

# INITIATION À LA STATISTIQUE DESCRIPTIVE

## La cartographie statistique

Maître de Conférences, Université de Pau  
Laboratoire Société Environnement Territoire  
UMR 5603 du CNRS et Université de Pau  
Domaine Universitaire, IRSAM, 64000 PAU  
Tél : 05 59 92 31 23 Fax : 05 59 80 83 39  
Mail : [dominique.laffly@univ-pau.fr](mailto:dominique.laffly@univ-pau.fr)

### 5. La cartographie en statistique descriptive

La description d'une variable en analyse des données – statistique descriptive – a pour but de résumer au mieux ces caractéristiques en se fondant sur des données et des graphes de synthèse.

Les valeurs mathématiques couramment retenues sont, pour une variable quantitative :

- des indicateurs de *position* tels que la moyenne arithmétique, le minimum, le maximum et les quantiles d'ordre  $k$  ;
- des indicateurs de *dispersion* tels que la variance, sa racine carrée l'écart type (exprimé dans l'unité de la variable) et le coefficient interquartile ;
- des indicateurs de « forme » tels que les coefficients d'asymétrie et d'aplatissement.

Les graphes couramment utilisés sont les boîtes à moustaches et la courbe de fréquences cumulées pour les variables quantitatives. Les données qualitatives ne se prêtent pas à des calculs, on se contente au stade de l'analyse univariée de dresser un tableau de fréquences des différentes modalités et de les représenter graphiquement par un diagramme en bâtons ou en secteurs.

La carte constitue un graphe particulier de l'analyse des données. Notons dès à présent que la réalisation d'une carte statistique ne peut s'envisager que lorsque les individus traités renvoient à des portions d'espace dont on connaît les caractéristiques – le fond de carte en quelque sorte.

On distingue différents types de carte selon la nature des données (figure 1) et les règles de la sémiologie graphique. Retenons néanmoins que le type de carte que nous proposons ici a pour but de nous permettre – après avoir observé la forme de la distribution – de visualiser la répartition des différentes valeurs dans l'espace. Ainsi peut-on observer des distributions aléatoires ou au contraire réparties selon des tendances plus ou moins fortes qu'il nous incomberait par la suite de tenter d'explicitier. **Comprenons bien qu'il ne s'agit pas d'analyse spatiale même si nous interprétons des distributions dans l'espace.**

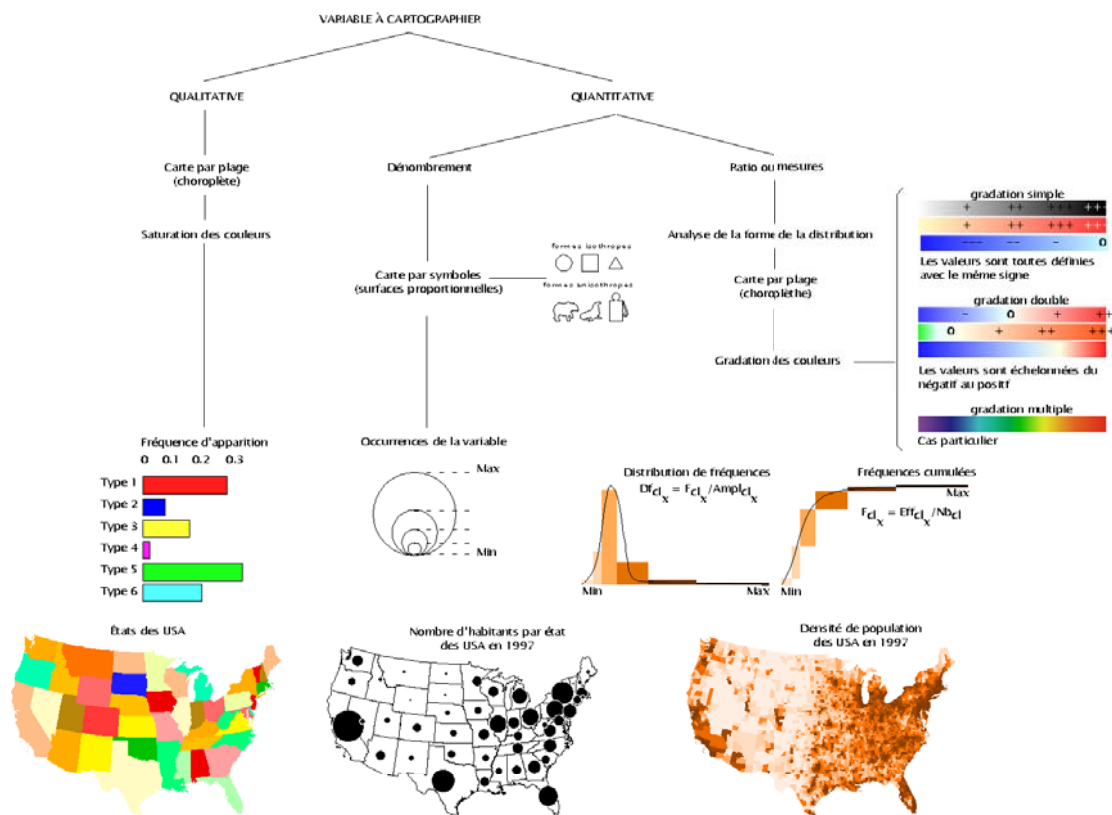


Figure 1 : Les différents modes de cartographie statistique

Par exemple, une carte des densités de population par commune (figure 2) fait ressortir des auréoles de valeurs décroissantes autour des nœuds urbains. L'interprétation de ce fait permet d'émettre l'hypothèse que la distance – entre autres facteurs – aux nœuds urbains est un élément de compréhension et d'explication des répartitions des densités de population. Un modèle – linéaire, gravitaire, polynomial... – fondé sur une équation qui permettrait de déduire les densités de population en ne connaissant *a priori* que la distance à des nœuds urbains serait véritablement de l'analyse et de la statistique spatiales.

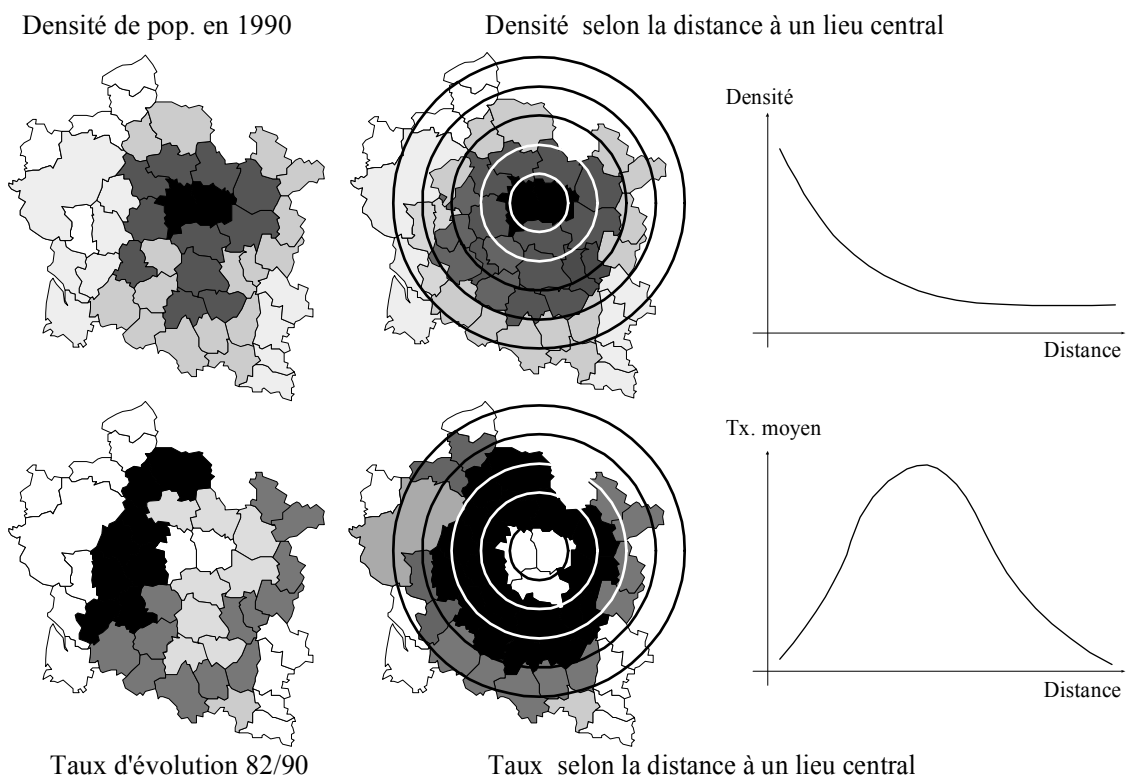


Figure 2 : Population observée et population simulée par un modèle gravitaire

### 5.1. Les cartes par surfaces proportionnelles

Les règles de sémiologie graphique imposent une représentation par surfaces proportionnelles pour des variables quantitatives exprimant un dénombrement, un effectif. La population par commune, les suffrages pour un candidat à une élection, un comptage à un point donné... Le fait que la surface soit proportionnelle permet de reconstituer – en théorie – toutes les valeurs à partir de l'échelle de la carte.

Les rapports de proportion sont fondés sur ceux des valeurs à cartographier. Généralement on fixe une surface maximale à la valeur la plus forte puis on déduit les autres surfaces selon le rapport de chaque valeur à celle la plus forte. Le tableau 1 présente un exemple à partir d'une population fictive de cinq individus et une surface maximale fixée à 250. Ainsi l'individu a dont la valeur est vingt fois inférieure ( $100/5 = 20$ ) à celle de *e* se voit attribué une surface vingt fois plus faible ( $250/12.5 = 20$ ).

Individus	Population	Surface	$V_i / V_{\max}$	côté carré	côté triangle	Rayon cercle
a	5.00	12.50	0.05	3.54		1.99
b	75.00	187.50	0.75	13.69		7.73
c	35.00	87.50	0.35	9.35		5.28
d	89.00	222.50	0.89	14.92		8.42
e	100.00	250.00	1	15.81		8.92

Tableau 1 : exemple théorique de calcul de surfaces proportionnelles

Reste à déterminer une forme pour la représentation graphique. On choisit généralement des formes pour lesquelles il n'y a qu'une et unique solution pour une surface donnée : le carré, le triangle équilatéral et le cercle. Ces formes offrent l'avantage d'être isotropiques, c'est-à-dire de proposer les mêmes caractéristiques quel que soit la direction d'allongement. En d'autres termes il ne peut y avoir de distorsions qui pourraient fausser l'interprétation, comme avec un rectangle, par exemple.

Connaissant la surface maximale, il reste à déterminer la valeur de la composante qui permet de dessiner la forme retenue. Pour le cercle il faut procéder en deux étapes. Déterminer tout d'abord la valeur du rayon maximal pour la surface maximale :

$$R_{\max} = \sqrt{\frac{S_{\max}}{\pi}}$$

Puis pour chaque individu *i* on détermine le rayon selon le rapport de proportionnalité suivant :

$$R_i = \sqrt{\frac{V_i}{V_{\max}}} * R_{\max}$$

- Rq. 1. Les remarques et formules précédentes supposent que l'on fonctionne d'une manière linéaire dans un espace euclidien ;
2. pour passer à la réalisation de la carte il est nécessaire que chaque individu soit localisé en ligne et colonne ou en latitude et longitude. La forme retenue est centrée sur ces coordonnées – ou centroïdes lorsqu'il s'agit du « centre » d'une forme complexe (une commune, par exemple).
3. Les logiciels de dessin permettent d'obtenir facilement des rapports de surfaces exacts avec n'importe quelle le forme (figure3). Celles-ci ne respectent cependant pas les conditions d'isotropie.

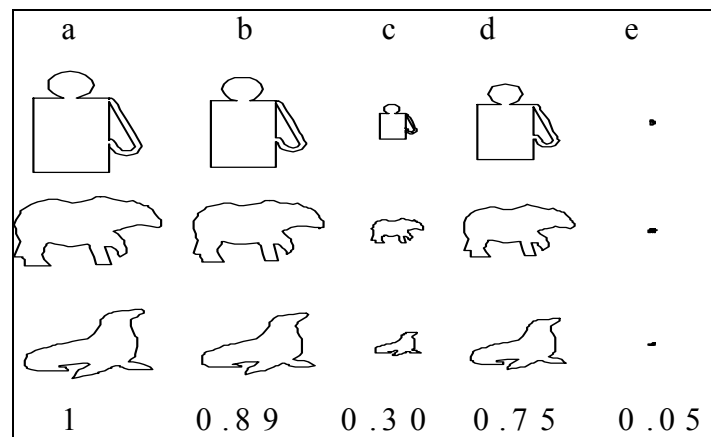


Figure 3 : Surfaces proportionnelles avec des formes complexes

4. Les cercles doivent être dessinés du plus grand au plus petit sans omettre de leur attribuer une couleur de trait différente de celle du remplissage de manière à ne pas masquer un ou plusieurs cercles au cours de la réalisation de la carte.
5. En cas de trop forte disparité des valeurs, la carte peut être illisible : quelques très « gros cercles » et d'autres trop « petits » pour lesquels on ne distingue plus les différences de surfaces (par exemple, la population des régions françaises). On peut choisir alors deux échelles de surfaces que l'on signalera en légende par des cercles de

couleurs différentes. Il est également possible de garder la même échelle mais d'éviter les cercles trop volumineux (figure 4).

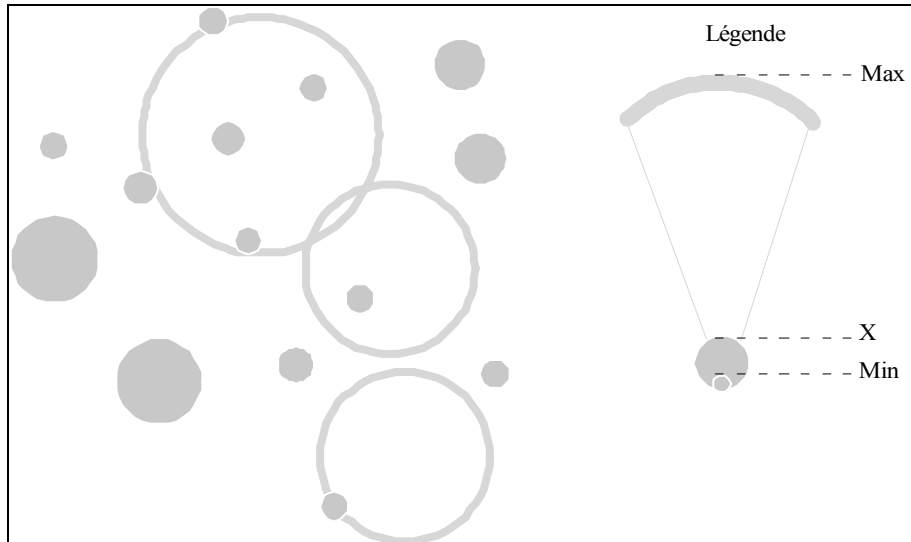


Figure 4 : Exemple de cercles pleins et évidés

6. À condition de le justifier et pour des cas précis, il est préférable de ne pas représenter par des surfaces proportionnelles certaines données même si les règles énoncées plus haut le laissent présager. C'est le cas notamment de la cartographie de relevés ponctuels de composition végétale. Sans cela les cartes seraient illisibles.

7. Il est possible de réaliser des cartes de flux en dessinant un rectangle entre deux points dont la largeur est proportionnelle à la valeur de l'effectif ou dénombrement du flux entre ces points. Il ne s'agit alors plus de rapports de surfaces mais d'un rapport uniquement sur la largeur du rectangle.

## 5.2. Les cartes par plage ou choroplètes

Avec des variables qualitatives ou quantitatives issues de combinaison de plusieurs variables (indices, ratios...) on applique une cartographie par plages de teinte, dite choroplète.

Dans le cas des variables qualitatives il convient d'appliquer aux différentes modalités des couleurs les plus saturées les unes par rapport aux autres de manière à éviter toute erreur d'interprétation liée à une gradation erronée des teintes. Supposons que les modalités soient numérotées de 1 à 10. Mathématiquement 2 est supérieur à 1, mais cela ne signifie rien du point de vue de la définition sémantique des classes. Ce qui était numéroté 1 pourrait tout aussi bien être 2, 3 ou 4 ou « a » ou « urbain dense »... Les teintes – ou trames – retenues doivent impérativement respecter cette indépendance des modalités.

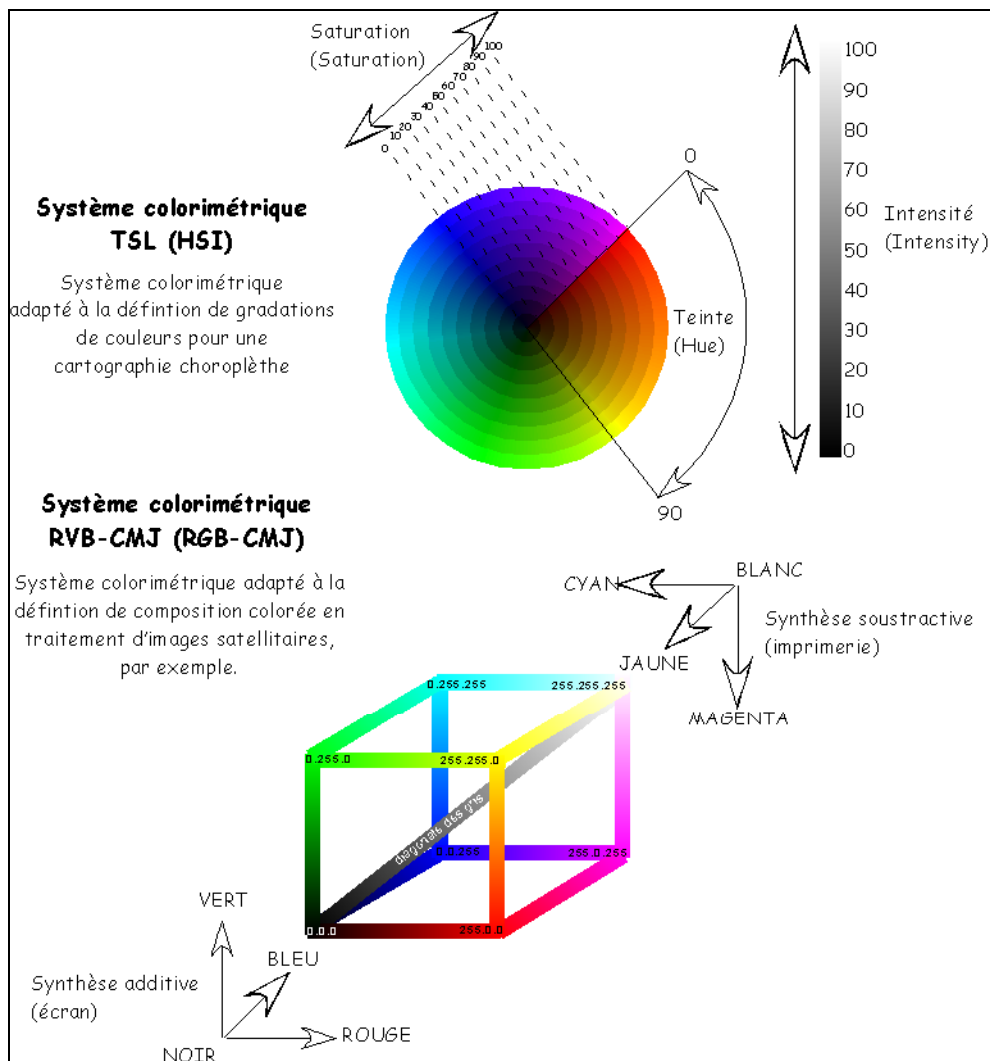


Figure 5 : Les espaces colorimétriques

La figure 5 présente différents modes colorimétriques couramment utilisés. La synthèse additive est celle de l'écran de l'ordinateur, soustractive celle de l'imprimerie. Toutes deux

sont définies dans un espace cubique dont les sommets définissent les couleurs primaires (Cyan, Magenta, Jaune, Rouge, Vert, bleu), le noir et le blanc. Entre ces deux dernières se dessine une diagonale composée de gris (même « quantité » de chaque couleur primaire). Il est cependant plus commode d'utiliser un autre espace colorimétrique pour répondre aux exigences de la cartographie statistique : saturation maximale des teintes ou au contraire, gradation pour une teinte donnée. Il s'agit de l'espace TSL défini par la **teinte** (équivalent des teintes du spectre électromagnétique, gradation de 0 à 360°), la **saturation** (pourcentage de saturation affecté au couple teinte/intensité) et **l'intensité** (pourcentage d'intensité affecté au couple teinte/saturation) sied parfaitement à nos contraintes.

Pour saturer au maximum les teintes, il suffit d'optimiser le pas angulaire entre chacune. Pour obtenir une gradation dans une teinte, il suffit de varier régulièrement le couple intensité/gradation d'un minimum vers un maximum. Notons que les logiciels de graphisme offrent des menus où tous les modes colorimétriques sont facilement accessibles. Reportez-vous à la carte de la variable qualitative présentée à la figure 1 pour un exemple de saturation des couleurs.

Une autre logique prévaut à la cartographie d'une variable quantitative. L'accent est mis sur le respect de la forme de la distribution. Les paramètres calculés ainsi que les graphes obtenus au cours de la phase initiale d'analyse univariée permettent une estimation correcte de cette forme. Le tableau 2 et la figure 6 présentent une telle synthèse obtenue à partir de données sur les départements français issues du RGP 1990.

	Migration (75-82)/75	Densité
minimum	-7.875	0.1615836
décile 1	-3.7845	0.31331672
quantile 1	-0.6345	0.48349458
médiane	1.704	0.91549743
quantile 3	4.141	2.25313965
décile 9	6.7365	5.61071489
max	13.128	42.3893063
moyenne	1.63075	3.23221097
écart type	4.204923563	7.68475101
asymétrie	0.235268826	4.08408707
nb. Classe	7	7

Tableau 2 : statistique descriptive

Où : - Quantile d'ordre 10 ou décile :

$$F_{\text{déc1}}=0.1 \text{ et } R_{\text{déc1}} = (F_{\text{déc1}} * N) + 0.5$$

et

$$Val_{\text{déc1}} = Val_{|R_{\text{déc1}}|} + \left\{ (R_{\text{déc1}} - |R_{\text{déc1}}|) * (Val_{|R_{\text{déc1}}|+1} - Val_{R_{\text{Déc1}}}) \right\}$$

et ainsi de suite pour les différentes fréquences des quantiles.

- Moyenne arithmétique :

$$\bar{x} = \frac{1}{N} \sum_{i=1}^n x_i$$

- Variance et écart type :

$$Var = \frac{1}{N} \sum_{i=1}^n (x_i - \bar{x})^2 \text{ et } \sigma_x = \sqrt{Var}$$

- Coefficient d'asymétrie : cette fonction caractérise le degré d'asymétrie d'une distribution par rapport à sa moyenne. Une asymétrie positive indique une distribution unilatérale décalée vers les valeurs les plus positives. Une asymétrie négative indique une distribution unilatérale décalée vers les valeurs les plus négatives.

$$Asym = \frac{n}{(n-1) * (n-2)} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{\sigma_x} \right)^3$$

- Coefficient d'aplatissement ou Kurtosis : le kurtosis caractérise la forme de pic ou l'aplatissement relatifs d'une distribution comparée à une distribution normale. Un kurtosis positif indique une distribution relativement pointue, tandis qu'un kurtosis négatif signale une distribution relativement aplatie.

$$Kur = \left\{ \left( \frac{n(n+1)}{(n-1)(n-2)(n-3)} \right) \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{\sigma_x} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

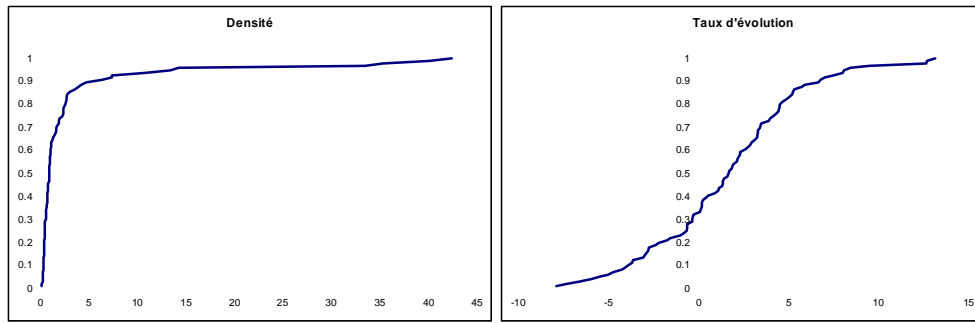


Figure 6 : Courbes de fréquences cumulées

Les deux variables retenues se distribuent de manière radicalement différentes. On ne saurait donc leur appliquer les mêmes traitements pour la réalisation de la carte. Le plus important consiste à définir des seuils de manière à discrétiser en classes la série étudiée. On ne peut représenter une carte composée de 96 départements avec 96, ou de 36 779 communes françaises avec autant de teintes. De plus, le seuillage permet un regroupement thématique facilitant l'interprétation des données.

Il existe différentes méthodes de seuillage, mais n'oublions cependant jamais qu'il est impératif de respecter prioritairement la forme de la distribution initiale et de trouver le meilleur compromis entre l'homogénéité des individus au sein d'une même classe et la plus grande hétérogénéité entre les classes.

Un nombre de classes  $Nb_{cl}$  indicatif est donné par la formule suivante (Sturge) :

$$Nb_{cl} = \left\lceil 1 + 3.3 * \text{Log}(N) \right\rceil$$

Ou bien encore par celle de Yule :

$$Nb_{cl} = \left\lceil 2.5 * N^{0.25} \right\rceil$$

où :  $N$  est le nombre d'individus

Chaque classe sera définie par des bornes, une amplitude, un effectif et une densité de fréquences. La borne minimale de la première classe est celle du minimum de la série, celle maximale de la dernière classe est celle du maximum de la série. La somme des effectifs est égale à  $N$ . Ces résultats peuvent être archivés dans une matrice.

La plupart des méthodes de seuillage sont fondées sur une forme de distribution normale ou plus généralement normale quelconque. La figure 6 présente la distribution de la variable « densité » retenue à titre d'exemple.

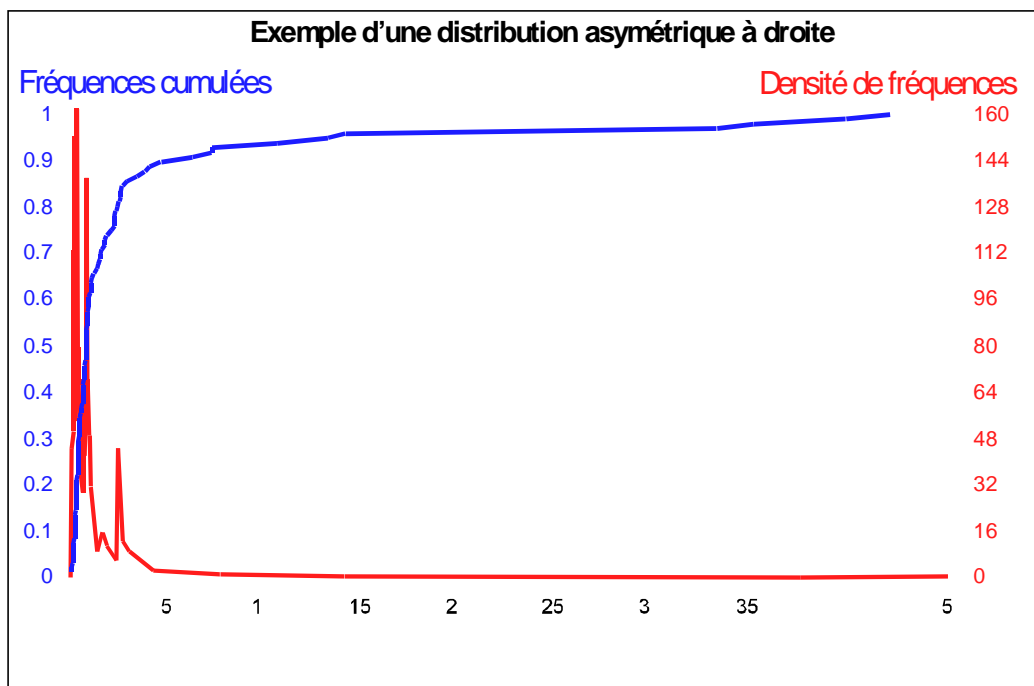


Figure 6 : distributions de la variable asymétrique « densité »

Il s'agit d'une distribution unimodale fortement asymétrique à droite, c'est-à-dire marquée par une plus longue queue de distribution à droite qu'à gauche (coefficient d'asymétrie = 4.08). Cette caractéristique de forme devra être respectée mais toutes les méthodes ne se prêtent pas à cette règle. On obtient alors des cartes aberrantes comme nous allons le voir dans ce qui suit.

### Équipopulation ou quantile d'ordre $Nb_{cl}$

Les limites des classes sont définies de manière à ce que chaque classe présente le même effectif.

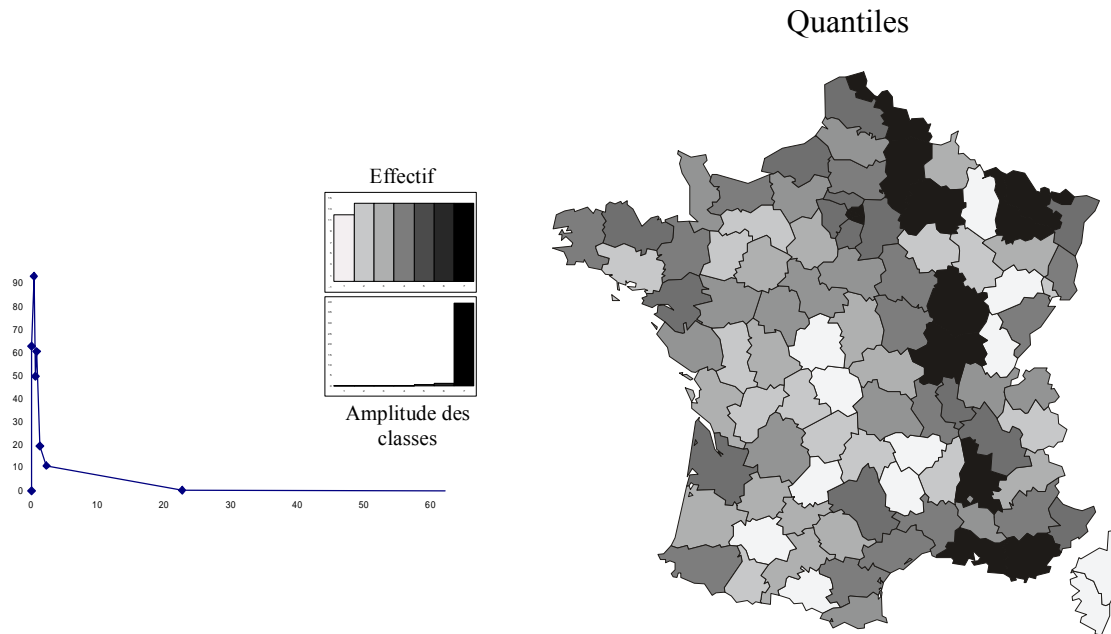


Figure 8 : Discretisation en quantiles

### Équivalence distributionnelle ou Équi-amplitude

Chaque classe est définie par une amplitude  $a_{cl} = a_{série} / Nb_{cl}$ .

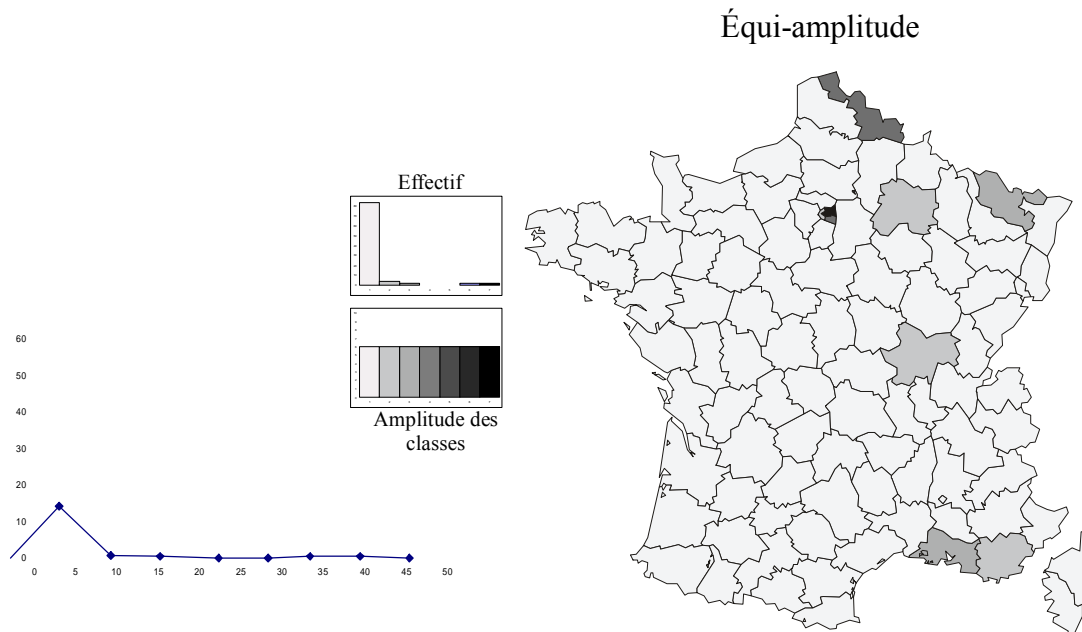


Figure 9 : Discretisation en équivalence distributionnelle

### Progression arithmétique et géométrique

Les amplitudes des classes augmentent au fur et à mesure que l'on incrémente les classes. La raison  $r$  de la croissance varie moins rapidement pour une progression arithmétique que pour une progression géométrique :

$$r_{\text{arithmétique}} = a_{\text{série}} / 2^{(\text{Nb}_{\text{cl}} - 1)} \rightarrow \text{l'amplitude de chaque classe est } a_{\text{cl } x} = r_{\text{arithmétique}} * 2^{(x-1)}$$

$$r_{\text{géométrique}} = (\text{Log}_{\text{max}} - \text{Log}_{\text{min}}) / \text{Nb}_{\text{cl}} \rightarrow \text{l'amplitude de chaque classe est } a_{\text{cl } x} = \text{max}_{\text{cl-1}} * r_{\text{géo.}}$$

$$\text{avec } \text{max}_{\text{cl1}} = \text{min} + r_{\text{géo.}}$$

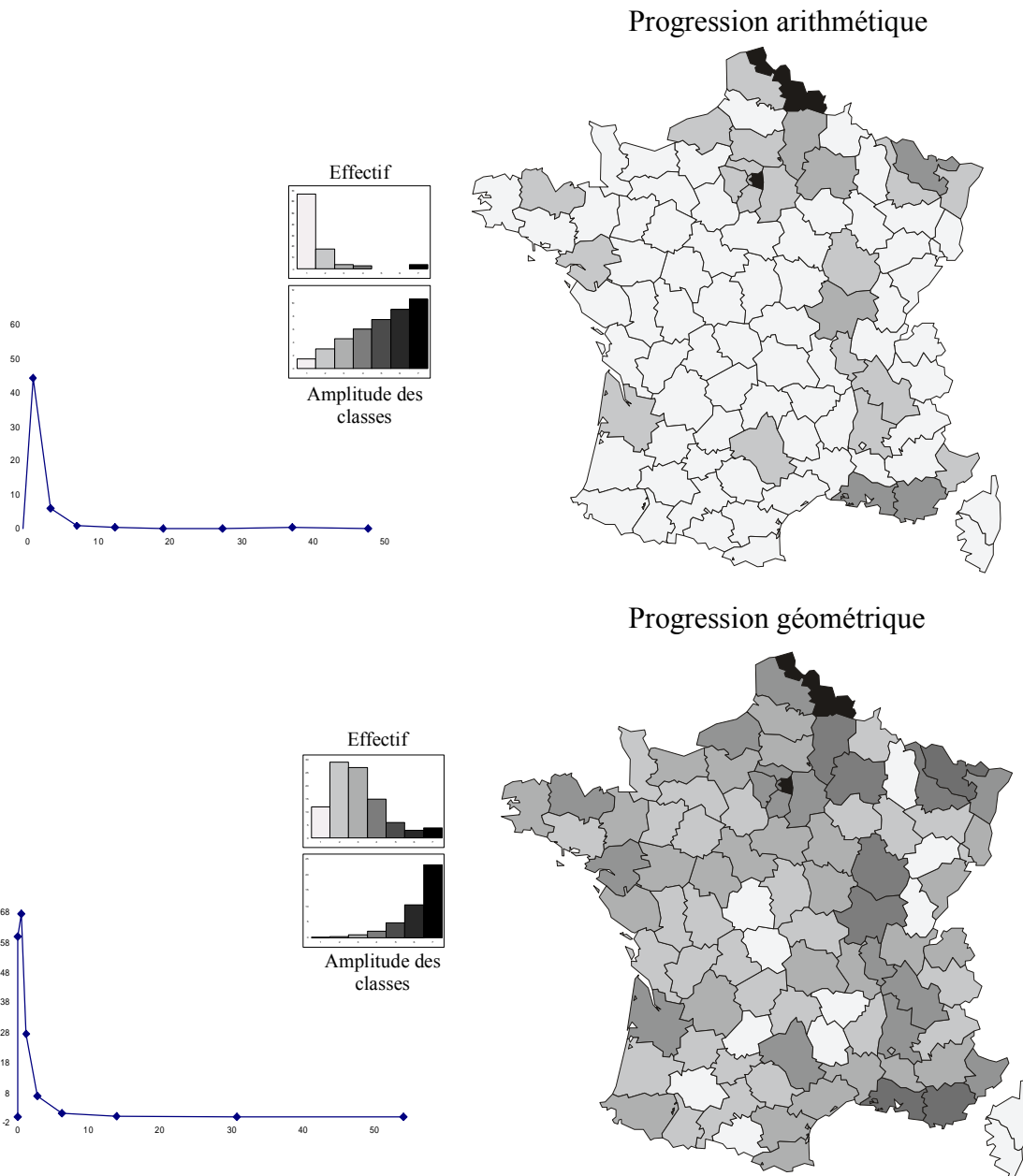


Figure 10 : Discretisation selon une progression arithmétique ou géométrique

### **Standardisation et Équiprobabilité**

Ces deux méthodes sont fondées sur l'hypothèse de normalité de la distribution.

La standardisation consiste à définir une première classe centrée sur la moyenne et d'une amplitude d'une valeur de 1 écart type. Les autres classes sont ensuite incrémentées selon un pas d'une valeur de 1 écart type de part et d'autre de cette classe centrale.

La définition des seuils selon une méthode d'équiprobabilité repose sur le même principe mais les bornes des classes sont lues dans une table de probabilité de la loi normale pour les pas d'écart type retenus.

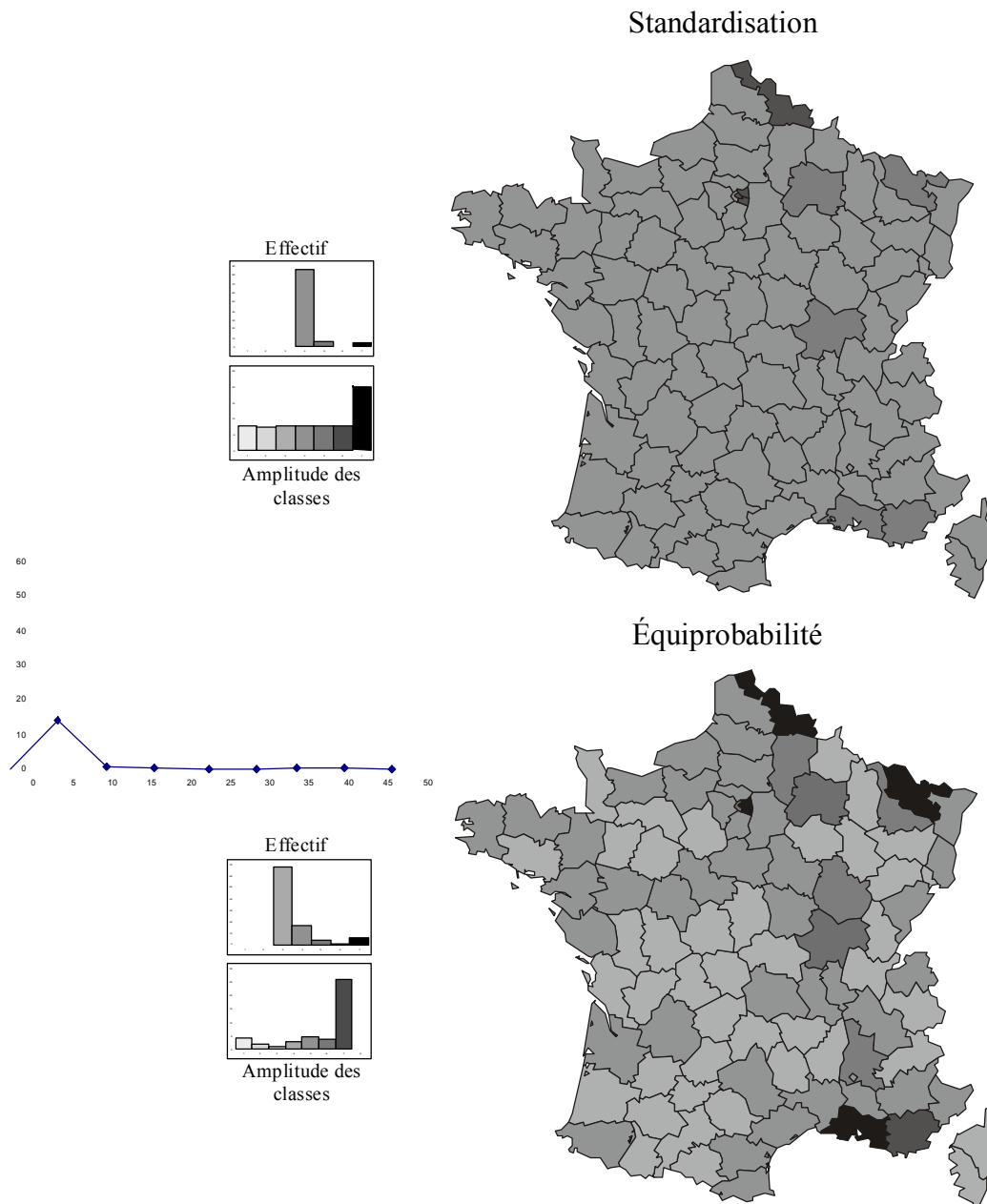


Figure 11 : Discretisation selon une hypothèse de normalité de la distribution

### Maximisation des variances interclasses et minimisation des variances intraclasses

Cette méthode – dite de Jenks – est fondée sur la décomposition de la variance :

$$Var_{totale} = Var_{intra} + Var_{inter} \text{ où la } Var_{intra} = \sum Var_{classe\ i} \text{ avec } \{i = 1 ; 2 \dots ; Nbcl\}$$

Les bornes retenues sont celles qui vérifient l'hypothèse où les  $N$  individus sont regroupés dans  $N_{cl}$  classes pour lesquelles la somme des variances intraclasses est minimale et celle interclasses maximale. Autrement dit, les individus au sein d'une même classe sont le moins

dispersés possible (homogénéité) tandis que les classes sont les plus éloignées les unes des autres (hétérogénéité).

Cette méthode donne généralement de bons résultats, les calculs peuvent être long avec un nombre important d'individus (processus récursif). Lorsque la distribution est trop hétérogène la méthode de Jenks n'est pas conseillée.

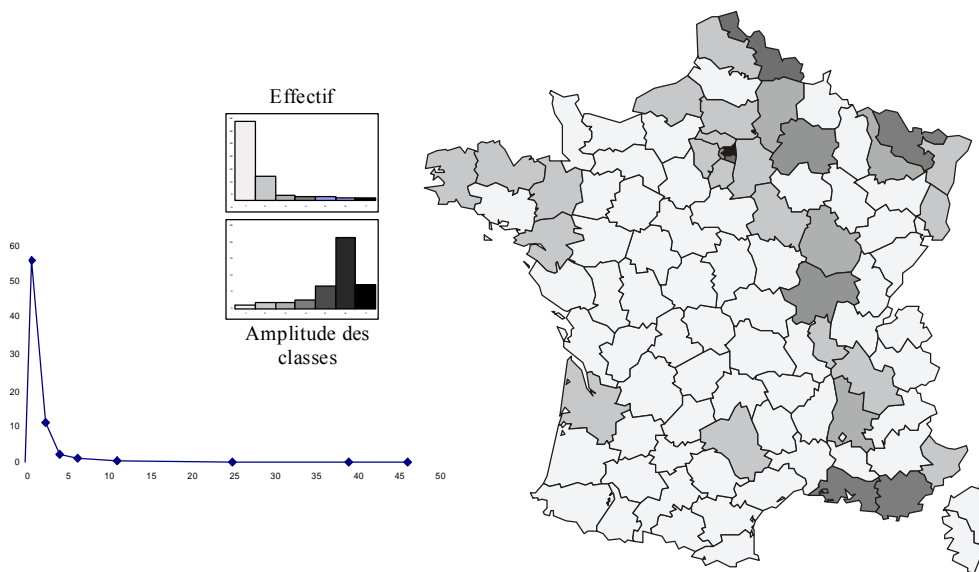


Figure 11 : Discretisation selon la décomposition de la variance

### Seuils issus du tableau de statistique descriptive

Cette méthode fixe les bornes selon les quantiles retenus lors de la description de la variable, généralement les déciles 1 et 9 et les quartiles 1, 2 (médiane) et 3. Elle permet de faire ressortir les 10% des individus au score les plus faibles et les plus forts (les « extrêmes ») et de regrouper les autres par ensemble de 25% de l'effectif. C'est une méthode relativement universelle qui fonctionne avec toutes les formes de distribution.

## Statistique descriptive

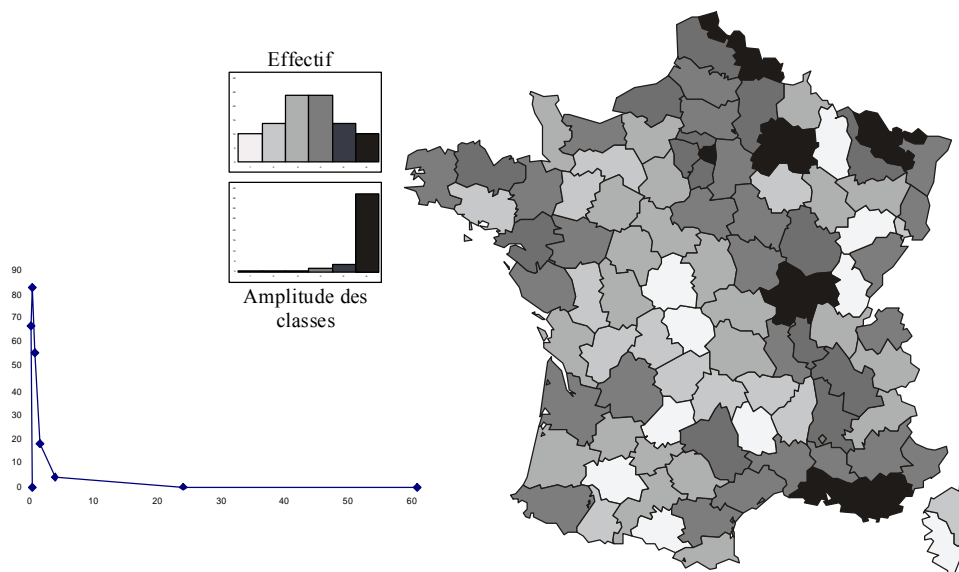


Figure 12 : Seuils issus de la statistique descriptive

Il existe bien d'autres méthodes de seuillage, nous n'avons présenté que les plus couramment utilisées. Les cartes obtenues soulignent l'importance du choix de la méthode de discrétisation. On peut constater que pour des distributions normales quelconques plus ou moins asymétriques toutes ne conviennent pas. Le choix de l'utilisateur sera alors fondé sur la lisibilité de la carte et le respect de la forme de la distribution de la série statistique. Quoi qu'il en soit, la carte doit toujours être accompagnée d'une légende présentant les bornes des classes, les effectifs de celles-ci et le graphe de densité de fréquences de manière à ce que le choix retenu soit justifiable... il est si facile de « fausser » des statistiques !

Dans certains cas, soit parce que la forme de la distribution est plurimodale soit parce que l'utilisateur impose un traitement particulier, il est préférable de fixer manuellement les limites des classes. Par exemple, on pourrait être intéressé par la comparaison des seuils d'impôts avec des seuils respectant la forme de la distribution des revenus. Il est alors impératif que ce choix soit expliqué.

### 5.2. Les cartes de résidus (exemple d'une régression linéaire)

Nous présentons ici la carte des résidus à une droite de régression de manière à souligner la double gradation des couleurs à respecter et la présentation particulière de la légende. Lorsque les valeurs à cartographier varient du négatif au positif, il est nécessaire de faire ressortir cette caractéristique en jouant sur une double gradation dans les teintes froides et dans les teintes chaudes en passant par une teinte « neutre » centrée sur les valeurs nulles. Celles-ci ne sont d'ailleurs pas toujours situées à mi-chemin entre le minimum et le maximum.

Il est conseillé de présenter en légende le graphe du nuage de points flanqué de la droite de régression. Rappelons que les résidus correspondent à la distance entre le point observé et le point théorique perpendiculairement à la variable explicative (cf. cours sur la régression et la corrélation). La figure 13 présente les cartes des variables « taux d'évolution moyen annuel 62-90 » (variable explicative, notée  $X$ ) et « taux d'évolution moyen annuel 75-90 » (variable à expliquer, notée  $Y$ ) utilisées pour l'exemple.

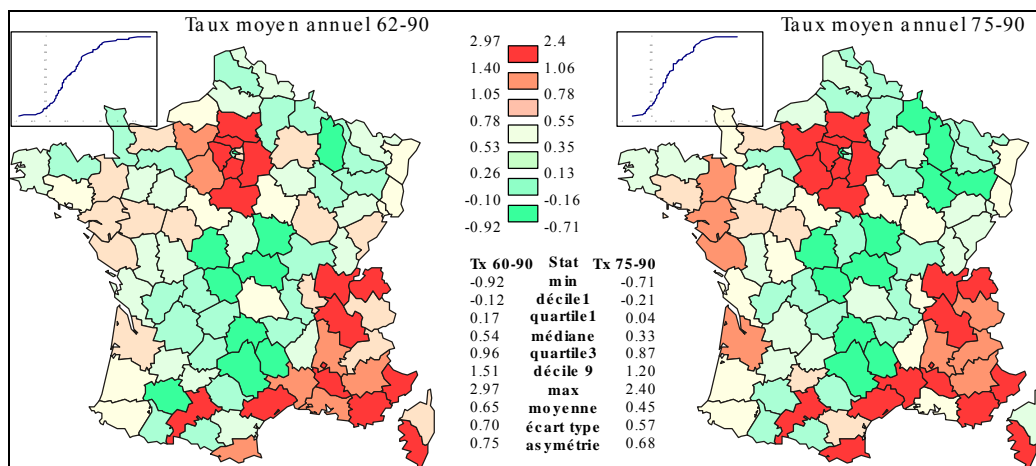


Figure 13 : Cartes des variables utilisées pour l'analyse des résidus

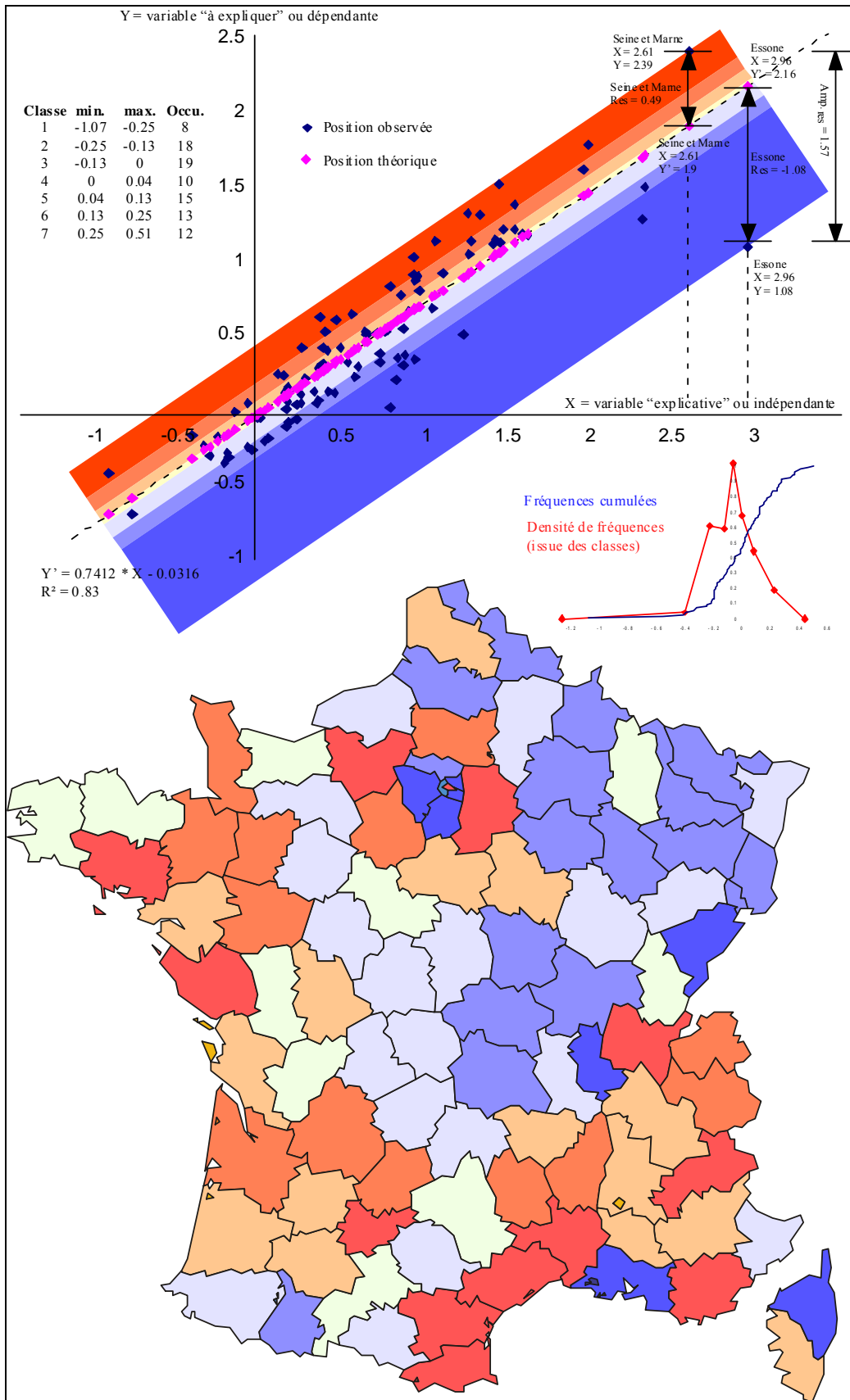


Figure 14 : Carte des résidus à une droite de régression

La figure 14 présente la cartographie des résidus. Les teintes *froides* correspondent aux départements où la valeur observée est inférieure à la valeur théorique. C'est-à-dire qu'il y a surestimation par le modèle. Inversement, les teintes *chaudes* présentent une sous-estimation par le modèle, les valeurs théoriques sont inférieures à celles observées. Notons que l'équation de la droite de régression peut être utilisée pour *modéliser* des données manquantes ou pour simuler un comportement. Il s'agit de ne pas confondre corrélation avec causalité...

## Conclusion

Les exemples traités dans ce cours ne constituent qu'un des aspects de la cartographie, celui couplé à l'analyse d'une variable. Il existe d'autres cartographies qui renvoient à l'analyse spatiale, elles sont présentées dans un support de cours particulier.