UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

GÉOGRAPHIE-CITÉS
UMR 8504

Geographic Information

DELHI GIS-R School

9-12th April 2019

Hadrien Commenges & Paul Chapron

hadrien.commenges@univ-paris1.fr

paul.chapron@ign.fr

Definitions

Geographical Information

Quantitative information, localized in 1, 2, 3 or n dimensions. This information is addressed from itls localization point of view.

Geographical information types :

1. Geographical objects (volcanos, railways, forest, etc.)
2. Event occurences (fires, crimes, etc.)
3. Measure points (altitude, temperature, etc.)
4. « Statistics » (population, unemployment rate, etc.)
5. Interaction measures (flows, catchment area, etc.)

Definitions

The question of nature

The nature of the geographical information is independent from the geographical object, it has to be set by the analyst, according to the research question.

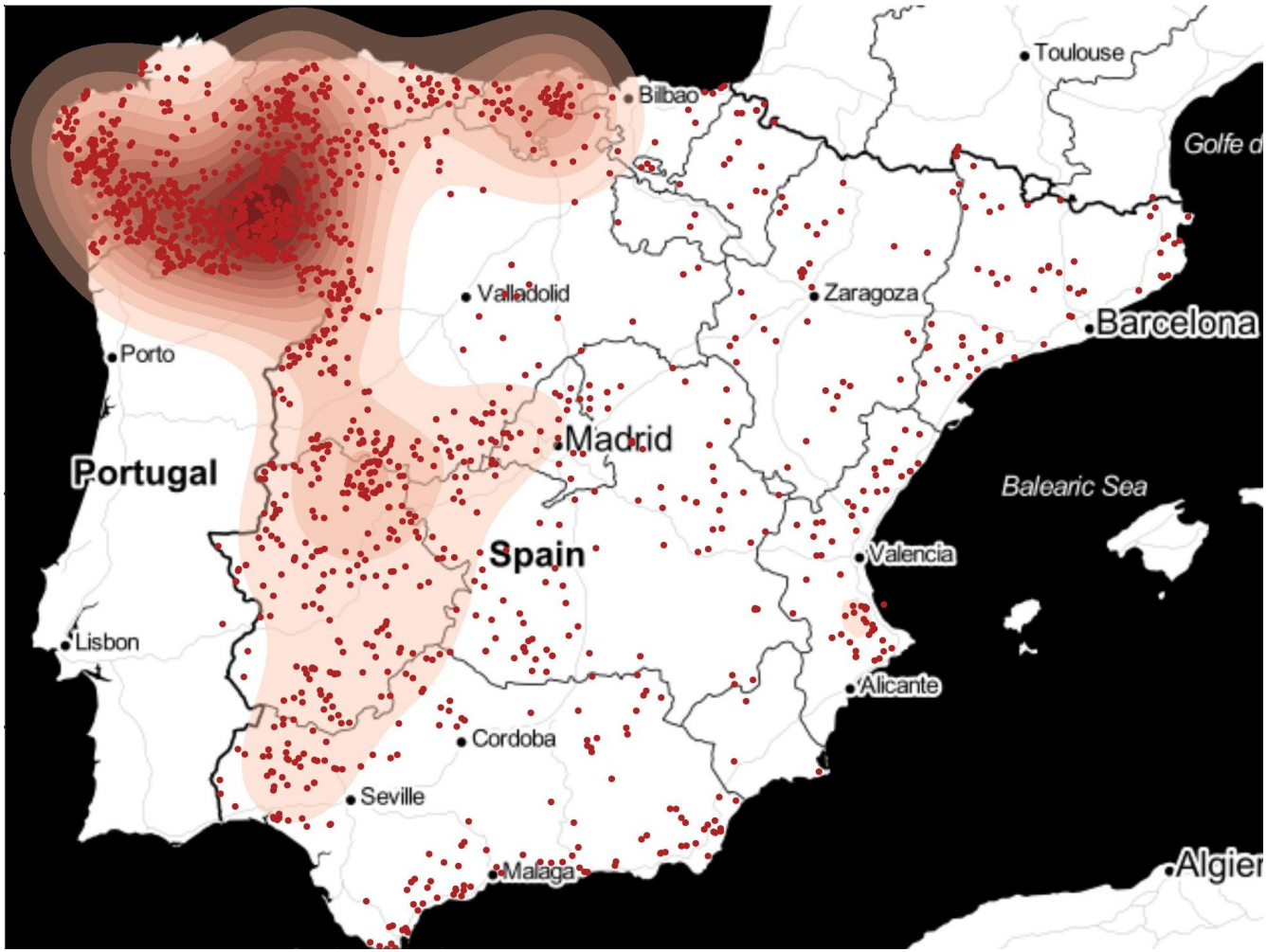
1. Dewellings point patterns (spatial object)
2. Dweelings sales (occurrences)
3. Dwellings prices (measure points)
4. Average price by district (« statistics »)

(1) Geographical Objects

Geographical objects come in three types : **points**, **lines** and **areas**.
Geographical data analysis focus on their **geometry** (e.g. length, morphology) and their **topology** (e.g. neighborhood , distance).

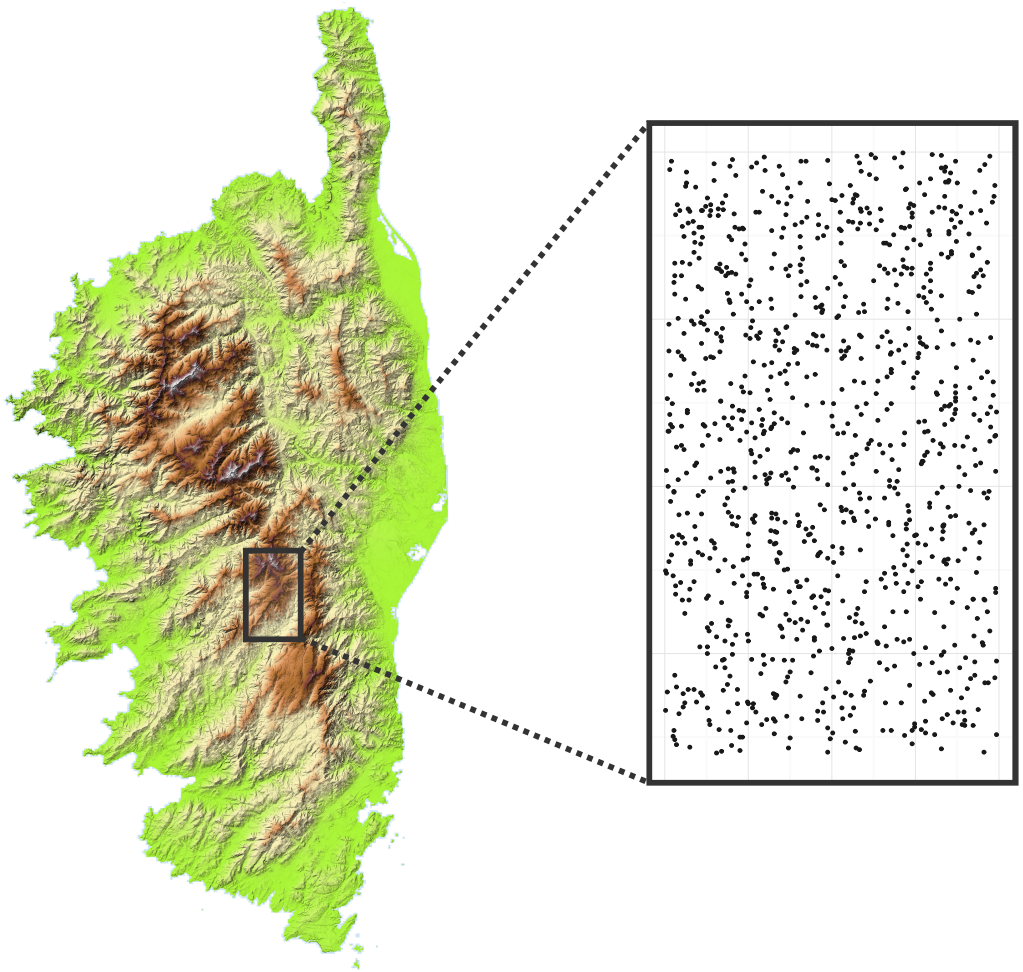
(2) Event occurence

Point data, **sampled or extensive**, whose localization is under study.
When it comes to model, localization is the **response variable**.



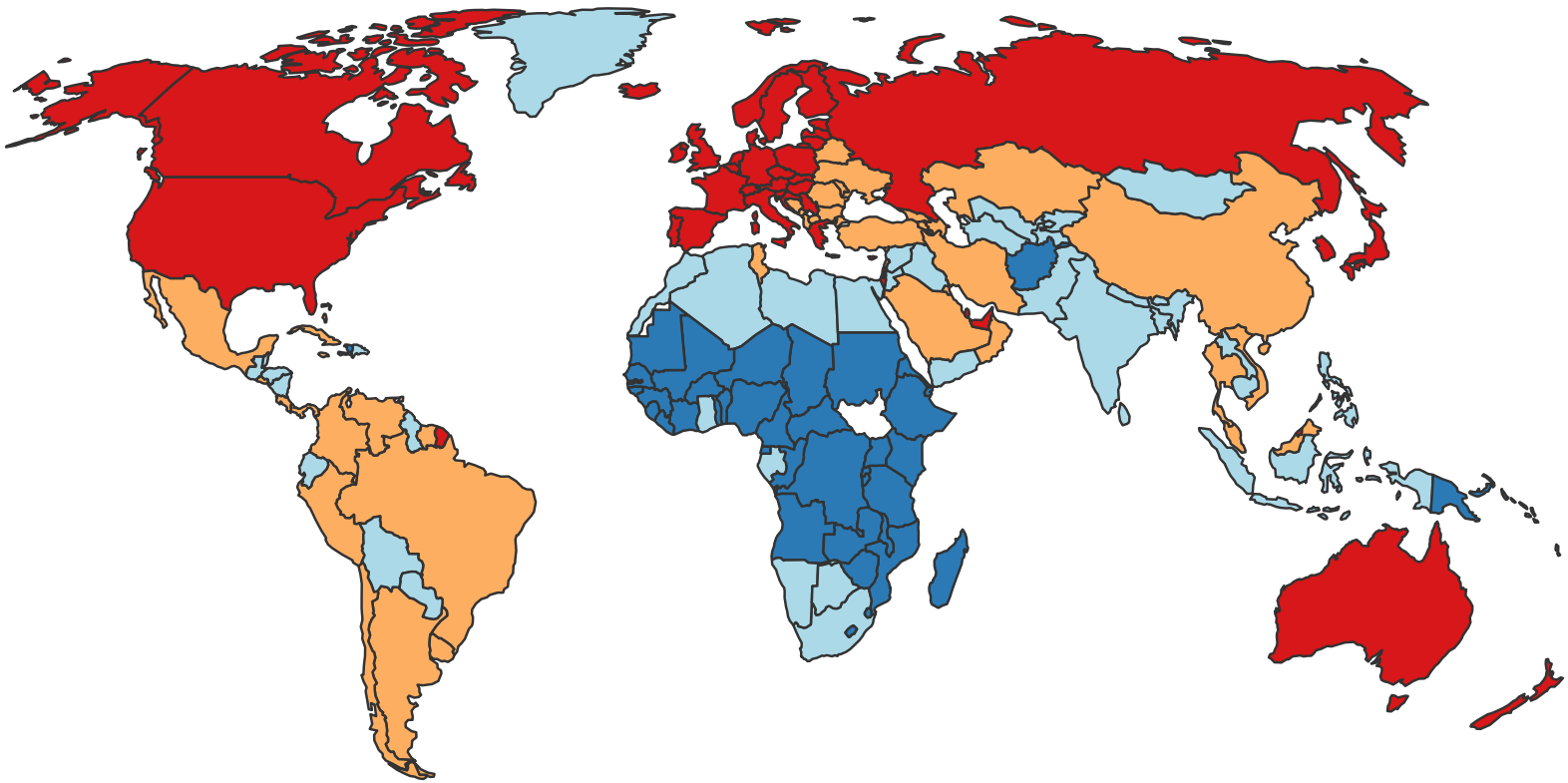
(3) Measure points

Point data, **sampled or extensive**, where a **value** is associated to each localization. Phenomenon under study is **the value variation according to the localization**.



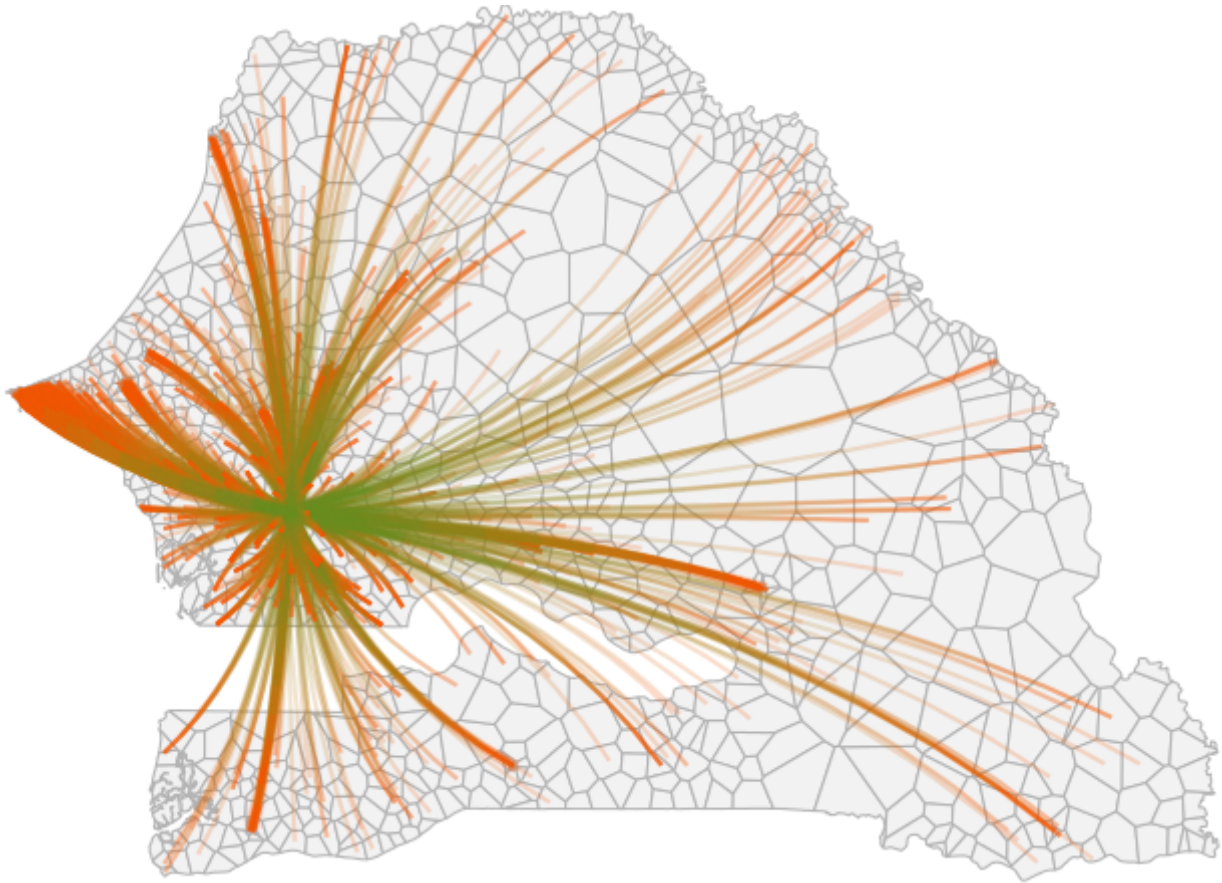
(4) Statistics

«Statistics», from *statista*, «state man» in italian.
Zonal extent variables created from census - **sampled or extensive** - for territorial management.
Such variables are attributes measured or computed **within spatial units**.

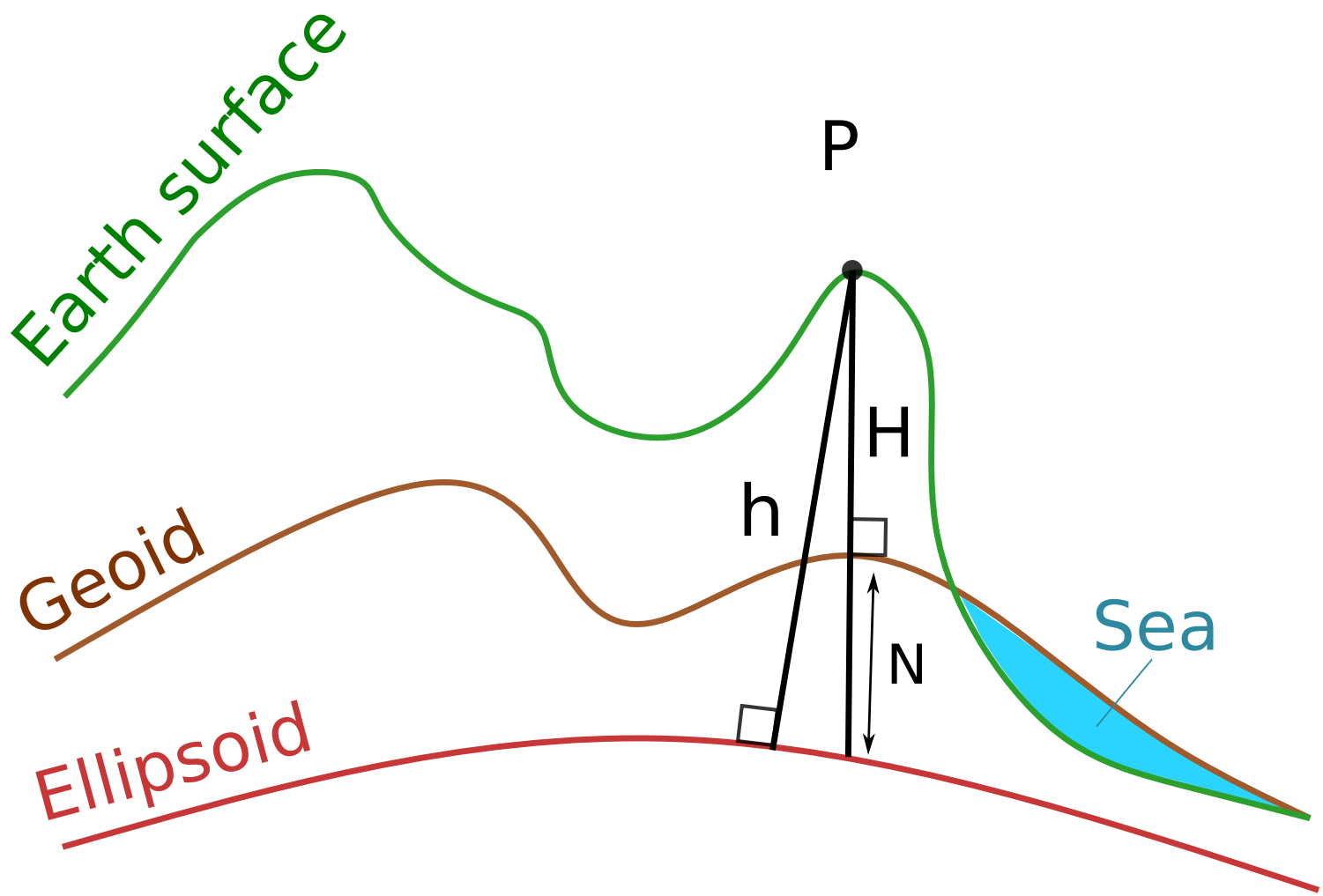


(5) Interactions

Interactions concern **geographical objects**, **occurrences** or **spatial units** and the **links** between them. The object under study is the **structure** and **dynamic** of these links (network analysis).



Coordinates, areas, distances



Source : ENSG, Les projections et référentiels cartographiques

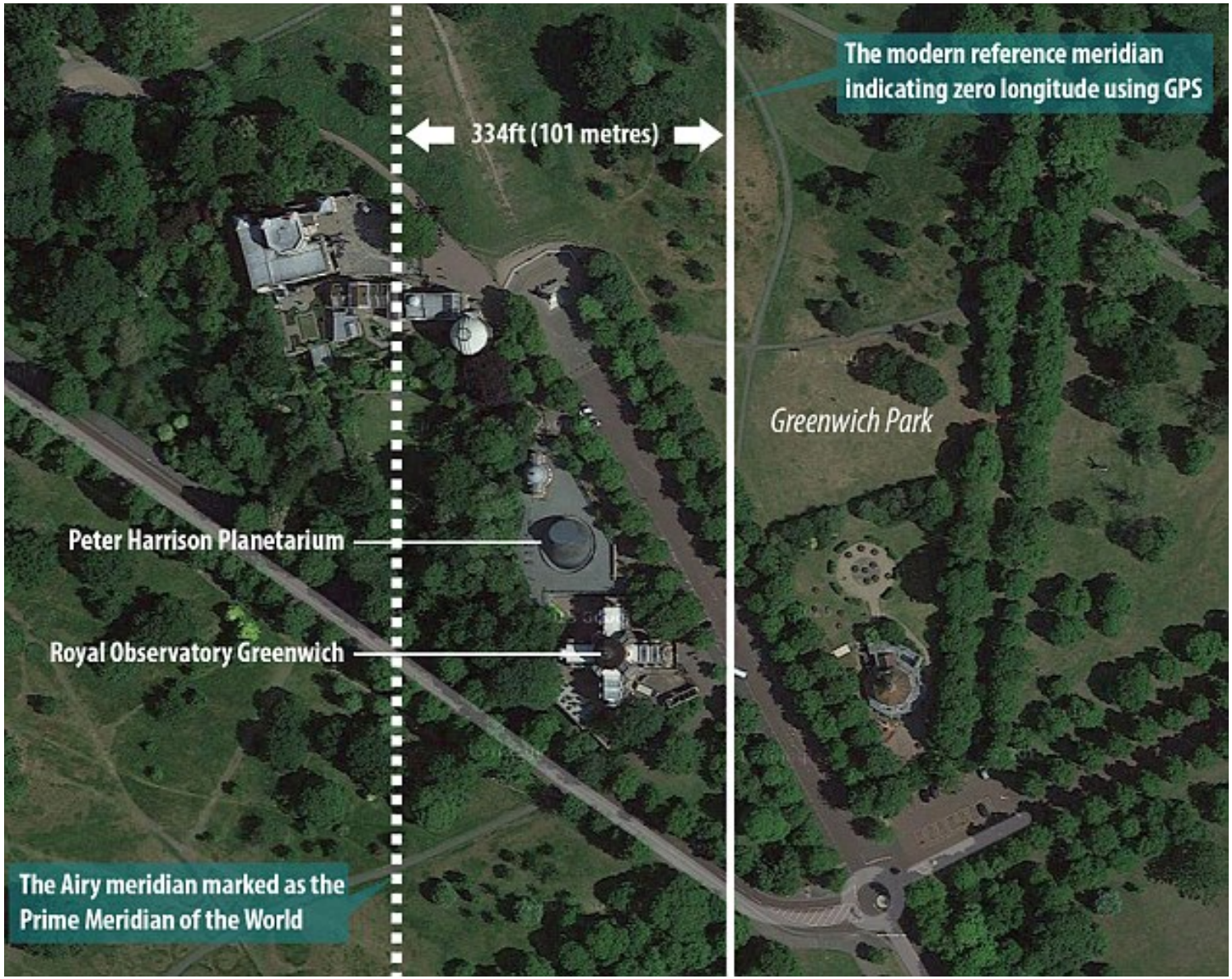
Coordinates, areas, distances

Geolocation of a spatial entity depends on :

- **Reference ellipsoid** : Clarke1880, Ellipsoide1909, IAG-GRS80
- **Geoid** : gravity field equipotential surface
- **Projection** : Mercator, Lambert, Mollweide, etc.

A **Geodetic system** (or datum) is the combination of these 3 elements (e.g. WGS84)

Coordinates, areas, distances



Coordinates, areas, distances

A sphere (globe) is a **non-developable** surface, i.e. cannot be represented as a plane (map) without **deformation**.

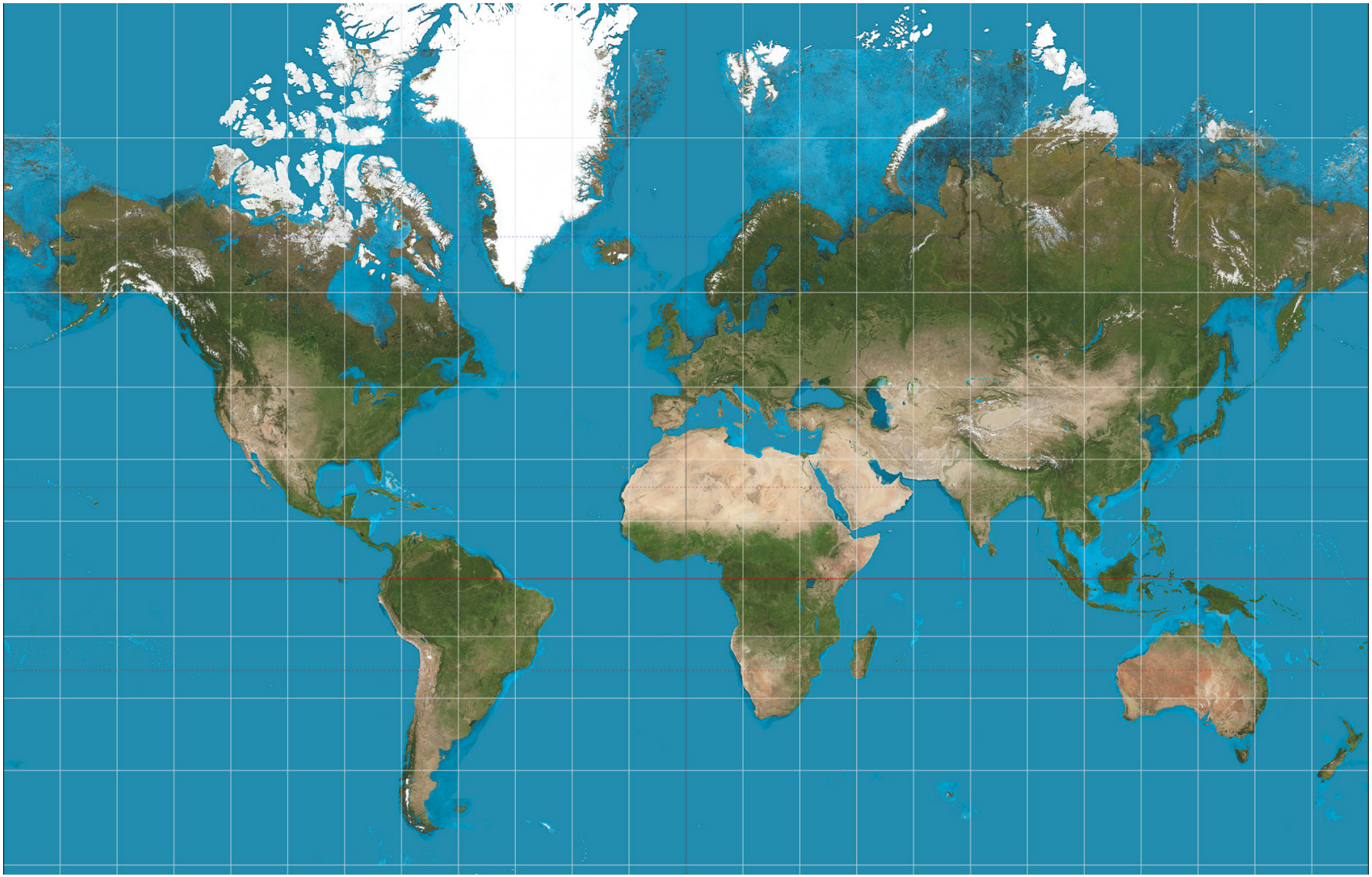
Some projections preserve some features

- **conformal** : conserve angles (shape)
- **equivalent** : conserve areas
- **equidistant** : conserve distances

Some others don't.

UTM (Universal Transverse Mercator, conformal) allows almost everywhere an acceptable projection.
What is the recommandation for India ? Specific ? Kalianpur 5 zones ?

Coordinates, areas, distances

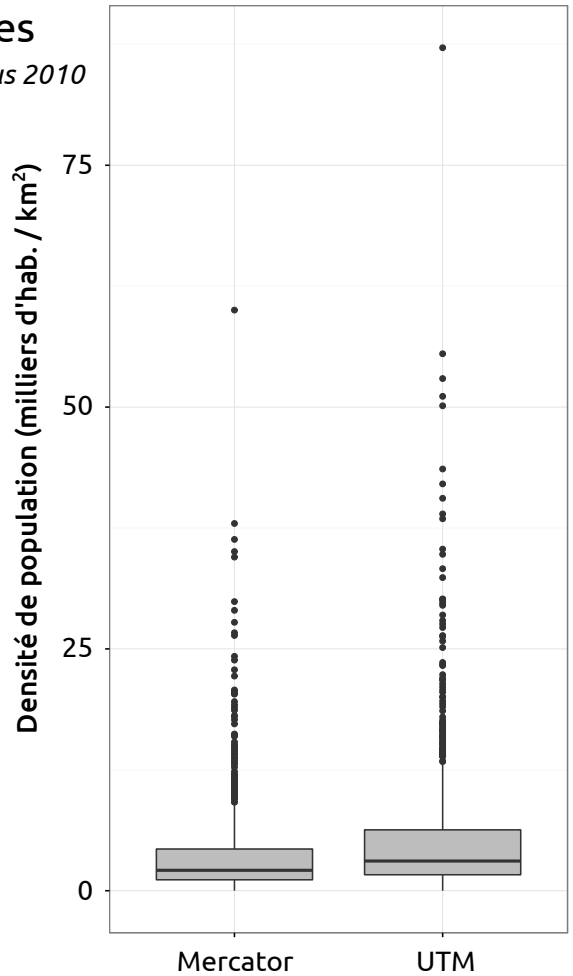


Source : Wikimedia, Mercator projection

Coordinates, areas, distances

Cape Town - Subplaces

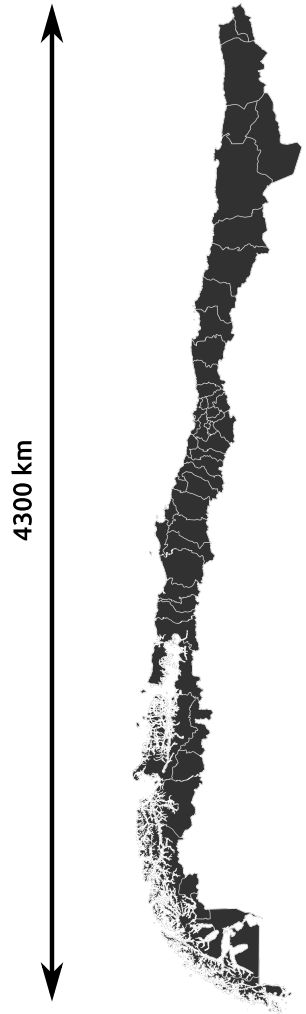
Source : Statistics South Africa, Census 2010



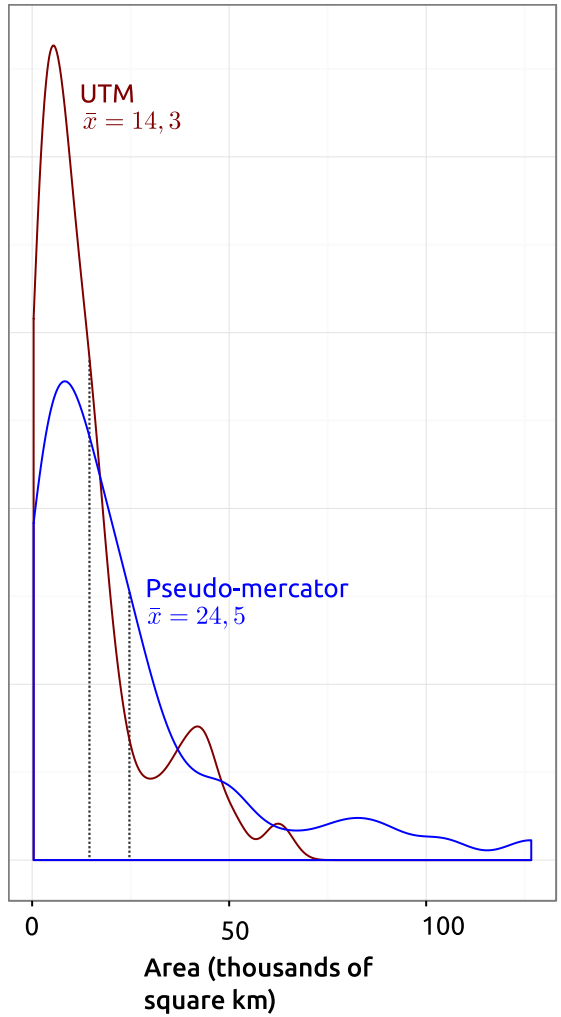
Med. 2111	**Med.** 3069
Moy. 3764	**Moy.** 5483

Paris : 22 000 hab./km²
Île-de-France : 1 000 hab./km²

Coordinates, areas, distances



Chile's provinces area



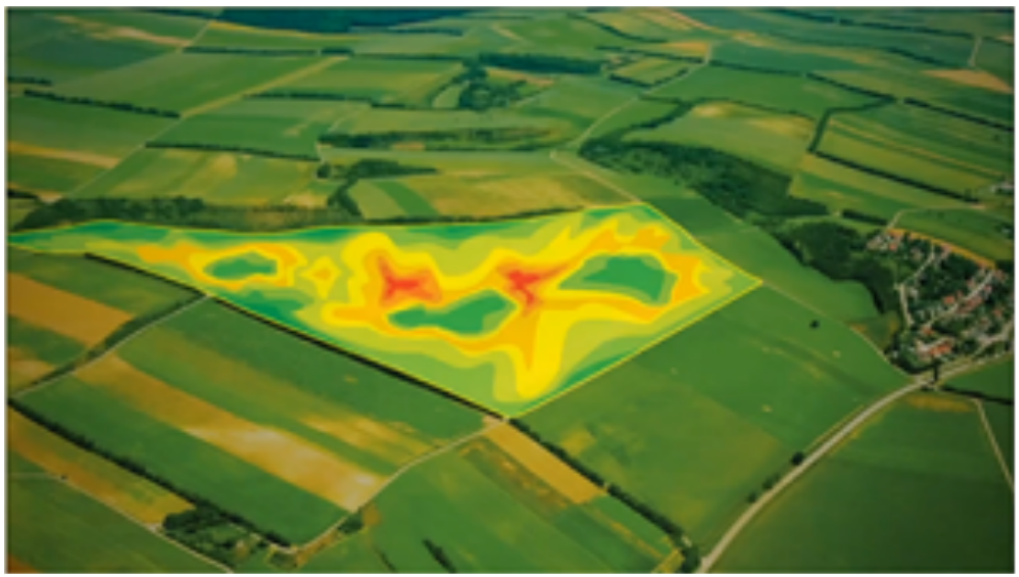
Coordinates, areas, distances

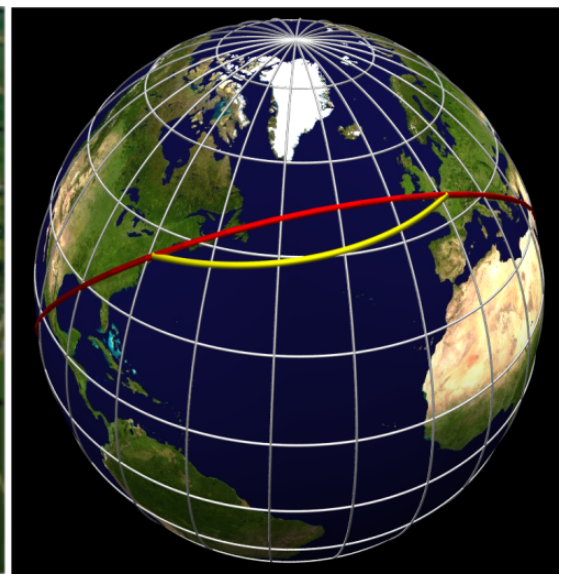
Basic precautions regarding projection :

- Density $\rightarrow$ any areas alteration ?
- Distance $\rightarrow$ any length alteration ?

Regarding measures : It depends on the scale ! (and the devices)

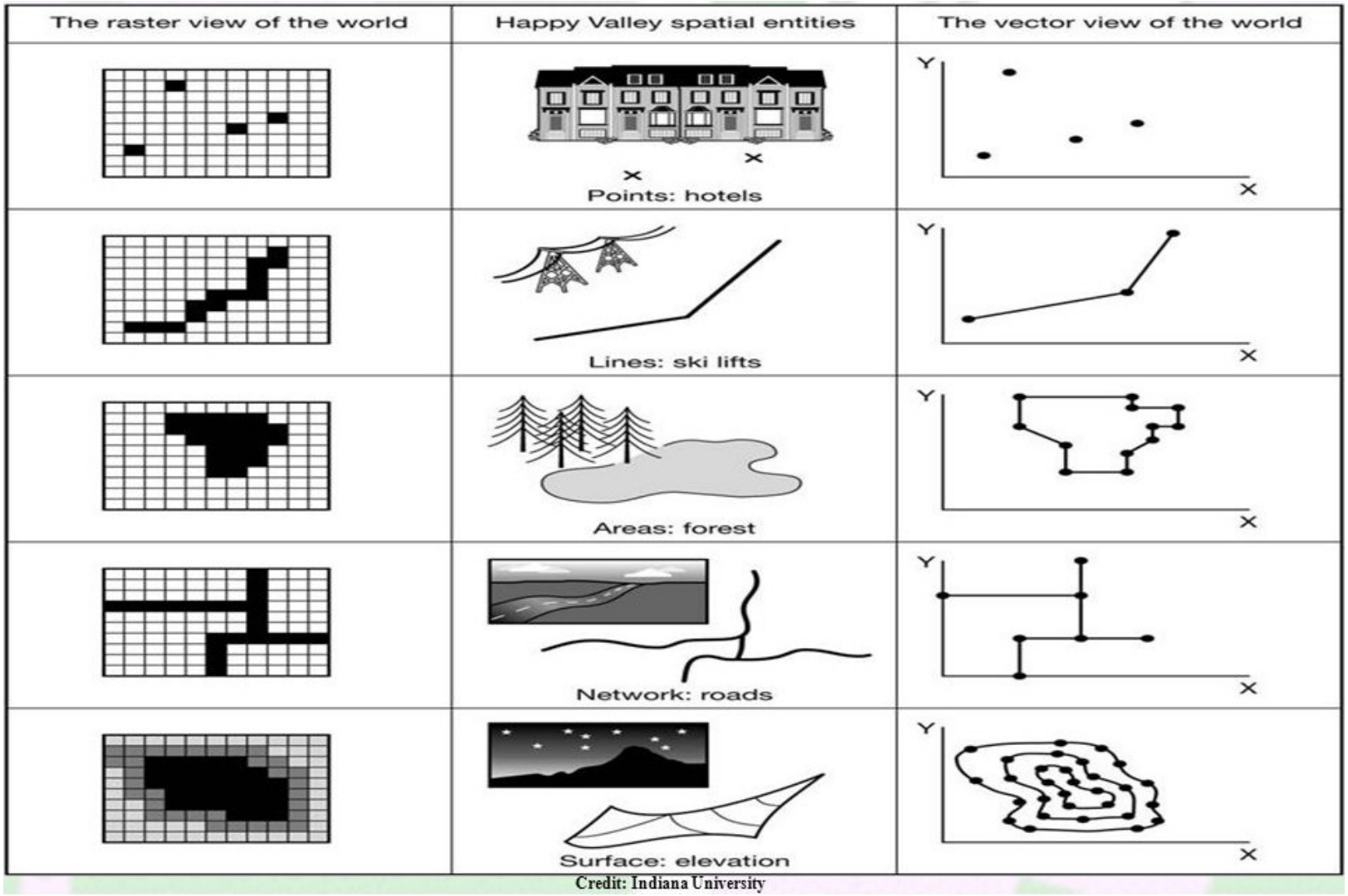
GPS-RTK (centimetric precision) or Great-circle distance ?





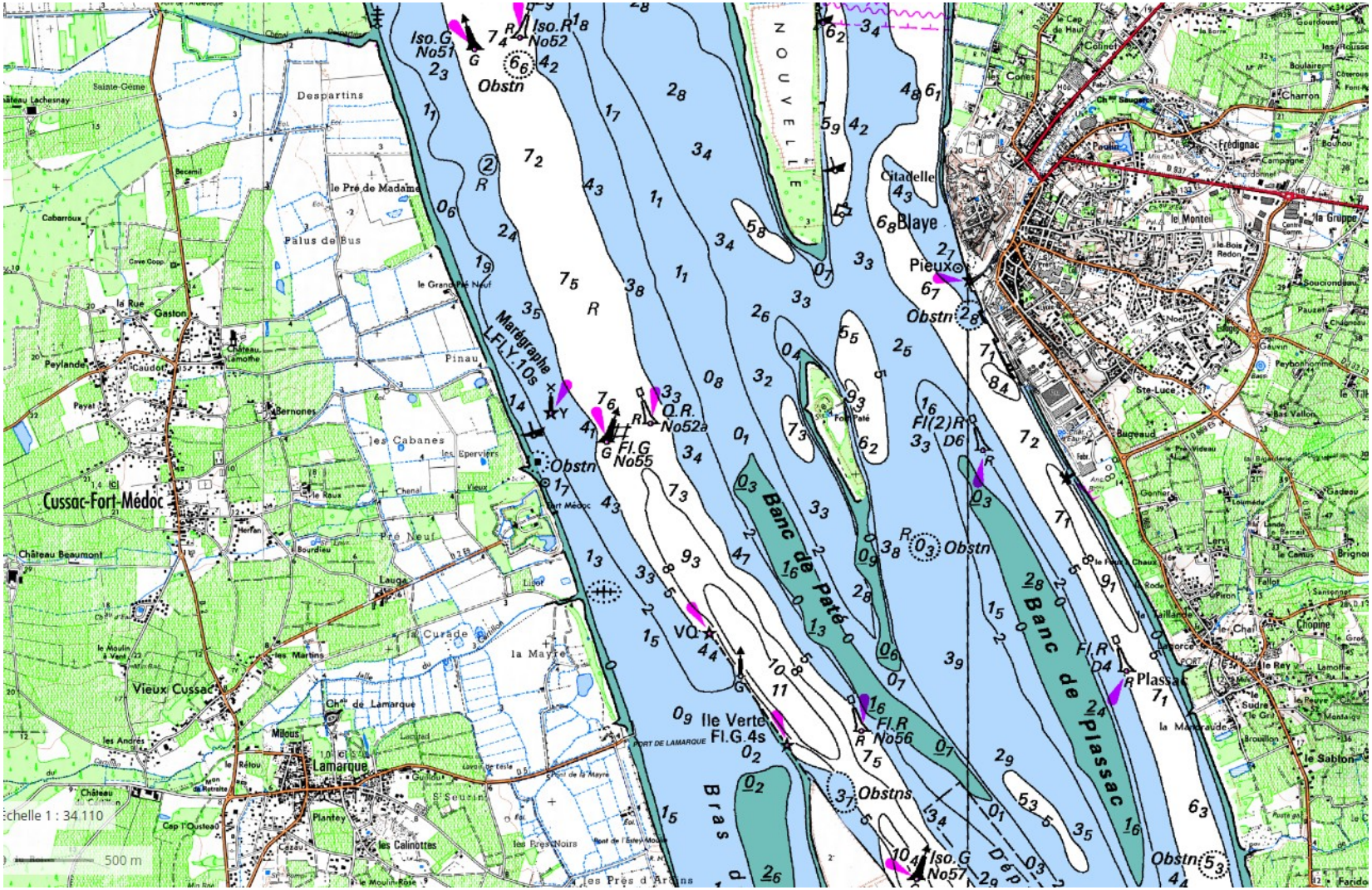
Modeling and representation

to model (here), is defining **categories** of objects **depicting** real-world objects (somehow linked to ontologies).



Modeling and representation

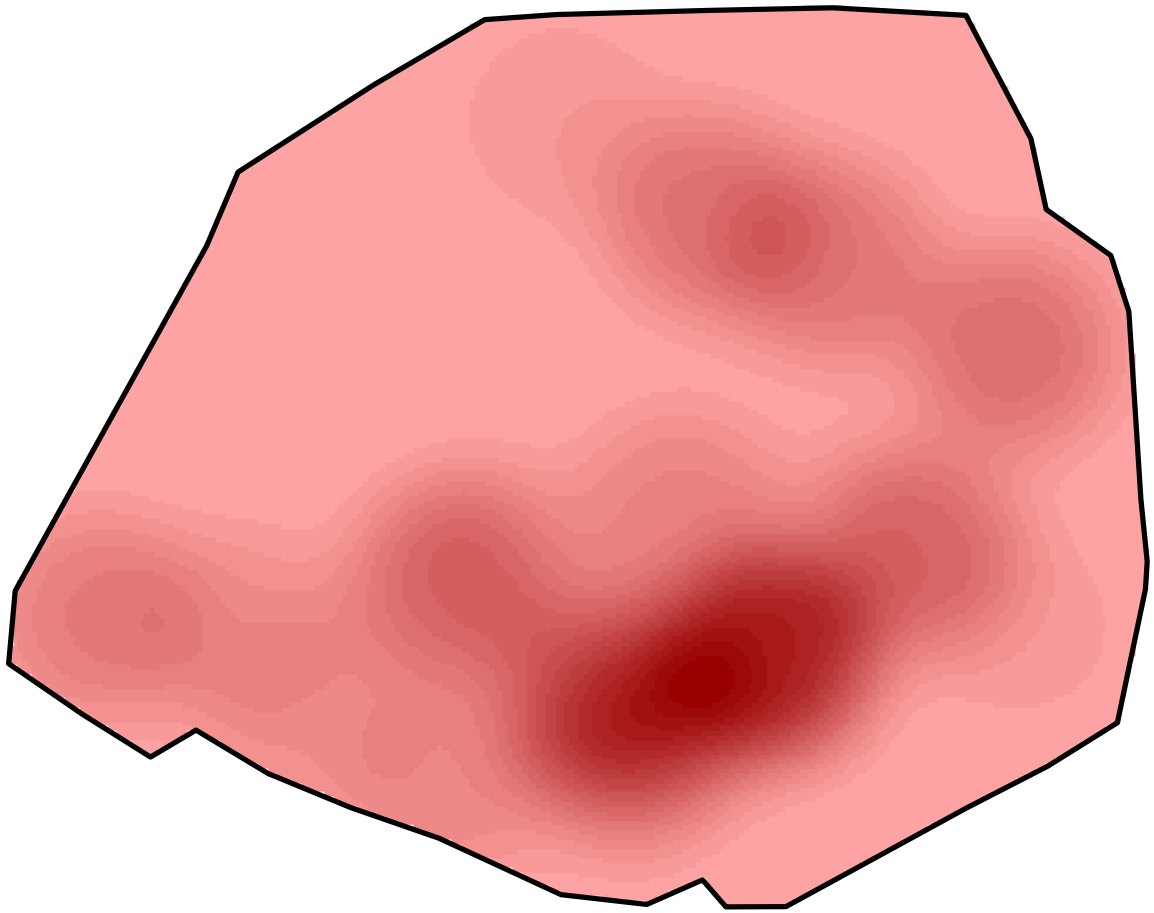
Objects and **Fields**.



Sources : IGN and SHOM

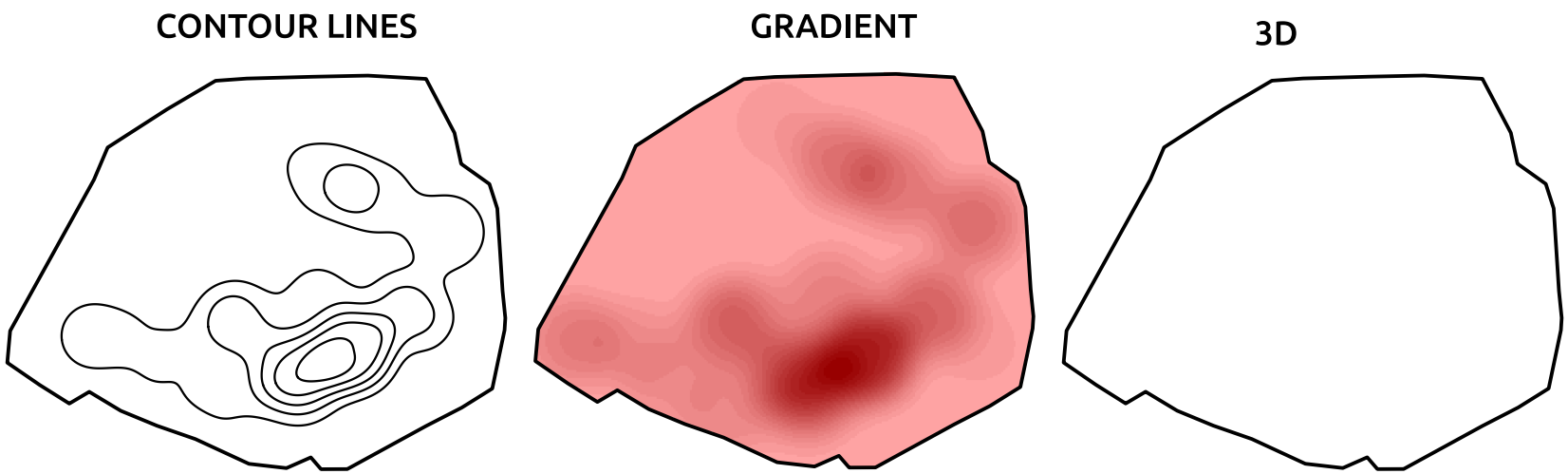
Fields

What does this field represent ?



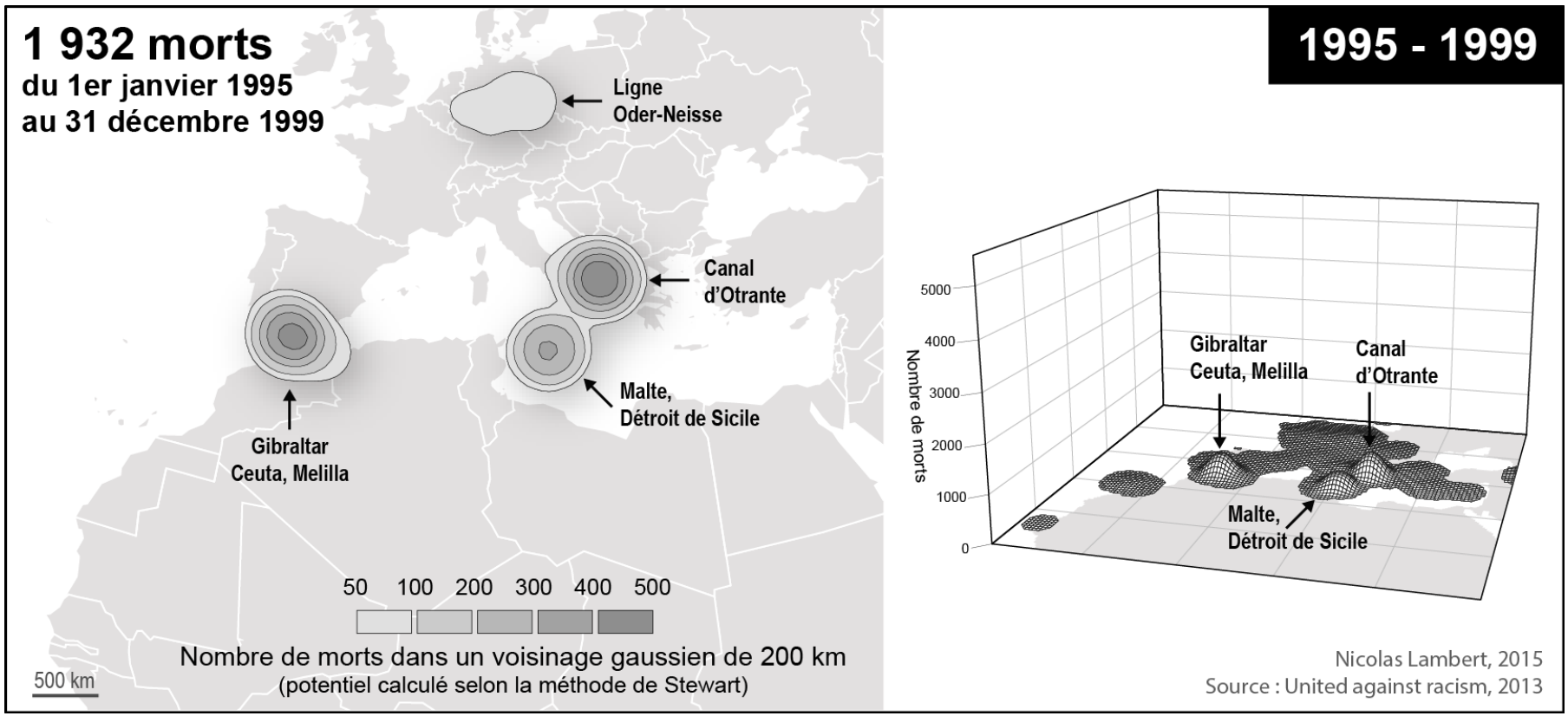
Fields

Representation modes



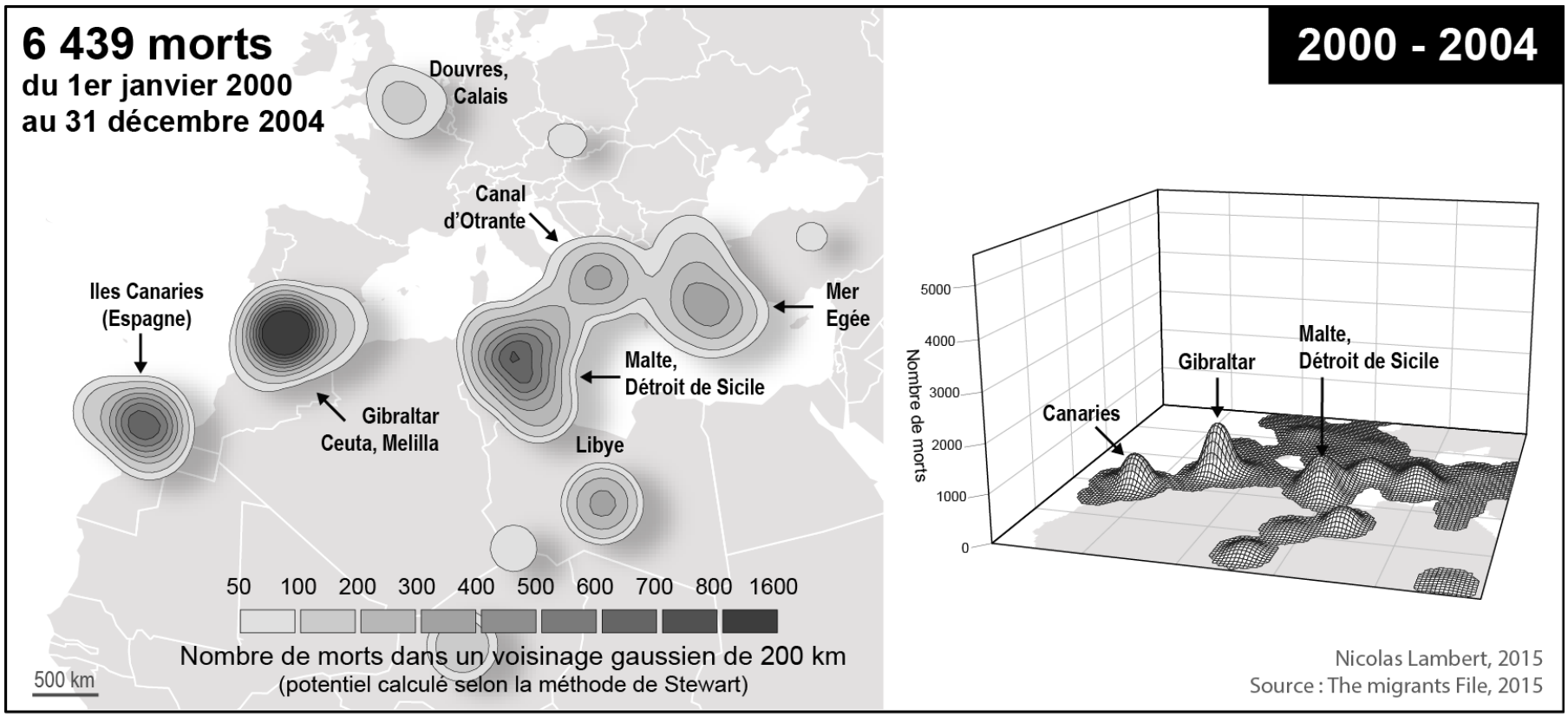
Fields

Representation modes examples



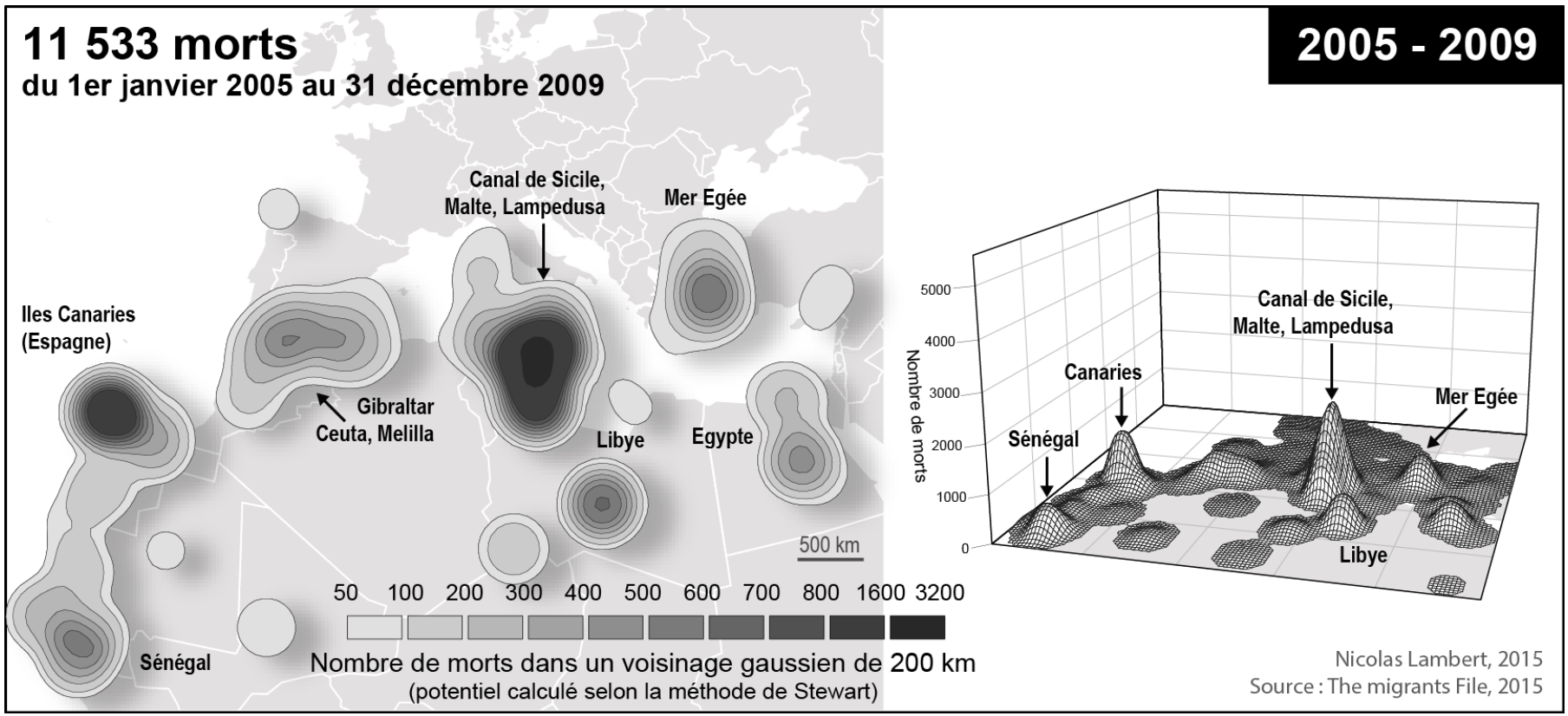
Fields

Representation modes examples



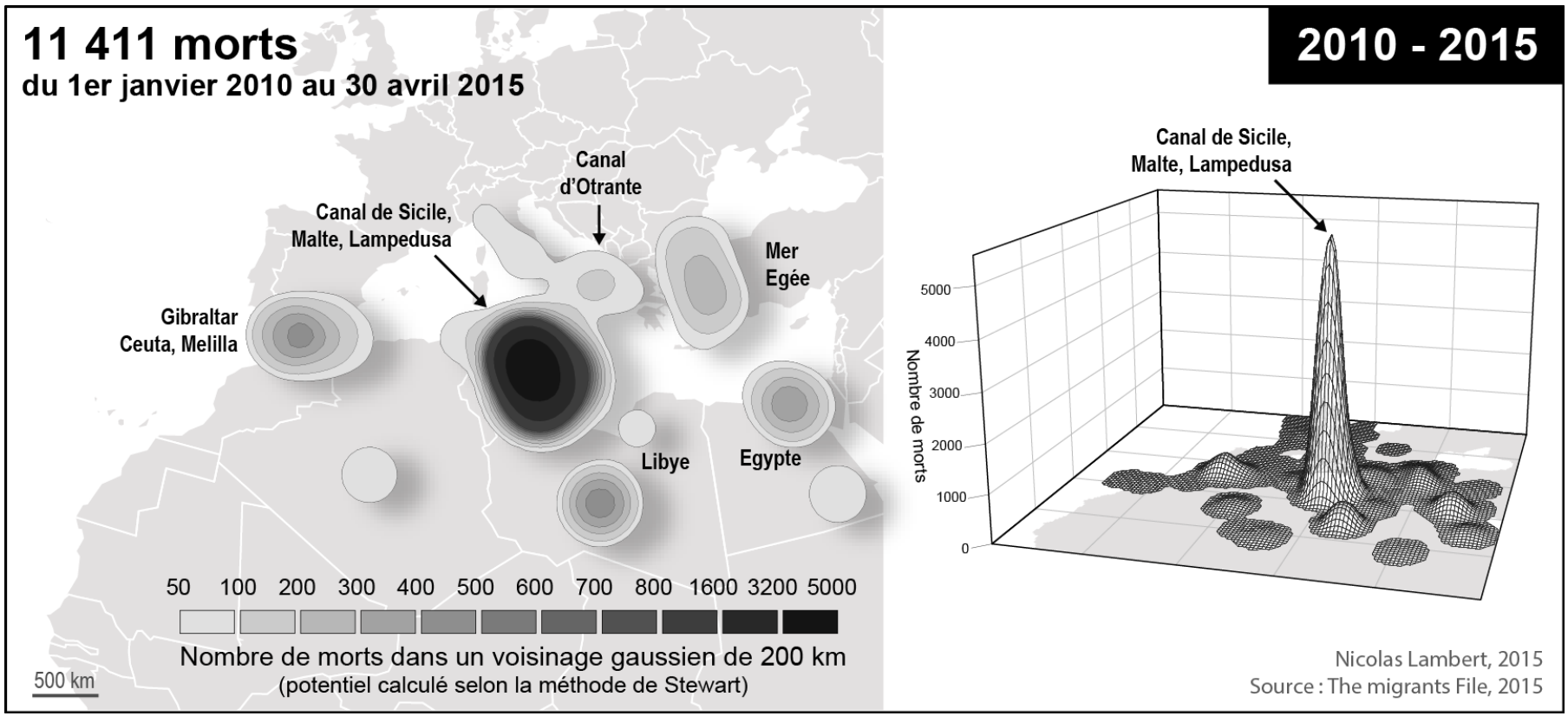
Fields

Representation modes examples



Fields

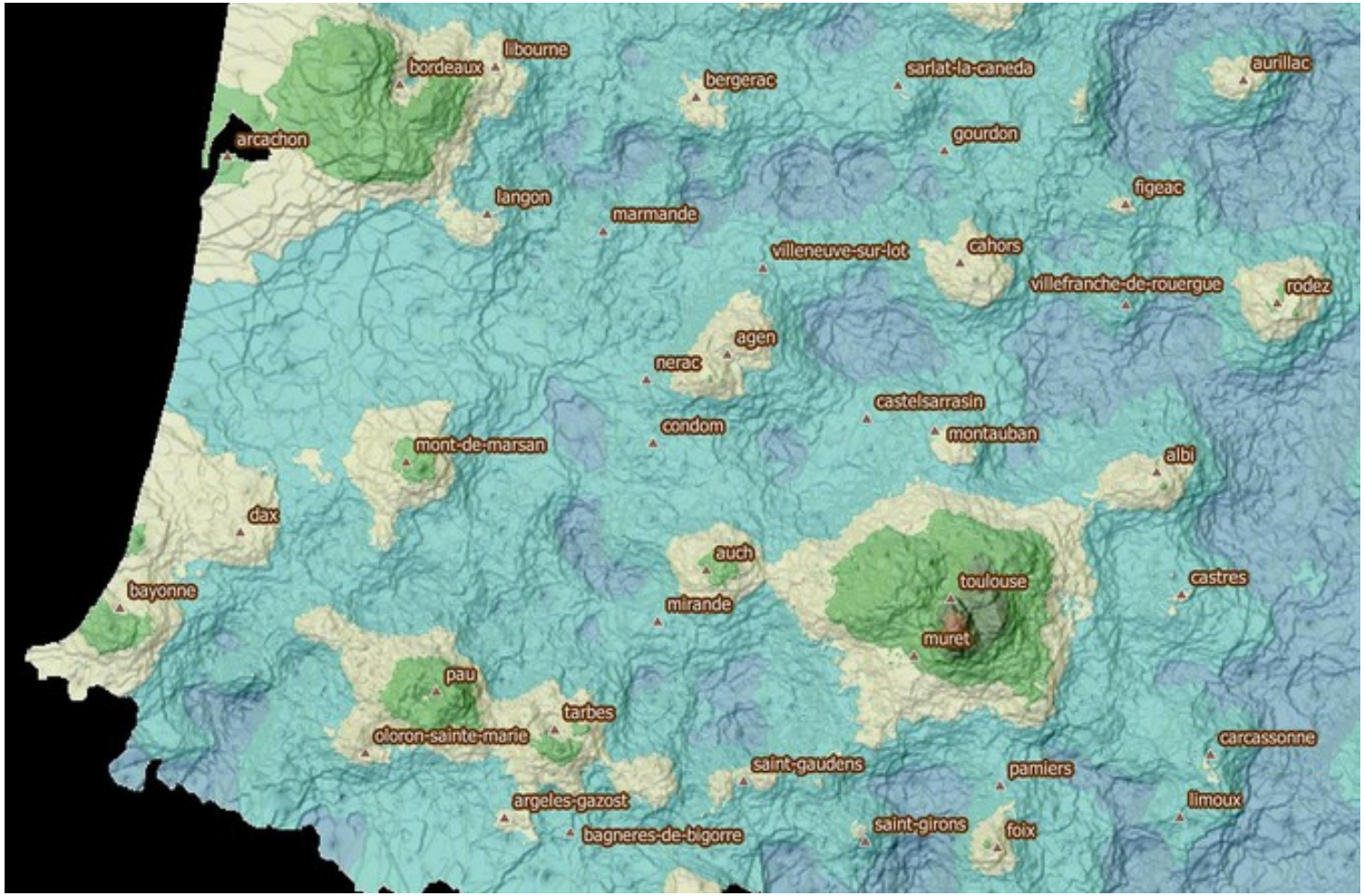
Representation modes examples



Fields

Source : Rajerison, Les archipels de la prospérité

Fields



Source : Rajerison, Les archipels de la prospérité

Simulation

DELHI GIS-R School

9-12th April 2019

Hadrien Commenges & Paul Chapron

hadrien.commenges@univ-paris1.fr

paul.chapron@ign.fr

Stochastic simulation

Stochastic simulation : data generation using **randomly drawn values**.

- **Monte Carlo :** generic term refering to process involving random process repetition.
- **Bootstrap :** re-sampling methods (usually to estimate distribution)
- **Permutation :** reordering elements of a set

Why ?

Most of the time : to approach a distribution

- (often) because the analytical way is hard
- because (sometimes) there is no analytical way
- because we look for robust estimation adequate to the use case data

Monte Carlo

→ **Example :** iterated dice rolls

→ **Goal :** examplify the Law of Large Numbers

Expected value μ of a dice roll :

$$\mu = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6$$

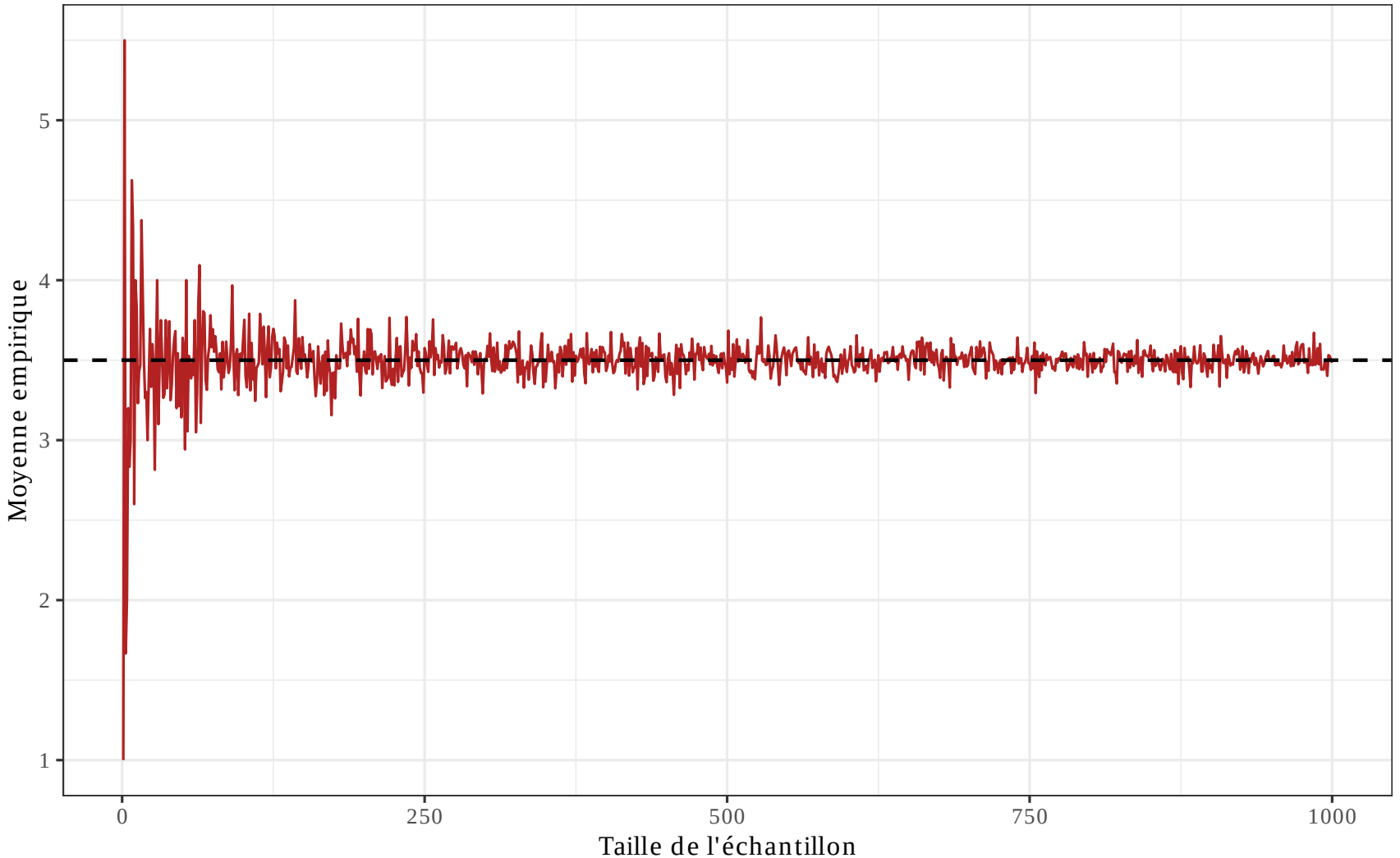
$$\mu = \frac{1+2+3+4+5+6}{6} = 3,5$$

(weak) Law of Large Numbers :

$$\bar{X}_n \to \mu \text{ when } n \to \infty$$

(sample average converges toward the expected value, for a sufficiently large sample)

Monte Carlo


Moyenne empirique
5
4
3
2
1
0
250
500
750
1000
Taille de l'échantillon

Bootstrap

Classical inference :

- Goal : approach μ and σ **parameters** of a distribution
- $\bar{X}$ and σ_X are computed on a **sample** X
- Sampling of X and $\bar{X}$ / σ_X computation are repeated

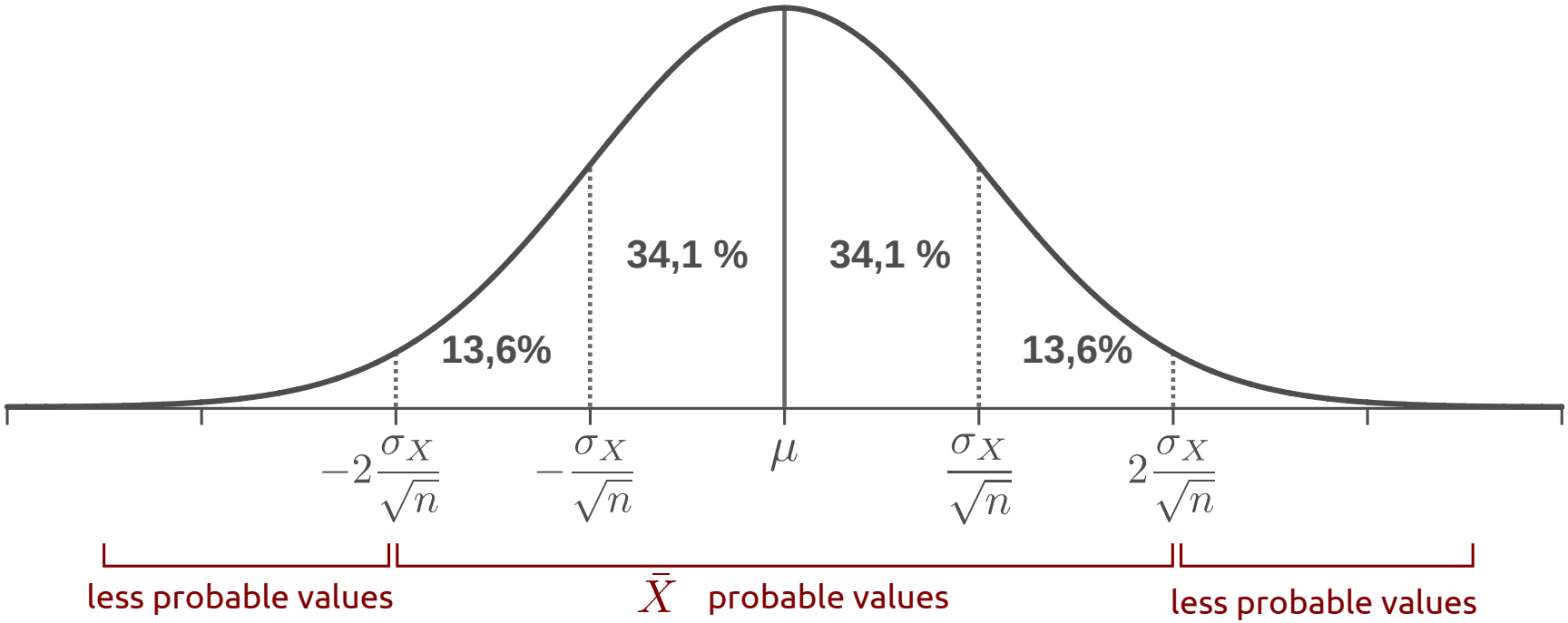
Example :

- For a 12M population (people from Île-de-France).
- People travel daily between 0 and 200 km.
- 500 samples are drawn, 100 people each.
- For each sample, the average travel distance ($\bar{X}$) is computed .

$\rightarrow$ these 500 average values form a *distribution* : the **sampling distribution of the mean**

Bootstrap

Sampling distribution of the mean (500 mean values) :



where μ et σ are the **parameters** – real mean and standard deviation of the population – and n is the **sample size**.

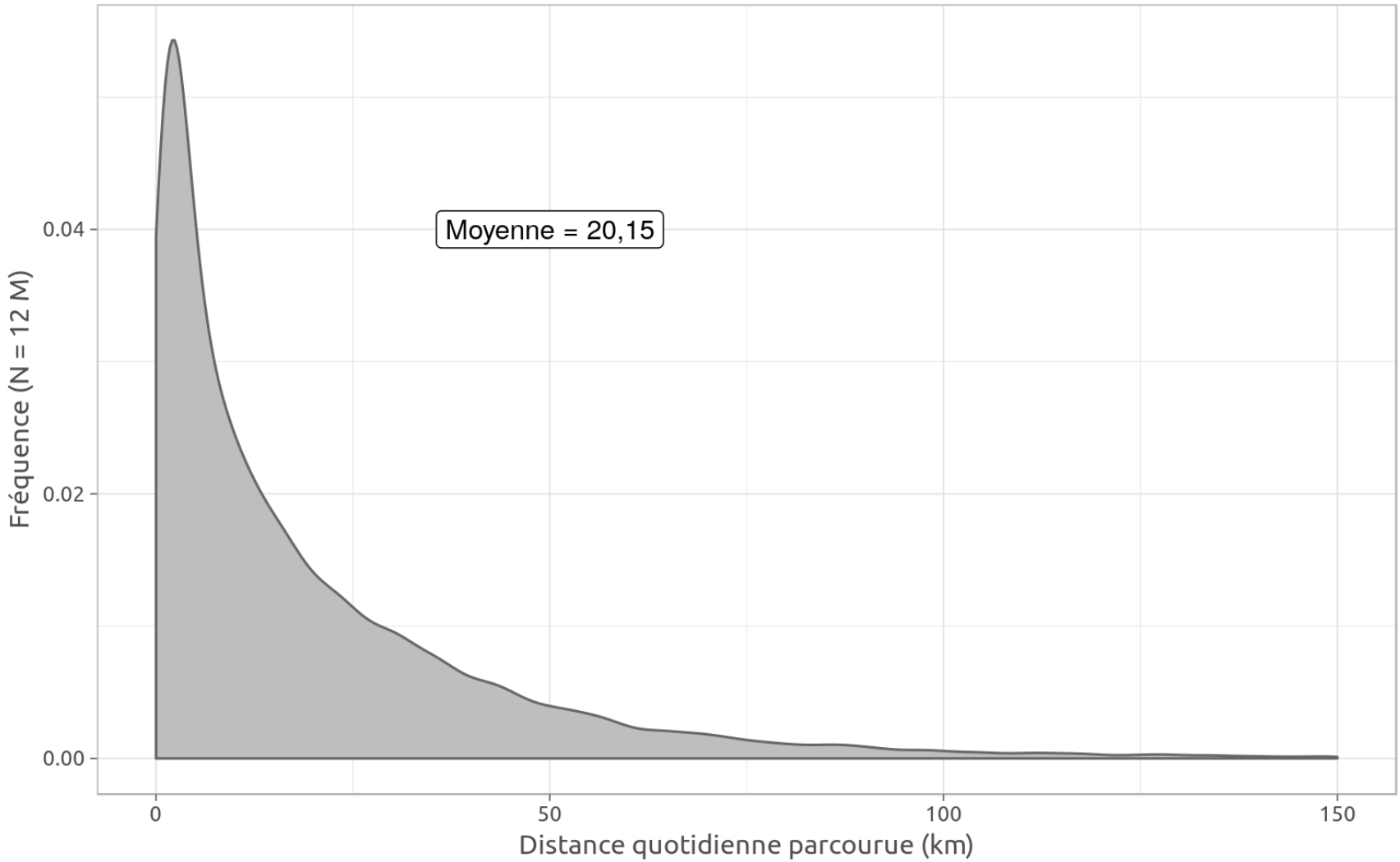
$\rightarrow$ This is **central limit theorem** (**CLT**).

Bootstrap

- **General inference idea :** the sample (which is known) allows to approach the parameters of the population distribution (which is unknown)
- **General bootstrap idea :** re-sampling (which is known) from the sample (which is known) give insights about what would sampling look like on the whole population.
- **Example :** to estimate the variance of the sampling distribution , we compute the variance of the re-sampling distribution of the mean.

Bootstrap

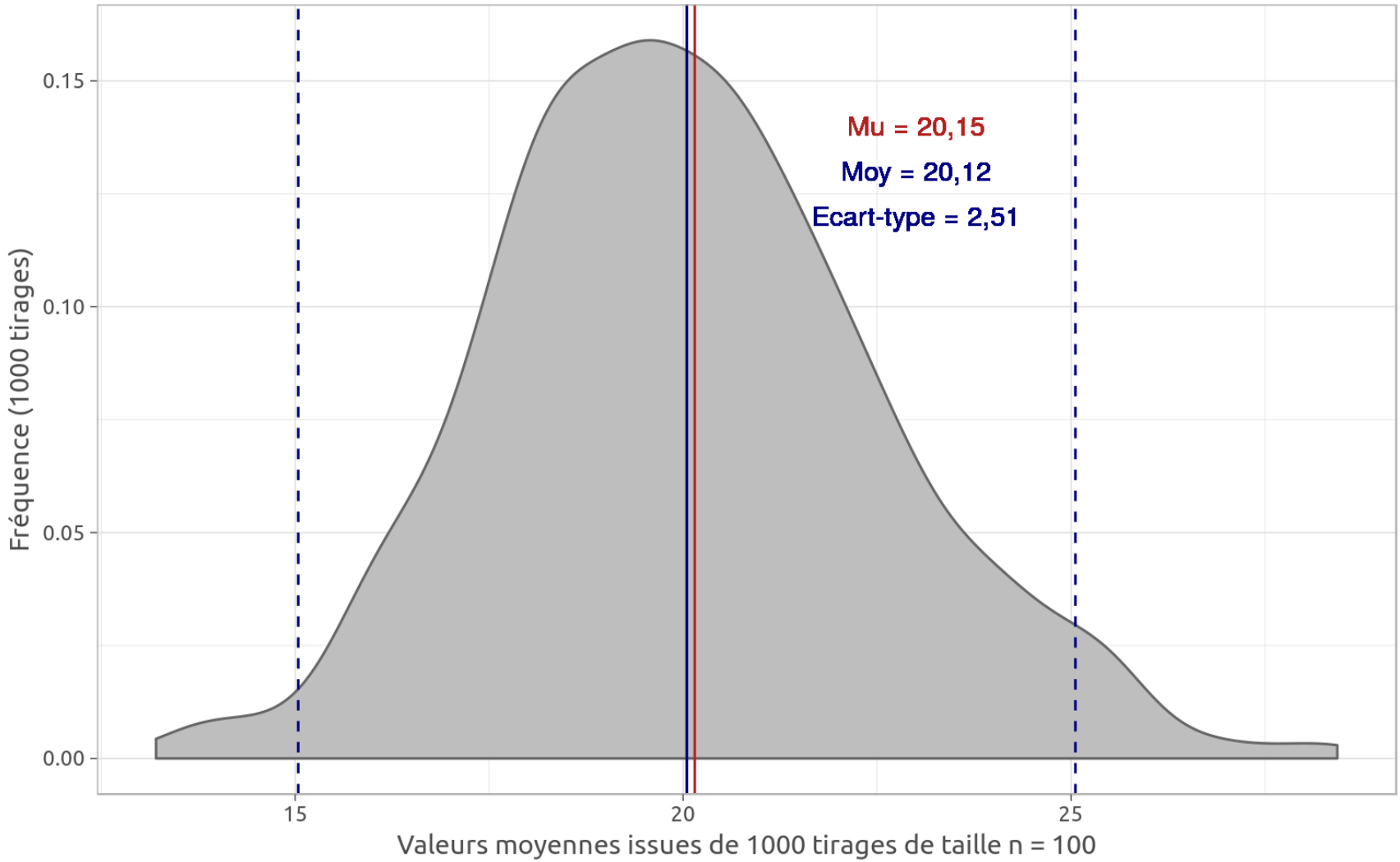
Daily travel distance for Île-de-France people



Source : Enquête Globale Transport 2010 (French transportation national census)

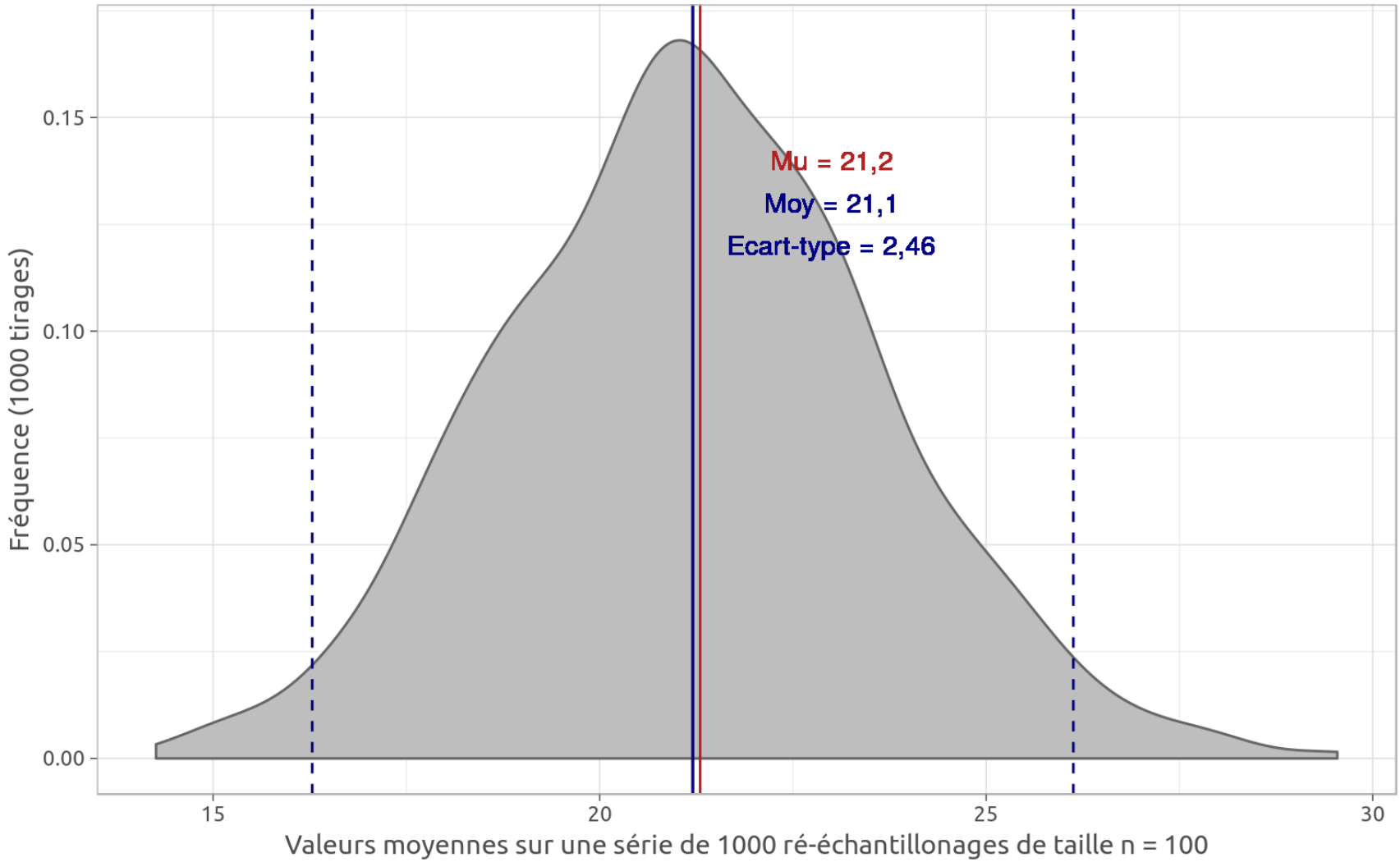
Bootstrap

Sampling distribution of the mean



Bootstrap

Re-sampling of the mean on a sample



Bootstrap

Power of the Bootstrap : this technique offers

- compute estimates (mean, variance) without any hypothesis or *a priori* knowledge on the population
- compute estimates variability (confidence interval) without any hypothesis or *a priori* knowledge on the population (*distribution-free confidence intervals*)
- assess the stability of some model results (*cross-validation*)

Density

DELHI GIS-R School

9-12th April 2019

Hadrien Commenges & Paul Chapron

hadrien.commenges@univ-paris1.fr

paul.chapron@ign.fr

Use case

Density

Density is a **spatial variable** depicting the **spatial variation** of some observations (concentration and dispersion) in 1, 2 or *n* dimensions.

- Mass / Volume ratio (volumetric mass), mass per volume unit, sometimes ratio between an object volumetric mass and a reference volumetric mass.
- Ratio between a **count** and its **extent** : a variable's density (1D), population density (2D, so **spatial extent**), etc.

What kind of geographical information is concerned ?

→ *TYPE 1 - Geographical Objects*

→ *TYPE 2 - Occurrences*

Goals

Main uses :

1. Describe a point pattern
2. Estimate the probability of an event to occur at a given point
3. Estimate the probability that a spatial distribution of events is random.

Spatial distribution parameters

Centrality and **dispersion** can be computed in a 1, 2 2 or n dimensions space.

Current analysis of these parameters :

- Description of a distribution : mean and standard deviation
- Evolution of these parameters over time
- Parameters weight

Spatial distribution parameters

2-Dimensions **mean** : **barycenter** (or **balancing point**).

$$x_g = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad y_g = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

$\rightarrow$ weights w_i might be constant or varying, depicting localized stocks variations.

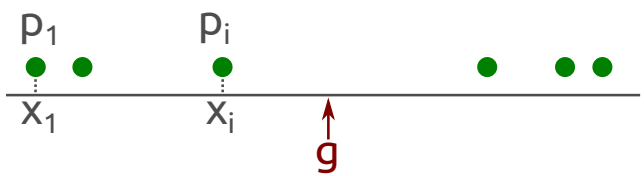
Spatial distribution parameters

2-Dimensions **variance** $\approx$ **inertia**.

General formula
-> squared distances mean

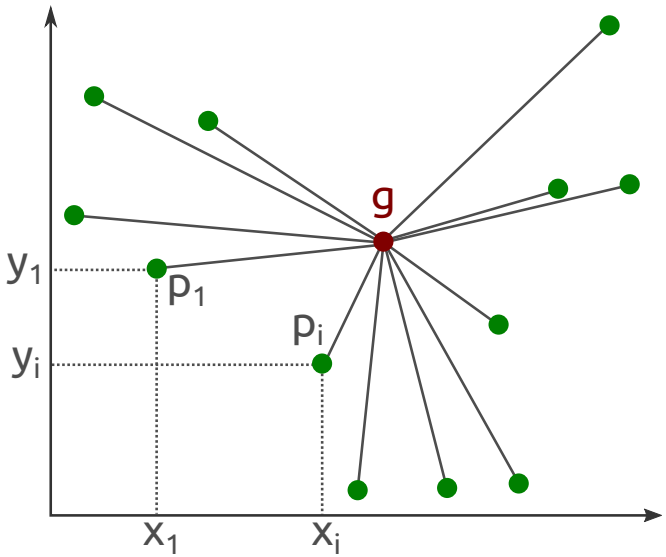
$$I = \frac{1}{n} \sum_{i=1}^{n} d^2(p_i, g)$$

1D



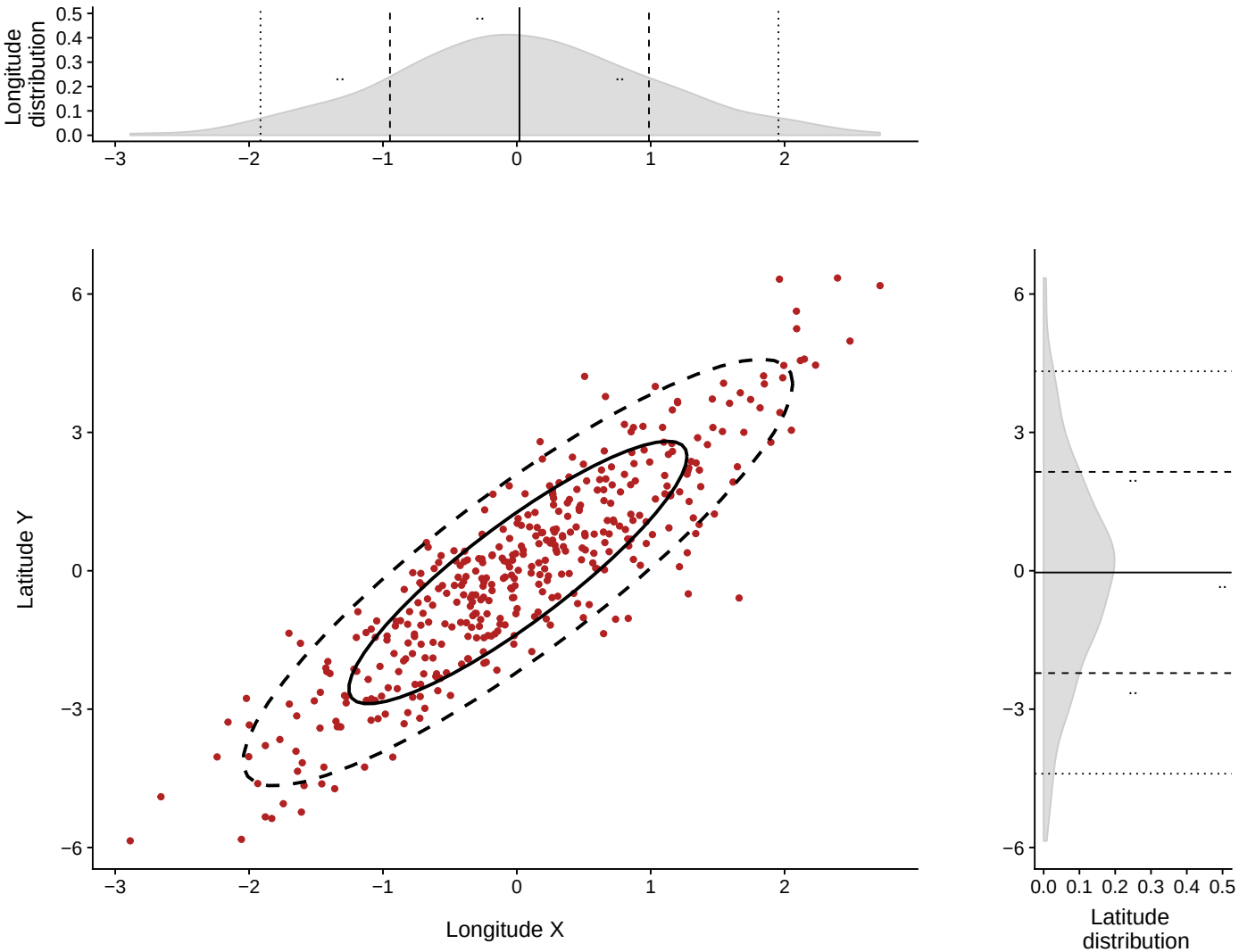
$$I = \frac{1}{n} \sum_{i=1}^{n} (x_i - x_g)^2$$

2D

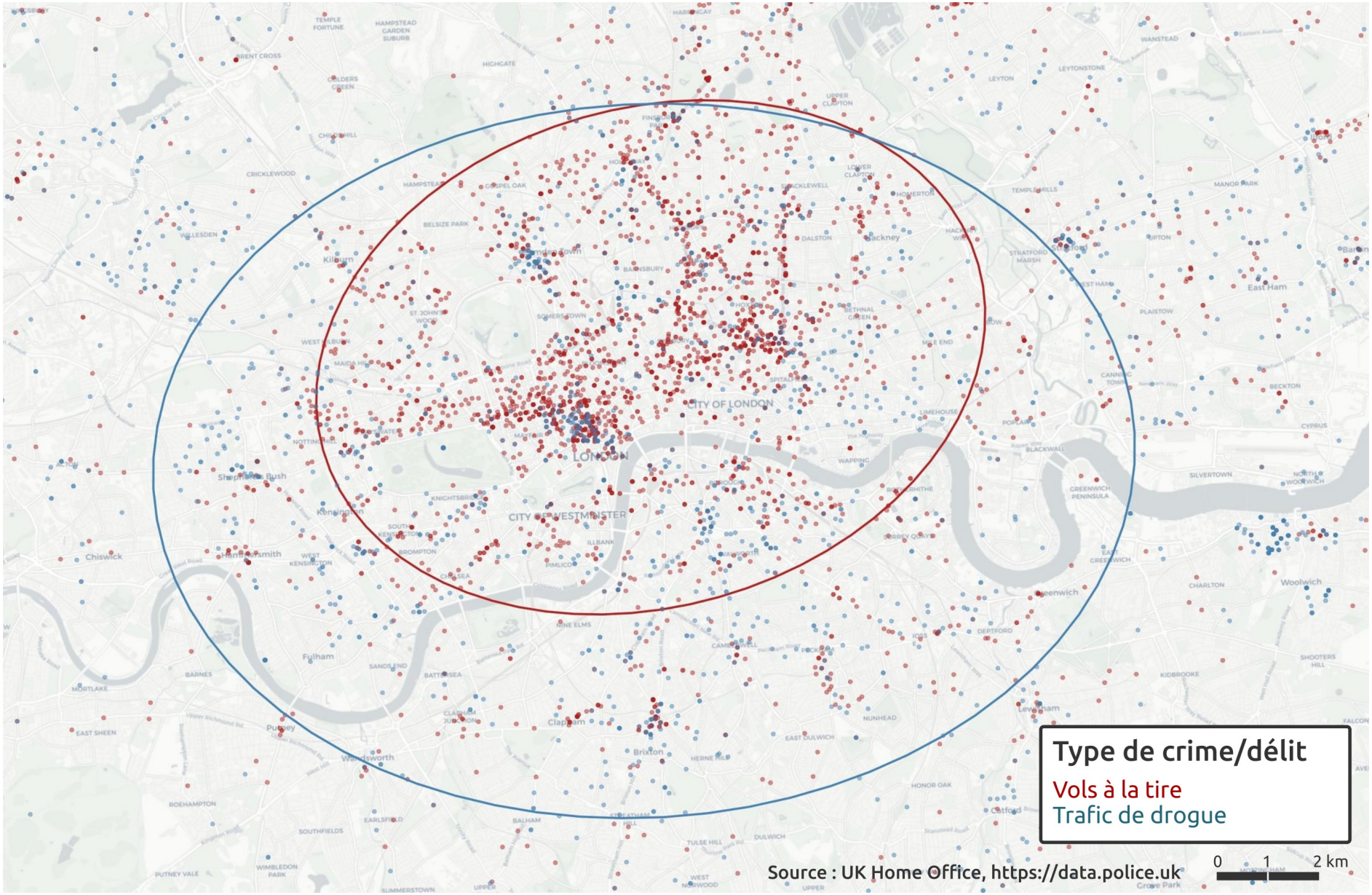


$$I = \frac{1}{n} \sum_{i=1}^{n} \left[(x_i - x_g)^2 + (y_i - y_g)^2 \right]$$

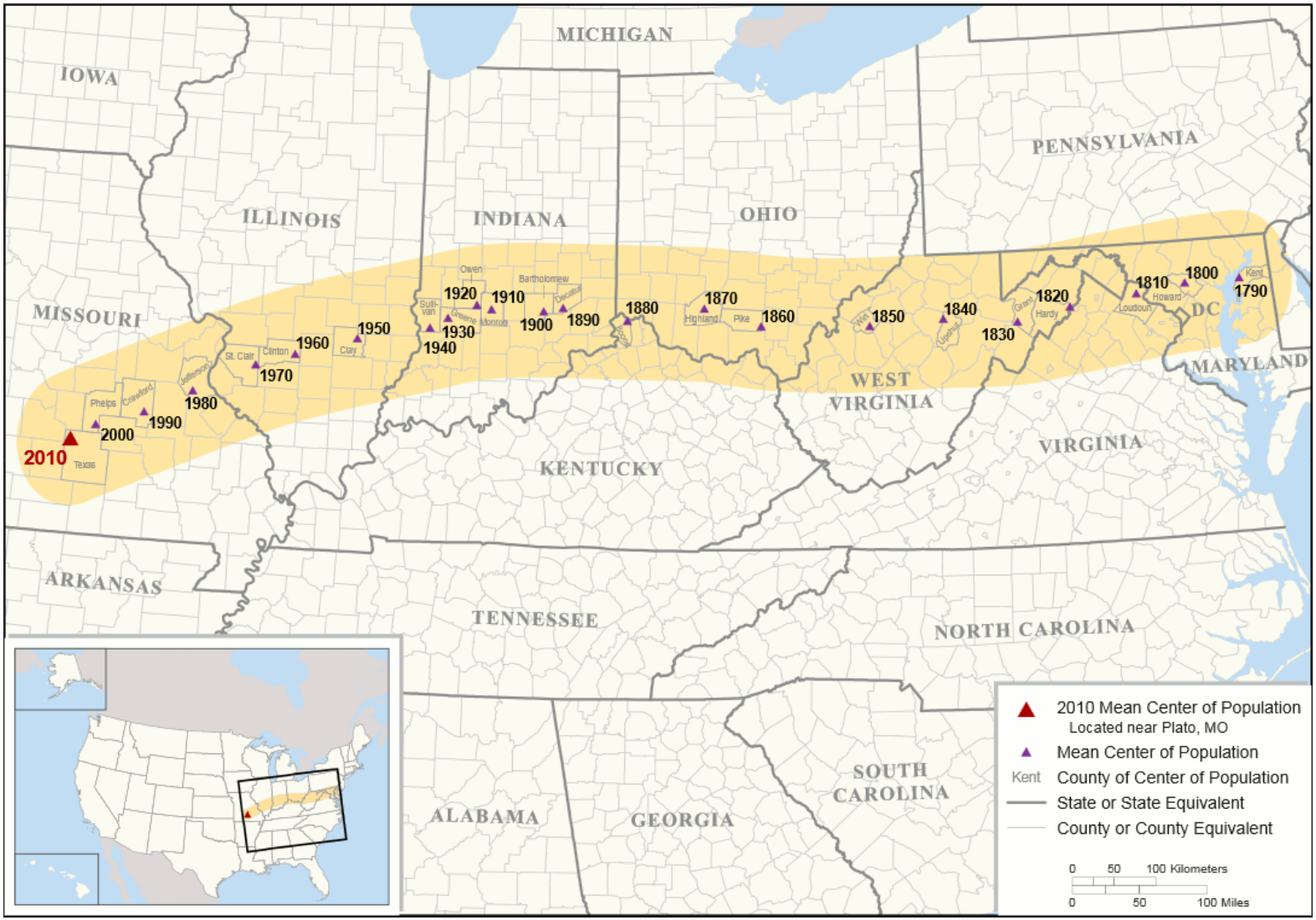
Spatial distribution parameters



Spatial distribution parameters



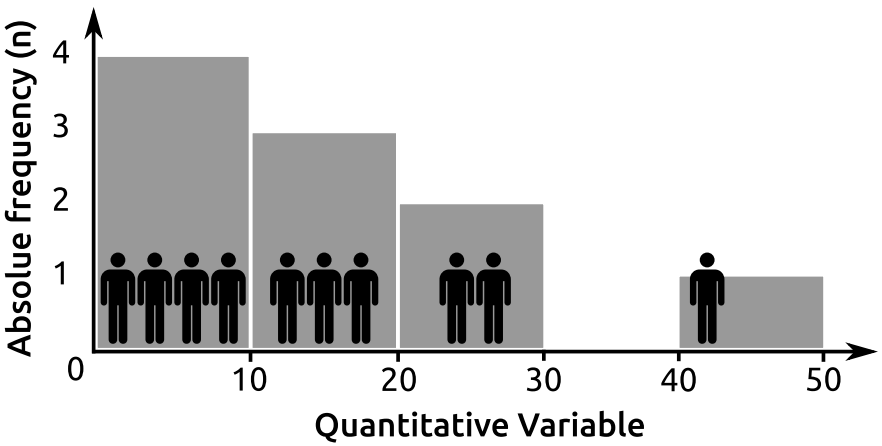
Spatial distribution parameters

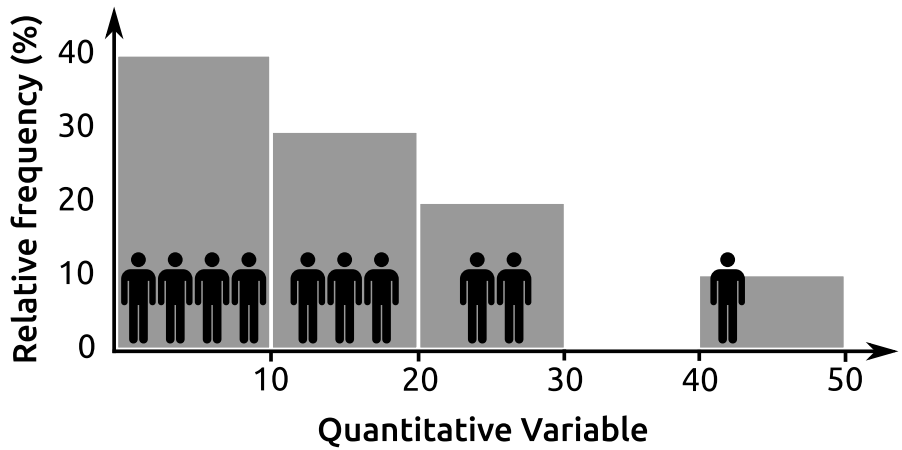


Source : US Census, `https://www.census.gov/geo/reference/centersofpop.html`

1D distribution graph (discrete)

Histogram

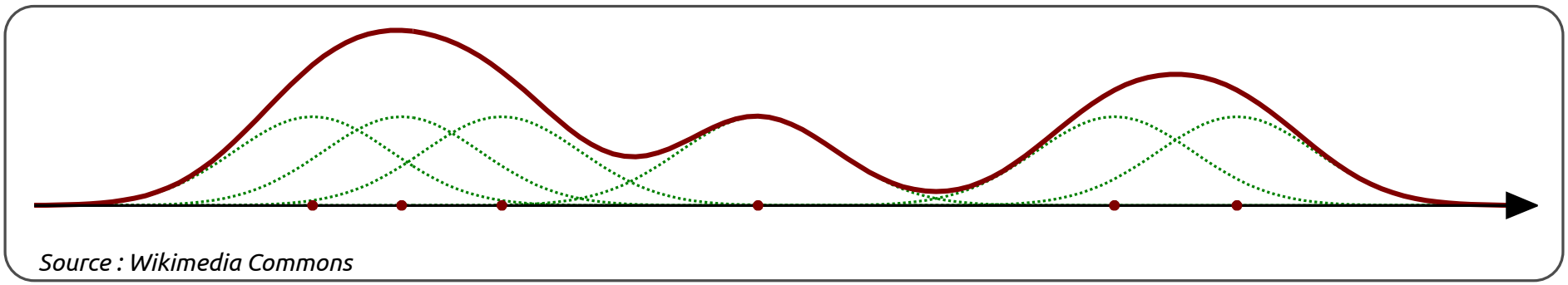




$\rightarrow$ histogram estimated density is **discrete** by construction.

1D distribution graph (continuous)

Kernel Density Estimate

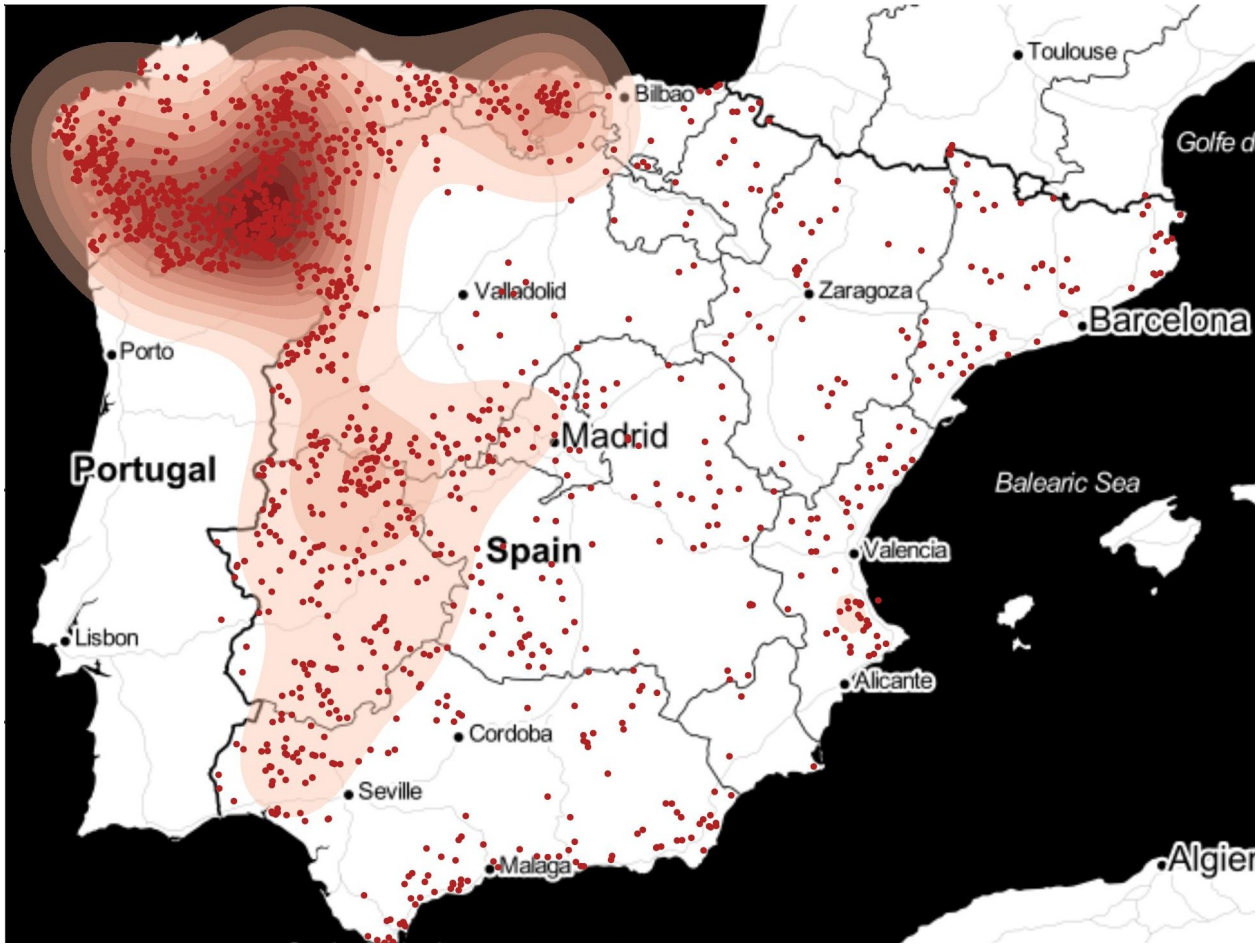


Source : Wikimedia Commons

Kernel Density Estimator) may be applied in 1 to *n* dimensions. For spatial nalysis, we use the **2D** version.

Cartographier une distribution

Densité par estimateur du noyau (KDE) en **deux dimensions**.



Tester une distribution

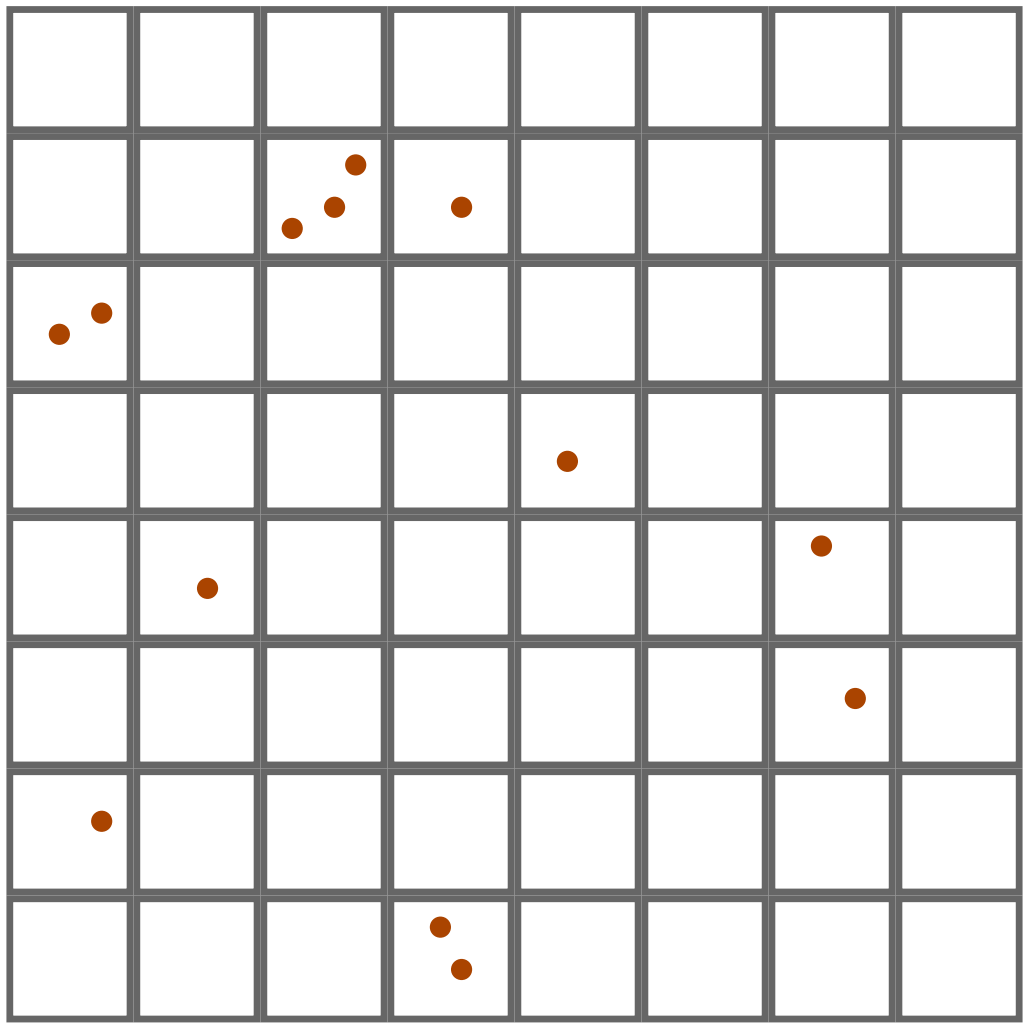
La distribution peut-elle être produite par un processus aléatoire ?

Number of times victimised	Respondents %	Incidents %
0	59.5	0.0
1	20.3	18.7
2	9.0	16.5
3	4.5	12.4
4	2.4	8.8
5+	4.3	43.5

Source : Farrell, Pease (1993) Once bitten, twice bitten, Police Research Group, London.

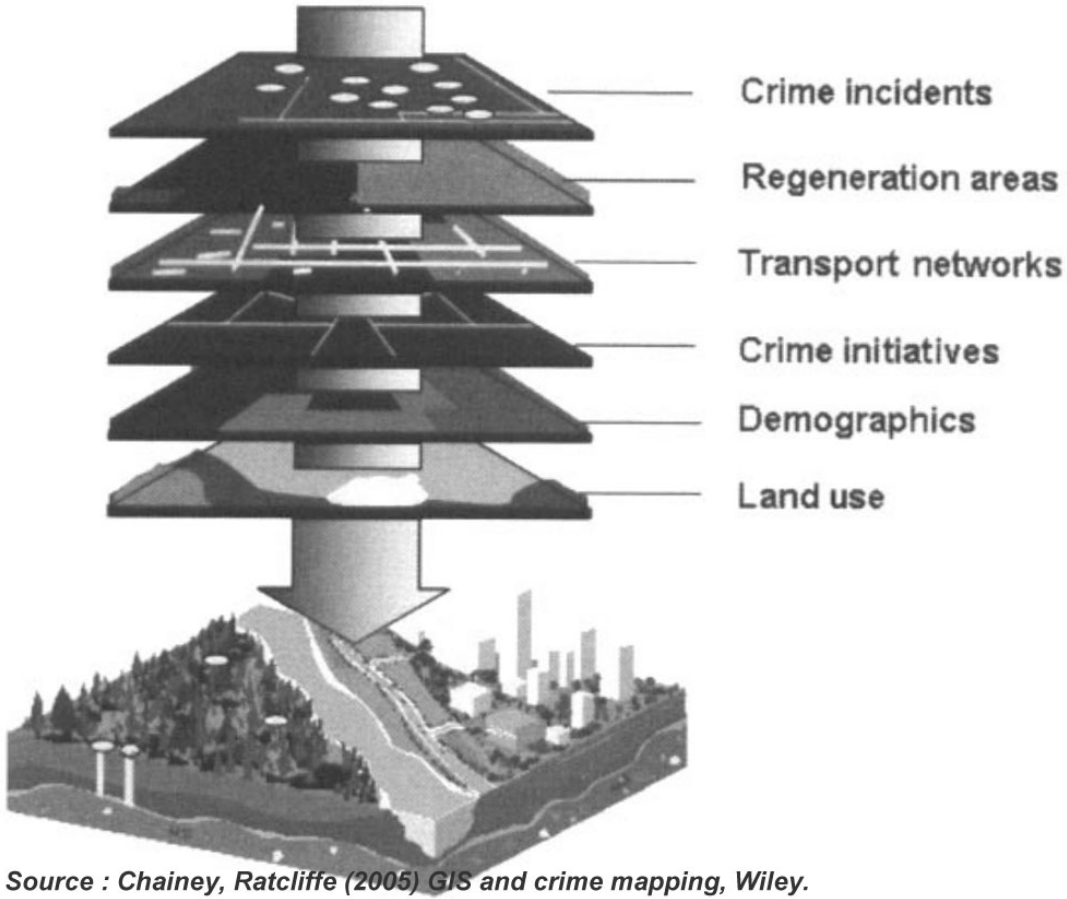
Tester une distribution

La distribution peut-elle être produite par un processus aléatoire ?



Tester une distribution

La distribution peut-elle être produite par un processus aléatoire ?

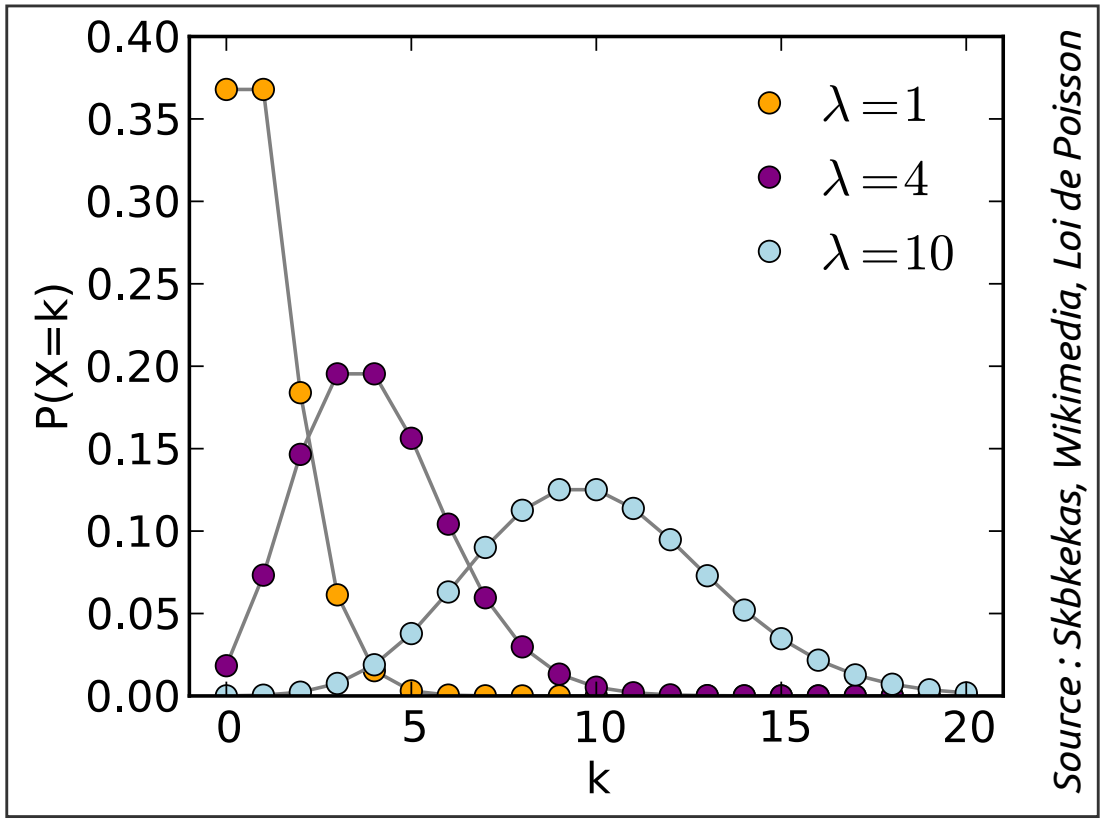


Source : Chainey, Ratcliffe (2005) GIS and crime mapping, Wiley.

Distribution statistique de Poisson

La distribution de Poisson se définit par un seul paramètre λ qui est à la fois la moyenne et la variance de la distribution.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



Distribution spatiale de Poisson

Un **processus de spatial de Poisson** est un processus **spatialement aléatoire**, en anglais on trouve les termes suivants :

- *spatial Poisson process*
- *homogeneous Poisson process*
- *complete spatial randomness* (CSR)

Dans un espace découpé en zones, on estime la probabilité d'un nombre d'occurrences (points) dans une zone par une distribution de poisson de moyenne $\lambda \times \mathit{surface}(\mathit{zone})$.

Indicateur de dispersion

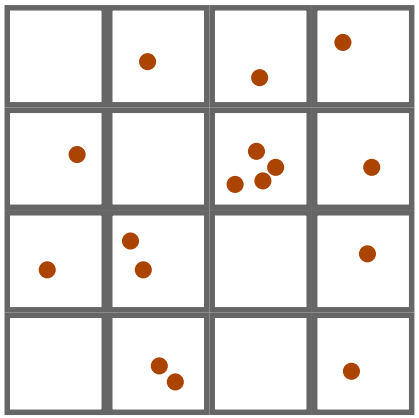
Calcul de l'indicateur VMR *Variance-to-Mean Ratio*

- Carroyer l'espace d'étude
- Dénombrer les occurrences
- Calculer la variance, la moyenne et le ratio variance/moyenne

Interprétation du VMR

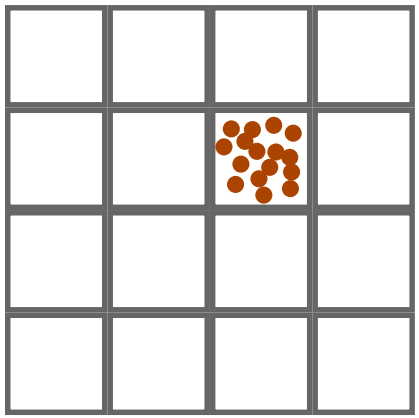
- $\mathrm{VMR} = 1$
  distribution possiblement produite par un processus de Poisson
- $\mathrm{VMR} < 1$
  distribution à tendance homogène
- $\mathrm{VMR} > 1$
  distribution à tendance concentrée

Indicateur de dispersion



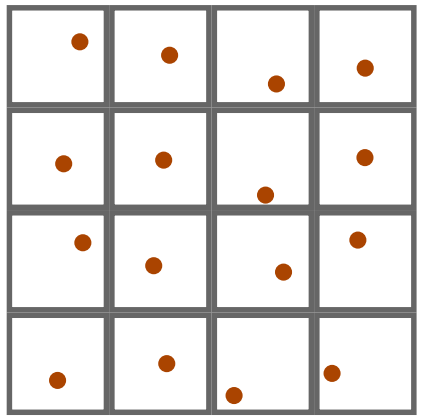
X (nbr. occurrences)
[0 1 1 1 1 0 4 1 1 2 0 1 0 2 0 1]

Var(X) = 1,1
$\overline{X}$ = 1
VMR = 1,1



X (nbr. occurrences)
[0 0 0 0 0 0 16 0 0 0 0 0 0 0 0 0]

Var(X) = 16
$\overline{X}$ = 1
VMR = 16



X (nbr. occurrences)
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]

Var(X) = 0
$\overline{X}$ = 1
VMR = 0

$\rightarrow$ **test :** le VMR est-il significativement différent de 1 ? (test de Student)

Méthode des quadrats

Calcul des quadrats

- Carroyer l'espace d'étude (quadrats).
- Le modèle de référence, modèle nul, qui donne les valeurs espérées dans la grille, est le processus spatial de Poisson.
- On obtient tableaux de contingence avec un compte d'occurrences (observé, espéré).

$\rightarrow$ **test :** les valeurs observées sont-elles significativement différentes des valeurs espérées (test du χ^2)