

Les données attributaires

Discrétisation cartographique (mise en classes)
Tableaux de données
L'information statistique

Christian Kaiser
Cartographie & SIG

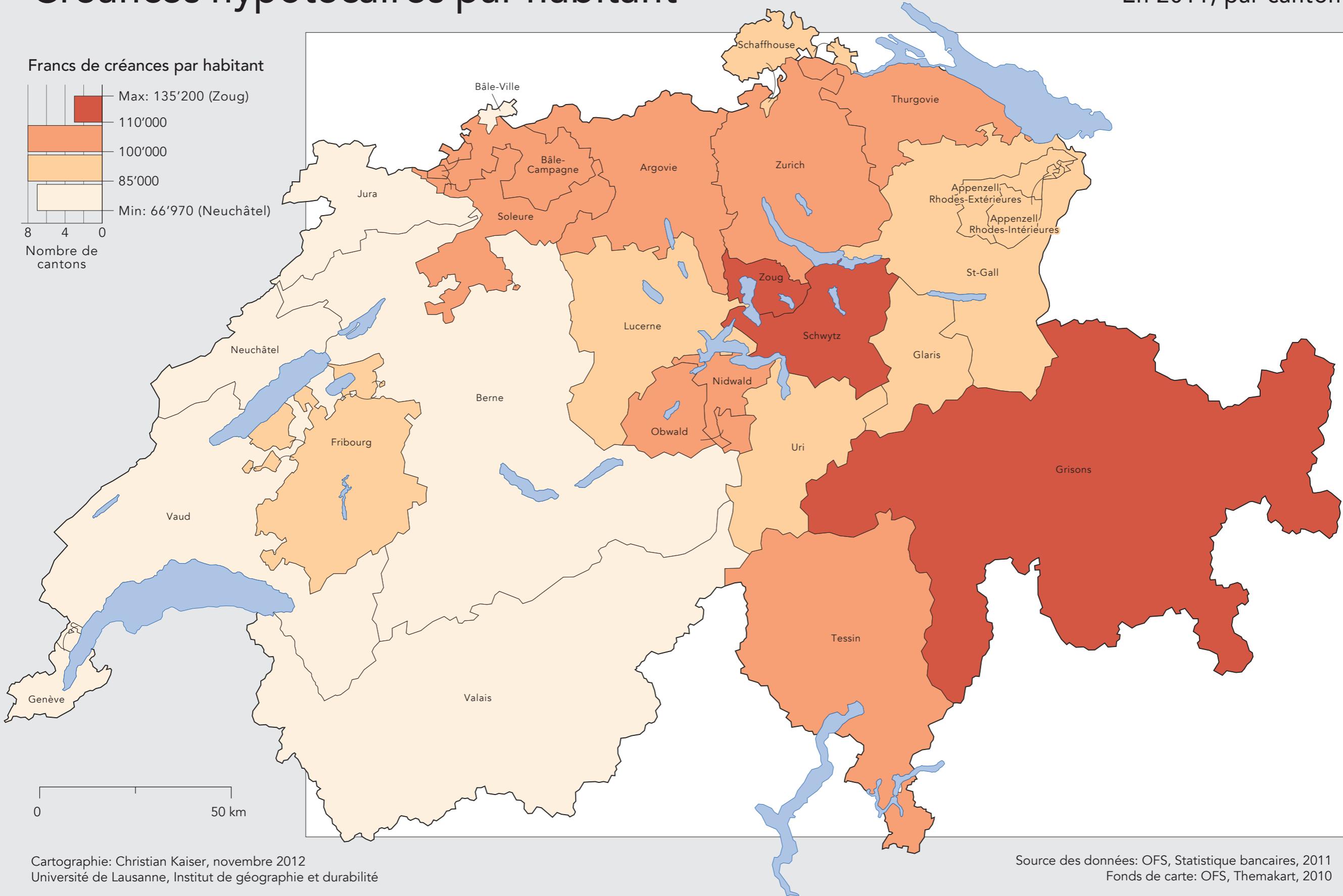
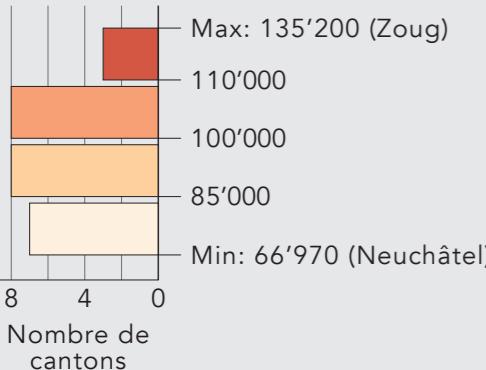
Discrétisation...

- .. Discrétisation (ou «mise en classes» ou «classification numérique» ou «généralisation des données»):
 - .. «*La discrétisation consiste à transformer la distribution d'une variable continue (c-à-d qui peut prendre n'importe quelle valeur), ou considérée comme telle, en une distribution discrète constituée de groupes ou classes voisines les unes des autres.*» (Raveneau 1997)
 - .. «*La discrétisation doit permettre de conserver au mieux l'information géographique contenue dans la série statistique, tout en permettant sa transmission par la carte, avec la meilleure lisibilité possible.*» (Béguin & Pumain 2003).

Créances hypothécaires par habitant

En 2011, par canton

Francs de créances par habitant



Discrétisation...

- .. Problème de mise en classe: comment trouver les valeurs limites
- .. But: entités dans la même classe sont similaires, et classes sont différentes
- .. Combien de classes?

Geocode	Canton	HypoPop2011
24	Neuchâtel	66.97
12	Basel-Stadt	80.13
25	Genève	81.45
26	Jura	82.45
22	Vaud	83.5
23	Valais	84.19
2	Bern	84.27
10	Fribourg	85.49
4	Uri	86.75
8	Glarus	87.37
15	Appenzell Ausserrhoden	88.99
16	Appenzell Innerrhoden	89.11
17	St. Gallen	91.68
3	Luzern	93.08
14	Schaffhausen	96.17
11	Solothurn	102.75
6	Obwalden	103.19
7	Nidwalden	103.65
20	Thurgau	106.14
21	Tessin	106.23
13	Basel-Landschaft	106.67
1	Zürich	106.86
19	Aargau	108.61
5	Schwyz	121.1
18	Graubünden	129.04
9	Zug	135.2

Discrétisation...

- ou **mise en classe** en cartographie
- .. En statistique: **classification** ou **clustering**
 - .. Potentiellement en considérant plusieurs variables
- .. Dans tous les cas:
tentative de former des **groupes cohérentes**

Procédure de discrétisation...

1. Regarder les données!

- .. Seuil externe?

P.ex. 0% de croissance,
50% de oui dans une votation etc.

2. Combien de classes?

3. Où mettre les limites des classes?

Choix de la **méthode de discrétisation**

4. **Evaluer l'erreur** introduit par la classification

Combien de classes?

- .. Contraintes déterminant le nombre de classes:
 - .. Caractéristiques des données
 - .. Objectif de la carte (type de public)
 - .. Seuil perceptif:
 - .. 7 à 8 classes max (en général)
 - .. 5 à 6 classes souvent idéal
 - .. Effectif (nombre d'entités géographiques)

Geocode	Canton	HypoPop2011
24	Neuchâtel	66.97
12	Basel-Stadt	80.13
25	Genève	81.45
26	Jura	82.45
22	Vaud	83.5
23	Valais	84.19
2	Bern	84.27
10	Fribourg	85.49
4	Uri	86.75
8	Glarus	87.37
15	Appenzell Ausserrhoden	88.99
16	Appenzell Innerrhoden	89.11
17	St. Gallen	91.68
3	Luzern	93.08
14	Schaffhausen	96.17
11	Solothurn	102.75
6	Obwalden	103.19
7	Nidwalden	103.65
20	Thurgau	106.14
21	Tessin	106.23
13	Basel-Landschaft	106.67
1	Zürich	106.86
19	Aargau	108.61
5	Schwyz	121.1
18	Graubünden	129.04
9	Zug	135.2

2. Combien de classes?

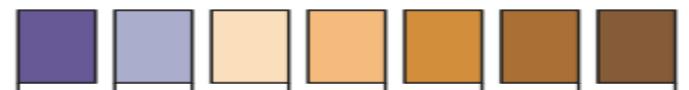
- .. En général pas de règle très précise
- .. Formules permettant de donner une approximation sur le **nombre de classes au niveau perceptif** (et non sur la base des données!):
 1. **Huntsberger** donne une approximation du nombre de classes K:
$$K_H = 1 + 3.35 \cdot \log_{10}N$$
 2. **Yule** donne l'estimation suivante:

2. Combien de classes?

N	K_h	K_y
10	4.4	4.3
20	5.4	4.6
30	5.9	4.8
40	6.4	5
50	6.7	5.2
60	7	5.3
70	7.2	5.4
80	7.4	5.5
90	7.5	5.6
100	7.7	5.7

2. Combien de classes?

- .. Mieux vaut être conservateur quant au nombre de classes > meilleure perception
- .. Nombre maximum de 7-8 classes peut être excédé si:
 - .. beaucoup d'entités et
 - .. légende divergente (bi-colore)
- .. En général: **privilégier 4-6 classes**



3. Quelle méthode de discréétisation?

1. Mises en classes cartographiques

- a. Seuils naturels (seuils observés)
- b. Jenks
- c. Amplitudes égales
- d. Effectifs égaux
- e. Discréétisation standardisée
- f. Progression géométrique

2. Classification statistique

- a. k-means
- b. Classification ascendante hiérarchique (CAH)
- c. ...

Quelle méthode de discréétisation?

Buts de la discréétisation:

- .. Limiter le nombre de couleurs
- .. **Regrouper les entités similaires et distinguer entités différentes**



<http://blog.musikexpress.de/wp-content/uploads/2011/10/Ursus-Wehrli.jpg>

Pour en voir plus, p.ex.:

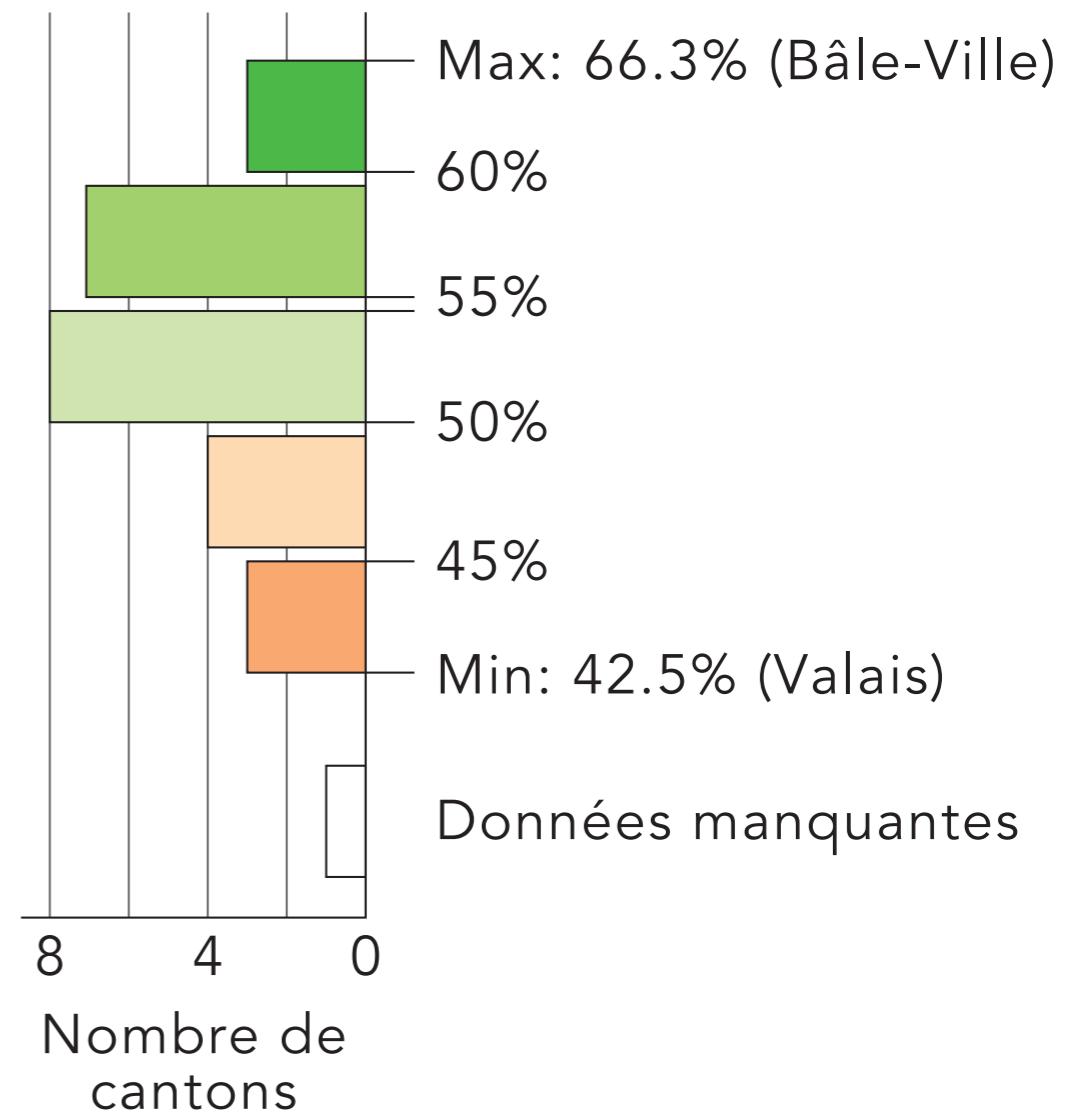
<http://www.kunstaufraeumen.ch/de/videos/ursus-wehrli-live-auf-der-ted-conference-en>

Norme externe

- .. Norme externe, p.ex.
 - .. taux de croissance (0%)
 - .. votations (50%)
- .. La norme externe doit toujours être une limite de classe
 - .. On symbolisera les entités en dessus ou en dessous de la norme différemment, p.ex. en utilisant des dégradés de couleurs différentes

Taux d'acceptation de l'initiative des Schtroumpfs

Moyenne suisse: 54.3%

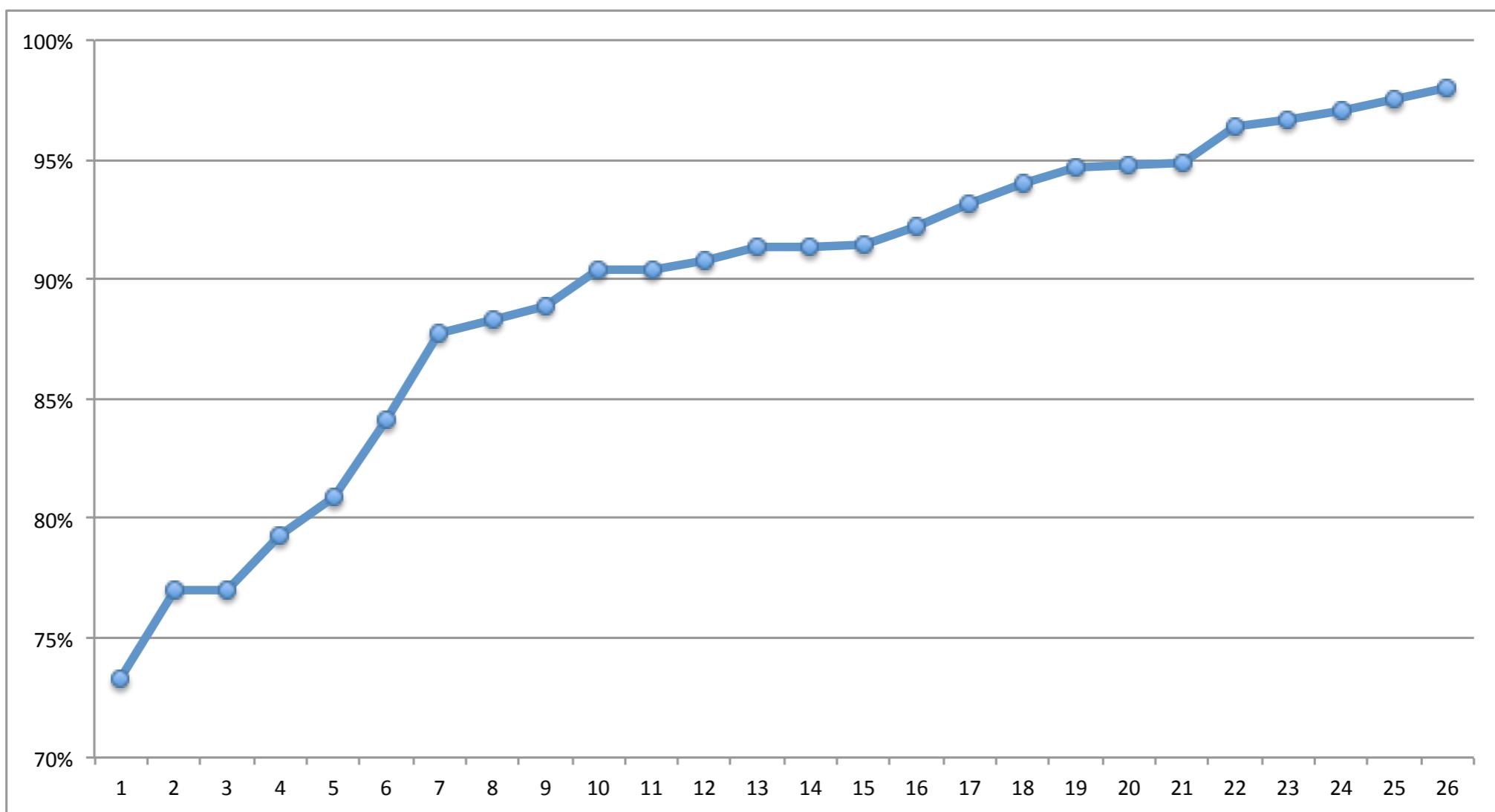


Seuils naturels (seuils observés)

- .. **Idiographique**: identifier les discontinuités
- .. Les **écart**s (=seuils) **les plus grands** définissent les classes
- .. Découpage à l'estime du relief de la série des valeurs
- .. Seule méthode cartographique (avec Jenks) qui respecte le regroupement d'entités semblables
- .. Adaptée à toutes les situations, sauf si besoin de comparer plusieurs cartes

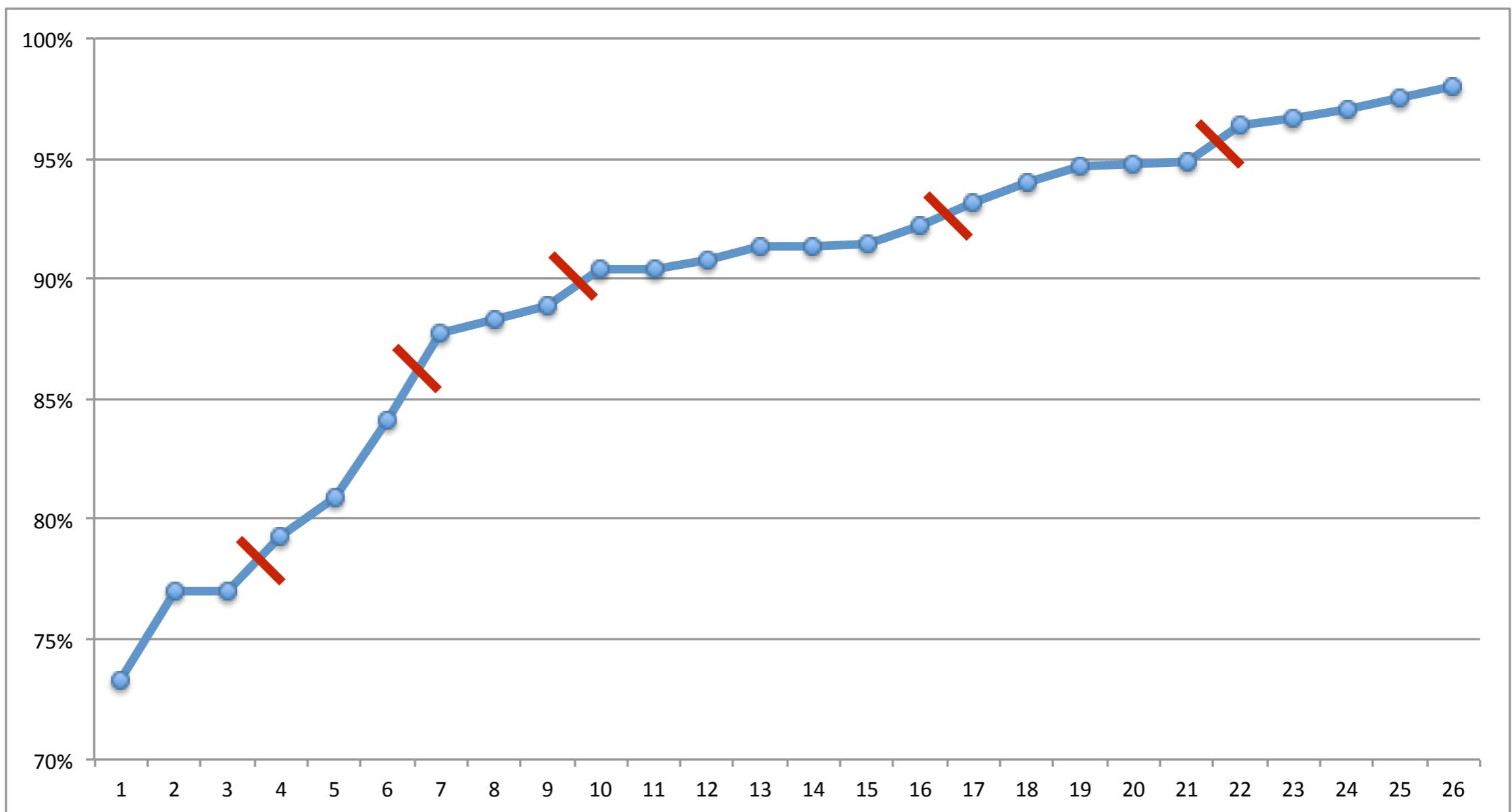
Procédure (manuelle)

- .. Copier les valeurs dans Excel, faire un tri croissant, et faire un graphique des valeurs



Procédure (manuelle)

- Identifier les seuils, regrouper valeurs semblables

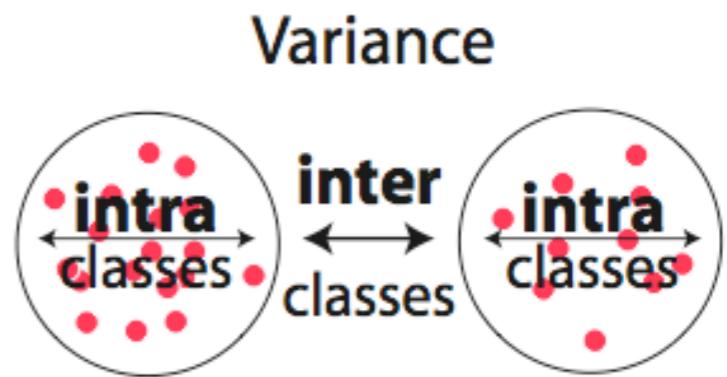


Procédure (manuelle)

- .. Alternative:
 - .. Calculer les différences entre les entités
(dans tableau Excel trié)
 - .. Sélectionner les plus grandes différences
→ limites de classes
 - .. Equilibrer un peu la taille des classes
 - .. Eviter de faire p.ex. 4 classes avec 1 entité et tout le reste dans l'autre entité

Classes optimales de Jenks

- .. Objectif:
 - .. Minimiser la variance intra-classes
 - .. Maximiser la variance inter-classes
- .. Correspond en gros à la méthode des seuils naturels
- .. Algorithme utilisé dans la plupart des logiciels de cartographie et SIG (p.ex. QGIS, ArcGIS)
- .. Méthode passe-partout, mais non adaptée si besoin de comparer plusieurs cartes



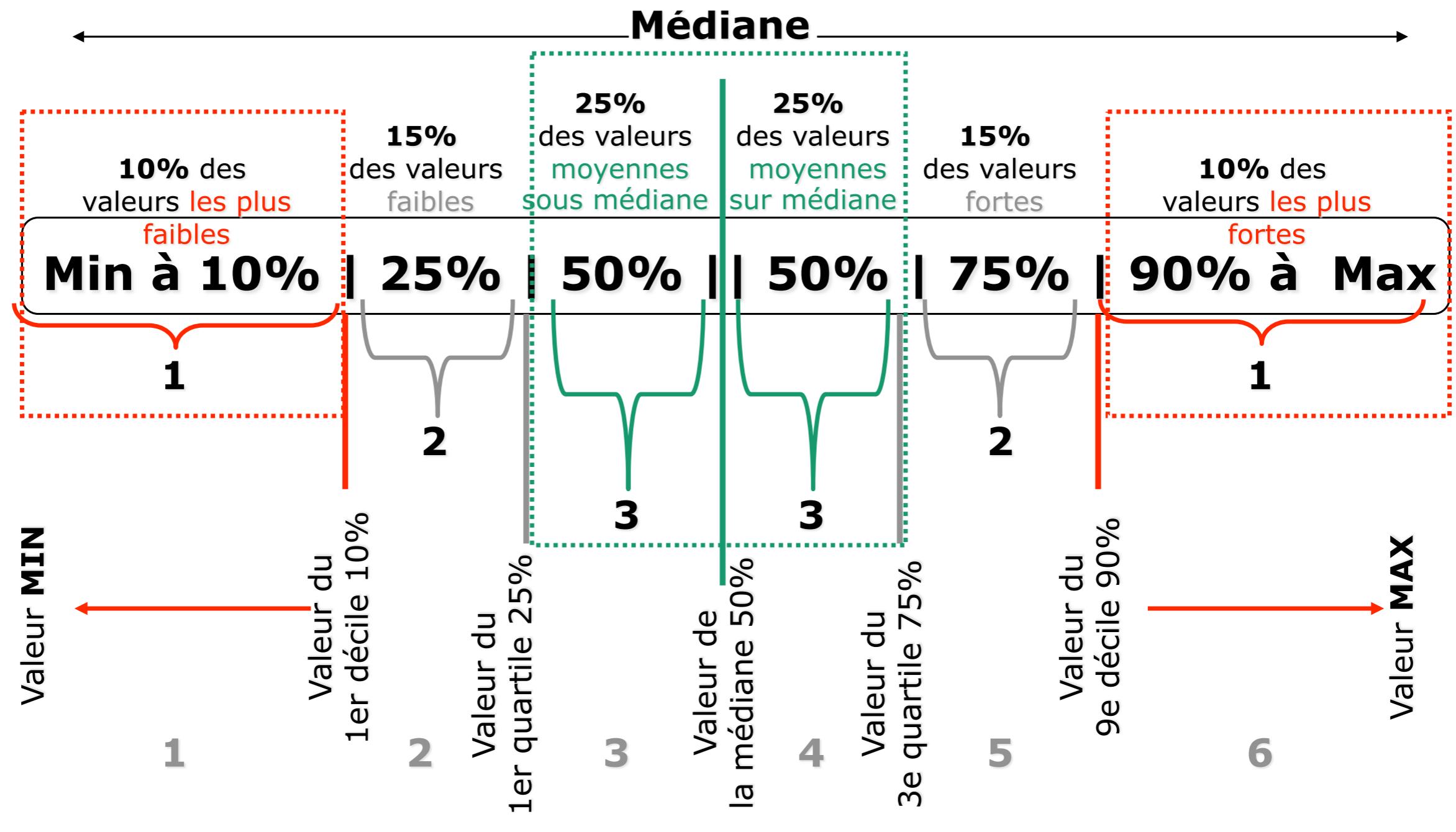
Amplitudes égales

- .. = equal range, equal size, des pas constant, ou même intervalle de valeurs
- .. **Attention:**
 - .. Entités avec valeurs proches peuvent être séparées
 - .. Ne permet pas d'atteindre le but de grouper des entités similaires et séparer entités différentes
- .. Typiquement utilisé pour cartes en courbes d'égales valeurs (isolines, isarithmes) avec accent sur les gradients
 - .. P.ex. courbes de niveau, cartes de pression atmosphérique

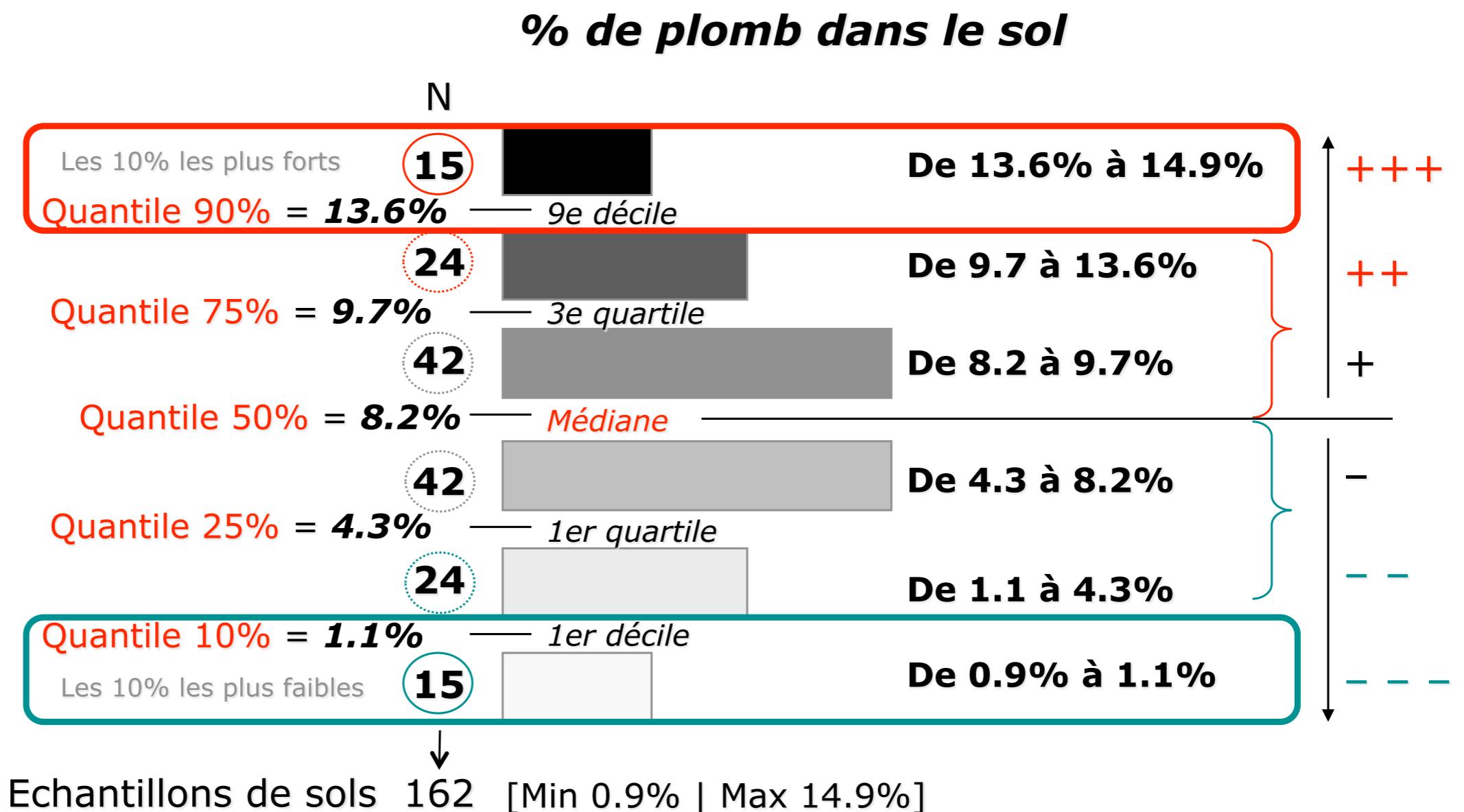
Effectives égaux (percentiles, quantiles)

- .. **Quantiles = percentiles** = égales fréquences = equal count
- .. **Mise en classes selon les fréquences des N**
(`hit-parade» des effectifs)
- .. Exemple **classification par quantiles avec 4 classes**:
limites = 25^{ème}, 50^{ème}, 75^{ème} percentiles (quartiles)
- .. Exemple **classification par quantiles avec 6 classes**:
limites = 10^{ème}, 25^{ème}, 50^{ème}, 75^{ème}, 90^{ème} percentiles
- .. **Attention.** Ne garantit pas une bonne séparation des entités différentes, mais permet la comparaison entre plusieurs cartes.

3d. Classification par quantiles



3d. Classification par quantiles



Discrétisation standardisée (par z-scores)

- .. D'abord, transformation des données:

$$Z = \frac{X - \mu}{\sigma}$$

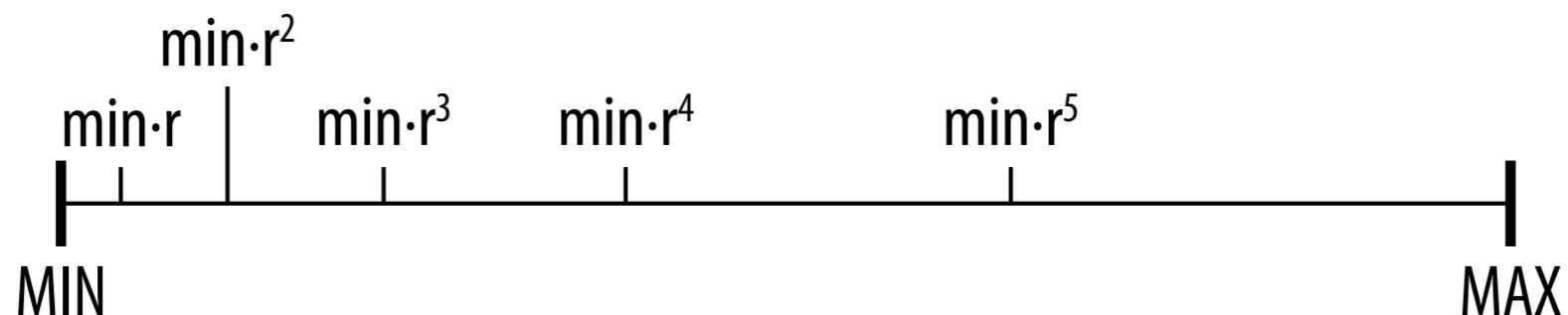
- .. $X - \mu$: centrer les valeurs (**moyenne = 0**)
- .. Division par σ : réduire les valeur (**écart-type = 1**)
- .. Z-scores = valeurs centrées-réduites =

Discrétisation standardisée (par z-scores)

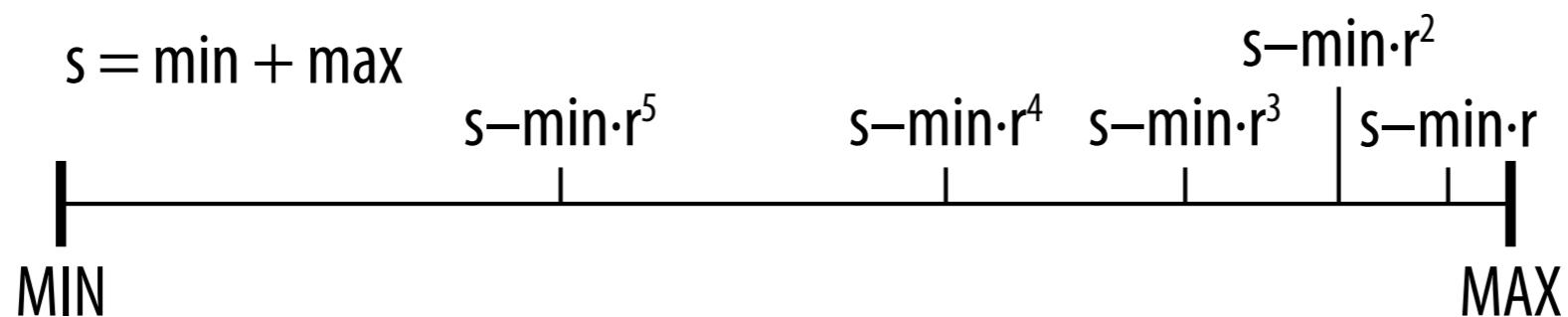
- .. Après standardisation, limites de classes:
égales amplitudes avec moyenne comme limite
- .. Donc p.ex.
 - .. $\mu - \sigma$ | $\mu - 0.5\sigma$ | μ | $\mu + 0.5\sigma$ | $\mu + \sigma$
- .. Ou
 - .. $\mu - 2\sigma$ | $\mu - \sigma$ | μ | $\mu + \sigma$ | $\mu + 2\sigma$
- .. **Attention:** ne permet pas de garantir le regroupement de valeurs semblables

Progression géométrique

Distribution dissymétrique à gauche:



Distribution dissymétrique à droite:



$$r = 10^{\frac{\log(\max) - \log(\min)}{k}}$$

k: nombre de classes

Progression géométrique

- .. Pour les distributions dissymétriques
- .. Équivalent à une discrétisation à pas constant sur données logarithmiques...
=> Il est possible de transformer les valeurs avant classification
- .. **Attention:** ne permet pas de garantir le regroupement de valeurs semblables

Classifications statistiques

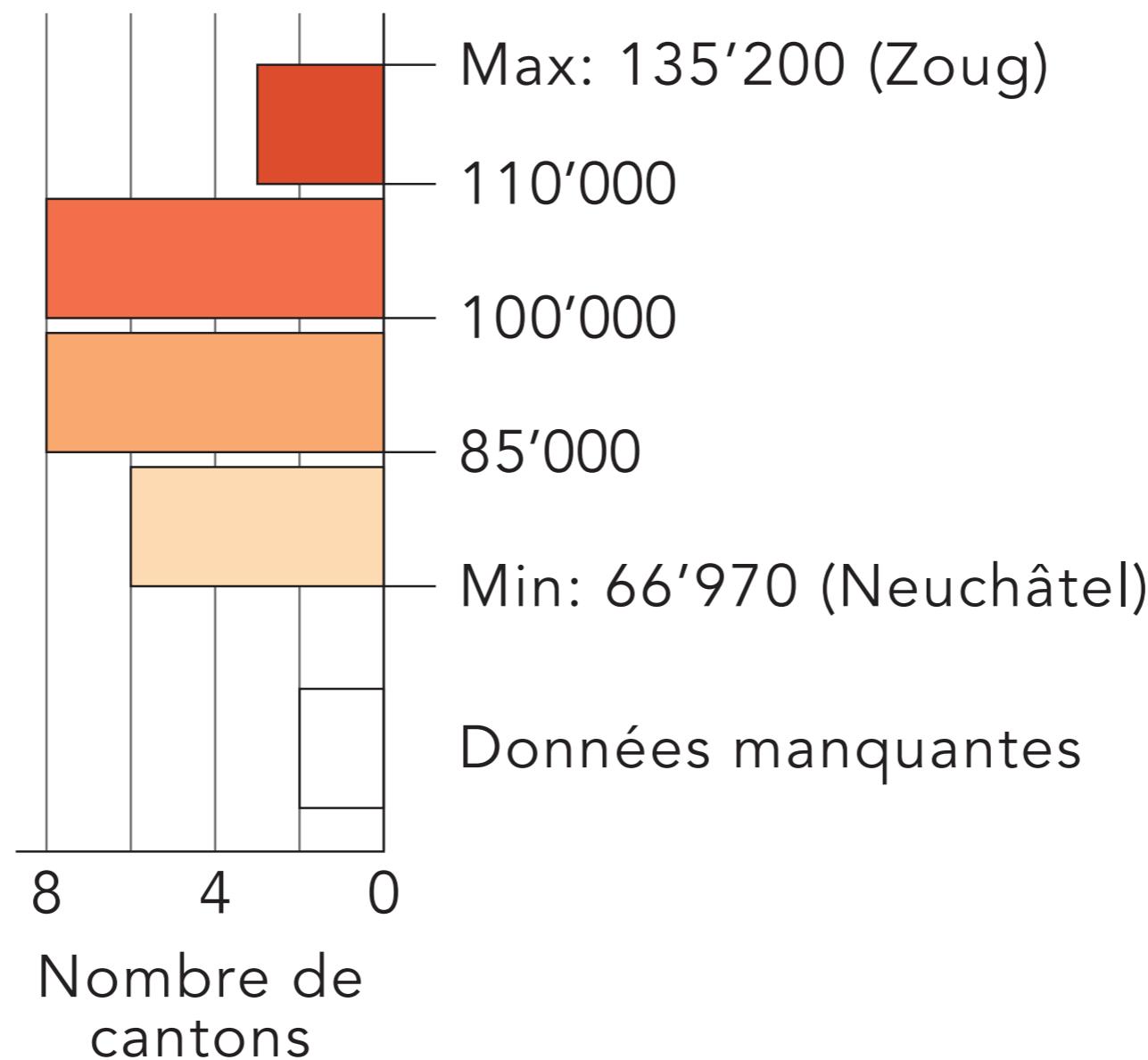
- .. Applicable à 1 variables ou plusieurs variables à la fois
- .. But: regrouper des entités similaires ensemble
- .. Beaucoup de méthodes disponibles, dont:
 - .. k-means, Classification ascendante hiérarchique, etc.
- .. Traditionnellement peu utilisées en cartographie, mais sur le principe rien n'empêche leur utilisation
 - .. Critères de séparation généralement bien définies, donc méthodes objectives
 - .. Utiliser une classification qui tient compte des poids des entités

Classifications statistiques

- .. Problèmes possibles:
 - .. Comparison de plusieurs cartes: procéder à une mise en classe avec l'ensemble des variables...
 - .. Il peut y avoir des classes avec un nombre d'unités très inégal (si presque toutes les entités sont similaires...)

Discrétisation: résultat...

Francs de créances par habitant



Tableaux de données

L'information statistique

Table d'attributs

Quantum GIS 1.7.4-Wroclaw

Attribute table - k4kant19970101gf :: 1 / 26 feature(s) selected

	geocode	name	abbr	pop	area
1	1	BERN	BE	320000	3200.0000
2	3	Luzern	LU	381966	1433.0007
3	4	Uri	UR	35382	1056.3607
4	5	Schwyz	SZ	147904	853.0118
5	6	Obwalden	OW	35885	485.0212
6	7	Nidwalden	NW	41311	237.6039
7	8	Glarus	GL	39217	675.9594
8	9	Zug	ZG	115104	207.3578
9	10	Fribourg	FR	284668	1600.3874
10	11	Solothurn	SO	256990	790.5308
11	12	Basel-Stadt	BS	186255	39.2022
12	13	Basel-Landschaft	BL	275360	521.2547
13	14	Schaffhausen	SH	77139	305.0722
14	15	Appenzell Ausserrhoden	AR	53313	243.2875
15	16	Appenzell Innerrhoden	AI	15743	172.8452
16	17	St. Gallen	SG	483156	1946.6903
17	18	Graubünden	GR	193388	7167.6115
18	19	Aargau	AG	618298	1396.467
19	20	Thurgau	TG	251973	865.6897
20	21	Tessin	TI	336943	2757.6774
21	22	Vaud	VD	725944	2826.3739
22	23	Valais	VS	317022	5261.6468
23	24	Neuchâtel	NE	173183	720.3714
24	25	Genève	GE	460534	258.9526
25	26	Jura	JU	70542	848.1599

Look for in name

Show selected only Search selected only Case sensitive ?

Table d'attributs

- Table d'attributs contient les données attributaires d'une couche vectorielle (p.ex. dans fichier Shape)
- Contient toujours un identifiant: le **géocode** ou **feature ID**
- Terme utilisé en relation avec un **système d'information géographique**
- Est partie intégrante dans une couche vectorielle; indissociable des géométries

Tableau d'Information Géographique (TIG)

- .. **Tableau d'information géographique** est similaire à une table d'attributs, mais sans lien explicite vers les géométries
- .. **Base de données** thématique
- .. Contient également un **géocode** permettant de **faire le lien** avec la couche vectorielle correspondante
- .. Tableau de **N lignes x M colonnes**
 - .. N unités géographiques $\Rightarrow i$
 - .. M caractéristiques (variables) $\Rightarrow j$

TIG: un peu de vocabulaire...

- .. Lorsqu'on **mesure** un phénomène, on affecte à chaque **unité géographique** une **modalité qualitative** ou une **valeur numérique** (une et une seule!)
 - .. Exemple de **modalités**: «gneiss», «calcaire», «basalte», etc.
 - .. Exemple de **valeurs**: 50 habitants par hectare, 20 hab./ha, 120 hab./ha, etc.
- .. L'ensemble des modalités ou valeurs possibles s'appelle une **variable**
 - .. Exemples de variables: «type de roche», «densité de population»

TIG: un peu de vocabulaire...

- .. L'application d'une variable à un ensemble d'individus ou d'unités géographiques s'appelle une **série statistique**
 - .. Exemple de série statistique: la densité de population des communes suisses
- .. Un **descripteur** est une mesure directe d'un phénomène
 - .. Exemple de descripteur: taux de mortalité
- .. Un **indicateur** est la mesure dérivée d'un phénomène
 - .. Exemple d'un indicateur: IDH (indice de développement humain)

Nature des données thématiques

- .. **Codes qualitatifs:** géocodes, modalités, types, catégories
- .. **Variables d'identification**, d'appartenance à un type, à une catégorie, à un groupe
- .. **Valeurs absolues**
 - .. **Mesures, comptages**
 - .. **Variables de taille** ou de masse qui donnent l'ordre de grandeur du type de... km, personnes, francs, degrés etc.
 - .. **Valeurs relatives**
 - .. **Dérivées de comptages: rapports, %, indices, densités, etc.**
 - .. **Variables de structure** entre une valeur au numérateur et une référence au dénominateur

Types de tableaux

- Trois types de tableaux de données pour décrire un espace géographique:
- **Géocode / FID = identifiant = fils rouge entre les deux colonnes**

Contenant	Contenu	Contenu																																																																																																									
<p>1. Fonds de carte</p> <p>Système spatial</p> <table><thead><tr><th>ID</th><th>X</th><th>Y</th></tr></thead><tbody><tr><td>1</td><td></td><td></td></tr><tr><td>2</td><td></td><td></td></tr><tr><td>.</td><td></td><td></td></tr><tr><td>j</td><td></td><td></td></tr><tr><td>.</td><td></td><td></td></tr><tr><td>N</td><td></td><td></td></tr></tbody></table> <p>Données géométriques</p>	ID	X	Y	1			2			.			j			.			N			<p>2. Régionalisations, zonages, ...</p> <p>Système spatial</p> <table><thead><tr><th>ID</th><th>1</th><th>2</th><th>...</th><th>i..</th><th>M</th></tr></thead><tbody><tr><td>1</td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>2</td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>.</td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>j</td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>.</td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>N</td><td></td><td></td><td></td><td></td><td></td></tr></tbody></table> <p>Données de maillage (qualitatif)</p>	ID	1	2	...	i..	M	1						2						.						j						.						N						<p>3. Mesures, comptages, indices, ...</p> <p>Système spatial</p> <table><thead><tr><th>ID</th><th>1</th><th>2</th><th>...</th><th>i..</th><th>M</th></tr></thead><tbody><tr><td>1</td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>2</td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>.</td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>j</td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>.</td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>N</td><td></td><td></td><td></td><td></td><td></td></tr></tbody></table> <p>Données statistiques (quantitatif)</p>	ID	1	2	...	i..	M	1						2						.						j						.						N					
ID	X	Y																																																																																																									
1																																																																																																											
2																																																																																																											
.																																																																																																											
j																																																																																																											
.																																																																																																											
N																																																																																																											
ID	1	2	...	i..	M																																																																																																						
1																																																																																																											
2																																																																																																											
.																																																																																																											
j																																																																																																											
.																																																																																																											
N																																																																																																											
ID	1	2	...	i..	M																																																																																																						
1																																																																																																											
2																																																																																																											
.																																																																																																											
j																																																																																																											
.																																																																																																											
N																																																																																																											

Du TIG vers la table d'attributs

- .. Malgré similarité: **TIG n'est pas une table d'attributs**
- .. Possibilité **d'attacher le TIG** à la table d'attributs à l'aide d'une **jointure**

Échelles de mesures

- .. L'échelle de mesure définit la **qualité** de l'information de la variable et les **opérations** qui peuvent être conduites sur les données
- .. Échelle **qualitative**, catégorielle:
 - .. Échelle **nominale**
 - .. Échelle **ordinale**
- .. Échelle **quantitative**
 - .. Échelle **d'intervalle**
 - .. Échelle **de rapports**

Type de donnée

Qualitative

Échelle de mesure

1° Nominale

- Concept de simple différenciation
- Variable discrète (distincte)
- Dans le cas de 2 modalités, dichotomie (présence/absence, oui/non)

Exemples:

- Types d'utilisation du sol
- Sexe
- Code d'identification (géocode)

2° Ordinale

- Concept d'ordre
- Variable en rang
- Permet des comparaisons <, > et =

Exemples:

- Rangs (1^{er}, 2^e, 3^e, ...)
- Grand, moyen, petit
- Directeur, sous-directeur, employé

Quantitative

3° Intervalle

- Variable mesurée avec un origine («0») arbitraire
- Opère des différences

Exemples:

- Latitude, longitude
- Dates, températures
- Écart à une valeur moyenne

4° De rapport

- Variable mesurée avec un origine («0») absolu
- Opère des rapports (%), ratio
- Fonctions arithmétiques + - x ÷
- Quantités bien définies

Exemples:

- Nombre de personnes
- Quantités de production
- Taux d'évolution

Type de donnée

Qualitative

Échelle de mesure

1° Nominale

- Concept de simple différenciation
- Variable discrète (distincte)
- Dans le cas de 2 modalités, dichotomie (présence/absence, oui/non)

2° Ordinale

- Concept d'ordre
- Variable en rang
- Permet des comparaisons <, > et =

s:

(1^{er}, 2^e, 3^e, ...) , moyen, petit eur, sous-directeur, employé

«Échelles NOIR»

Quantitative

3° Intervalle

- Variable mesurée avec un origine («0») arbitraire
- Opère des différences

Exemples:

- Latitude, longitude
- Dates, températures
- Écart à une valeur moyenne

4° De rapport

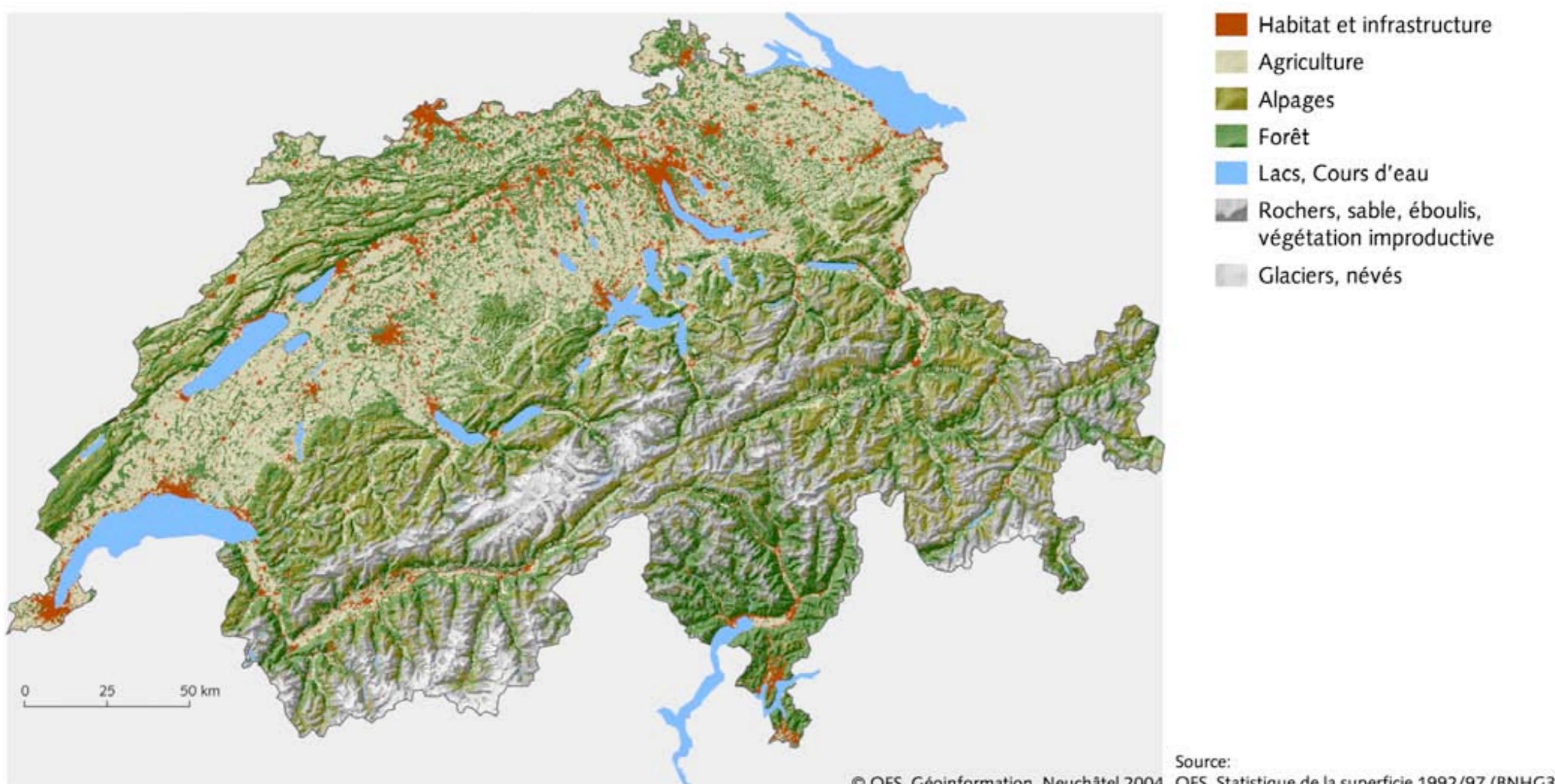
- Variable mesurée avec un origine («0») absolu
- Opère des rapports (%), ratio
- Fonctions arithmétiques + - x ÷
- Quantités bien définies

Exemples:

- Nombre de personnes
- Quantités de production
- Taux d'évolution

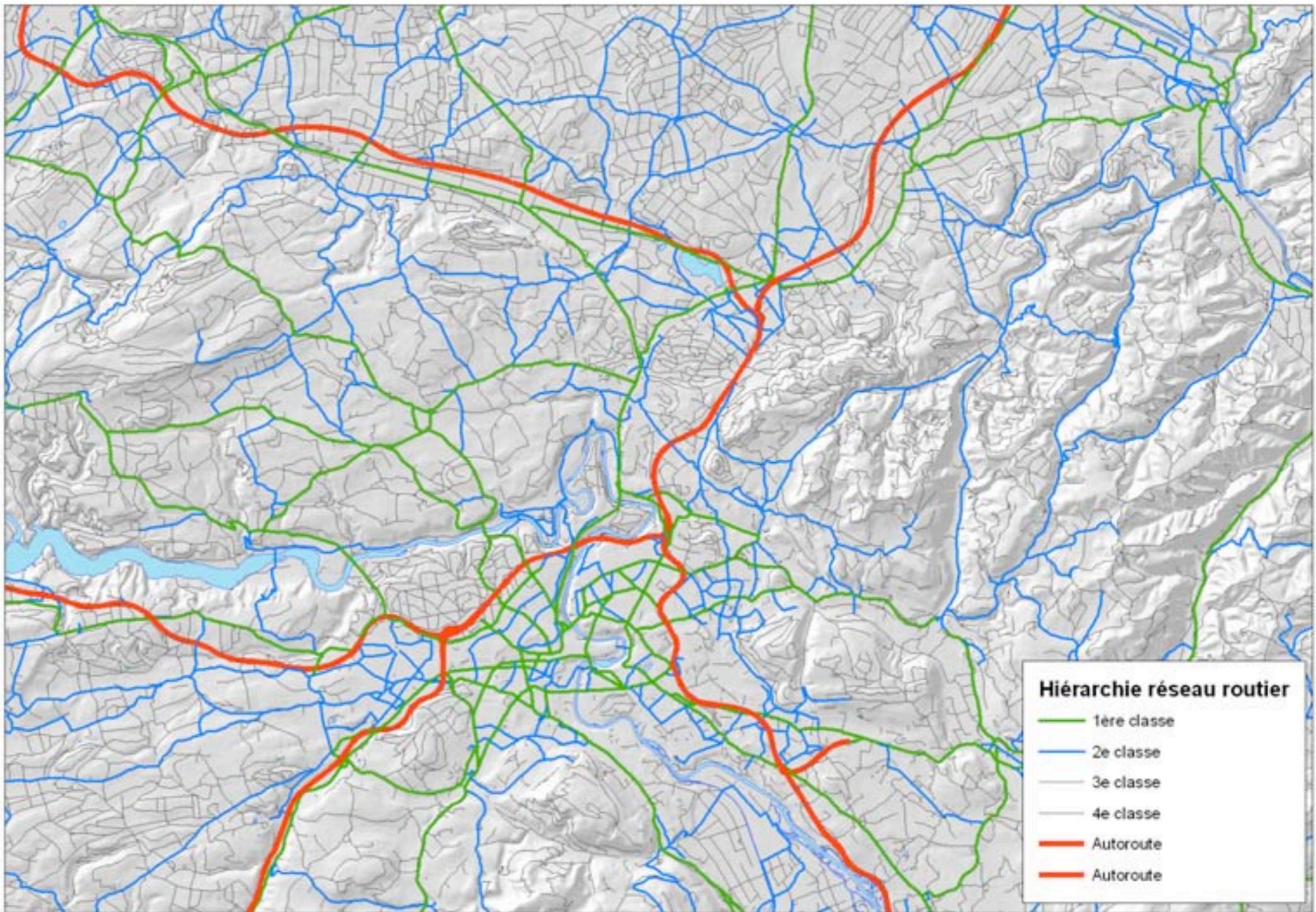
Échelle nominale: exemple

L'utilisation du sol en Suisse



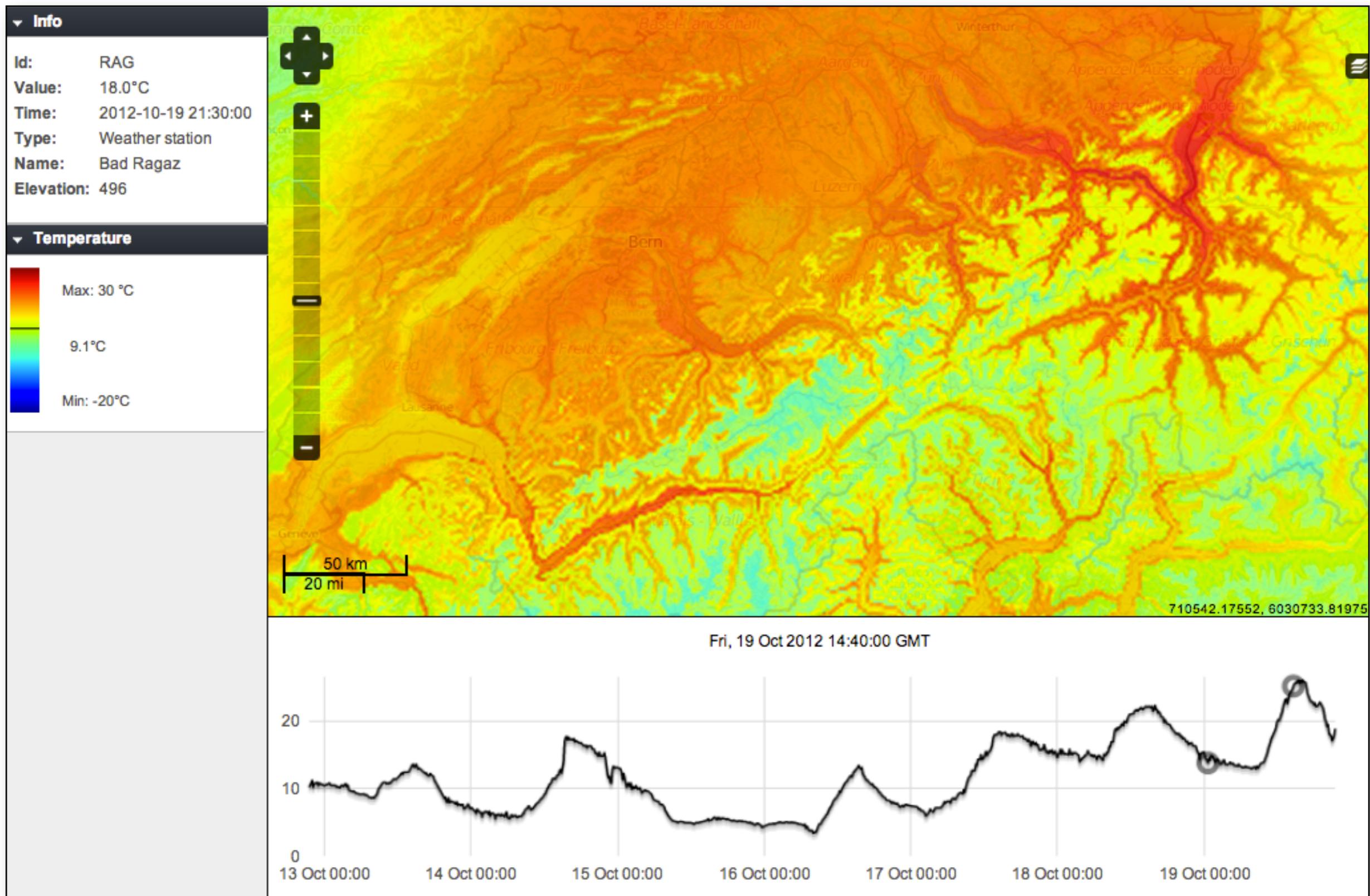
Échelle ordinaire: exemple

**Hiérarchie du
réseau routier**
Région de Berne



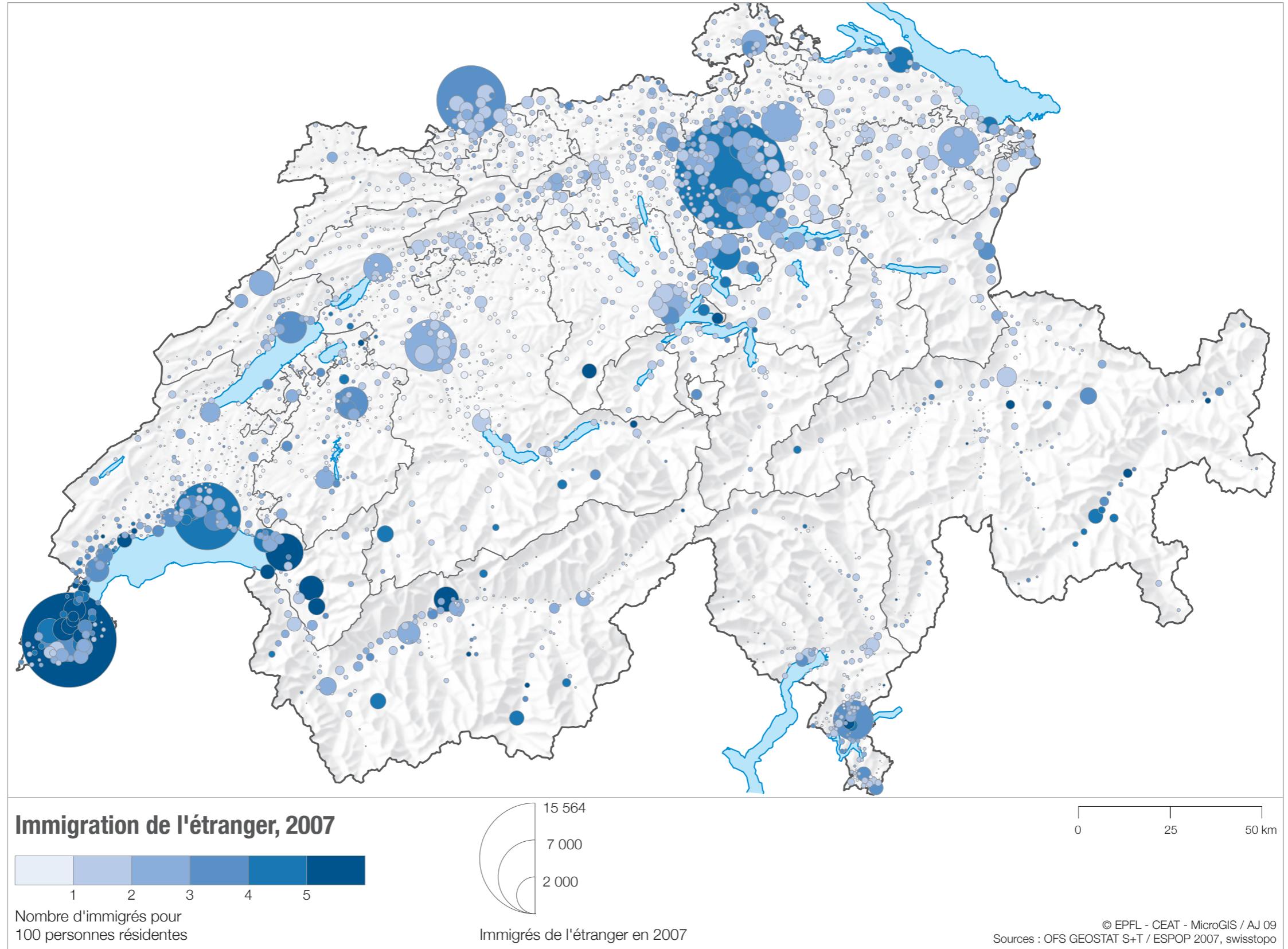
Échelle d'intervalle: exemple

Température de la Suisse
Carte interactive créée avec i2maps



Échelle de rapport: exemple

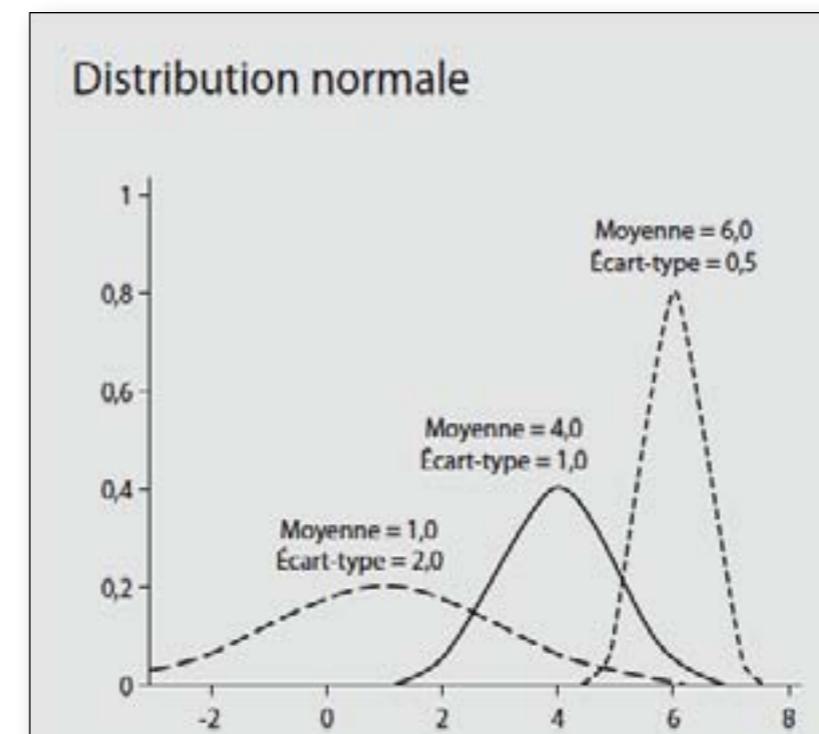
Pierre Dessemontet, Alain Jarne, Martin Schuler (2009)
Suisse romande. Les facettes d'une région affirmée.



Statistique descriptive

- .. Ensemble de méthodes qui permettent de **classer**, de **décrire**, de **représenter** graphiquement et de résumer des séries d'informations
- .. Réduction de l'information
- .. Méthodes mathématiques / méthodes graphiques
- .. Avant la cartographie!

$$s_n = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

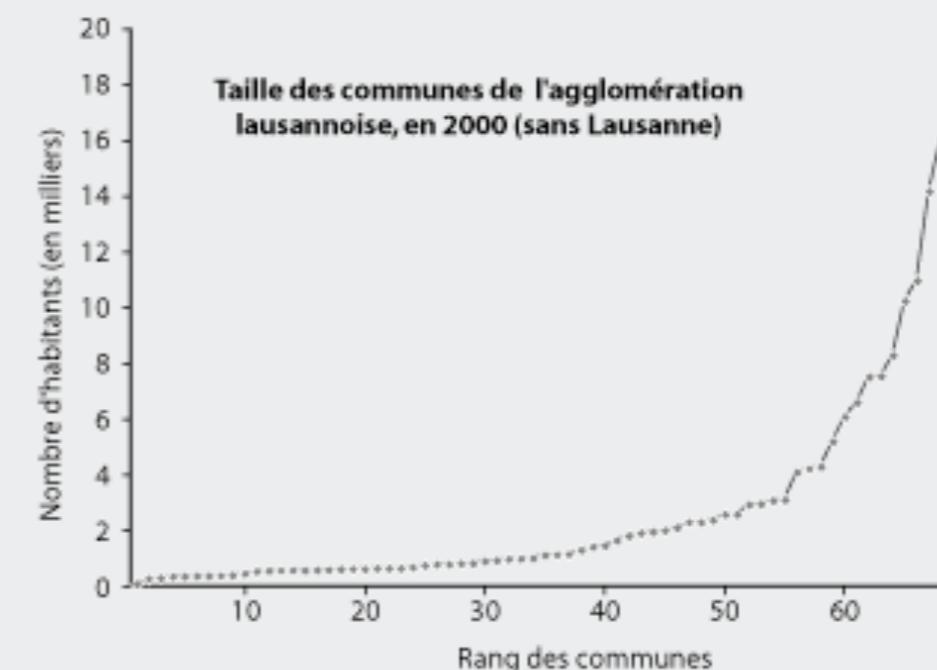


Méthodes graphiques

Histogramme des fréquences



Courbe de répartition



Scalogramme



Composantes de l'information statistique

1. Ordres de grandeur, valeurs centrales
2. Dispersion
3. Forme de la distribution
4. Cas particuliers

Composante 1: valeurs centrales

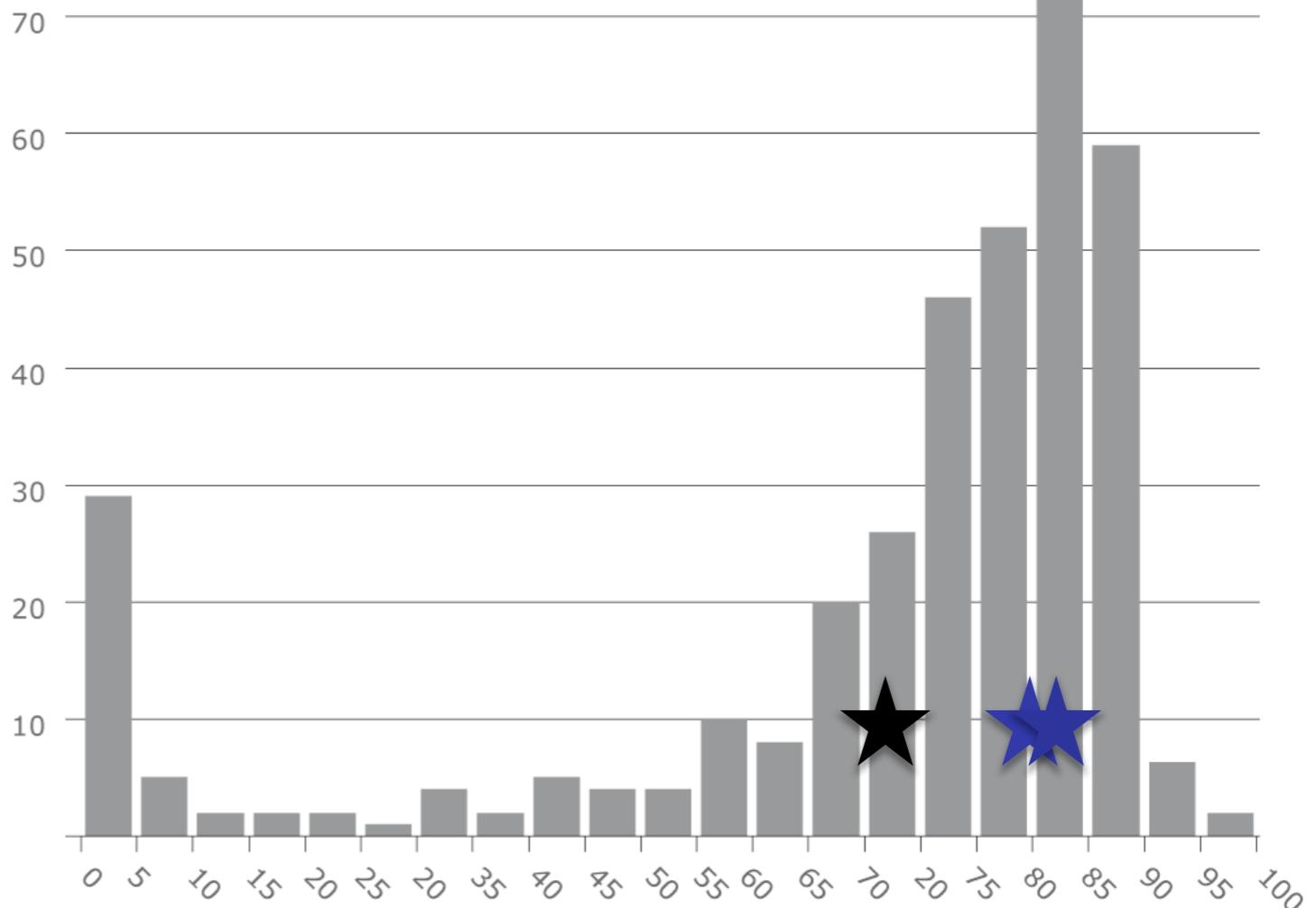
Ordre de grandeur donné par les **valeurs centrales**

- ★ **Mode** : la valeur (ou classe) la mieux représentée
- ★ **Médiane** : valeur située au milieu d'une série ordonnée de valeurs
- ★ **Moyenne** : somme des valeurs divisée par l'effectif

Grouped Frequency Histogramm

Albania – by communes, 2001

% Active people in agriculture sector



Composante 2: dispersion

- Etendue (max-min) et **écart-type**

Etape 1

2, 4, 4, 4, 5, 5, 7, 9.

Etape 2

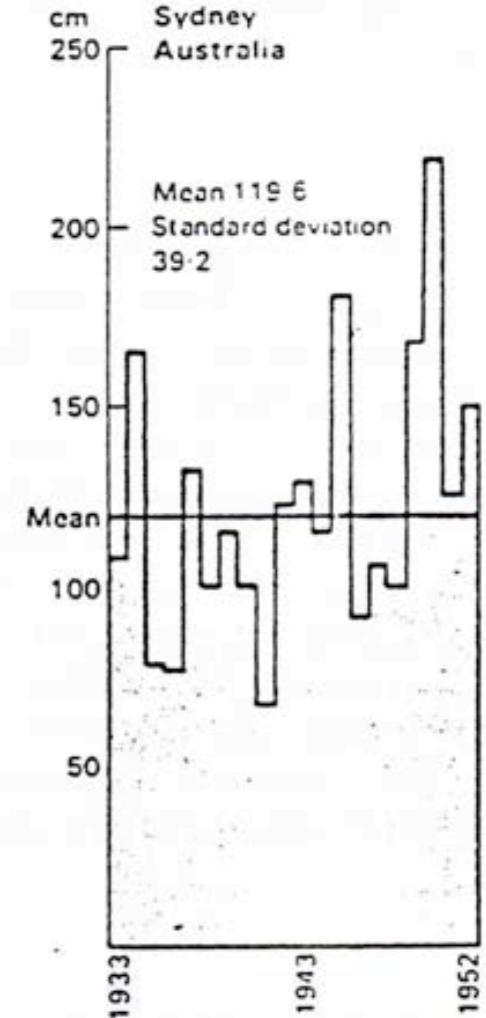
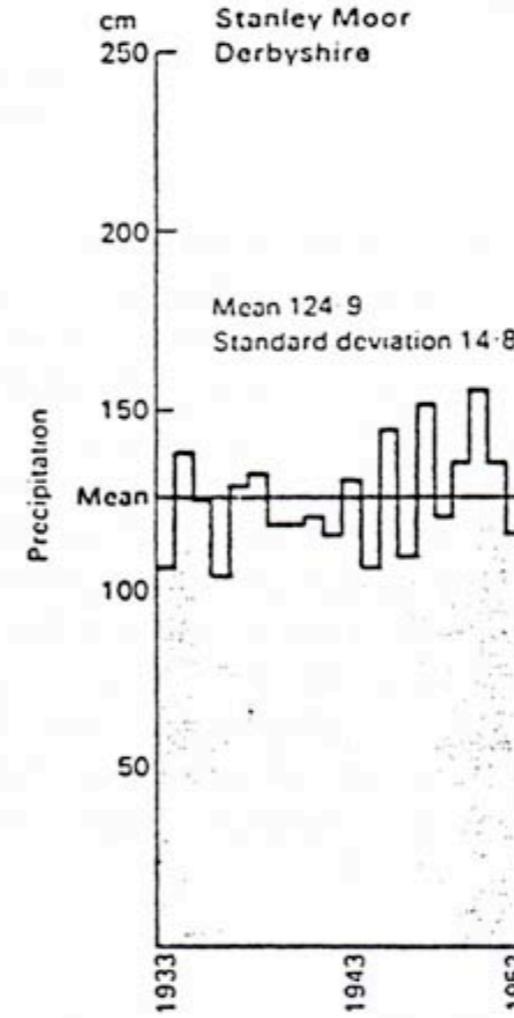
$$\frac{2 + 4 + 4 + 4 + 5 + 5 + 7 + 9}{8} = 5.$$

Etape 3

$$\begin{aligned}(2 - 5)^2 &= (-3)^2 = 9 & (5 - 5)^2 &= 0^2 = 0 \\(4 - 5)^2 &= (-1)^2 = 1 & (5 - 5)^2 &= 0^2 = 0 \\(4 - 5)^2 &= (-1)^2 = 1 & (7 - 5)^2 &= 2^2 = 4 \\(4 - 5)^2 &= (-1)^2 = 1 & (9 - 5)^2 &= 4^2 = 16\end{aligned}$$

Etape 4

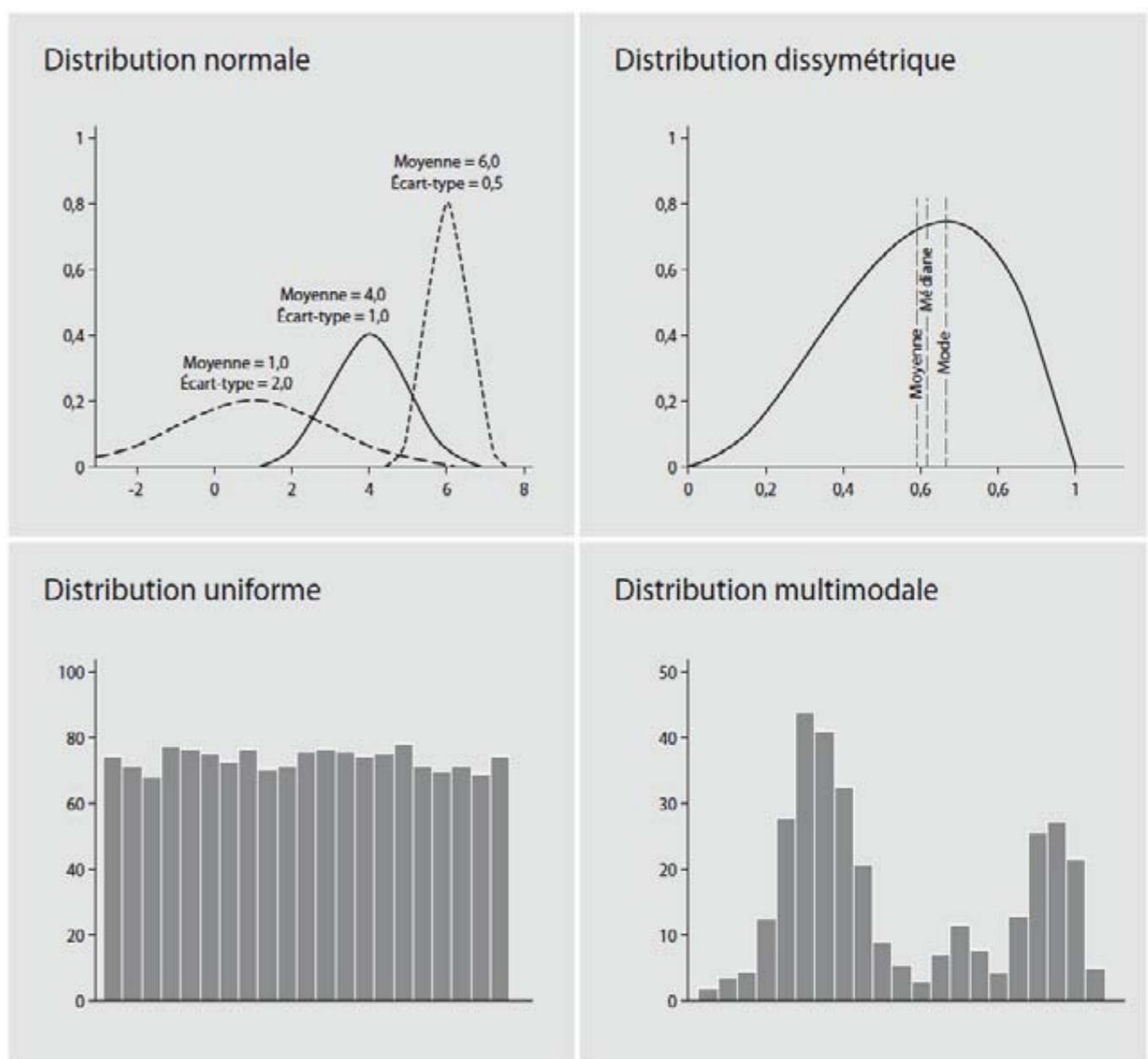
$$\sqrt{\frac{9 + 1 + 1 + 1 + 0 + 0 + 4 + 16}{8}} = \sqrt{4} = 2.$$



Composante 3: forme de la dispersion

Représentation graphique :
histogramme ou courbe des fréquences

Souvent très **dissymétrique** :
variables absolues (ou de stock)
comme population, superficie,
production, etc., et certains ratios
(densité)



Tendance **gaussienne** :
nombreuses **variables relatives**
(effet de taille disparaît lors du
calcul du rapport), comme taux de
croissance, de natalité, etc.

Composante 4: **cas particuliers**

- .. Une série statistique peut présenter d'importantes discontinuités et des **valeurs extrêmes** (très petites ou très grandes)
- .. Déterminer s'agit d'une **valeur aberrante** (erreur de mesure, situation conjoncturellement anormale) ou d'un cas présentant un **intérêt véritable**

Jusqu'à la semaine prochaine...

- .. Devoirs:
 - .. Lecture dans Lambert & Zanin 2016:
 - .. Chapitre 2: les données attributaires
 - .. Exercice 5
 - .. Calcul d'indicateur dans Excel
 - .. Discrétisation