Géoinformation – Cours 7

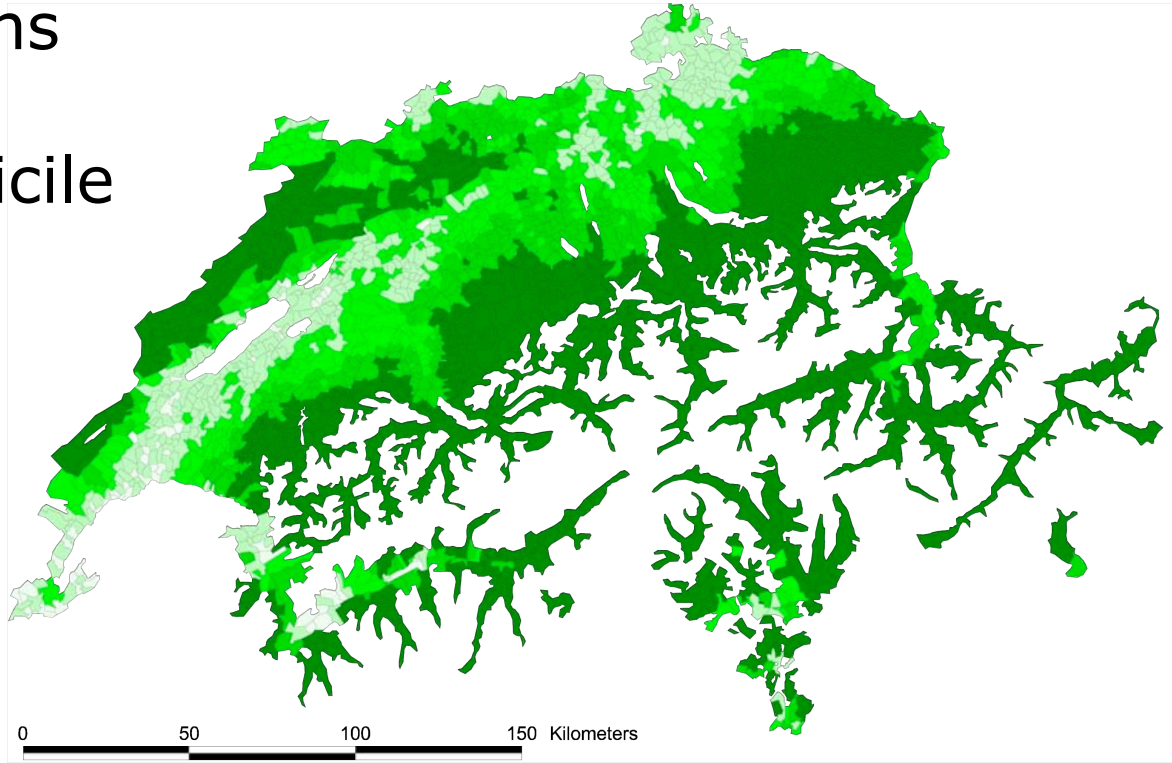
Reduction et discrétisation de l'information

Référence:

Béguin & Pumain, chapitre 5, 6

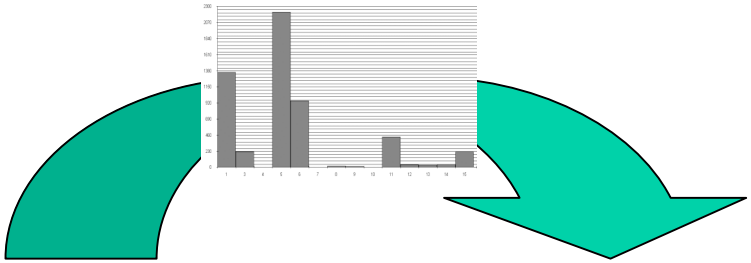
Pourquoi réduire l'information ?

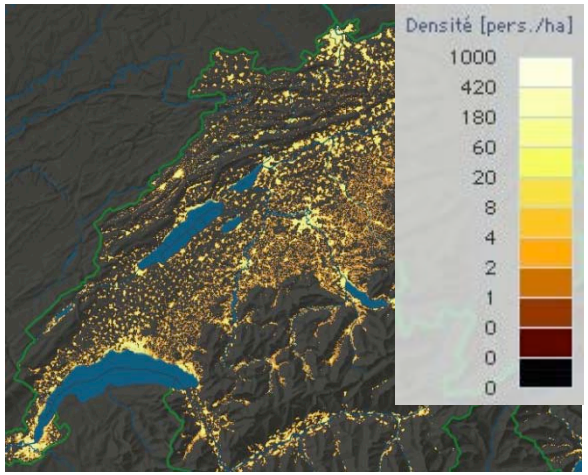
Représentation directe des valeurs d'une variables mesurées dans l'espace géographique souvent inefficace, difficile à interpréter



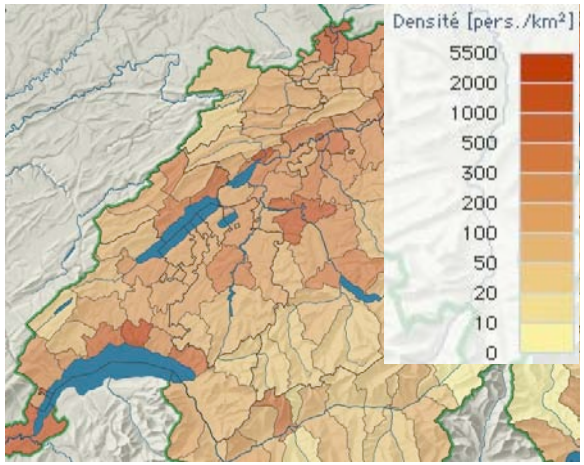
***Exemple**:*
représentation d'une variable (déjà discrétisée ici !) par commune suisse

Traitements pour la réduction de l'information





Densité de population 1990,
en personnes / ha



Densité de population 1990,
en personnes / km^2,
agrégé par région de mobilité spatiale

Source: atlas numérique de la Suisse, © Swisstopo

Réduction / discrétisation d'une série quantitative

Un objectif de la représentation cartographique est de présenter les informations avec autant de contenu (aussi peu de pertes de sens) que possible.

Les statistiques et la cartographie se heurtent cependant à des exigences contradictoires:

- d'un point de vue statistique, il serait judicieux de conserver l'entier du domaine de valeurs des variables
- d'un point de vue cartographique, la lisibilité est d'autant plus grande que le nombre de symboles (ici: de classes) est faible

On appelle *réduction* ou *discrétisation* d'une série quantitative l'opération qui consiste à ramener l'ensemble des valeurs ordonnées à un nombre limité de classes de valeurs, à l'intérieur desquelles les différences ne sont plus perceptibles.

Principes

- Maximiser l'information véhiculée
  → Selon la théorie de la communication, ce critère est satisfait si on transfert un nombre égal de chaque symbole – pour une carte choroplèthe, s'il y le même nombre d'entités géographiques dans chaque classe.
- Informer sur les caractéristiques de la distribution statistique:
  - ordre de grandeur
  - forme de la distribution
  - dispersion (variance inter-classes)
  - irrégularités & valeurs extrêmes

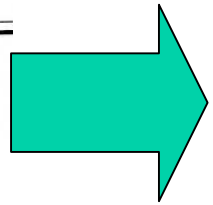
A nouveau, ces deux principes sont contradictoires !
Le cartographe / l'analyste devra donc choisir un processus de discrétisation conforme au message qu'il veut communiquer.

Base de l'analyse: le tableau de valeurs

- Un *tableau de valeurs* rattache à chaque unité géographique la valeur du phénomène considéré
- Les données nominales pourront être traitées une fois dénombrées dans un tableau de dénombrement

ensemble E des unités géographiques	caractère qualitatif J
nom de chaque unité	modalité j
.....	
.....	
.....	
.....	

Tableau élémentaire



caractère qualitatif J	effectif des classes
nom des modalités j	n_j
.....	
.....	
.....	
	$N = \sum_p n_j$

Tableau de dénombrement

code	communes	%pop CH
6601	Aire-la-Ville	80.37
6602	Anières	68.76
6603	Avully	82.99
6604	Avusy	86.73
6605	Bardonnex	79.37
6606	Bellevue	61.18
6607	Bernex	81.16
6608	Carouge	64.61
6609	Cartigny	77.43
6610	CÚligny	70.8
6611	Chancy	82.19
6612	Chêne-Bougerie	69.51
6613	Chêne-Bourg	65.22
6614	Choulex	81.06
6615	Collex-Bossy	71.07
6616	Collonge-Belleri	71.96
6617	Cologny	63.17
6618	Confignon	83.83
6619	Corsier	72.21
6620	Dardagny	75.55

Variables quantitatives

Les variables quantitatives peuvent être de différents types:

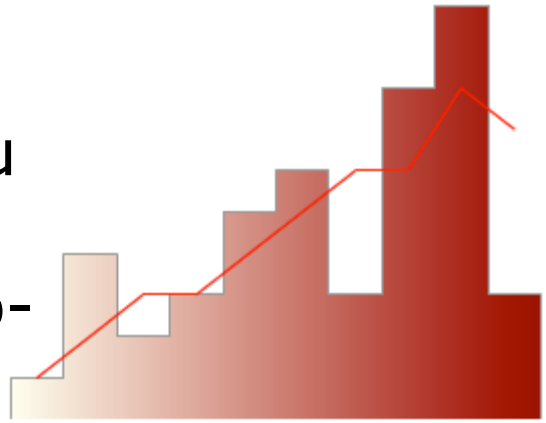
- décomptes et sommes
  (nombre d'habitants, précipitation annuelle en cm, …)
  agrégation par surface: sommation ← ***carte choroplèthe: NON***
- rapports (ratios) → relation entre 2 quantités
  agrégation par surface: moyenne ← ***carte choroplèthe: OK***
  - moyennes (averages) → rapport entre 2 variables différentes
    ex: nombre d'habitants par foyer
  - proportions → rapport entre la valeur mesurée et le total des valeurs
    ex: % de la population entre 20 et 30 ans
  - densités → rapport entre la valeur mesurée et la superficie correspondante
    ex: nombre d'habitants par hectare

On appelle *normalisation* le processus de dérivation d'une variable en rapport, afin qu'elle puisse être cartographiée

Analyse des valeurs

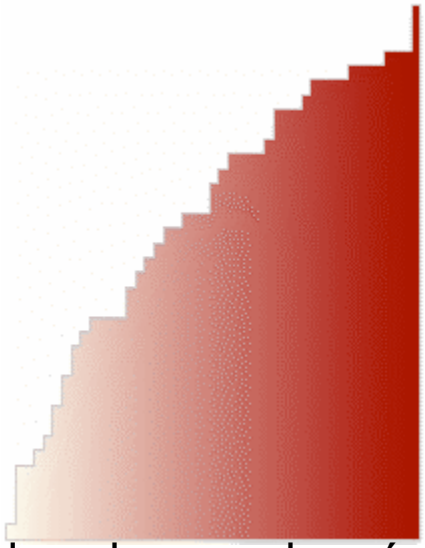
code	communes	%pop CH
6601	Aire-la-Ville	80.37
6602	Anières	68.76
6603	Avully	82.99
6604	Avusy	86.73
6605	Bardonnex	79.37
6606	Bellevue	61.18
6607	Bernex	81.16
6608	Carouge	64.61
6609	Cartigny	77.43
6610	CÚligny	70.8
6611	Chancy	82.19
6612	Chêne-Bougeri	69.51
6613	Chêne-Bourg	65.22
6614	Choulex	81.06
6615	Collex-Bossy	71.07
6616	Collonge-Beller	71.96
6617	Cologny	63.17
6618	Confignon	83.83
6619	Corsier	72.21
6620	Dardagny	75.55

Les données quantitatives ou les effectifs dénombrés sont présentés sous forme d'histo-gramme aux fins d'analyse



Histogramme de distribution
(nbre d'unités par plage de valeurs)

caractère qualitatif J		effectifs	fréquences	pourcentages
modalités	1 2 ... i j ... ... p	n_j	$f_j = \frac{n_j}{N}$	$p_j = \frac{n_j}{N} \times 100$
total		N	1	100



Histogramme de valeurs ordonnées
(classées par valeur croissante)

Méthodes de discrétisation de séries statistiques quantitatives

référence: Béguin&Pumain §5

- en classes d'égale étendue / amplitude

  le domaine de valeurs (entre valeur min. et valeur max.) est divisé en n classes d'égale étendue. Méthode simple bien adaptée à une répartition homogène des valeurs de la série considérée.

- selon la moyenne et l'écart-type

  variante de la précédente, pour lequel la moyenne de la série est prise comme valeur centrale et l'écart-type comme amplitude de classe. Méthode permettant de faire mettre en relation les caractéristiques statistiques et spatiales de la série considérée (pour série gaussienne).

- (selon une progression géométrique)

  adaptée à des distributions dissymétriques résultant de fonctions de type multiplicatif ($y = x^k$: distribution lognormale).

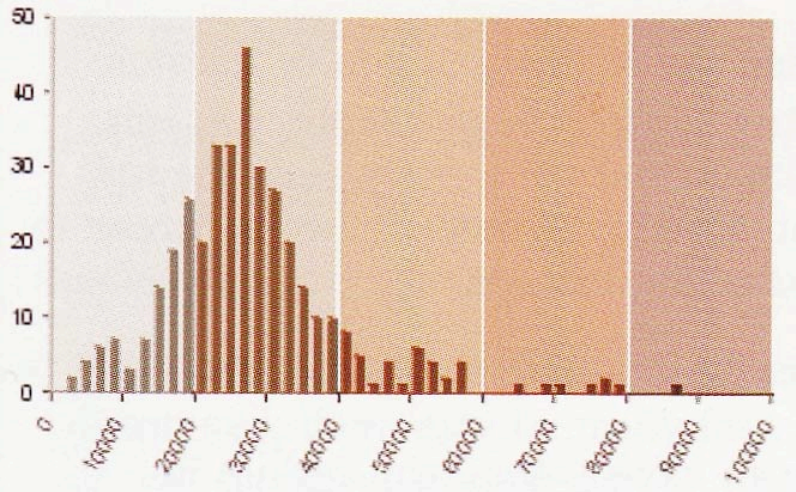
- selon les quantiles (effectifs égaux)

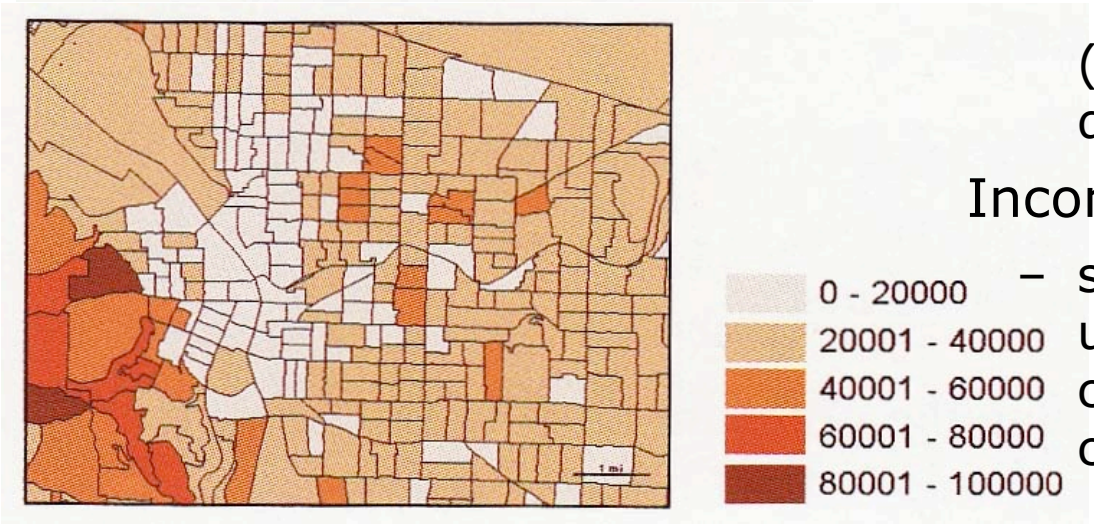
  contruite autour de la médiane, quartiles, …, autres quantiles, cette méthode permet de constituer des classes regroupant le même nombre de valeurs, et maximisant donc le contenu informatif de la discrétisation.

- méthode des seuils naturels

  pour séries irrégulières (plurimodales) sans loi reconnaissable. Les classes sont délimitées empiriquement pour regrouper des plages bien identifiables de valeurs de la série.

Classes d'égale étendue / amplitude





Avantages:

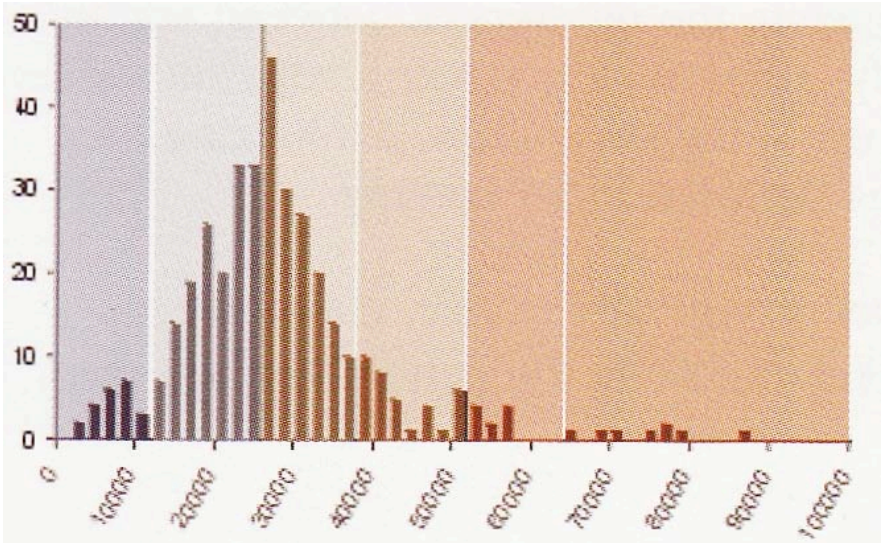
- facile à comprendre et à interpréter
- Particulièrement adapté à une distribution uniforme

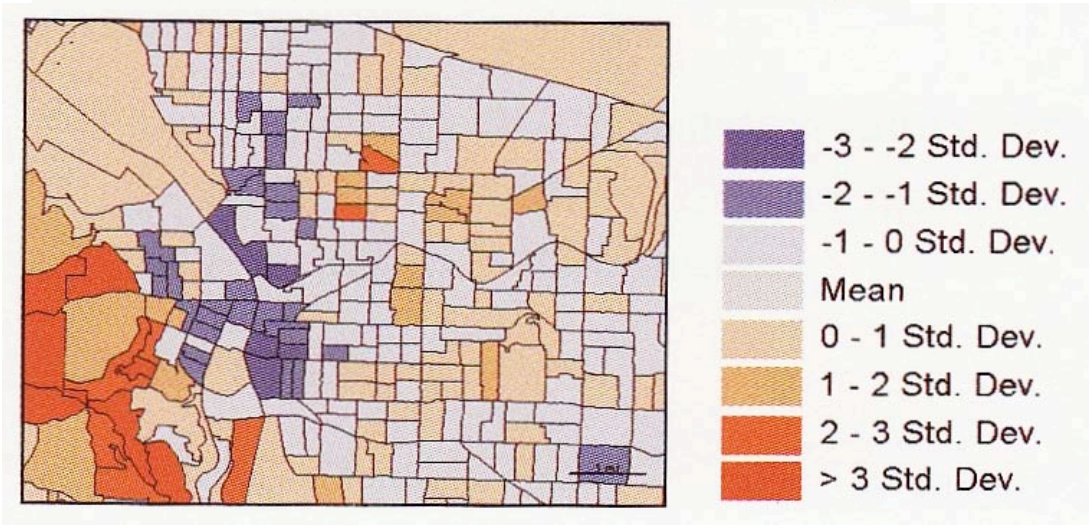
(ici: mise en évidence des valeurs extrêmes !)

Inconvénients

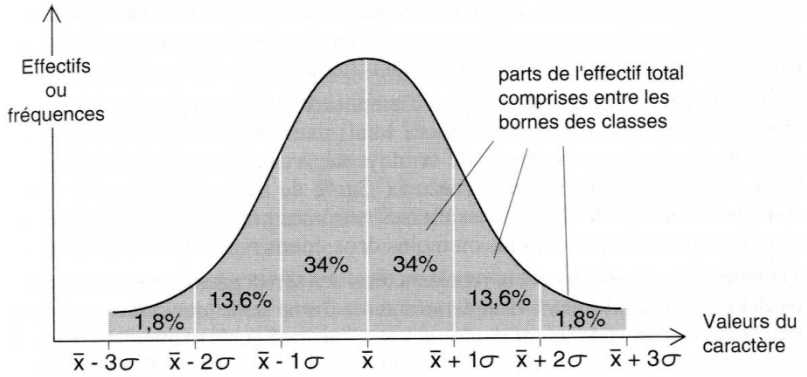
- si la distribution n'est pas uniforme, les valeurs sont concentrées dans quelques classes seulement

Moyenne et écart-type



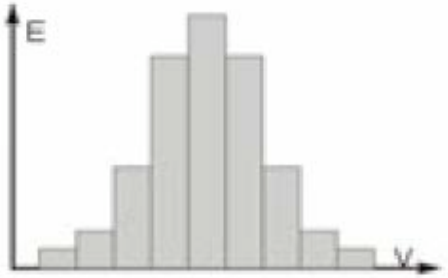


Chaque classe est définie par sa distance à la moyenne, en tenant compte de la variance



Atouts:

- bonne mise en évidence d'une distrib. normale

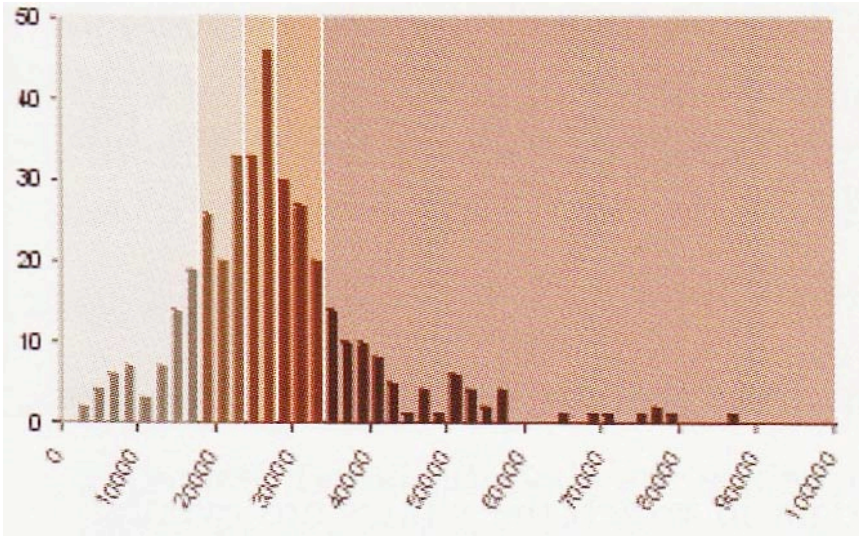


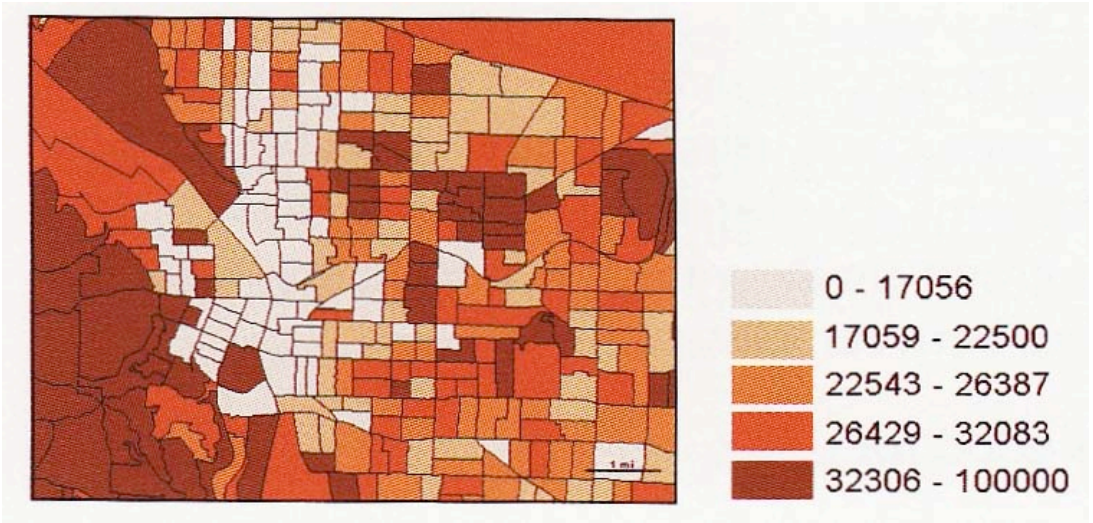
- calage par rapport à une valeur moyenne, « neutre »

Limites:

- mal adapté à des distributions non-normales

Selon les quantiles (effectifs égaux)





Chaque classe contient un même nombre de valeurs

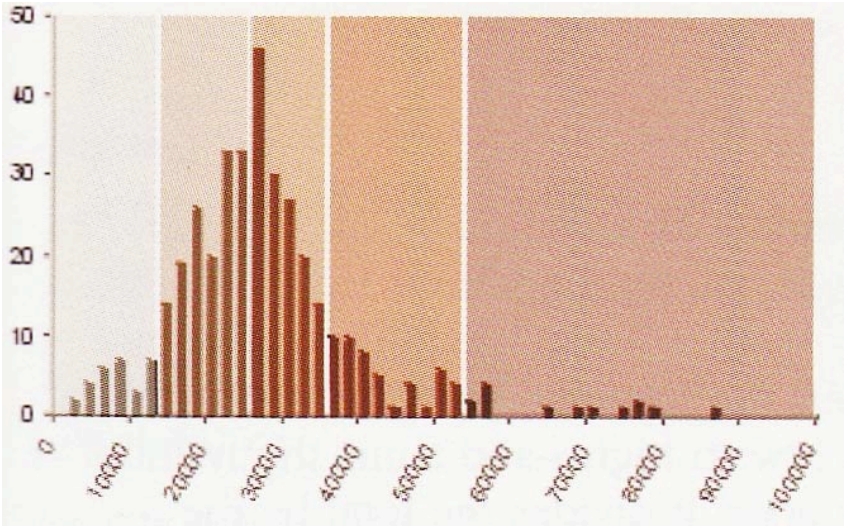
Avantages:

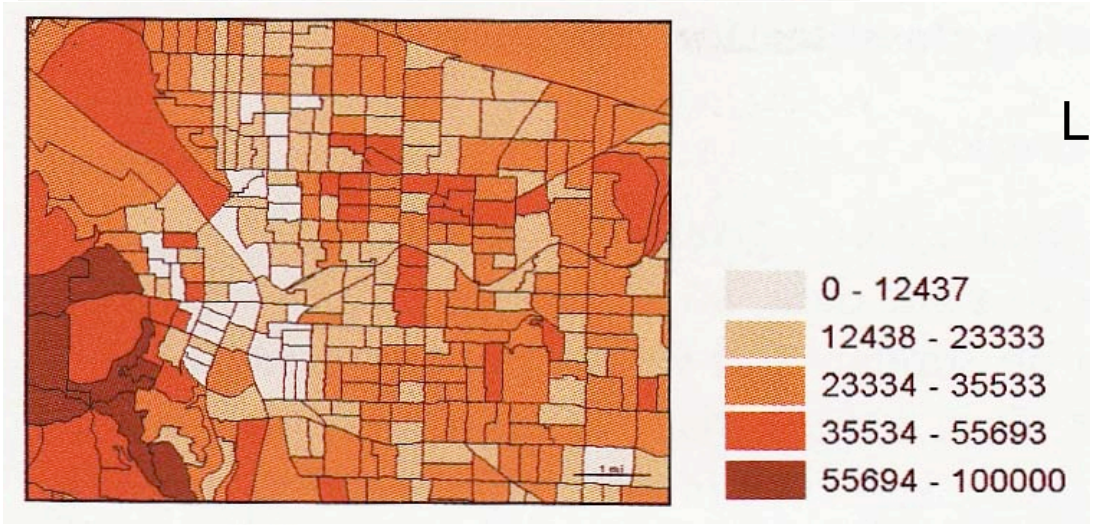
- toutes les classes ont un nombre égal de valeurs
- porteur d'un maximum d'information (entropie max)
- adapté à une distribution uniforme

Limites:

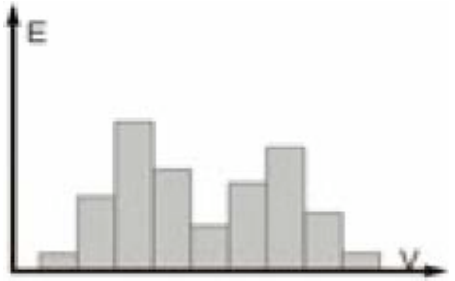
- ne reflète pas les originalités d'une distribution

Natural breaks





Classification basée sur un regroupement naturel, empirique des valeurs



Avantages:

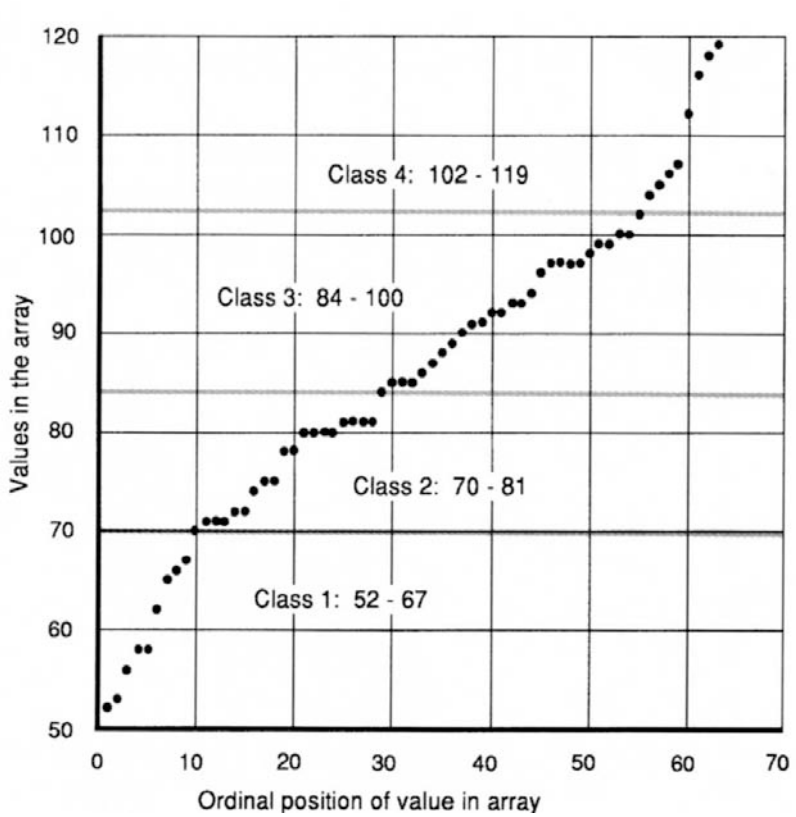
- distributions polycentriques bien représentées

Limites:

- difficile à mettre en œuvre
- difficile à interpréter
- pas de comparaison possible entre cartes

Method of optimized breaks (Jenk's)

52	75	86	97
53	75	86	98
56	78	88	99
58	78	89	99
58	80	90	100
62	80	91	100
65	80	91	102
66	80	92	104
67	81	92	105
70	81	93	106
71	81	93	107
71	81	94	112
71	84	96	113
72	85	97	116
72	85	97	118
74	85	97	119



Jenk's algorithm:

- maximize inter-class variance
- minimize intra-class variance

1. choose number of classes
2. heuristic, iterative process to maximize the *Goodness of Variance Fit* (GVF) index

Jenk's optimized breaks algorithm

SDAM = Squared Deviations, Array Mean
SDCM = Squared Deviations, Class Means
GVF = Goodness of Variance Index

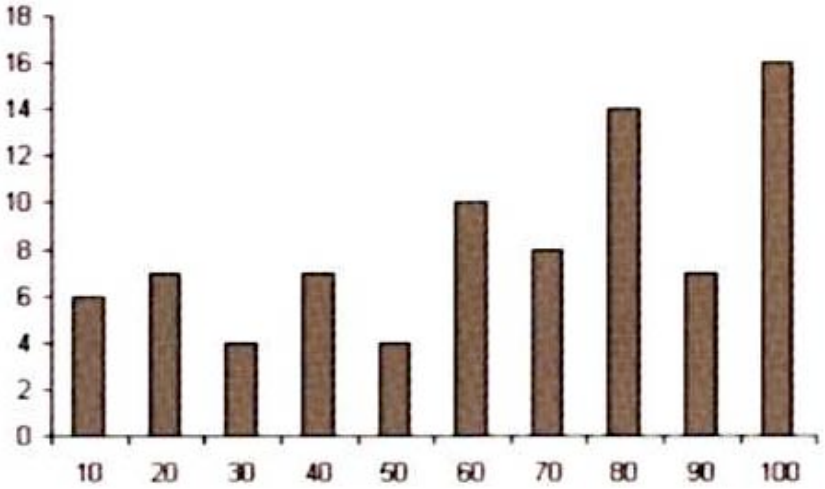
Array Values	Deviations from Array Mean	Squared Deviations from Array Mean	Solution 1 (Special Case with a Class for Each Value)			Solution 2			Solution 3		
	$(x-\bar{X})$	$(x-\bar{X})^2$		$(x-\bar{X}c)^2$	$(x-\bar{Z}c)^2$		$(x-\bar{Z}c)$	$(x-\bar{Z}c)^2$		$(x-\bar{Z}c)$	$(x-\bar{Z}c)^2$
2	–4.54	20.61	$Z_1 = 2$	0	0	$Z_1 = 3.0$	–1.0	1.0	$Z_1 = 2.5$	.5	.25
3	–3.54	12.53	$Z_2 = 3$	0	0		0	0		5	.25
4	–2.54	6.45	$Z_3 = 4$	0	0		1.0	1.0		–1.0	1.0
5	–1.54	2.37	$Z_4 = 5$	0	0	$Z_2 = 5.5$	.5	.25	$Z_2 = 5.0$	0.0	0
6	–.54	.29	$Z_5 = 6$	0	0		.5	.25		1.0	1.0
7	.46	.21	$Z_6 = 7$	0	0	$Z_3 = 7.0$	0	0	$Z_3 = 7.33$	–.33	.11
7	.46	.21	$Z_7 = 7$	0	0		0	0		.33	.11
8	1.46	2.13	$Z_8 = 8$	0	0		1.5	2.25		.67	.45
9	2.46	6.05	$Z_9 = 9$	0	0	$Z_4 = 9.5$	–.5	.25	$Z_4 = 10.0$	–1.0	1.0
10	3.46	11.97	$Z_{10} = 10$	0	0		.5	.25		0.0	0
11	4.46	19.89	$Z_{11} = 11$	0	0		1.5	2.25		1.0	1.0

$\bar{X} = 6.54$ $\quad$ SDCM $= \Sigma\Sigma(x-\bar{Z}_c)^2 = 0$ $\quad$ SDCM $= \Sigma\Sigma(x-\bar{Z}_c)^2 = 7.5$ $\quad$ SDCM $= \Sigma\Sigma(x-\bar{Z}_c)^2 = 5.17$

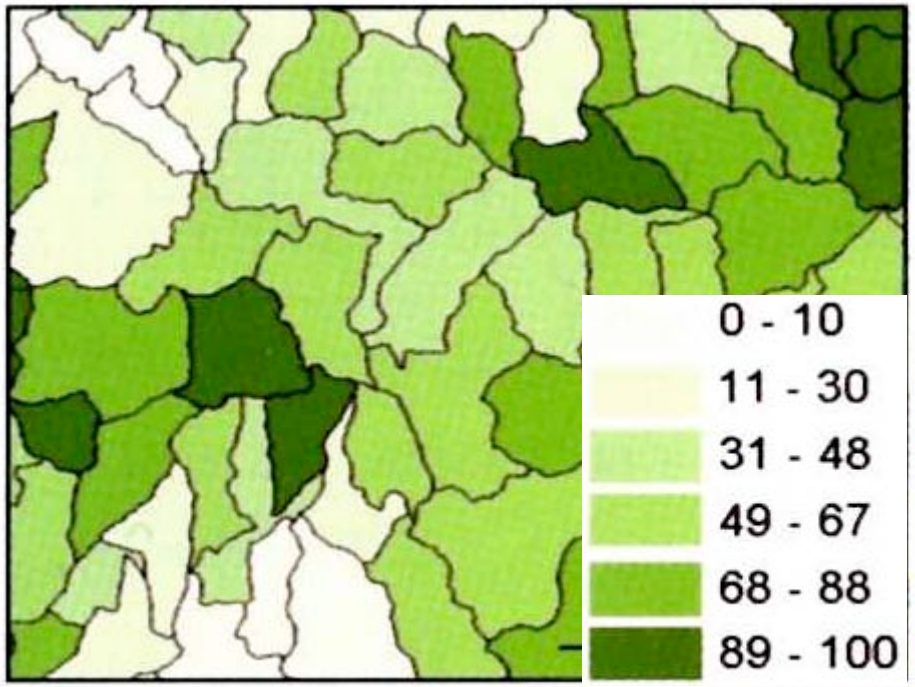
SDAM $= \Sigma(x-\bar{X})^2 = 82.72$ $\quad$ GVF $= \dfrac{\text{SDAM} - \text{SDCM}}{\text{SDAM}} = 1.0$ $\quad$ GVF $= \dfrac{\text{SDAM} - \text{SDCMS}}{\text{SDAM}} = .909$ $\quad$ GVF $= \dfrac{\text{SDAM} - \text{SDCM}}{\text{SDAM}} = .937$

Mise en œuvre : « best practices » (1)

Si les données sont distribuées de manière irrégulière, privilégiez la méthode des seuils naturels (natural breaks) ou de Jenk's

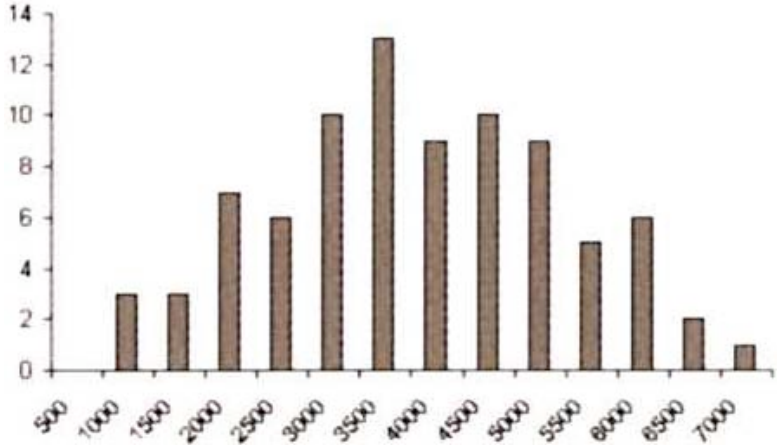


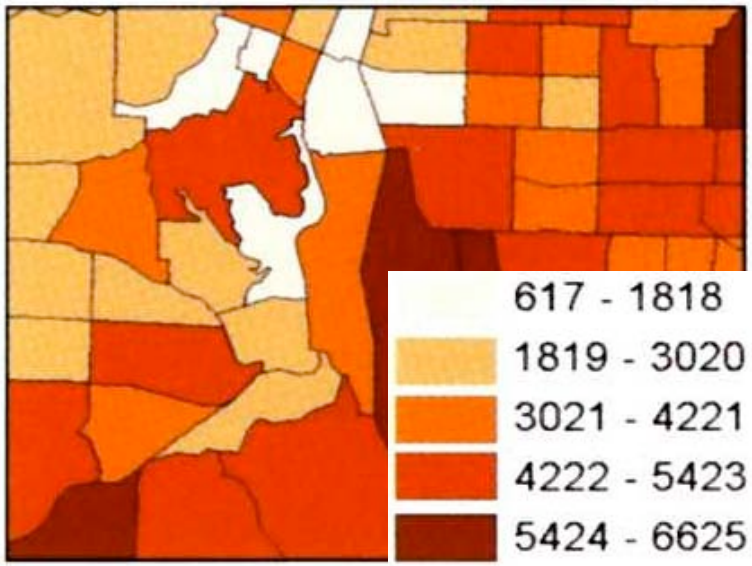
Pourcentage de forêt sur des bassins versants (doc ESRI)



Mise en œuvre : « best practices » (2)

Pour des données réparties régulièrement, privilégiez les quantiles (carte de droite) ou les intervalles égaux (carte de gauche) pour mettre en évidence les valeurs extrêmes

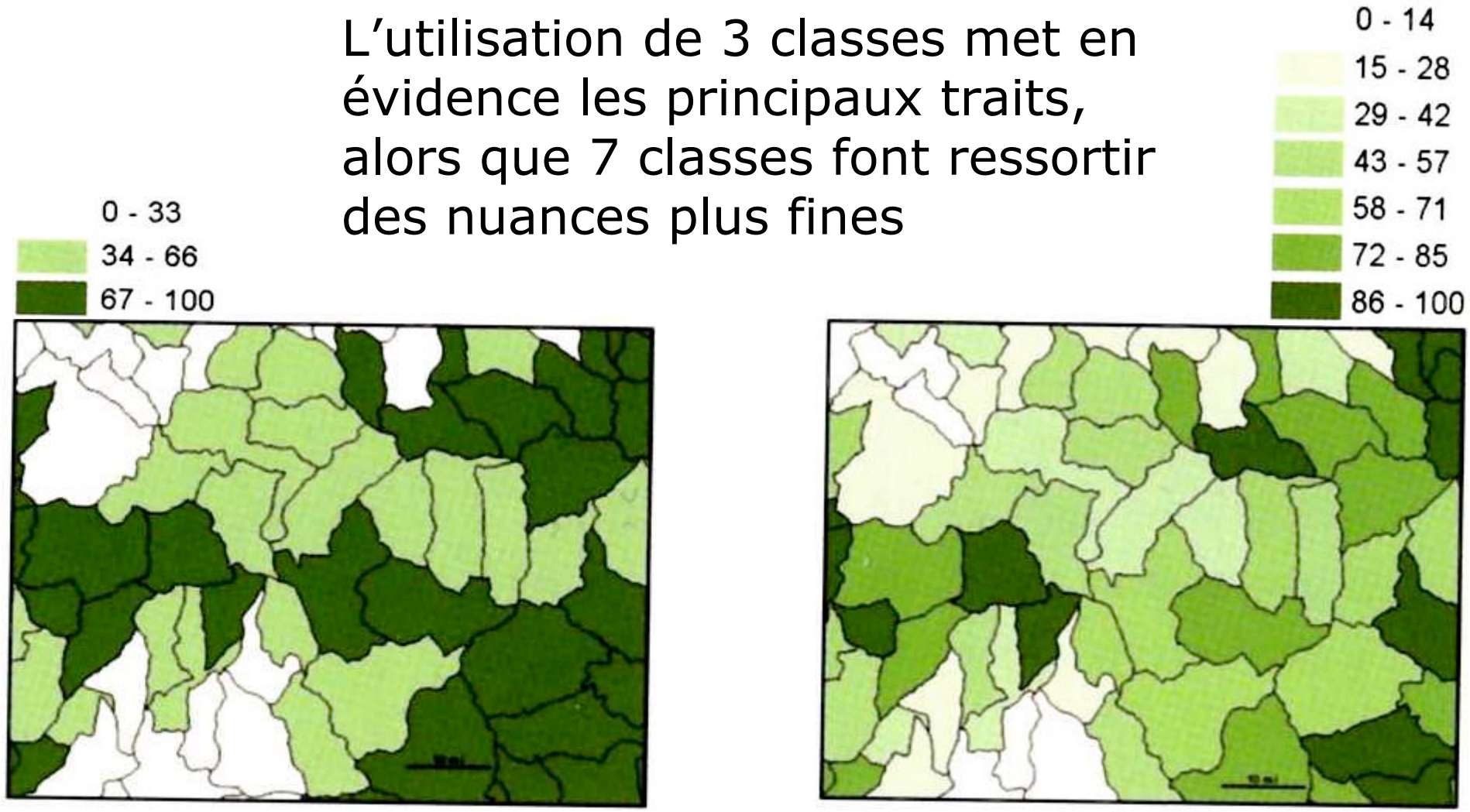




*Population par secteur de recensement (doc ESRI) – **Attention: valeurs non normalisées !***

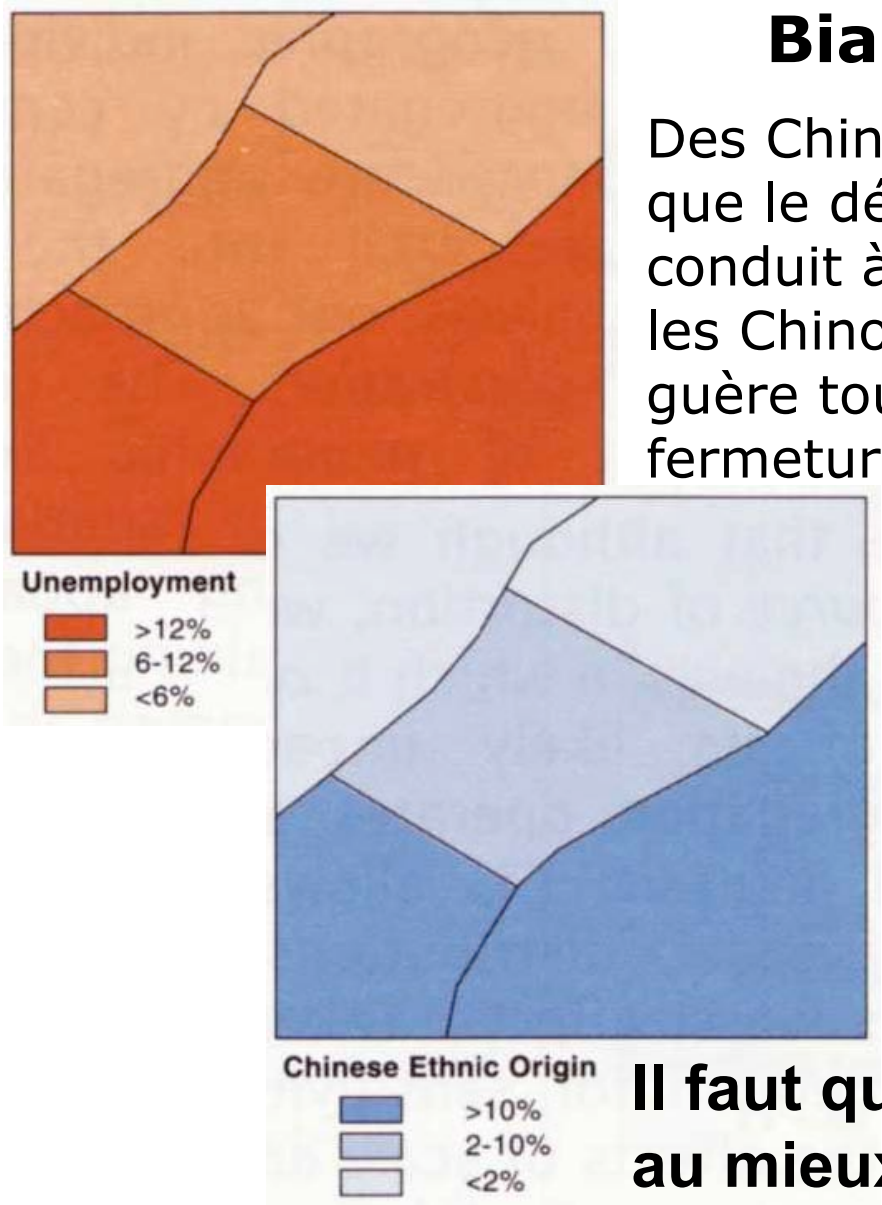
Mise en œuvre : « best practices » (3)

L'utilisation de 3 classes met en évidence les principaux traits, alors que 7 classes font ressortir des nuances plus fines

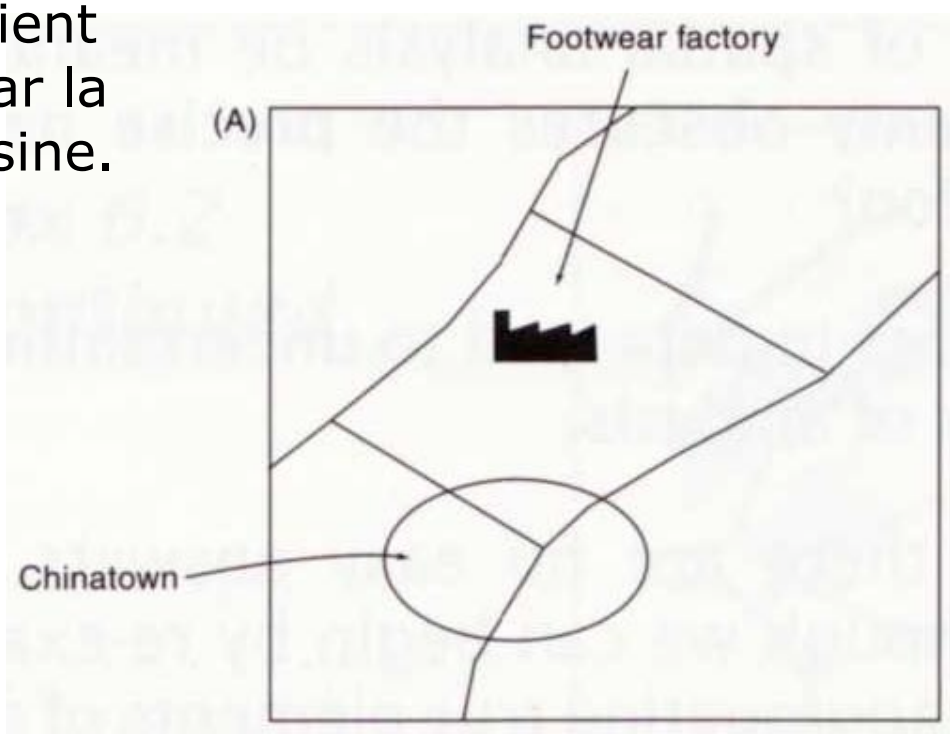


Pourcentage de forêt sur des bassins versants (doc ESRI)

Biais d'échelle et d'agrégation



Des Chinois au chômage ? La réalité montre que le découpage inadéquat des zones conduit à des conclusions erronées: les Chinois n'étaient guère touchés par la fermeture de l'usine.

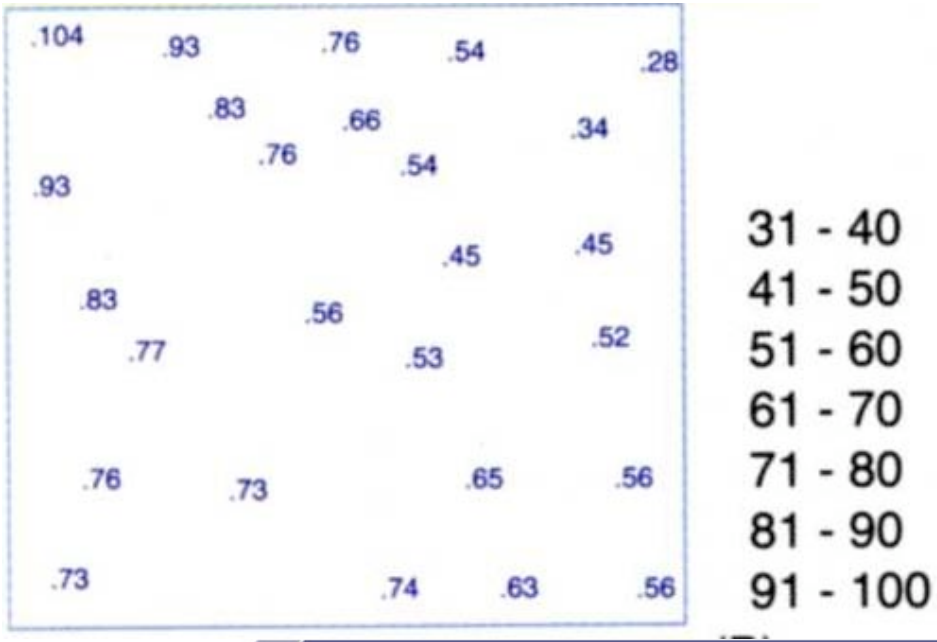


Il faut que les limites de zones correspondent au mieux aux limites du phénomène !

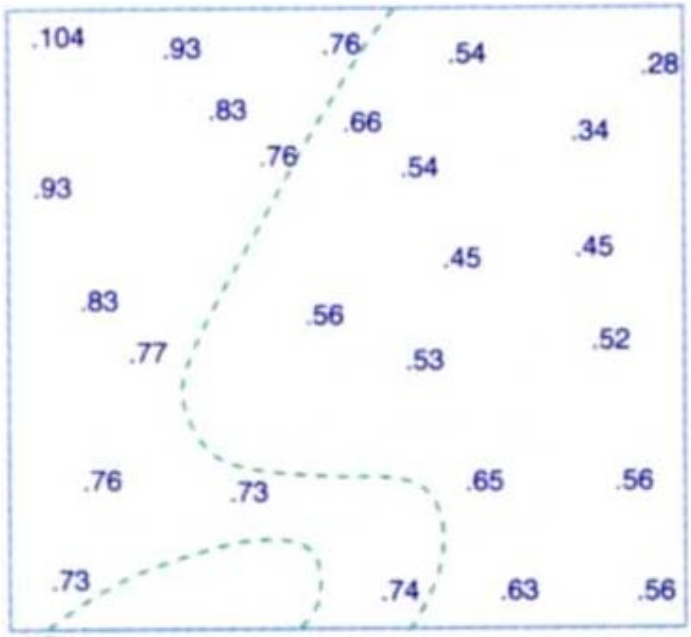
Elaboration d'une carte d'isolignes / isoplèthe (1)

Dans la mesure du possible, en ajustant au mieux le découpage des zones au phénomène à illustrer, on peut éviter ces biais d'échelle et d'agrégation.

Exemple de méthode de découpage du territoire par isoligne:



31 - 40
41 - 50
51 - 60
61 - 70
71 - 80
81 - 90
91 - 100



Elaboration d'une carte d'isolignes / isoplèthe (2)

Les isolignes découpent des espaces (isoplèthes) selon les ruptures naturelles du phénomène. Les isolignes mettent plus particulièrement en évidence les gradients.