CARTHAGEO PRO

# Introduction aux méthodes de traitement des données géographiques.

Hélène Mathian
CNRS – Géographie-cités

## Le domaine de la Géomatique

- Domaine interdisciplinaire, réunissant des spécialistes
  - De l'espace (géographes, agronomes, géologues, hydrologues…)
  - De l'informatique et plus précisément des bases de données
  - Du traitement des données (statisticiens)
- La matière commune est *l'information géographique* qui s'organise en *Système d'Information Géographique*

## SIG et logiciel de SIG

- Système d'information géographique:
  *« Ensemble de données repérées dans l'espace, structuré de façon à pouvoir en extraire commodément des synthèses utiles à la décision » M.Didier*
- Le logiciel SIG, 5 fonctionnalités
  *Abstraction :* concevoir un modèle qui organise les données
  *Acquisition :* alimenter la géométrie , les attributs et leur relation
  *Archivage :* stocker les données
  *Analyse :* mettre en relation les attributs sémantiques, géométriques et topologiques
  *Affichage :* production cartographique, perception des relations spatiales, visualisation des données

## Analyse spatiale

thématique

- Analyse formalisée de la configuration et des propriétés de l'espace géographique, tel qu'il est produit et vécu par les sociétés humaines (Pumain, Saint-Julien, 1997)
- Ensemble de techniques et de modèles qui appliquent des structures formelles, généralement quantitatives, à des systèmes dans lesquels la principale variable évolue de façon significative à travers l'espace (Longley, Batty, 1996)
- Ensemble de fonctionnalités d'un SIG permettant d'analyser et de traiter une information géo-référencée (superposition, analyse de réseau, fonctions topologiques et géométriques, buffer..)

SIG

## Expliquer la localisation d'un phénomène en un lieu.
### Conjugaison de 3 dimensions

« Ecologique »
Caractéristiques humaines
Et physiques de ***ce*** lieu

Spatiale
Interactions avec
les ***autres*** lieux

localisation d'un
phénomène
en un lieu donné

*Composantes verticales*

*Composante horizontale*

Historique
Héritage, succession
d'événements en ***ce*** lieu

*D'après Durand-Dastès (Géopoint 1990) et Pumain, Saint-Julien (L'analyse spatiale, 1997, cursus, Colin)*

## Différentes formalisations possibles pour un même objectif

- décrire et comprendre l'évolution 1982-2000 de la distribution de la population au niveau communal dans la région de Montpellier ;
- proposer des prévisions pour 2020

**approche statistique**

**modèle dynamique de type logistique**

**Micro-simulation**

## Les méthodes statistiques au service de l'analyse spatiale

Ensemble de techniques et méthodes statistiques pour **décrire** et **expliquer** les répartitions spatiales.

- Approches exploratoires (Tukey)
- Approches descriptives (analyse des données)
- Approches explicatives (modèles linéaires, log-linéaires…)
- Approche de généralisation (lissage)

## L'Analyse Statistique de données spatiales

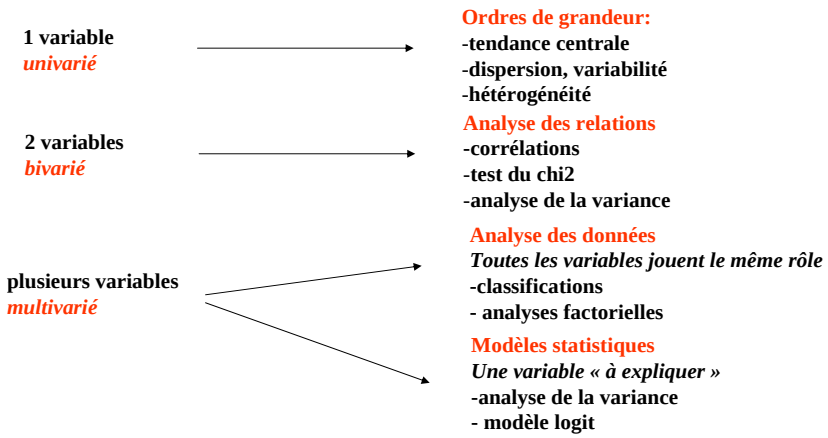
A donné lieu à 3 types de développements méthodologiques:

- l'analyse géostatistique
- l'économétrie spatiale
- l'analyse statistique spatiale et/ou analyse exploratoire des données spatiales (spatial data mining).

## Objectifs et moyens de l'analyse spatiale

- Objectifs: décrire et expliquer une organisation spatiale
  - –Analyse des localisations et structures
  - –Analyse des facteurs explicatifs
- Moyens: enchaînement de méthodes
  - –Identification et description de structures spatiales
  - –Tester la pertinence d'un modèle spatial
  - –Simuler un processus spatial

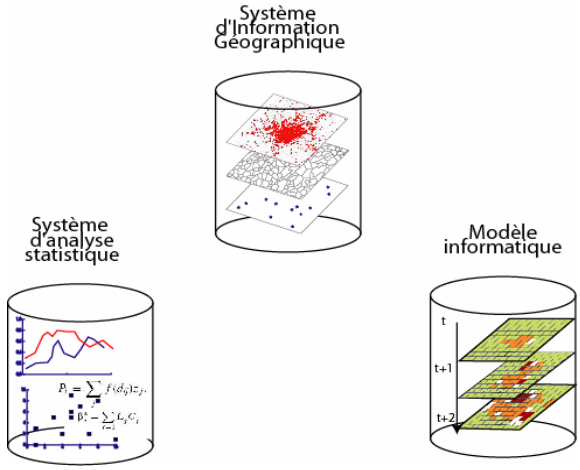
## Les méthodes statistiques
### dans une démarche d'analyse spatiale pour décrire et expliquer les répartitions spatiales.



## Le couplage SIG et outils de modélisation « thématiques »

- quelles formes de couplage ?

- le SIG en position centrale ou le SIG un outil parmi beaucoup d'autres ?




## Le traitement univarié de l'information

PRÉALABLEMENT À LA CONSTRUCTION DE LA CARTE

- **chercher l'information et la valider**
- **traiter l'information**
- **définir le mode de représentation le plus approprié**

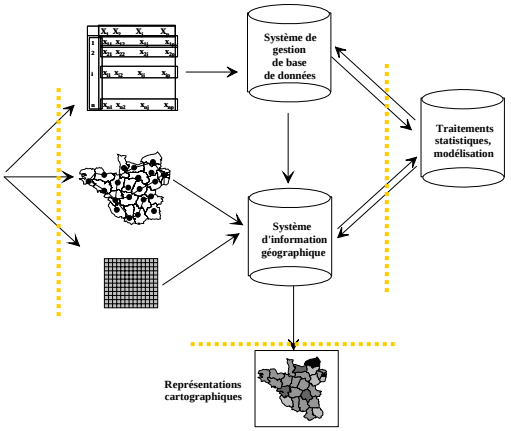
## De l'information à la représentation cartographique

La collecte des données et leurs traitements constituent des phases très importantes.

Des données fausses déguisées en jolie carte…

- Pourquoi ? *Quel est le but ?*
- Quoi ? *Quelle description ? Quelles mesures ?*
- Comment et où ? Quand et qui ?

## De la réalité au système d'information: 3 grandes phases de modélisations

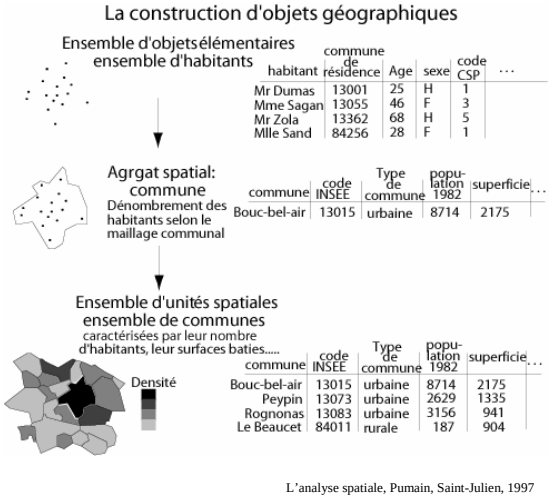
## Les données

- L'ensemble des *observations* de n *individus* décrits par p *variables* se présente sous la forme d'un tableau individus/caractères qui constitue les

| N° étudiant | Sexe | Âge | Résultat test | moyenne concours | Code INSEE | NOM |
|---|---|---|---|---|---|---|
| 99231 | F | 26 | Fort | 20.14 | 75119 | Paris 19e Arrondissement |
| 98123 | F | 31 | Faible | 38.82 | 92077 | VILLE-D'AVRAY |
| 99113 | F | 31 | Fort | 18.62 | 94022 | CHOISY-LE-ROI |
| 87133 | H | 29 | Moyen | 24.61 | 94058 | LE PERREUX-SUR-MARNE |
| 98328 | H | 29 | Fort | 36.40 | 94076 | VILLEJUIF |
| 99747 | F | 28 | Faible | 23.08 | 94056 | PERIGNY |
| 87621 | F | 26 | Faible | 45.02 | 94079 | VILLIERS-SUR-MARNE |
| 89821 | H | 28 | Faible | 13.19 | 94019 | CHENNEVIERES-SUR-MARNE |
| 98221 | F | 27 | Moyen | 17.97 | 93006 | BAGNOLET |
| 97321 | H | 29 | Moyen | 44.20 | 94081 | VITRY-SUR-SEINE |
| 98632 | H | 31 | Faible | 27.84 | 92004 | ASNIERES-SUR-SEINE |
| 99713 | F | 30 | Moyen | 10.59 | 94034 | FRESNES |
| 98263 | H | 28 | Faible | 32.02 | 75105 | Paris 5e Arrondissement |
| 99732 | F | 27 | Fort | 41.49 | 92012 | BOULOGNE-BILLANCOURT |

| | $X_1$ | $X_2$ | $X_j$ | $X_p$ |
|---|---|---|---|---|
| 1 | $x_{11}$ | $x_{12}$ | $x_{1j}$ | $x_{1p}$ |
| 2 | $x_{21}$ | $x_{22}$ | $x_{2j}$ | $x_{2p}$ |
| i | $x_{i1}$ | $x_{i2}$ | $x_{ij}$ | $x_{ip}$ |
| n | $x_{n1}$ | $x_{n2}$ | $x_{nj}$ | $x_{np}$ |

## Des données individuelles aux tableaux d'information géographique




L'analyse spatiale, Pumain, Saint-Julien, 1997

## La nature des objets

- 1. le niveau d'observation des objets :

- niveau *micro-géographique* :

entités *élémentaires* : individus, arbres, galets, pixels …

- niveau *meso-géographique* :

entités agrégées: quartiers, villes, départements, forêts…

- niveau *macro-géographique* :

ensemble d'entités agrégées: ville, système de ville, bassin hydrographique…

NB: Le choix d'un niveau d'observation et d'analyse dépend :

- du questionnement;
- de contraintes techniques et d'accès aux données

## Les variables
## (1) point de vue technique : variables quantitatives et qualitatives

Tableau : exemples de variables quantitatives et qualitatives pour des objets de différente nature

| ***Objets*** | ***variables quantitatives*** | ***variables qualitatives*** |
|---|---|---|
| individu | Age, revenu, distance à l'école la plus proche | Sexe, catégorie sociale |
| tronçon de rivière | quantité de sédiments transportés, teneur en produit toxique, pente, surface du bassin | type de couverture végétale |
| parcelle agricole | surface, production à l'ha | utilisation du sol, irrigué ou non, utilisation d'insecticide |
| commune | nombre d'habitants, surface, budget | statut administratif, couleur politique de la mairie |

## Les variables
## (2) point de vue conceptuel
## variables élémentaires et variables contextuelles

| *Objets* | *variables élémentaires* | *variables contextuelles* |
|---|---|---|
| individu | Age, revenu, catégorie sociale | nombre d'habitants, équipement scolaire, dans la commune de résidence, |
| tronçon de rivière | quantité de sédiments transportés, teneur en produit toxique, pente, | type de couverture végétale des versants, surface du bassin, caractéristiques amont du point de relevé |
| commune | Profil social, nombre d'habitants, statut administratif, existence d'une école, budget | caractéristiques du syndicat intercommunal auquel appartient la commune |

## Les variables
## (3) variables directes ou résultant d'un processus d'agrégation

*entités élémentaires / entités agrégées*

exemples : individus/communes ; communes/département ; buissons de buis/parcelle ; arbre/forêt

*. mesures déduites du niveau des entités élémentaires:*

. dénombrement (nombre de buissons, nombre total d'individus, nombre de personnes de telle catégorie sociale)

. moyenne (circonférence moyenne, revenu moyen)

. mode (espèce la plus représentée, idem CS)

. rapport (nombre buissons atteints/surface de la parcelle, nombre d'habitant/surface commune, taux de migration).

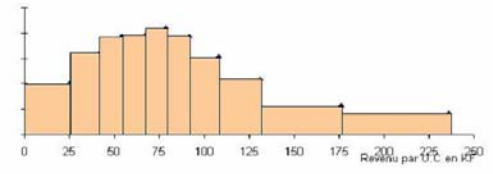
*. mesures directes :*

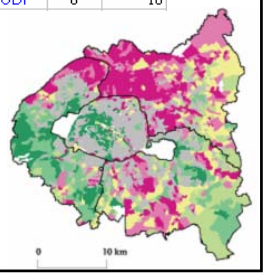
. surface, distance, date (ouverture d'une route, début ou fin d'un syndicat inter-communal)

# Un jeu … de données

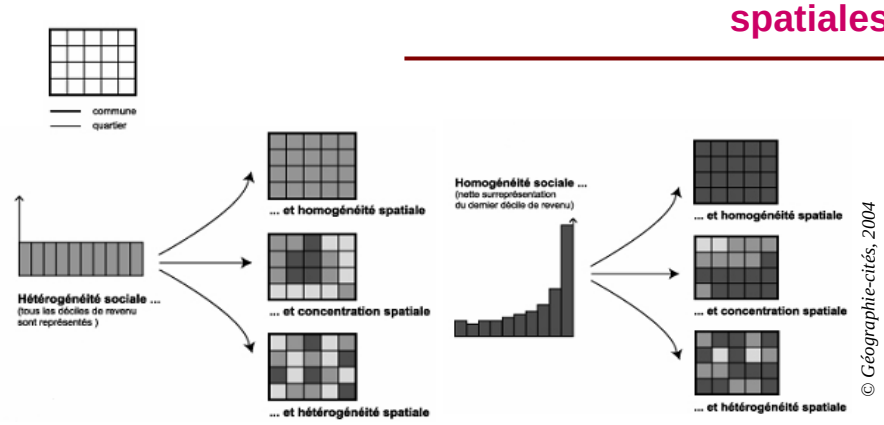
## Distribution statistique et distribution spatiale

| CTOT | NOM | PSDC99 | densité | tx82 99 | emp/act | %chômeurs | Revenu | Muni 2001 | Type Com. | nb crèches |
|---|---|---|---|---|---|---|---|---|---|---|
| 75101 | Paris 1er Arr. | 16888 | 7573 | -0.54 | 6.12 | 10.52 | 167906 | RPR | 2 | 3 |
| 75102 | Paris 2e Arr. | 19585 | 20191 | -0.47 | 4.83 | 13.34 | 109396 | GP | 5 | 2 |
| 75103 | Paris 3e Arr. | 34248 | 23458 | -0.31 | 1.51 | 11.67 | 134496 | GP | 2 | 4 |
| 92002 | ANTONY | 59855 | 6139 | 0.54 | 0.68 | 7.92 | 111026 | RPR | 2 | 6 |
| 92004 | ASNIERES-SUR-SEINE | 75837 | 13739 | 0.38 | 0.54 | 11.62 | 96344 | RPR | 5 | 13 |
| 92007 | BAGNEUX | 37252 | 9130 | -0.47 | 0.76 | 13.44 | 75792 | PC | 6 | 6 |
| 92009 | BOIS-COLOMBES | 23885 | 12571 | 0.03 | 0.40 | 9.93 | 103931 | RPF | 3 | 6 |
| 93015 | COUBRON | 4612 | 1083 | 0.42 | 0.26 | 7.40 | 103528 | DVD | 3 | 3 |
| 93027 | LA COURNEUVE | 35310 | 4209 | 0.30 | 0.89 | 24.34 | 54706 | PC | 7 | 7 |
| 93029 | DRANCY | 62263 | 8280 | 0.20 | 0.44 | 17.20 | 68974 | UDF | 6 | 9 |
| 93030 | DUGNY | 8641 | 1959 | 0.13 | 0.37 | 17.00 | 58843 | RPR | 7 | 5 |
| 93031 | EPINAY-SUR-SEINE | 46409 | 10971 | -0.47 | 0.41 | 18.89 | 66413 | UDF | 6 | 10 |




# Exemple: Disparités sociales / Inscriptions spatiales






On cherche à évaluer si les lieux proches ont plus tendance à se ressembler que des lieux éloignés et à mesurer cette relation.

# Exemple:

moyenne 7.4

std 4.7

Légende

0
5
10
15
20

moyenne 7.4

std 6.2

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 5 | 10 | 10 |
| 0 | 0 | 5 | 10 | 10 |
| 5 | 5 | 5 | 10 | 10 |
| 5 | 5 | 10 | 15 | 15 |
| 5 | 5 | 10 | 15 | 15 |

Var.T 22.2
Var.L 4.4

| | | | | |
|---|---|---|---|---|
| 0 | 15 | 5 | 10 | 15 |
| 5 | 10 | 5 | 10 | 5 |
| 10 | 5 | 15 | 0 | 10 |
| 15 | 5 | 10 | 5 | 5 |
| 0 | 10 | 5 | 5 | 0 |

Var.T 22.2
Var.L 25.6

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 5 |
| 0 | 0 | 5 | 5 | 10 |
| 0 | 5 | 5 | 10 | 15 |
| 5 | 5 | 15 | 15 | 15 |
| 10 | 10 | 15 | 15 | 20 |

Var.T 38.2
Var.L 8.4

| | | | | |
|---|---|---|---|---|
| 5 | 15 | 0 | 15 | 5 |
| 10 | 0 | 5 | 5 | 10 |
| 0 | 5 | 0 | 15 | 0 |
| 15 | 5 | 15 | 10 | 0 |
| 5 | 10 | 15 | 0 | 20 |

Var.T 38.2
Var.L 52.0

# Les résultats aux municipales de 2001: une variable qualitative

| Partis 2001 | nb com | fréquence |
|---|---|---|
| DL | 7 | 4.90 |
| DVD | 10 | 6.99 |
| RPF | 4 | 2.80 |
| RPR | 37 | 25.87 |
| UDF | 15 | 10.49 |
| DVG | 4 | 2.80 |
| GP | 12 | 8.39 |
| MDC | 1 | 0.70 |
| PC | 29 | 20.28 |
| PRG | 2 | 1.40 |
| PS | 17 | 11.89 |
| Verts | 1 | 0.70 |
| SE | 4 | 2.80 |
| | | |
| Total | 143 | |

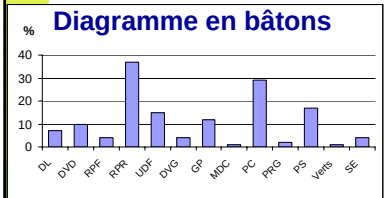
Agrégation des nomenclatures

| couleur | nb com | fréquence |
|---|---|---|
| droite | 73 | 51.05 |
| gauche | 66 | 46.15 |
| sans étiquett | 4 | 2.80 |

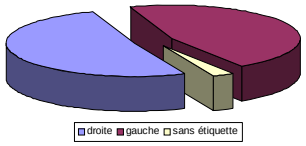
Tri à plat

| | Effectifs | Fréquences |
|---|---|---|
| $m_1$ | $n_1$ | $n_1/n$ |
| $m_2$ | $n_2$ | $n_2/n$ |
| $m_k$ | $n_k$ | $n_k/n$ |

## Diagramme en bâtons




## Diagramme circulaire

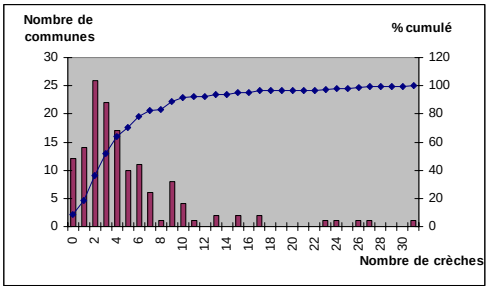


## Le nombre de crèches par commune: une variable quantitative discrète

### Tri à plat

| nb crèches | nb com. | % | % cumulé |
|---|---|---|---|
| 0 | 12 | 8.39 | 8.39 |
| 1 | 14 | 9.79 | 18.18 |
| 2 | 26 | 18.18 | 36.36 |
| 3 | 22 | 15.38 | 51.75 |
| 4 | 17 | 11.89 | 63.64 |
| 5 | 10 | 6.99 | 70.63 |
| 6 | 11 | 7.69 | 78.32 |
| 7 | 6 | 4.20 | 82.52 |
| 8 | 1 | 0.70 | 83.22 |
| 9 | 8 | 5.59 | 88.81 |
| 10 | 4 | 2.80 | 91.61 |
| 11 | 1 | 0.70 | 92.31 |
| 13 | 2 | 1.40 | 93.71 |
| 15 | 2 | 1.40 | 95.10 |
| 17 | 2 | 1.40 | 96.50 |
| 23 | 1 | 0.70 | 97.20 |
| 24 | 1 | 0.70 | 97.90 |
| 26 | 1 | 0.70 | 98.60 |
| 27 | 1 | 0.70 | 99.30 |
| 31 | 1 | 0.70 | 100.00 |
| Total | 143 | | |

### Diagramme en bâtons et courbe cumulée

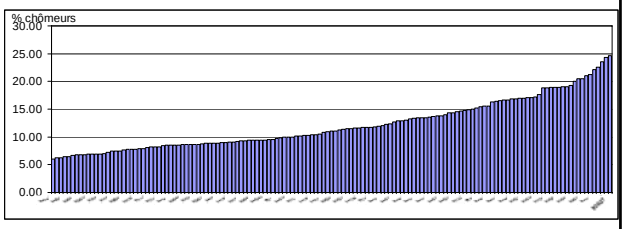



## Le pourcentage de chômeurs par commune: une variable quantitative continue

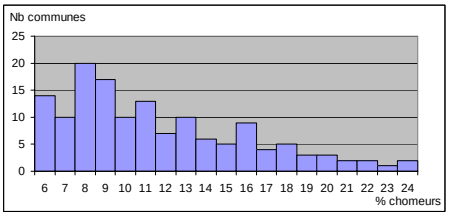
### Série ordonnée et classée

| NOM | % chômeurs | classe val |
|---|---|---|
| MARNES-LA-COQUETTE | 6.05 | 6 |
| RUNGIS | 6.25 | 6 |
| VAUCRESSON | 6.29 | 6 |
| MAROLLES-EN-BRIE | 6.45 | 6 |
| PERIGNY | 6.46 | 6 |
| BRY-SUR-MARNE | 6.68 | 6 |
| SAINT-CLOUD | 6.76 | 6 |
| BOURG-LA-REINE | 6.77 | 6 |
| VILLE-D'AVRAY | 6.78 | 6 |
| GARCHES | 6.85 | 6 |
| NOISEAU | 6.85 | 6 |
| MANDRES-LES-ROSES | 6.89 | 6 |
| SCEAUX | 6.89 | 6 |
| CHAVILLE | 6.97 | 6 |
| VILLECRESNES | 7.27 | 7 |
| COUBRON | 7.40 | 7 |
| FONTENAY-AUX-ROSES | 7.46 | 7 |
| CHATILLON | 7.46 | 7 |
| RUEIL-MALMAISON | 7.64 | 7 |
| SANTENY | 7.75 | 7 |
| GOURNAY-SUR-MARNE | 7.78 | 7 |
| VANVES | 7.83 | 7 |

### Représentation de la série « brute »

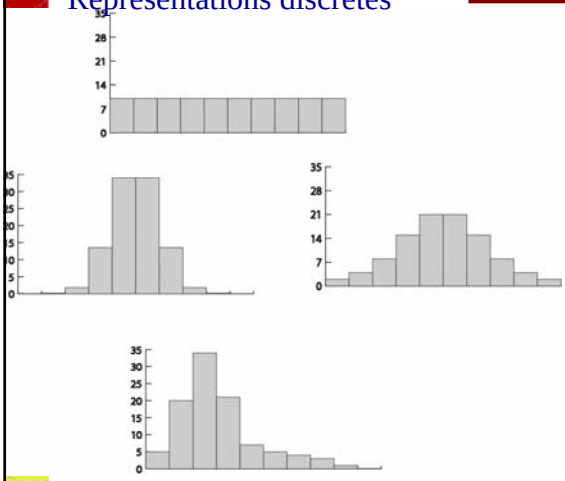



### Histogramme

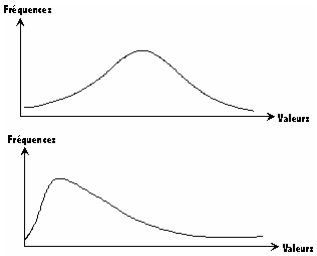


## Des distributions statistiques « types »

Représentations discrètes



Représentations continues

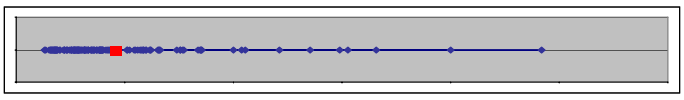






## Résumer les distributions: Les indicateurs de position
## 1- La moyenne

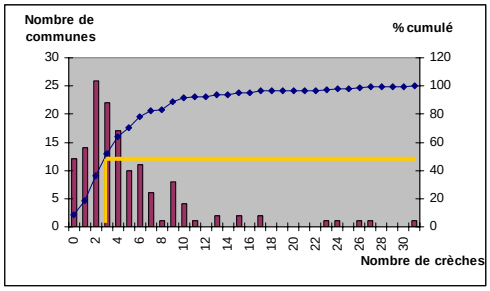
- La moyenne arithmétique: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$



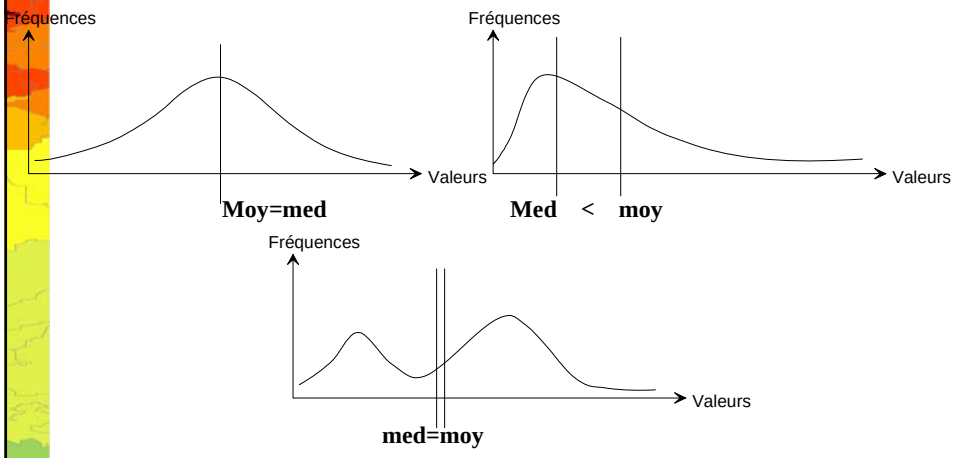
- La moyenne pondérée: $\overline{x} = \frac{\sum_{k=1}^{m} n_k x_k}{\sum_{k=1}^{m} n_k} = \frac{\sum_{k=1}^{m} n_k x_k}{n}$
- Propriétés: $\sum_i (x_i - \overline{x}) = 0$

## Résumer les distributions: les indicateurs de position
## 2- La médiane

- La médiane: Valeur qui partage la distribution d'une série d'observations en 2 parties égales.



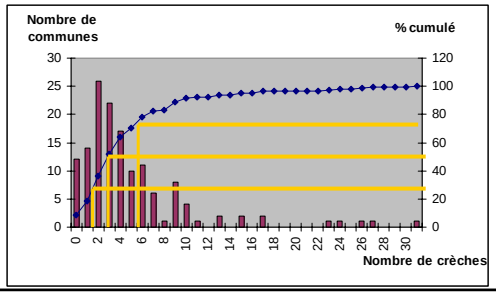

## Résumer les distributions: Comparaison des indicateurs de position



## Résumer les distributions: les indicateurs de position
## 3- Les quartiles

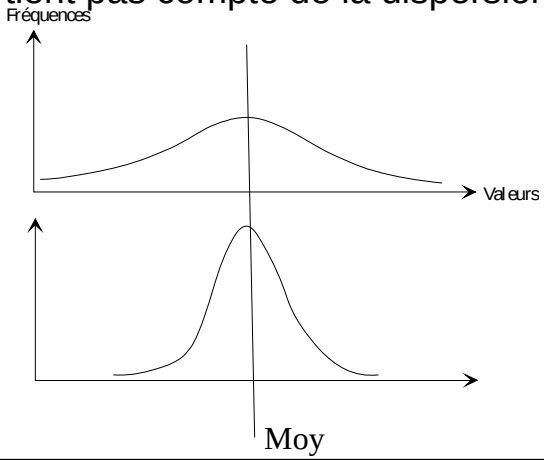
- Les quartiles: Valeurs (Q1,Q2,Q3) qui partagent la distribution d'une série d'observations en 4 parties égales.

  F(Q1)=0.25, F(Q2)=0.5, F(Q3)=0.75




## Résumer les distributions:
## Les indicateurs de dispersion

- La moyenne est un « pauvre » indicateur si l'on ne tient pas compte de la dispersion.



## Résumer les distributions: Les indicateurs de dispersion
### 1- Variance et écart-type

• La variance : *mesure de la dispersion moyenne des observations autour de la moyenne.*

$$V(X)=\frac{\sum_i (x_i-\bar{x})^2}{n}$$

• écart-type (*standard deviation*) $\boldsymbol{V(X)=S^2}$

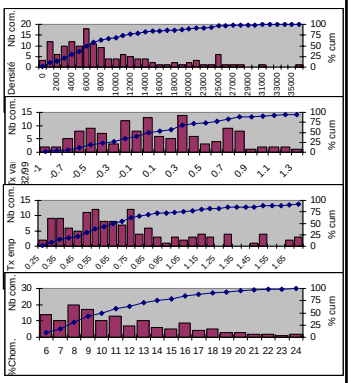
***S*** s'exprime dans l'unité de mesure de X

## Autres mesures de dispersion

- Etendue: *max-min*
- Ecart-moyen: $E.M=\frac{\sum_i |x_i-\bar{x}|}{n}$
- Ecart médian: $E.med=\frac{\sum_i |x_i-med|}{n}$
- Intervalle interquartile: *I.Q*=Q3-Q1
- Coefficient de variation: $cv=\frac{s}{\bar{x}}$

## Un ensemble d'indicateurs

| NOM | moy enne | écart-type | CV | min | Q1 | médiane | Q3 | max |
|---|---|---|---|---|---|---|---|---|
| densité | 9660 | 8477 | 0.88 | 321 | 4377 | 7033 | 12202 | 64827 |
| tx82_99 | 0.33 | 0.99 | 3.03 | -0.99 | -0.21 | 0.13 | 0.71 | 7.57 |
| emp/act | 1.02 | 1.16 | 1.14 | 0.26 | 0.53 | 0.71 | 1.09 | 8.59 |
| %chômeurs | 12.13 | 4.45 | 0.37 | 6.05 | 8.62 | 11.01 | 14.94 | 24.62 |
| Revenu moyen/UC en 1999 | 103900 | 45697 | 0.44 | 54213 | 75745 | 92455 | 113558 | 320943 |



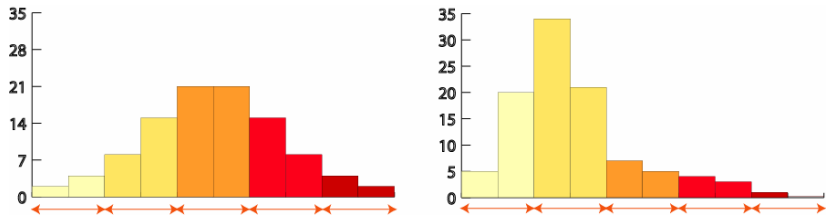

## Discrétiser pour mieux cartographier

- Cartographier une série quantitative nécessite de discrétiser les valeurs
- Discrétiser = découper une série en classes
- Grand compromis statistique - cartographie
  - –Résumer au mieux la distribution (vérité)
    Optimum: le plus grand nombre de classes
  - –Construire une carte efficace (lisibilité)
    Optimum: un faible nombre de classes et d'effectifs égaux.

## Les méthodes de discrétisation

- Classes d'amplitudes égales
- Classes basées sur la moyenne et l'écart-type
- Classes d'effectifs égaux (ou quantiles)
- Classes en progression géométrique
- Méthode des seuils naturels

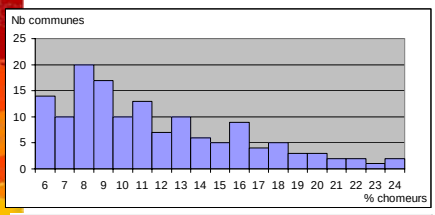
## Discrétisation en classes d'amplitudes égales

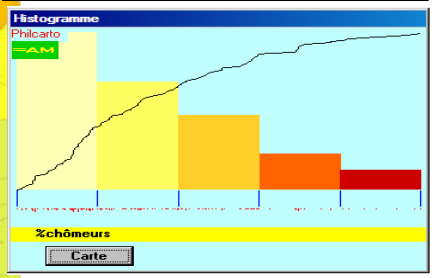


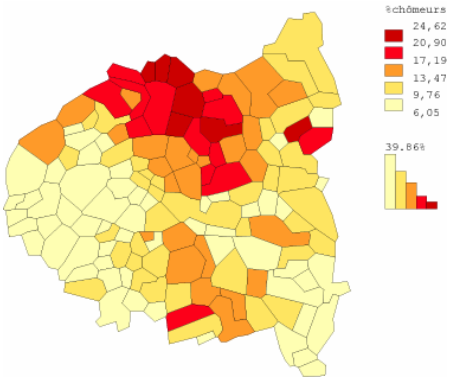

**+ Facile à construire, facile à lire, conserve la forme de la distribution**

**- Optimum carto pour distribution uniforme (rare)**
**-- pour les distributions asymétriques**

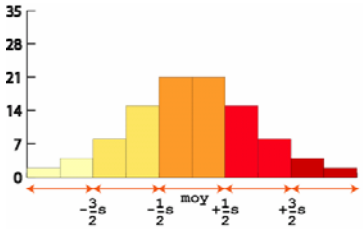
## Classes d'amplitudes égales

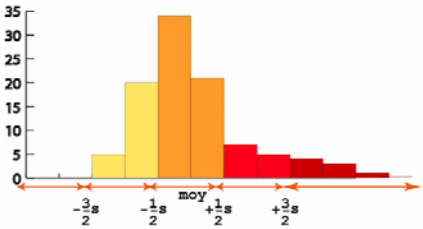


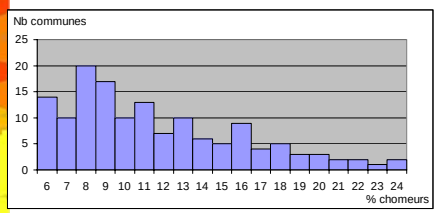

## Discrétisation basée sur la moyenne et l'écart-type

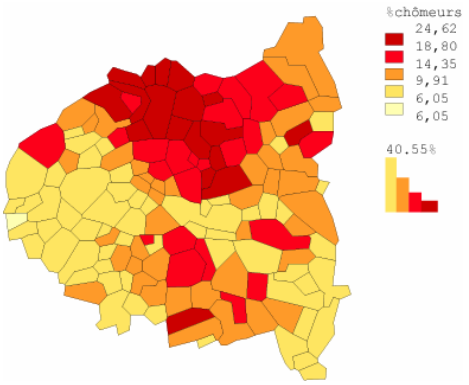


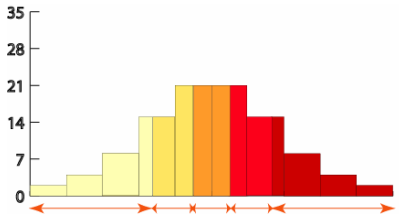

+ Fait référence aux caractéristiques de la distribution (moy, s)
+ permet la comparaison dans une unité commune
- Convient mal aux distributions asymétriques

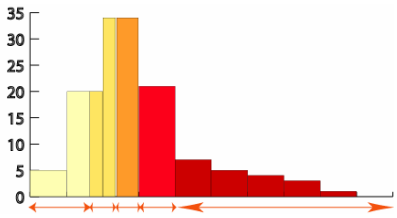
## Classes basées sur la moyenne et l'écart-type




## Discrétisation en classes d'effectifs égaux (méthodes des quantiles)



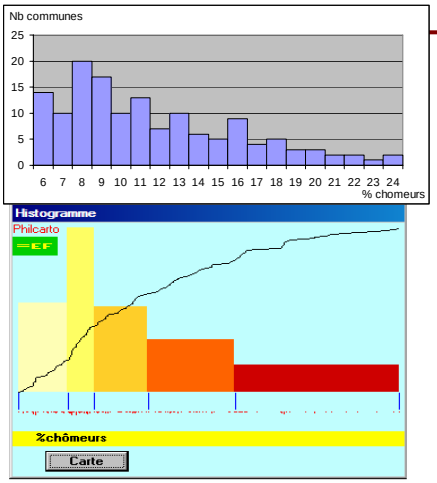



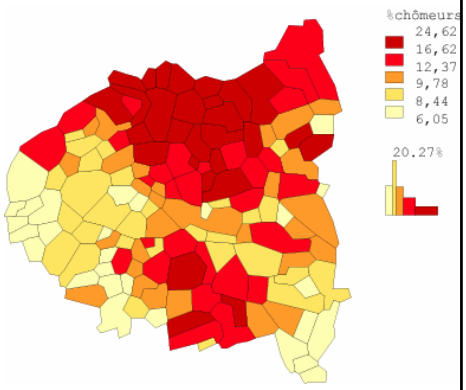

**+ Transmet une information maximale**

**+ Privilégie les rangs plus que les valeurs**

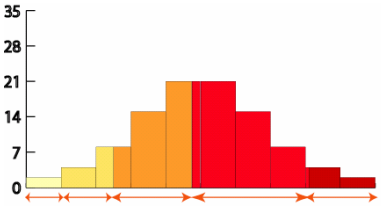
**- Transforme complètement la distribution**

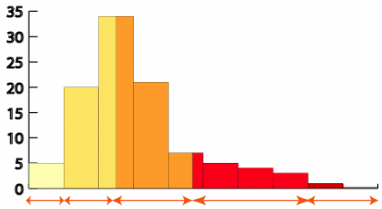
## Classes d'effectifs égaux



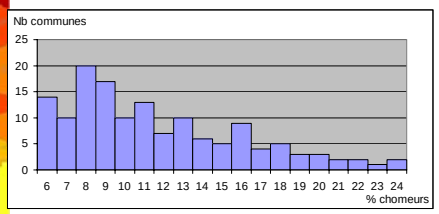

## Discrétisation en classes en progression géométrique

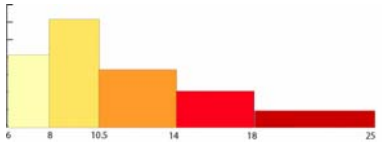


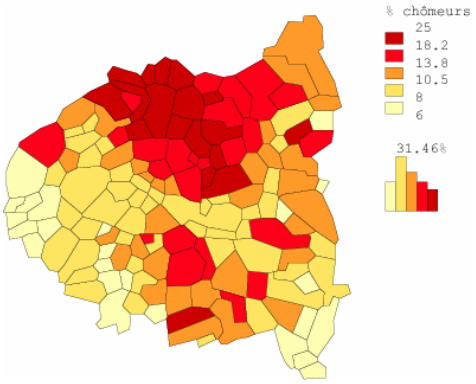




**+ Convient aux distributions asymétriques produites par exemple par des processus multiplicatifs**

## Classes en progression géométrique



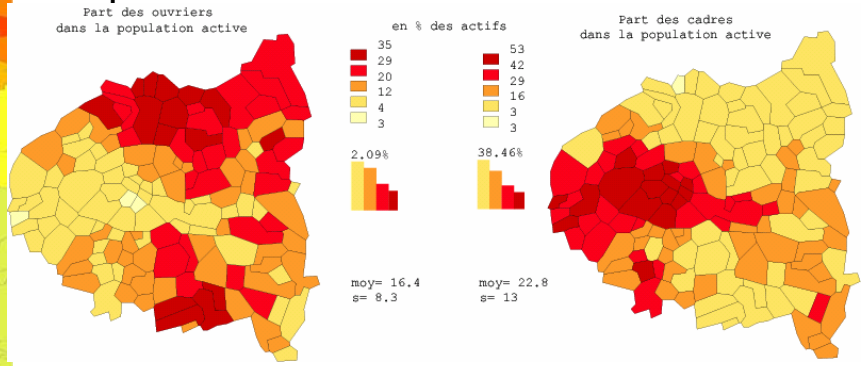

## Discrétiser pour comparer

- La comparaison est un objectif spécifique de la cartographie
- Comparaison des positions relatives ou des positions absolues ?
- Comparaison
  - Des distributions de 2 phénomènes sur un même espace
  - Dans le temps (cartographie d'une évolution)
  - Des distributions d'un même phénomène sur 2 espaces différents ou à 2 échelons différents.

## Comparaison des positions relatives

- 2 phénomènes sur un même espace:

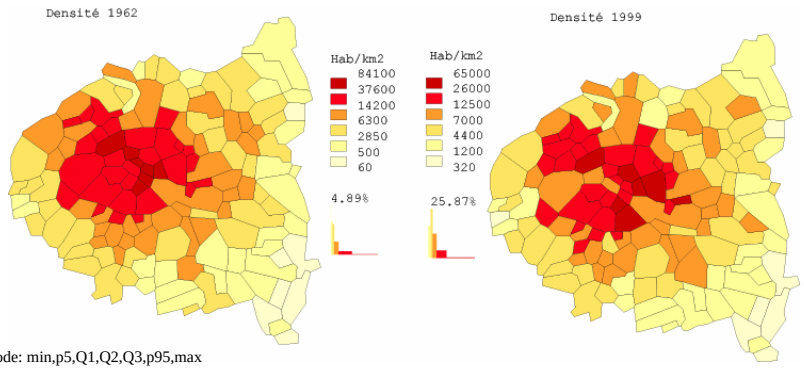
La part des cadres et des ouvriers.




## Comparaison des positions relatives

- Évolution d'un phénomène:

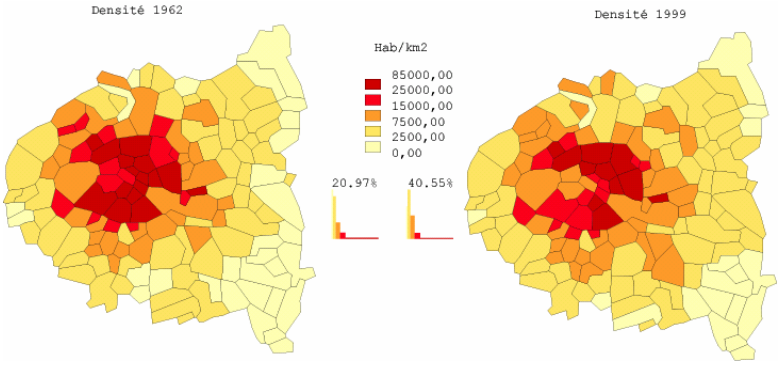
les densités communales entre 1962 et 1999




Méthode: min,p5,Q1,Q2,Q3,p95,max

## Comparaison des positions absolues

- Évolution d'un phénomène: les densités communales entre 1962 et 1999




## Et enfin… les problèmes que posent le territoire aux statistiques et à la cartographie

- Problèmes de comparabilité:
  - Effet de contour
  - Effet de bordure
  - Effet de définition

L'exemple du découpage en Iris:

•Homogénéité de définition sémantique

•Hétérogénéité de définition spatiale