - 2.5 s (ou min)	-229.66
Borne 1	x - 1.5 s	-139.46
Borne 2	x - 0.5 s	40.96
Borne 3	x + 0.5s	221.37
Borne 4	x + 1.5s	401.78
Borne 5	x + 2.5s (ou max)	887.59

Discrétisation manuelle par seuils observés

Plusieurs discrétisations sont possibles. La solution proposée individualise les deux régions les plus densément peuplées séparément. Il en résulte six classes.

Discrétisation manuelle des seuils observés			
		valeur	nombre de régions
Borne 0	min	28.77	2
Borne 1		50	8
Borne 2		77	6
Borne 3		103	4
Borne 4		200	1
Borne 5		400	1
Borne 6	max	887.59	

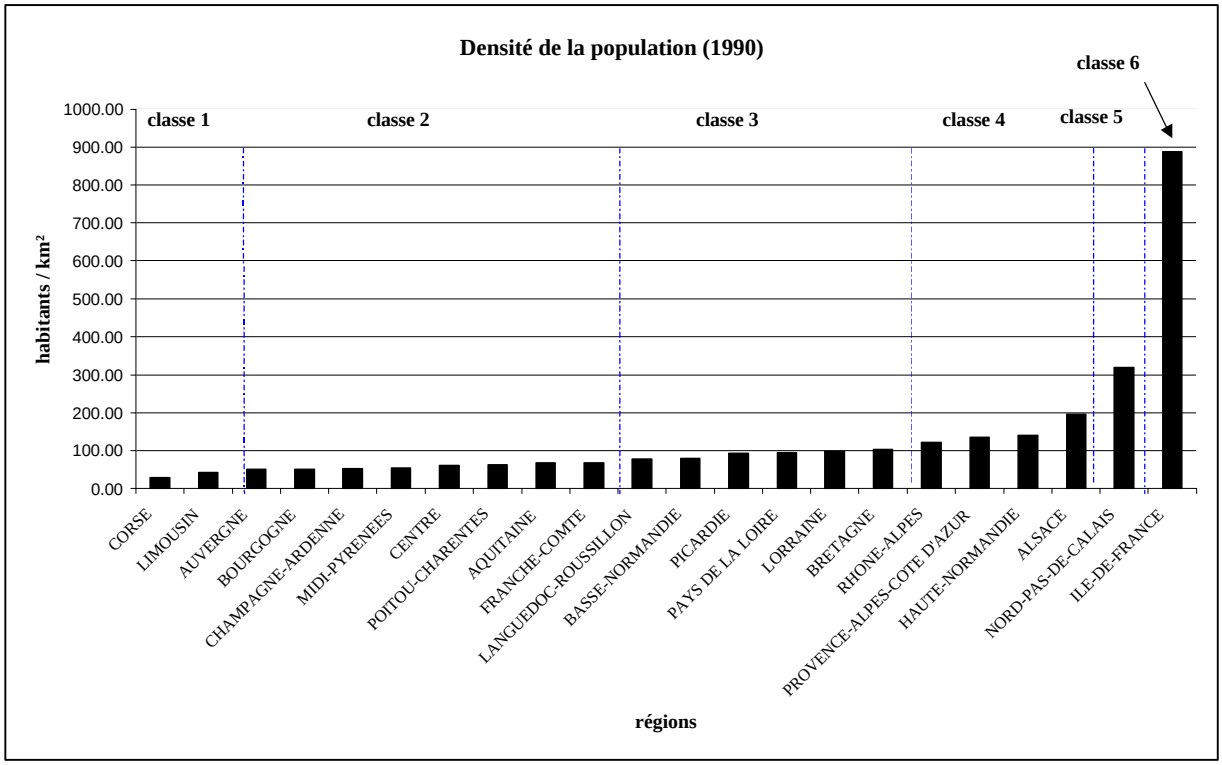
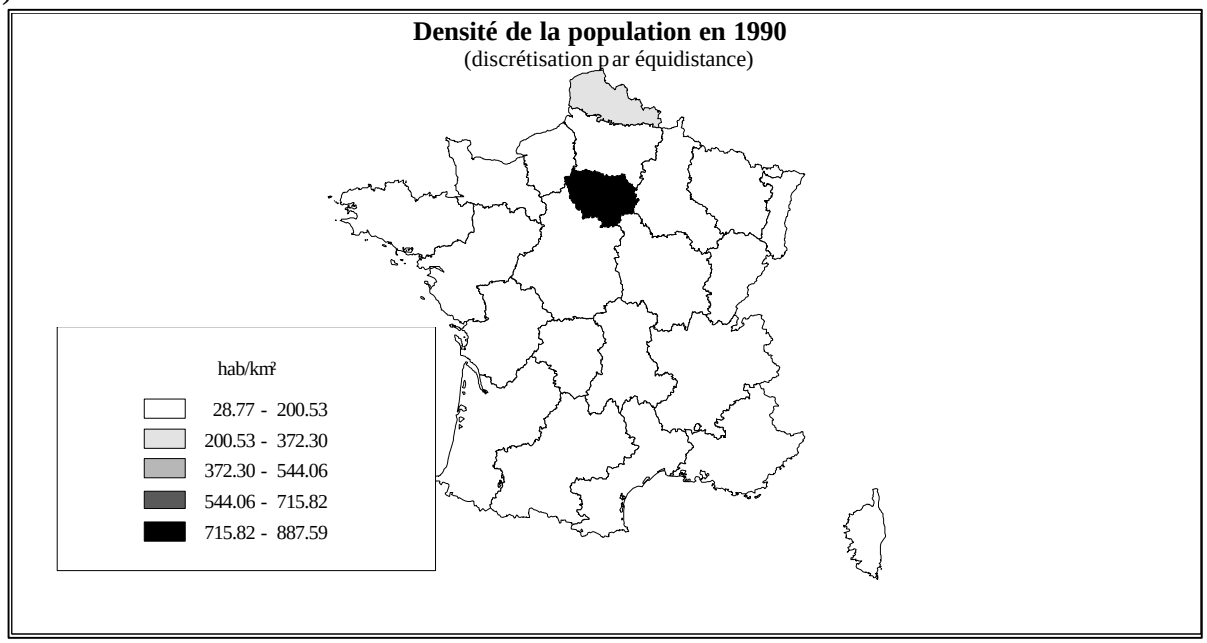
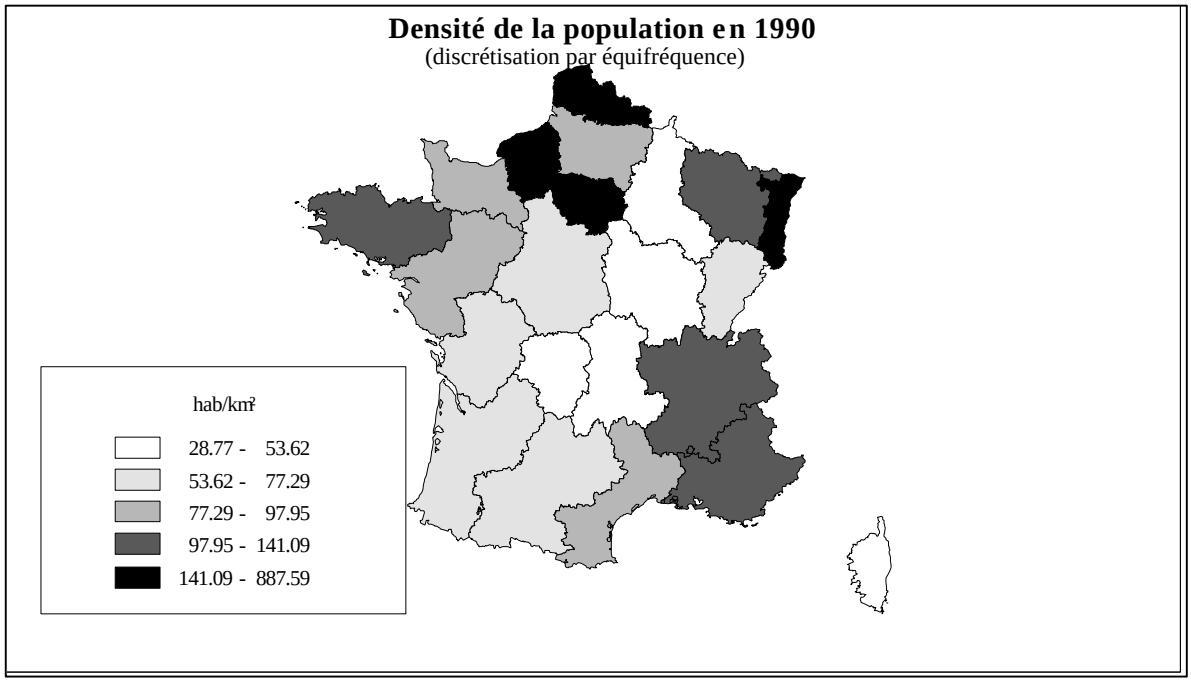


Fig. 85: Densité de la population par région en 1990 : série de données, descripteurs statistiques, discrétisations, histogramme de fréquence et seuils observés

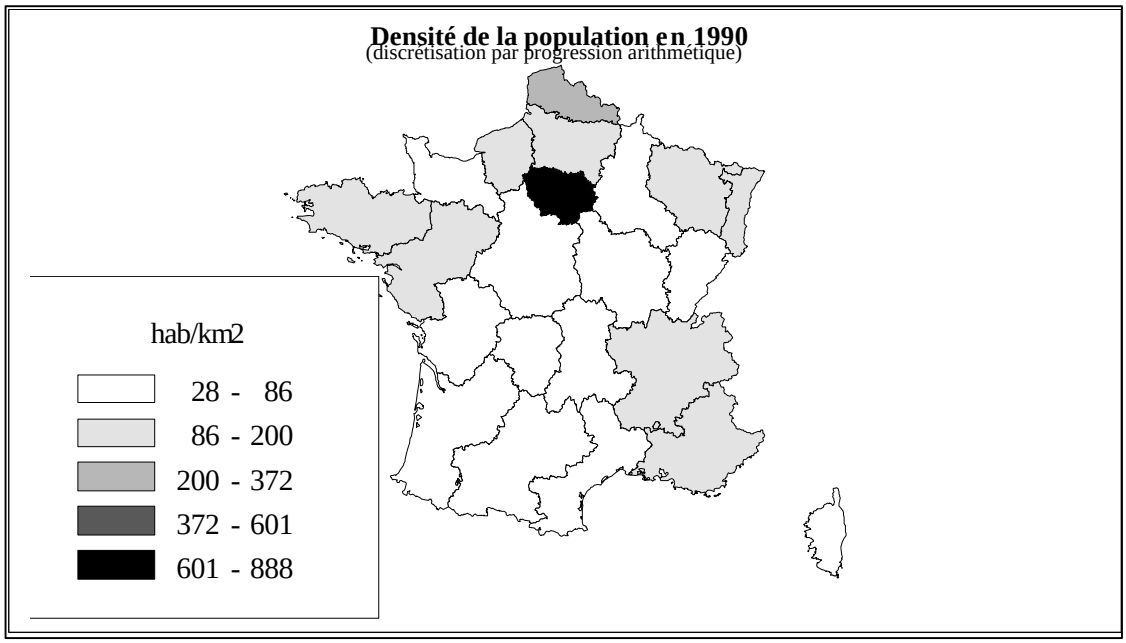
a)


Densité de la population en 1990
(discrétisation par équidistance)
hab/km²
28.77 - 200.53
200.53 - 372.30
372.30 - 544.06
544.06 - 715.82
715.82 - 887.59

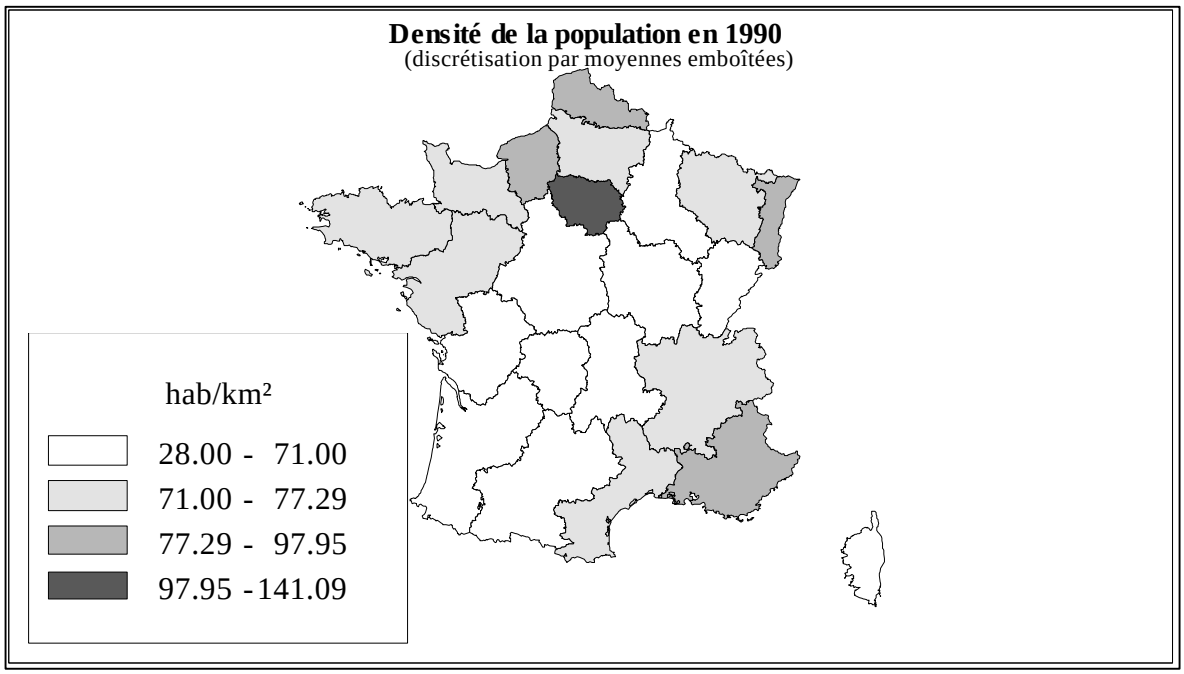
b)


Densité de la population en 1990
(discrétisation par équifréquence)
hab/km²
28.77 - 53.62
53.62 - 77.29
77.29 - 97.95
97.95 - 141.09
141.09 - 887.59

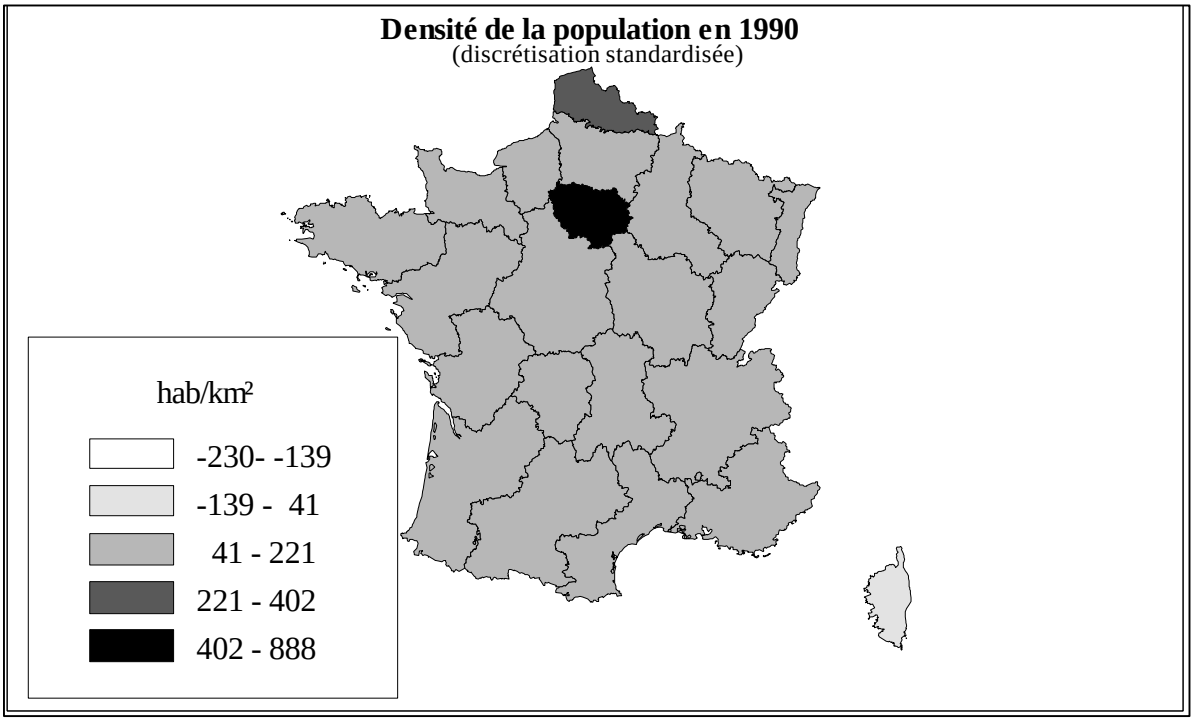
c)


Densité de la population en 1990
(discrétisation par progression arithmétique)
hab/km2
28 - 86
86 - 200
200 - 372
372 - 601
601 - 888

d)


Densité de la population en 1990
(discrétisation par moyennes emboîtées)
hab/km²
28.00 - 71.00
71.00 - 77.29
77.29 - 97.95
97.95 - 141.09

e)



f)

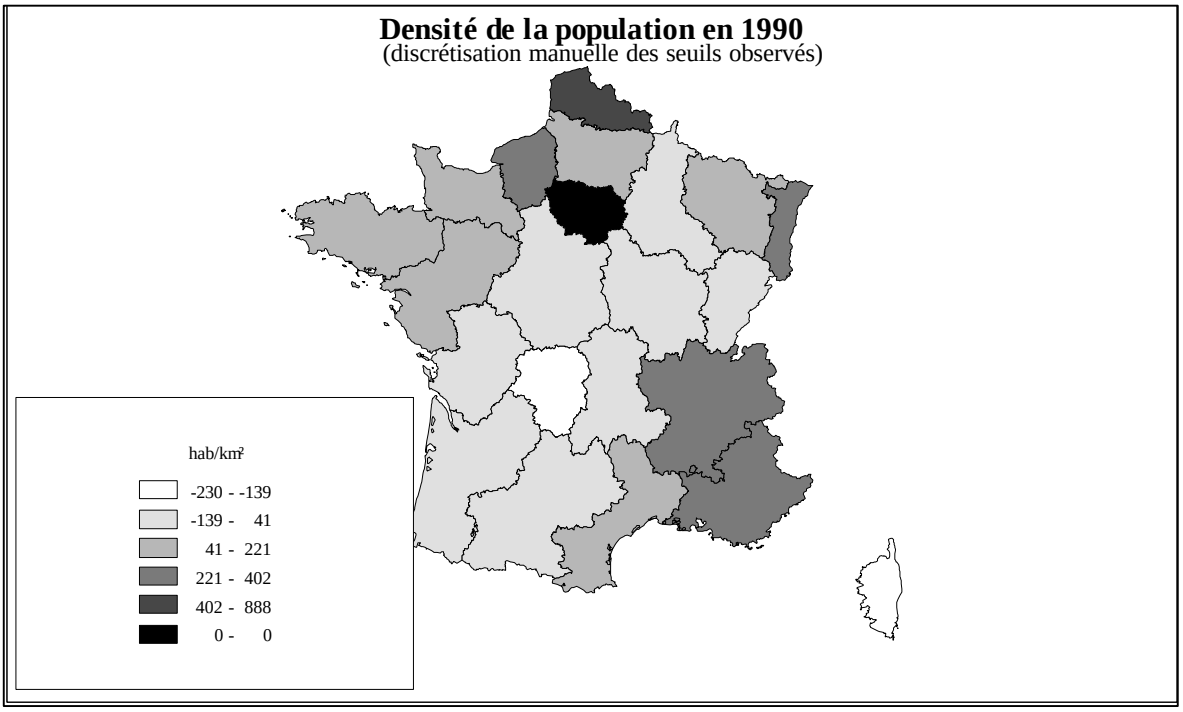


Fig. 86: Cartes de la densité de la population par région en 1990

A chaque méthode une image de la France. On est tenté de croire qu'il n'y a pas de cartes justes - il y a juste des cartes. Pourtant la fig. 86 nous enseigne que:

- la discrétisation par équidistance produit trop de classes vides, - l'image est uniforme
- la discrétisation par équifréquence nivelle trop les données – l'image n'est pas significative
- la discrétisation par moyennes emboîtées n'est pas adaptée pour cause de série asymétrique
- la discrétisation standardisée n'est pas adaptée non plus (la série ne suit pas une distribution normale des valeurs)
- que la discrétisation manuelle par seuils observés et par progression arithmétique sont certainement les deux méthodes les plus adaptées à la distribution des valeurs de la série si étendue et asymétrique.

Enfin, l'essentiel est d'être honnête, c'est-à-dire avoir une démarche transparente et indiquer la méthode employée.

C. Exercice: Discrétisation

1. A partir du tableau (fig. 87 et du fond de carte :fig. 88), réaliser, pour chaque méthode présentée, une discrétisation ainsi que la carte correspondante. Nombre de classes souhaitées: 5. Exprimer la densité de routes nationales en km/km² en 1990 par la valeur en implantation surfacique.

Densité des routes nationales en 1990	
Région	**km/km²**
ALSACE	481
AQUITAINE	1 314
AUVERGNE	987
BASSE-NORMANDIE	809
BOURGOGNE	1 237
BRETAGNE	1 135
CENTRE	1 566
CHAMPAGNE-ARDENNE	1 083
CORSE	569
FRANCHE-COMTE	828
HAUTE-NORMANDIE	758
ILE-DE-FRANCE	1 148
LANGUEDOC-ROUSSILLON	1 290
LIMOUSIN	657
LORRAINE	1 257
MIDI-PYRENEES	1 482
NORD-PAS-DE-CALAIS	848
PAYS DE LA LOIRE	1 668
PICARDIE	1 360
POITOU-CHARENTES	1 064
PROVENCE-ALPES-COTE D'AZUR	1 018
RHONE-ALPES	2 399

Indicateurs		
Nombre d'individus	n	22
Maximum	max	2 399
Minimum	min	481
Moyenne arithmétique	x	1 134
Ecart-type	σ	414

Nombre de classes souhaité k	5

Source : selon E. Blin & J.-P. Bord, Initiation géo-graphique, SEDES, 1993

Fig. 87 : Densité des routes nationales (km/km²) en 1990 par régions (France métropolitaine) et descripteurs clé

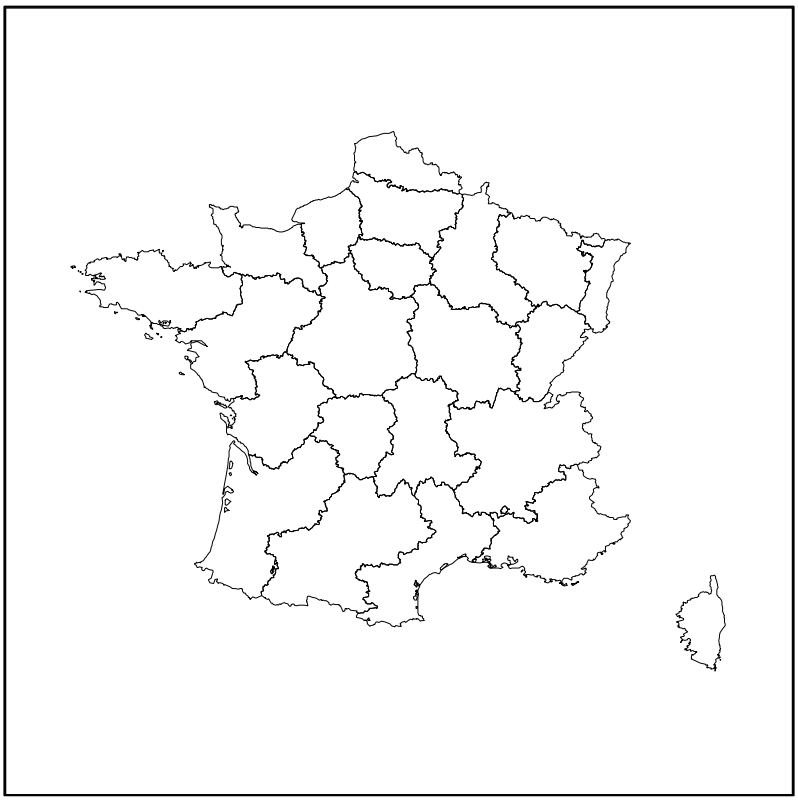


Fig. 88 : Découpage régional de la France