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.

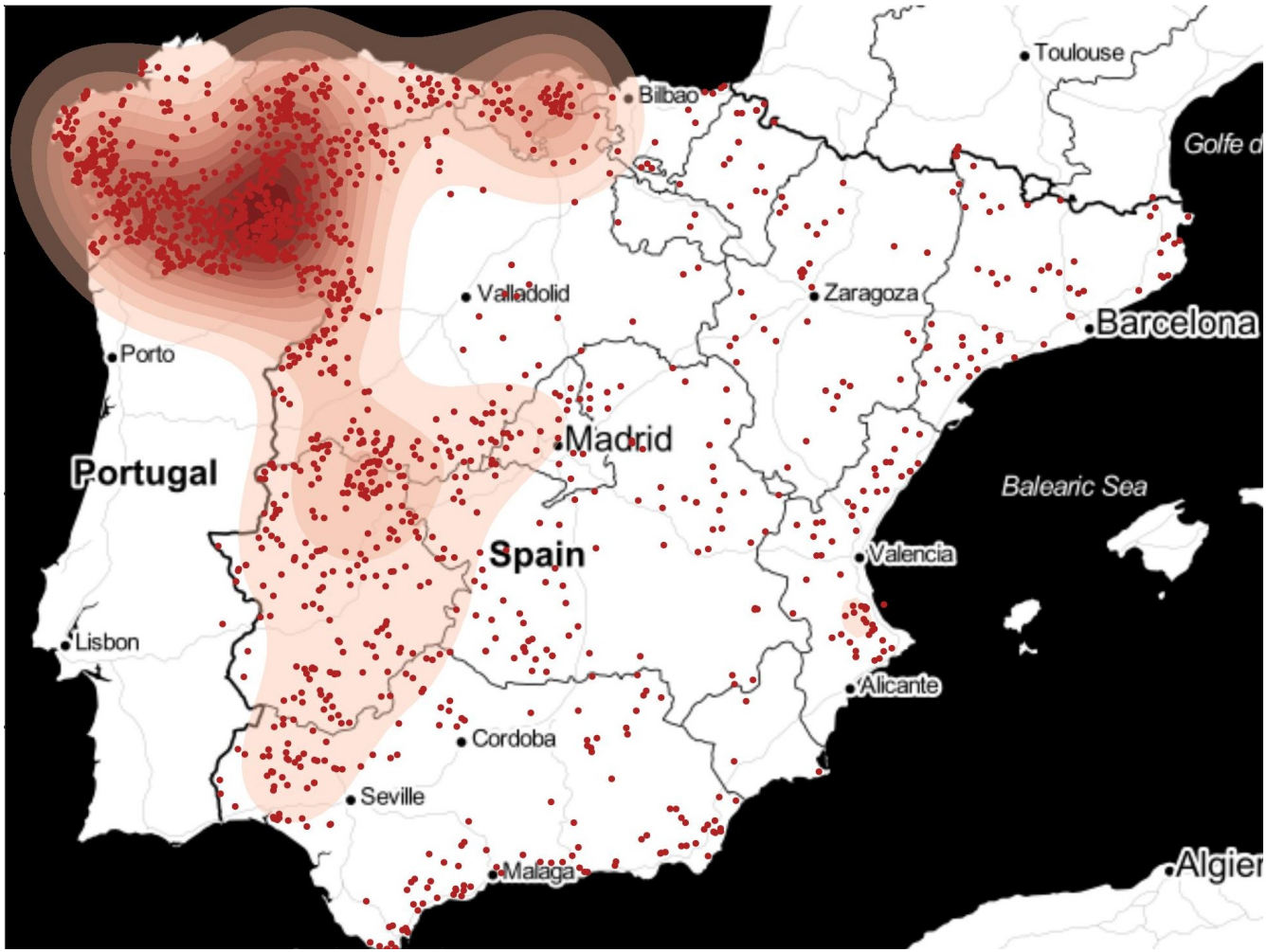
1. Dewellings point patterns (spatial object)
2. Dweelings sales (occurrences)
3. Dwellings prices (measure points)
4. Average price by district (« statistics »)

(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).

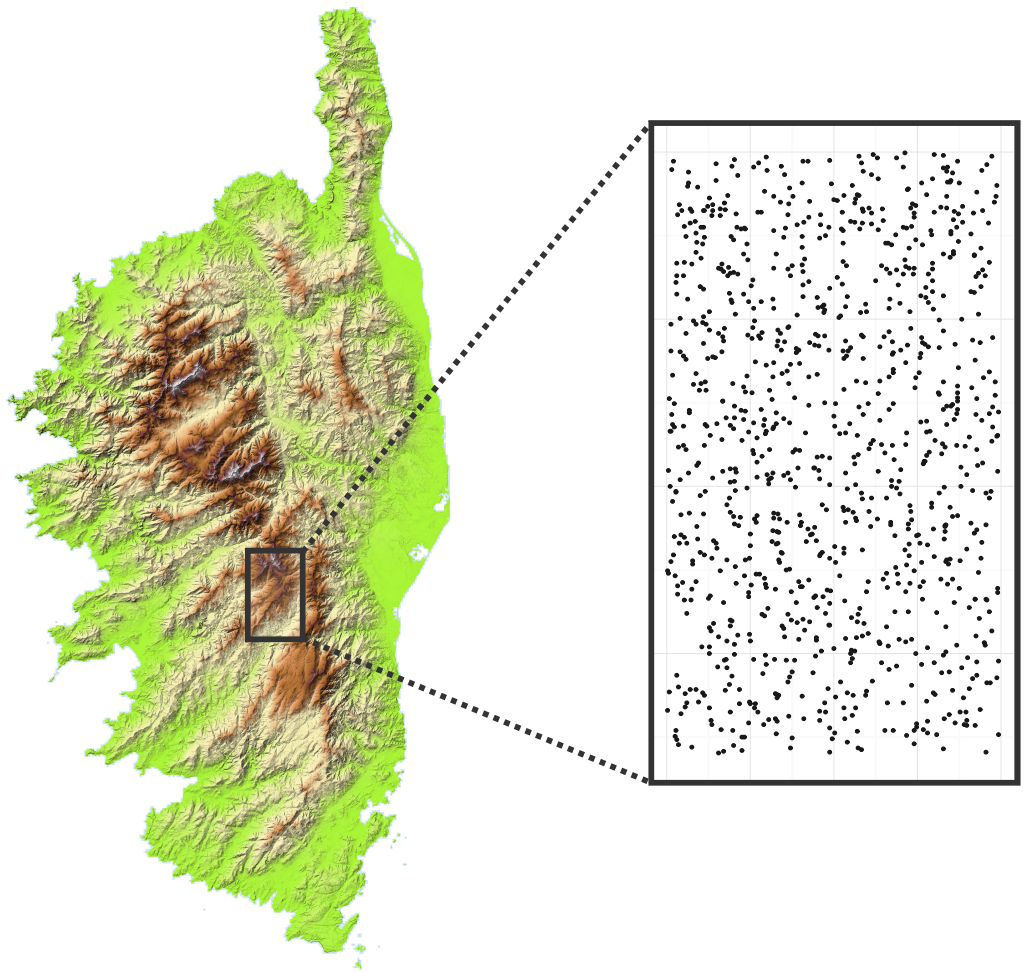
(2) Event occurence

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



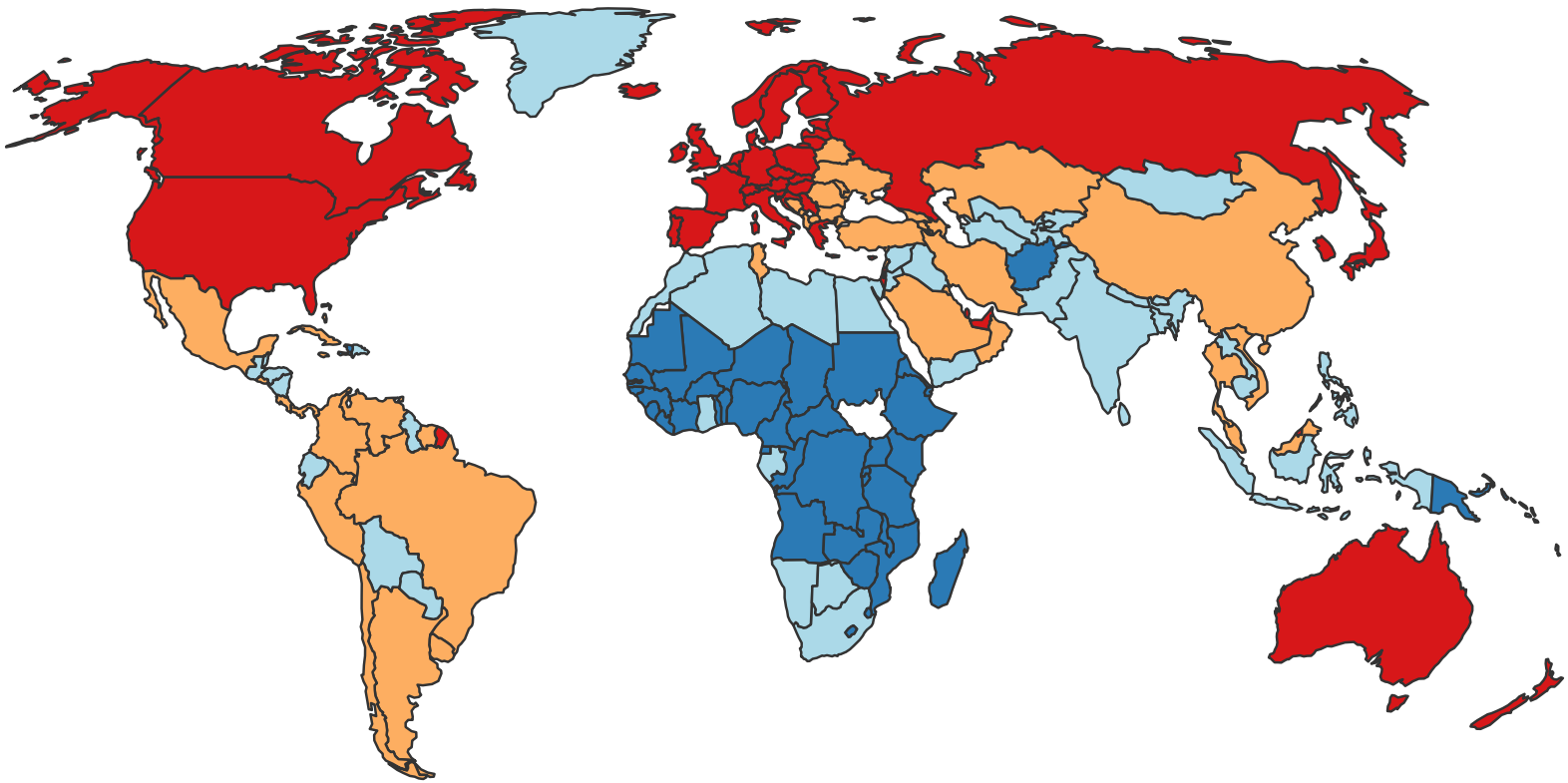
(3) Measure points

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



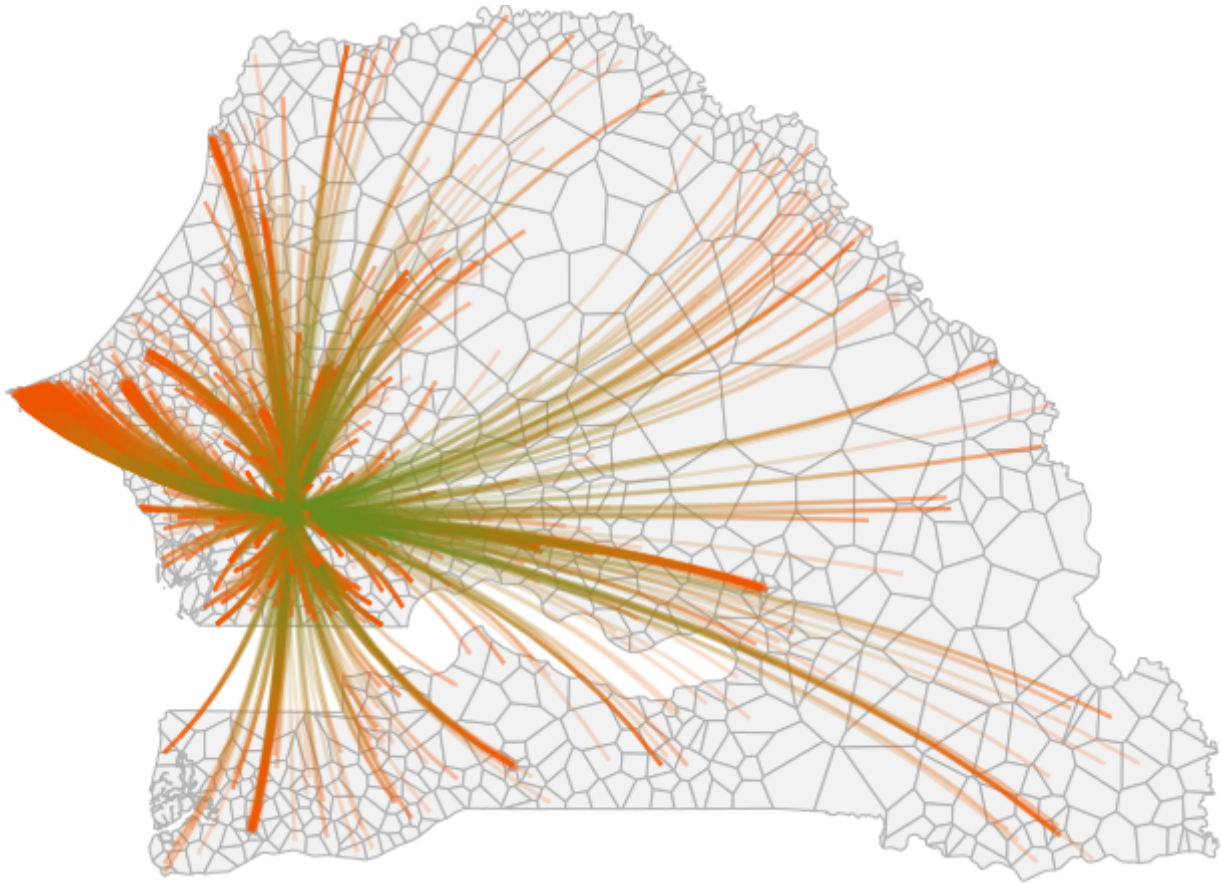
(4) Statistics

«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**.

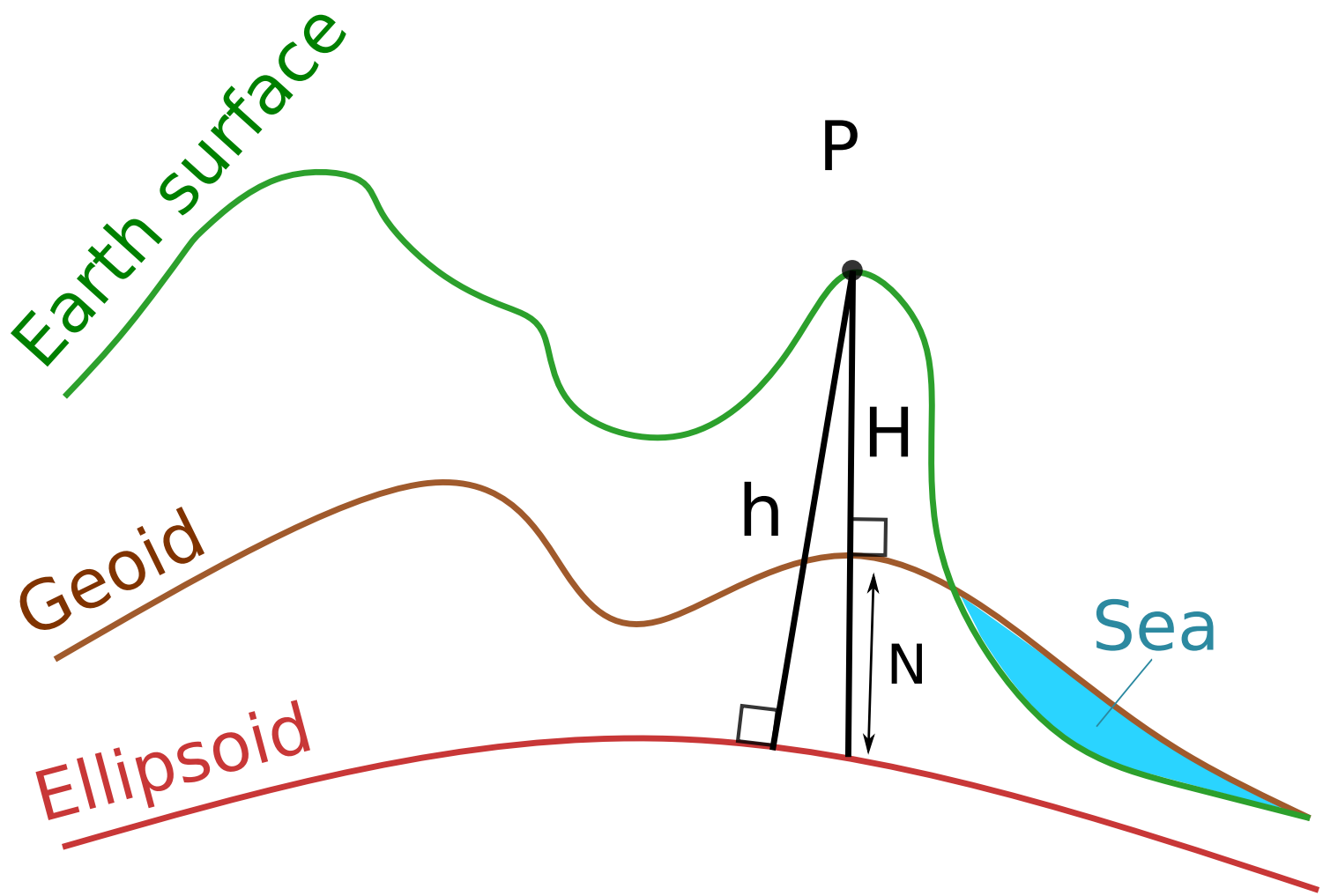


(5) Interactions

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

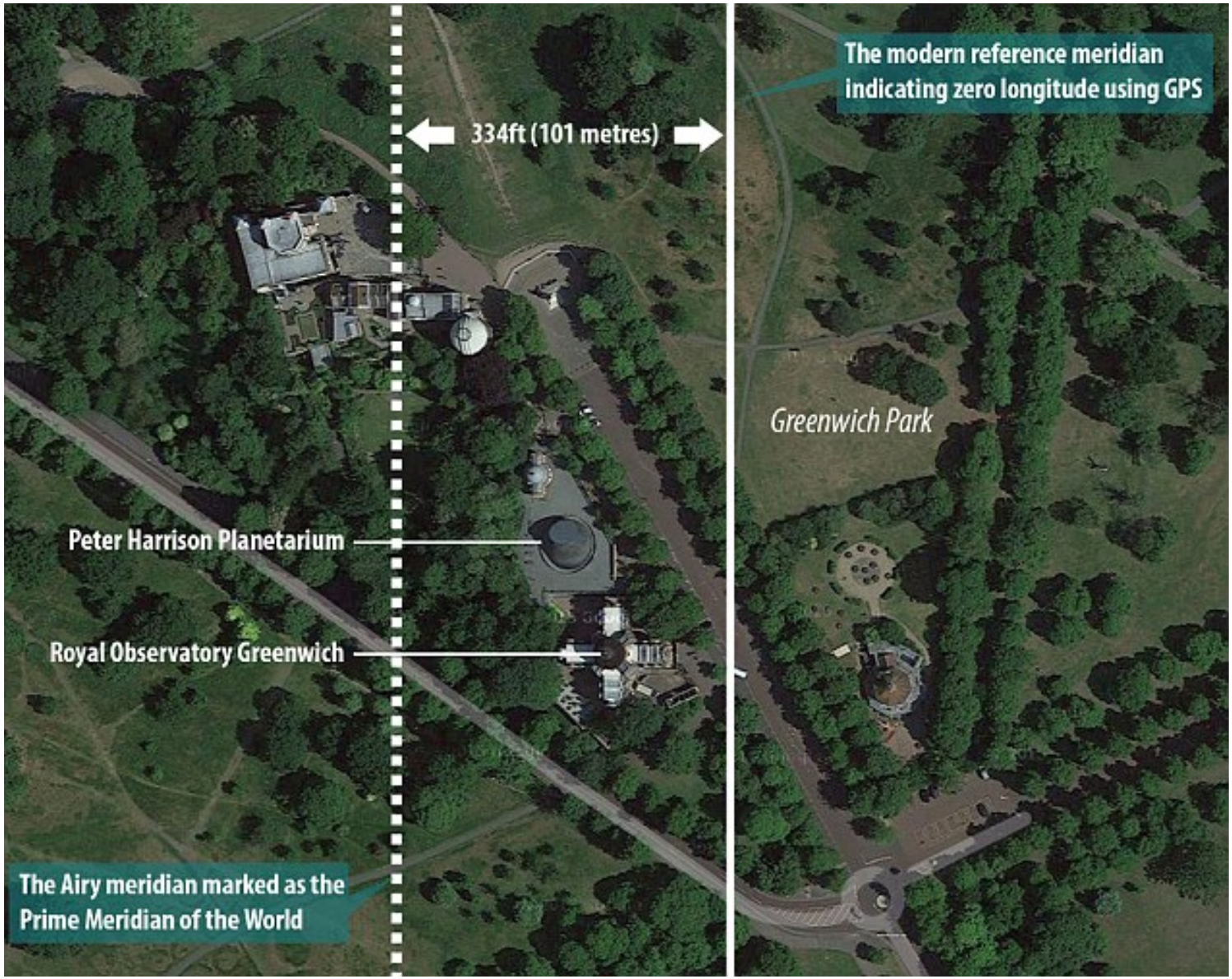
Coordinates, areas, distances

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**.

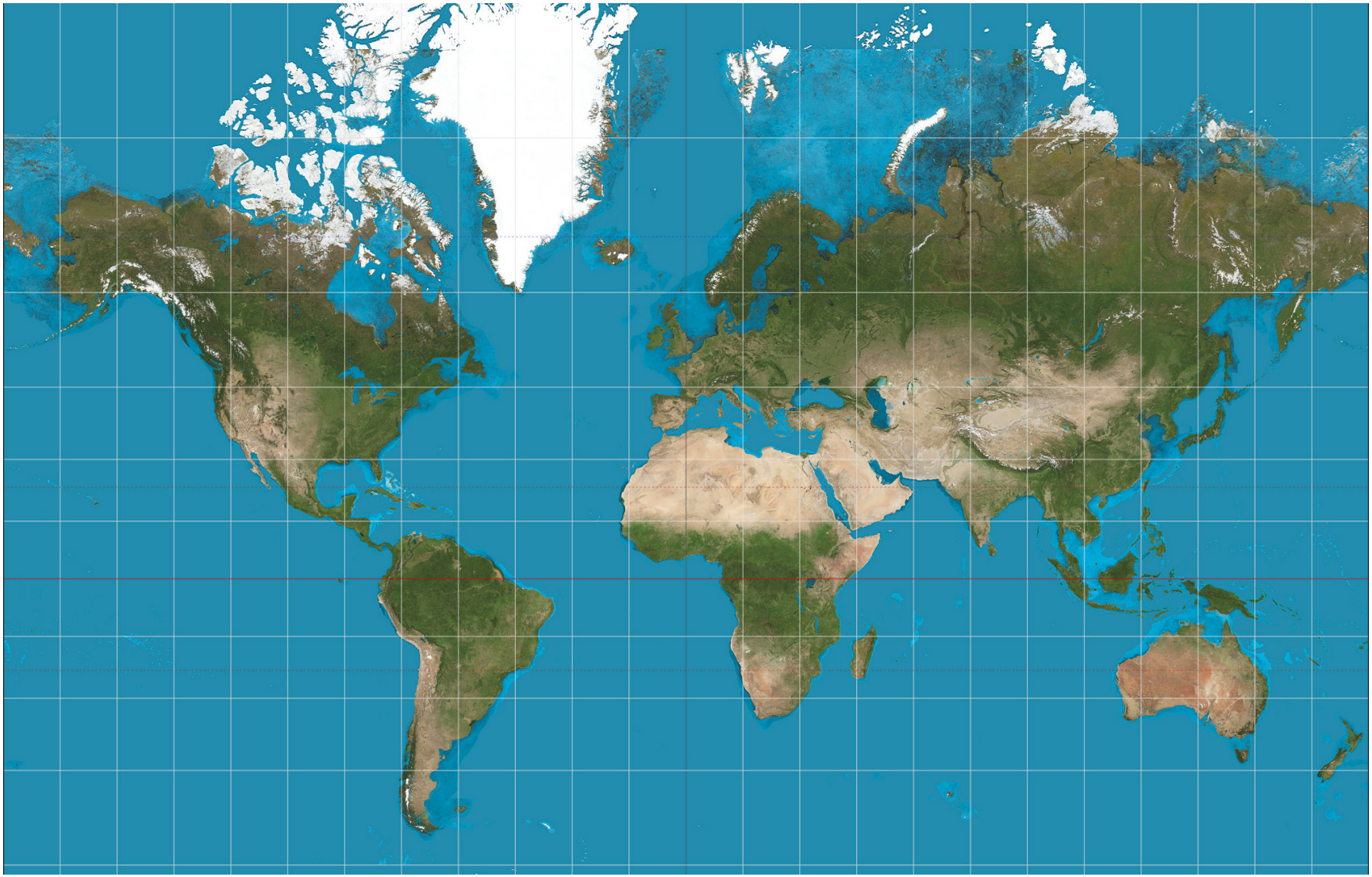
Some projections preserve some features

- **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

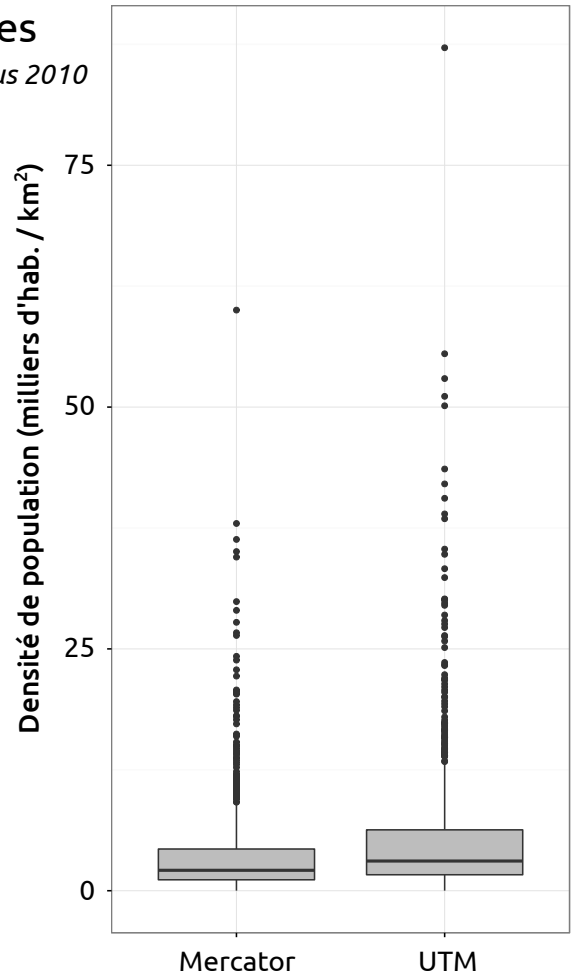


Source : Wikimedia, Mercator projection

Coordinates, areas, distances

Cape Town - Subplaces

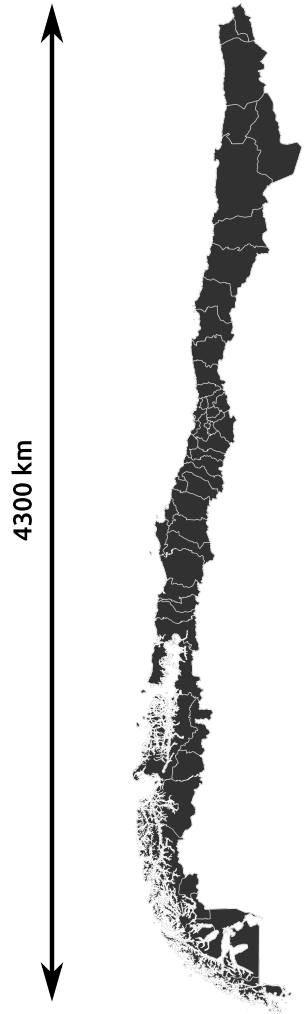
Source : Statistics South Africa, Census 2010



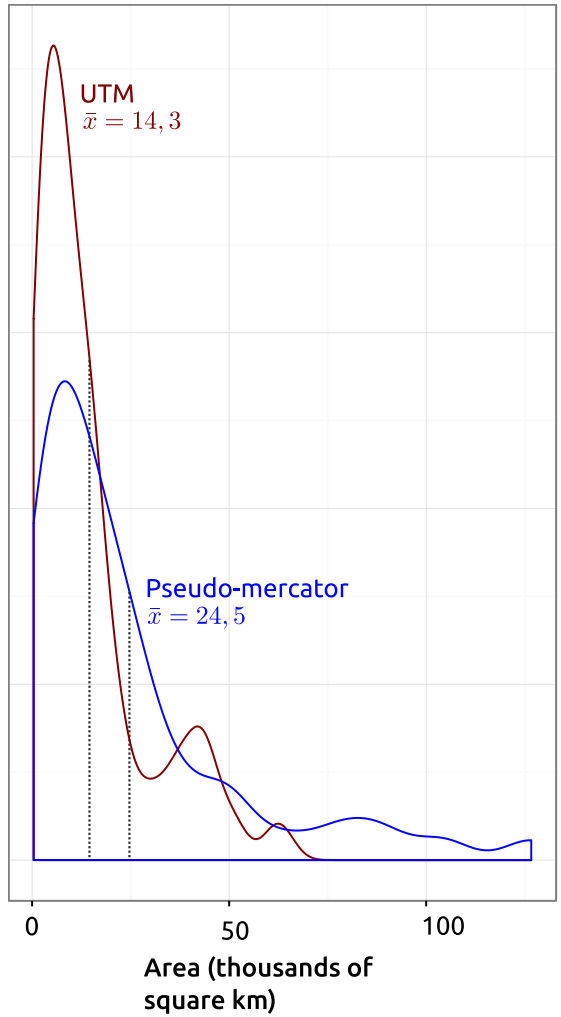
Med. 2111	**Med.** 3069
Moy. 3764	**Moy.** 5483

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

Coordinates, areas, distances



Chile's provinces area



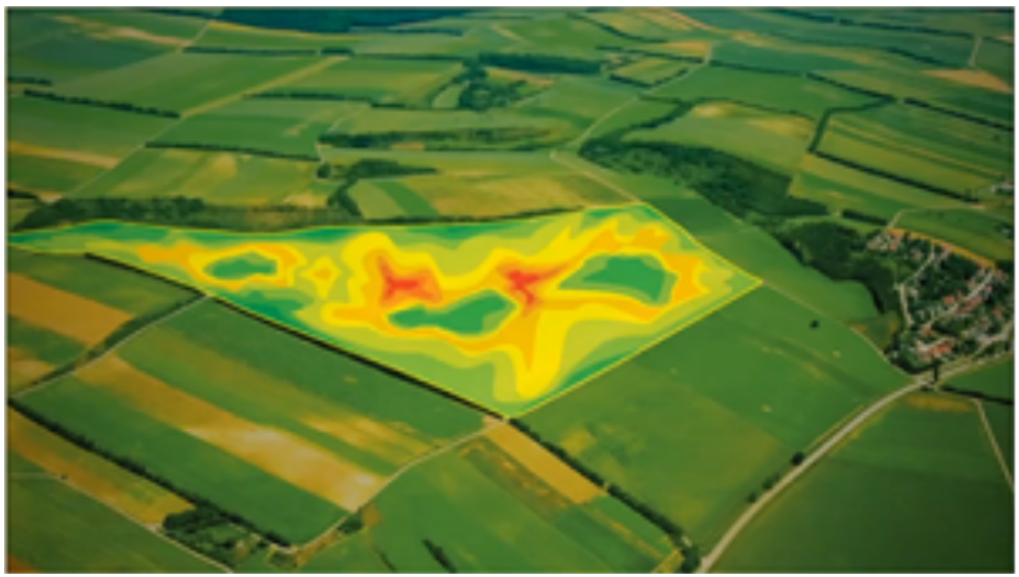
Coordinates, areas, distances

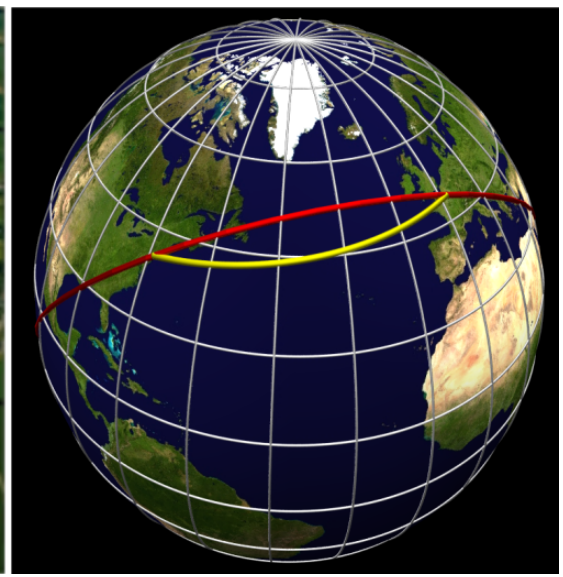
Basic precautions regarding projection :

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

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

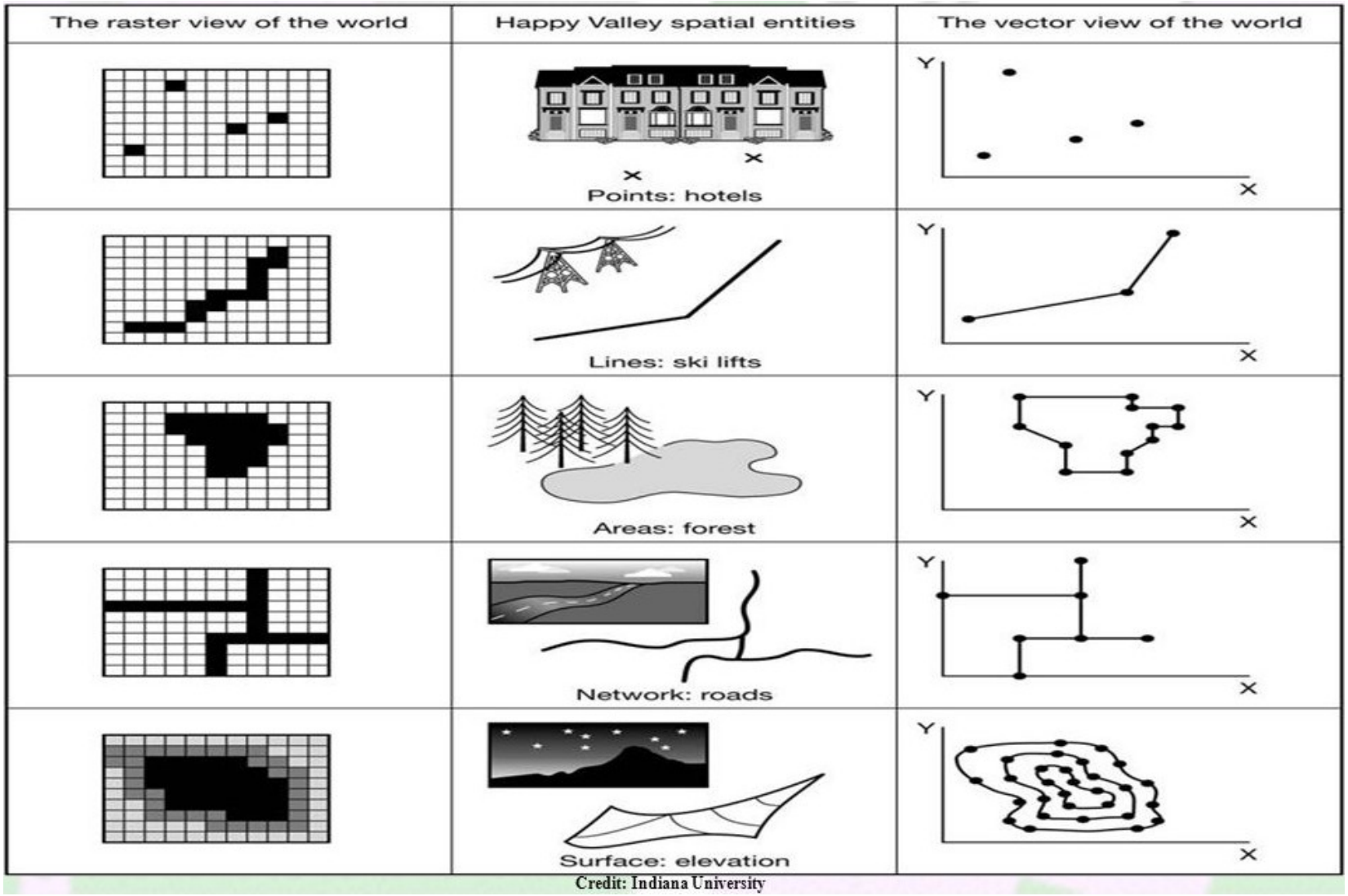
GPS-RTK (centimetric precision) or Great-circle distance ?





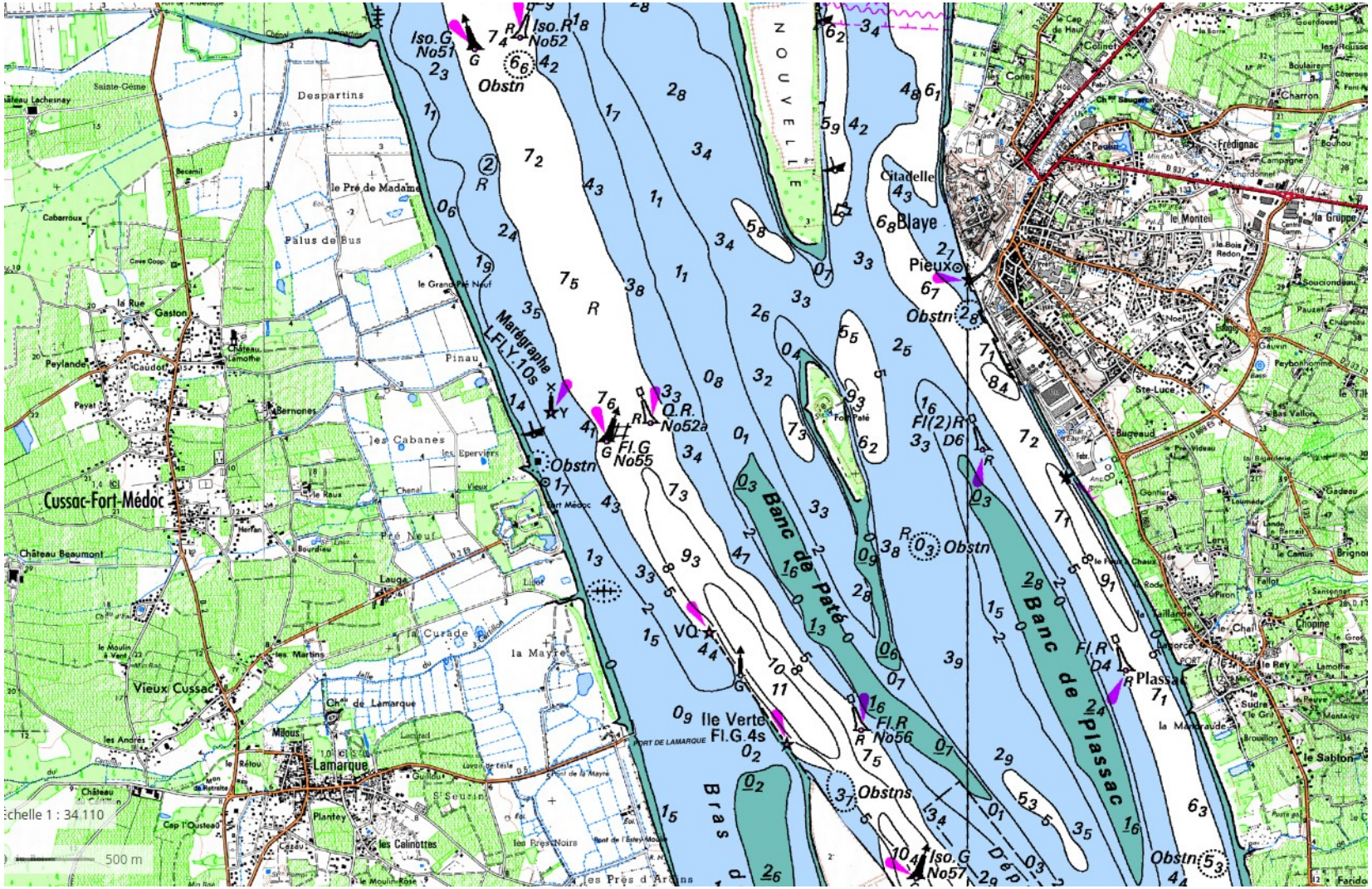
Modeling and representation

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



Modeling and representation

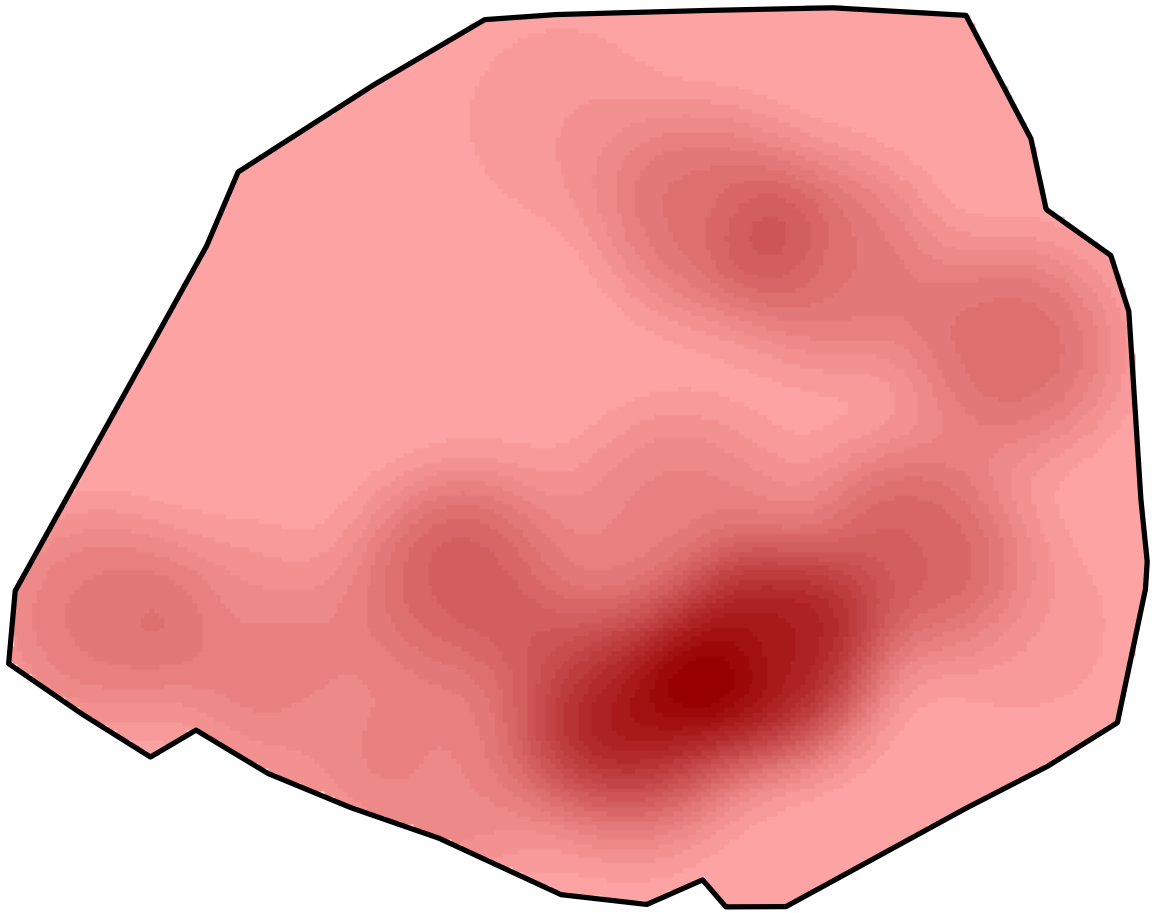
Objects and **Fields**.



Sources : IGN and SHOM

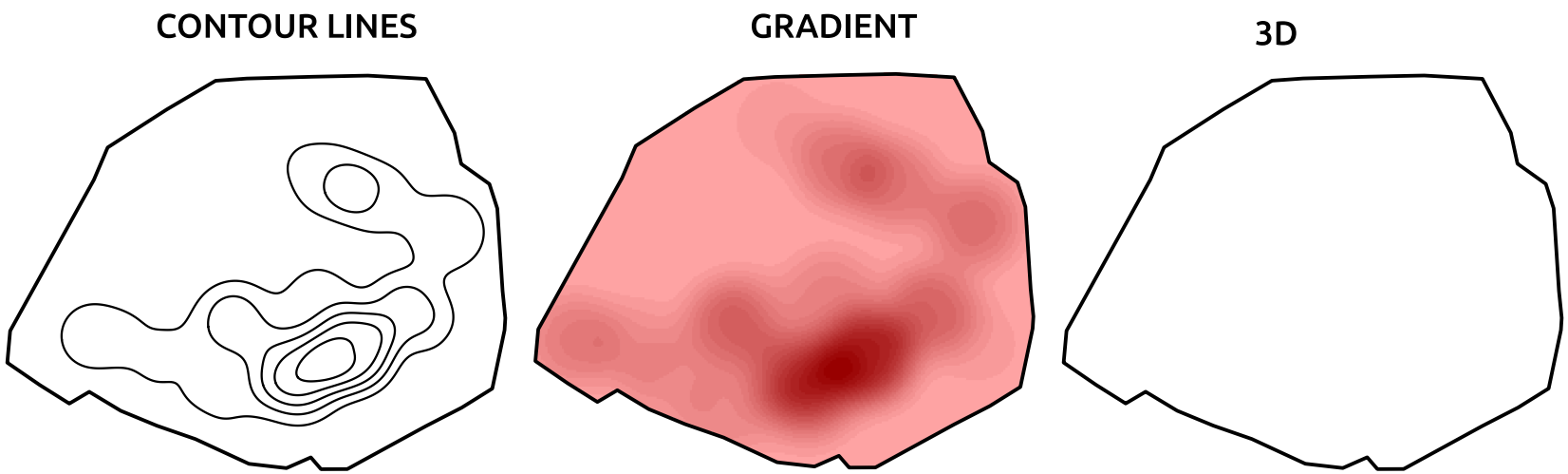
Fields

What does this field represent ?



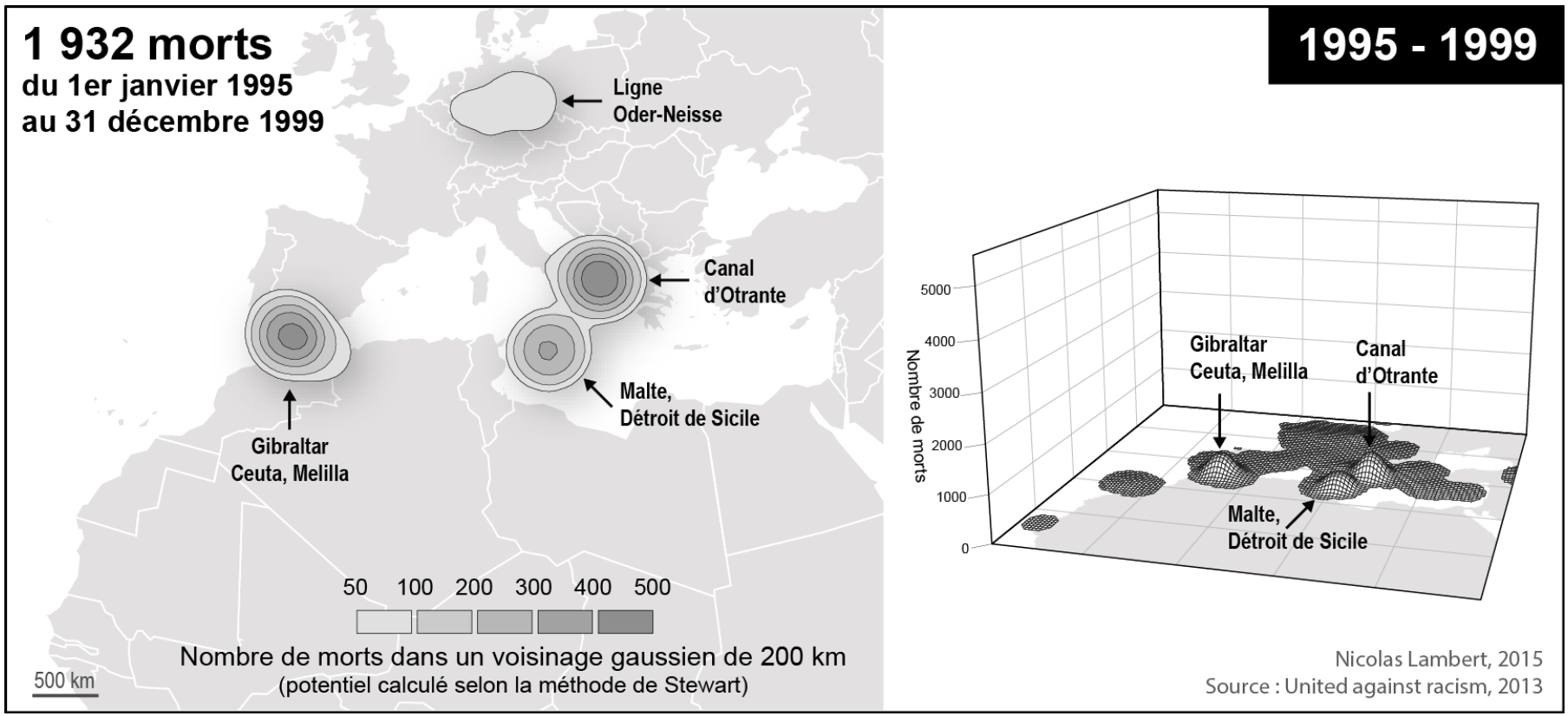
Fields

Representation modes



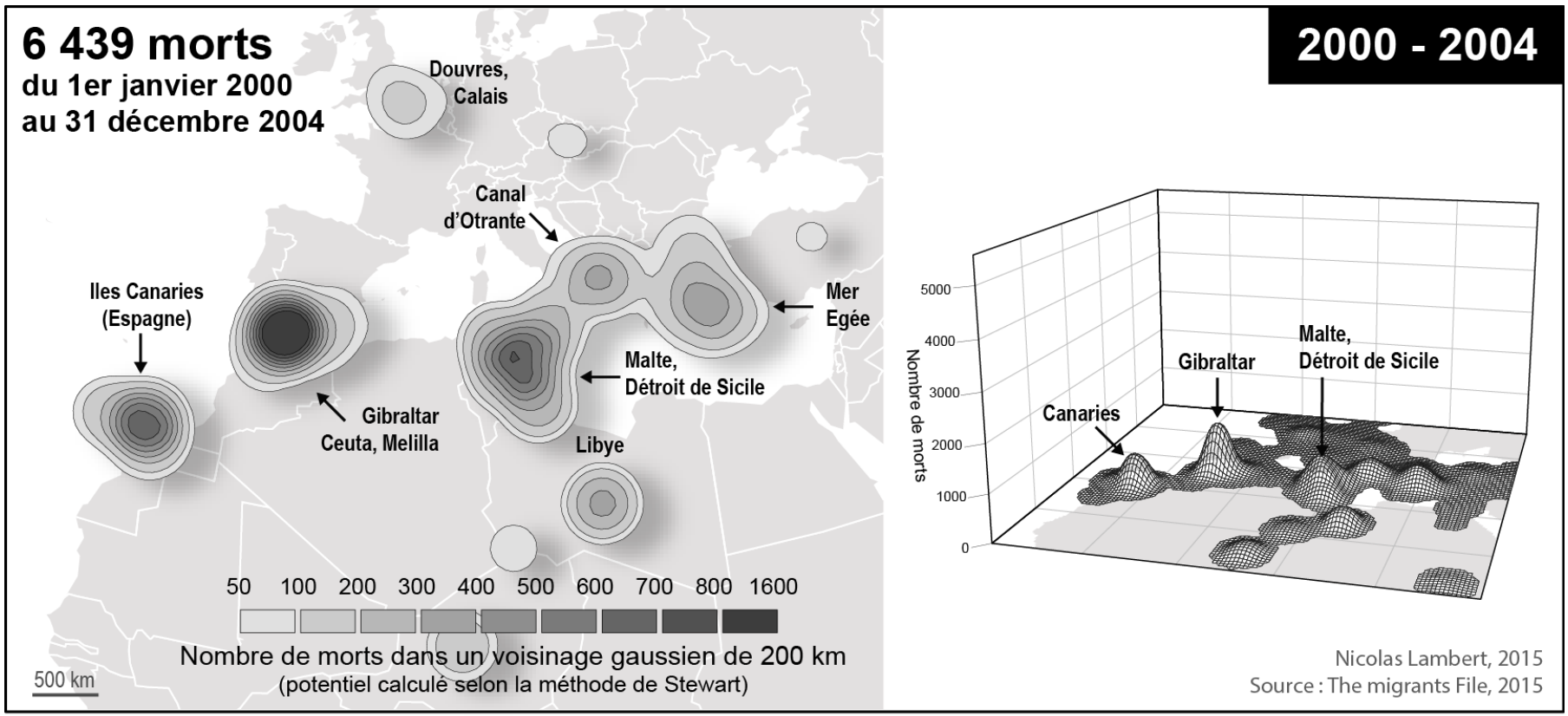
Fields

Representation modes examples



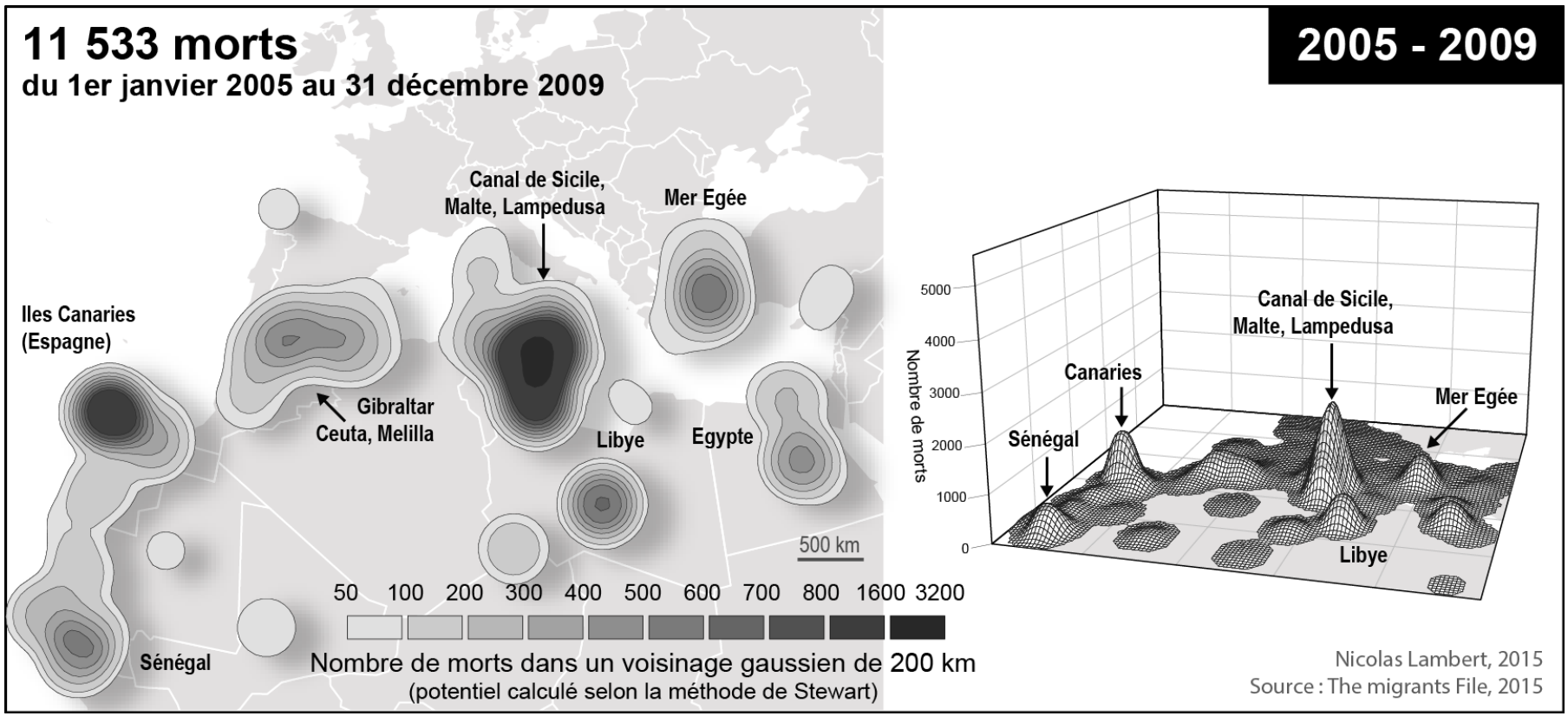
Fields

Representation modes examples



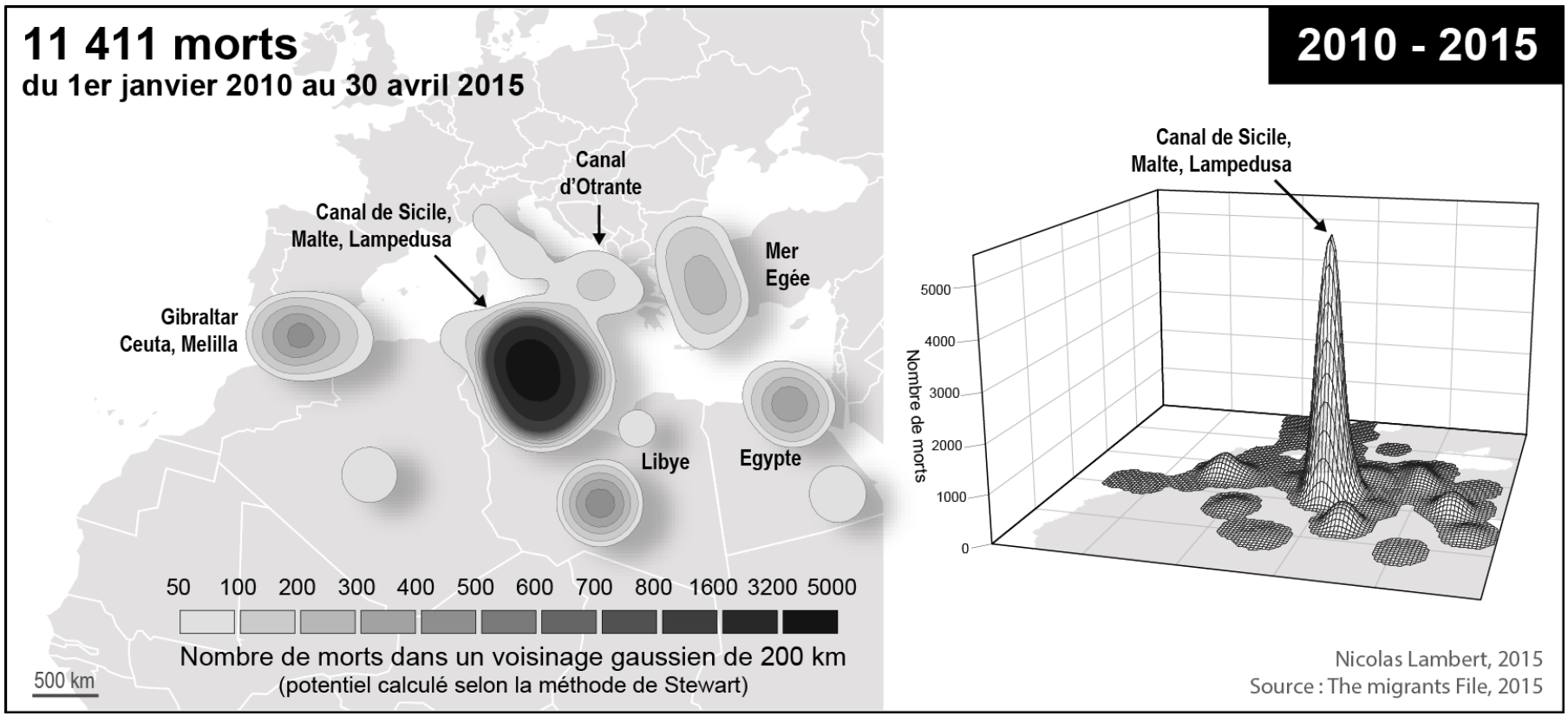
Fields

Representation modes examples



Fields

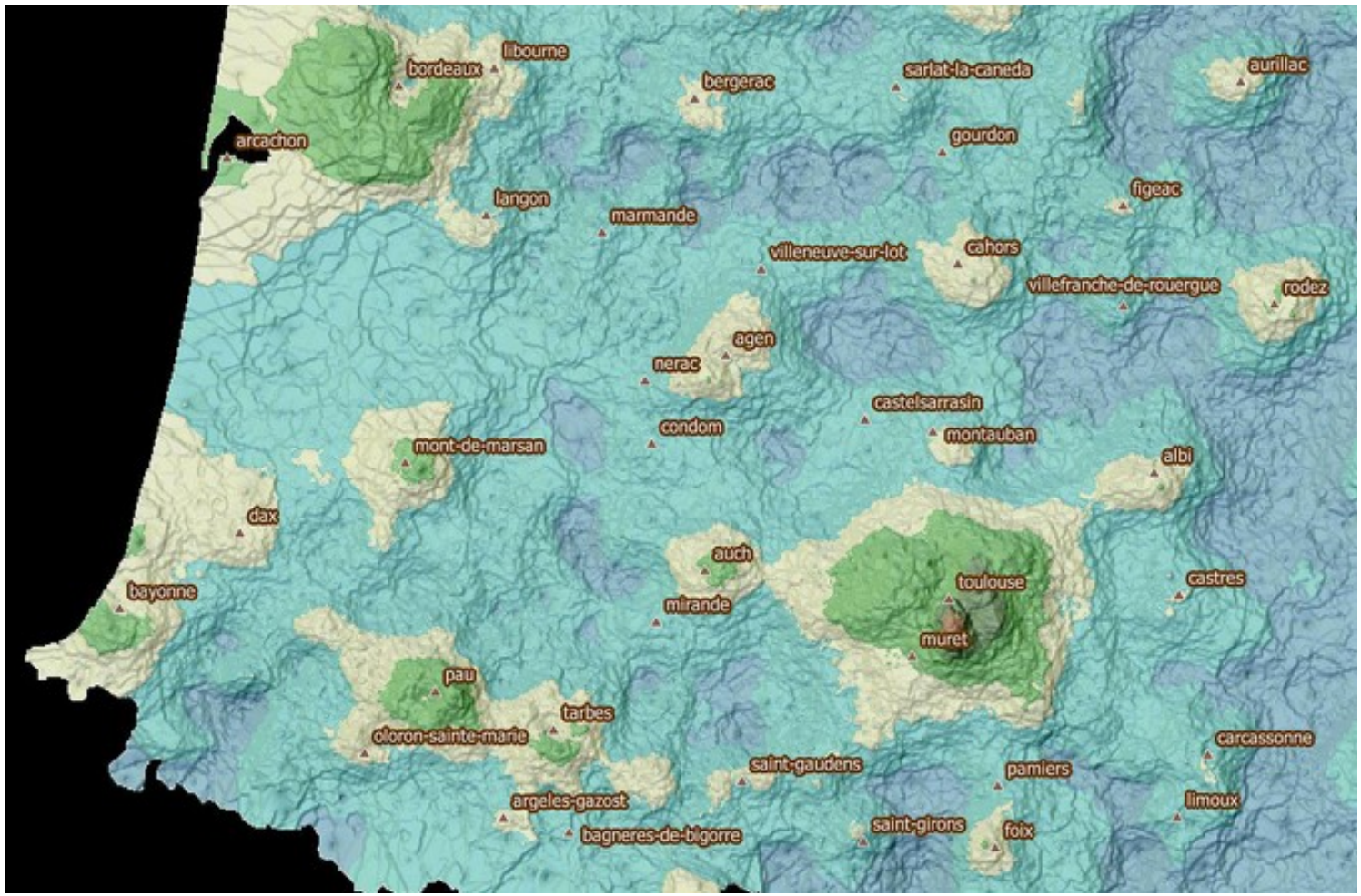
Representation modes examples



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 :

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

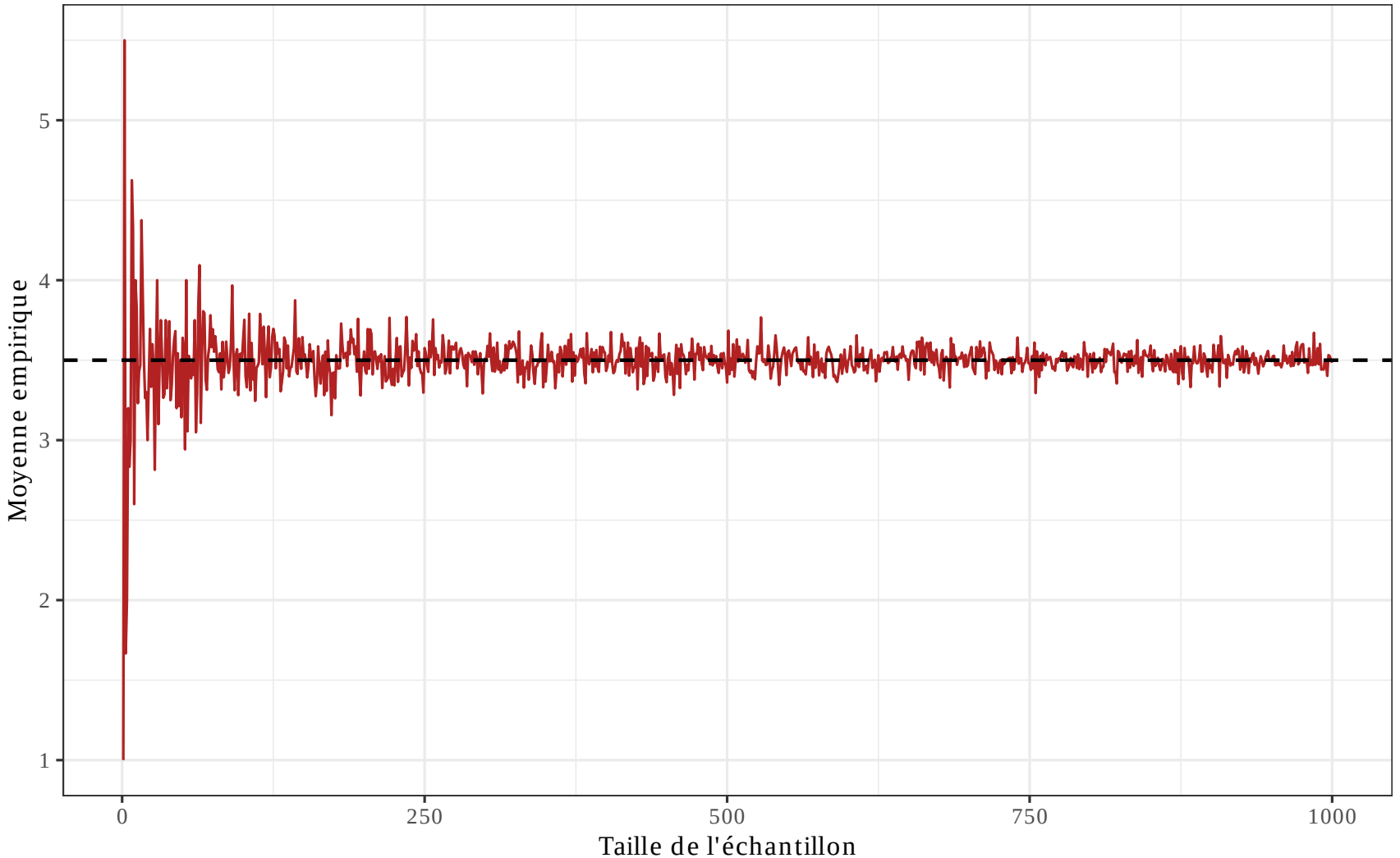
$$\mu = \frac{1+2+3+4+5+6}{6} = 3,5$$

(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

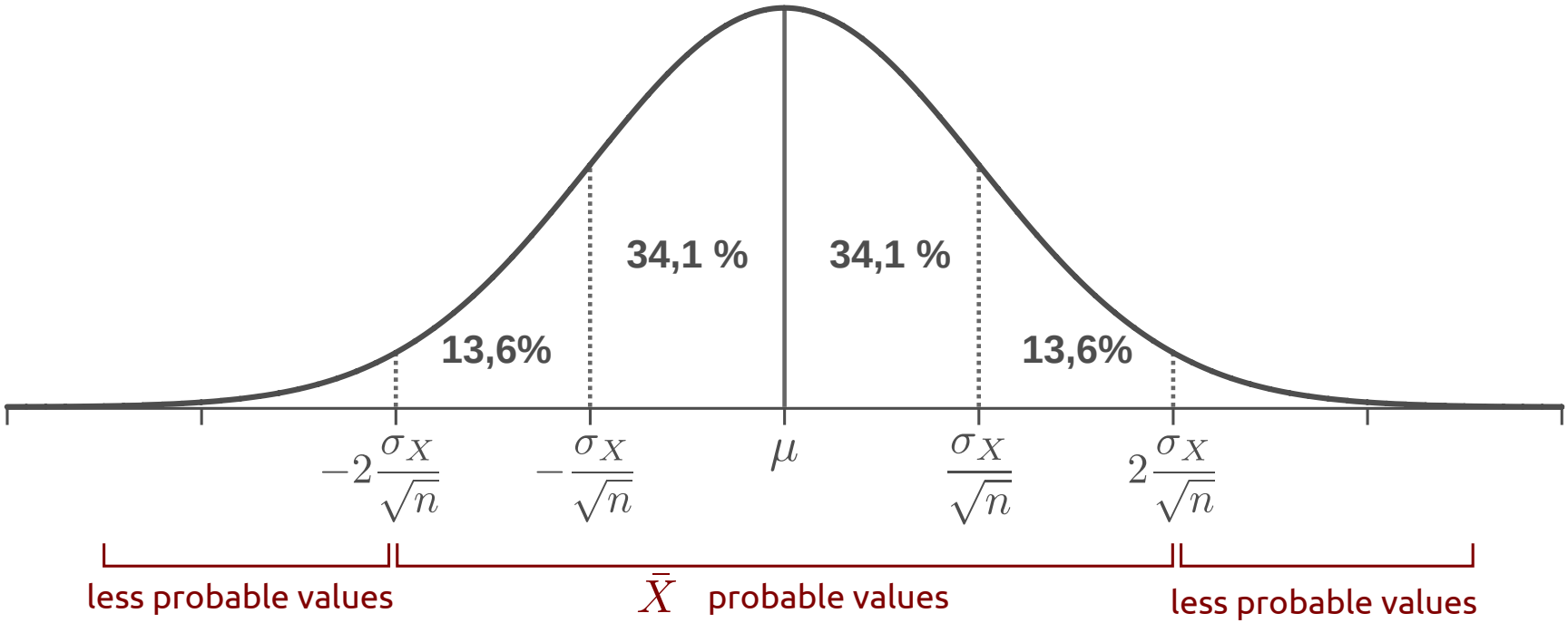
Example :

- 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**.

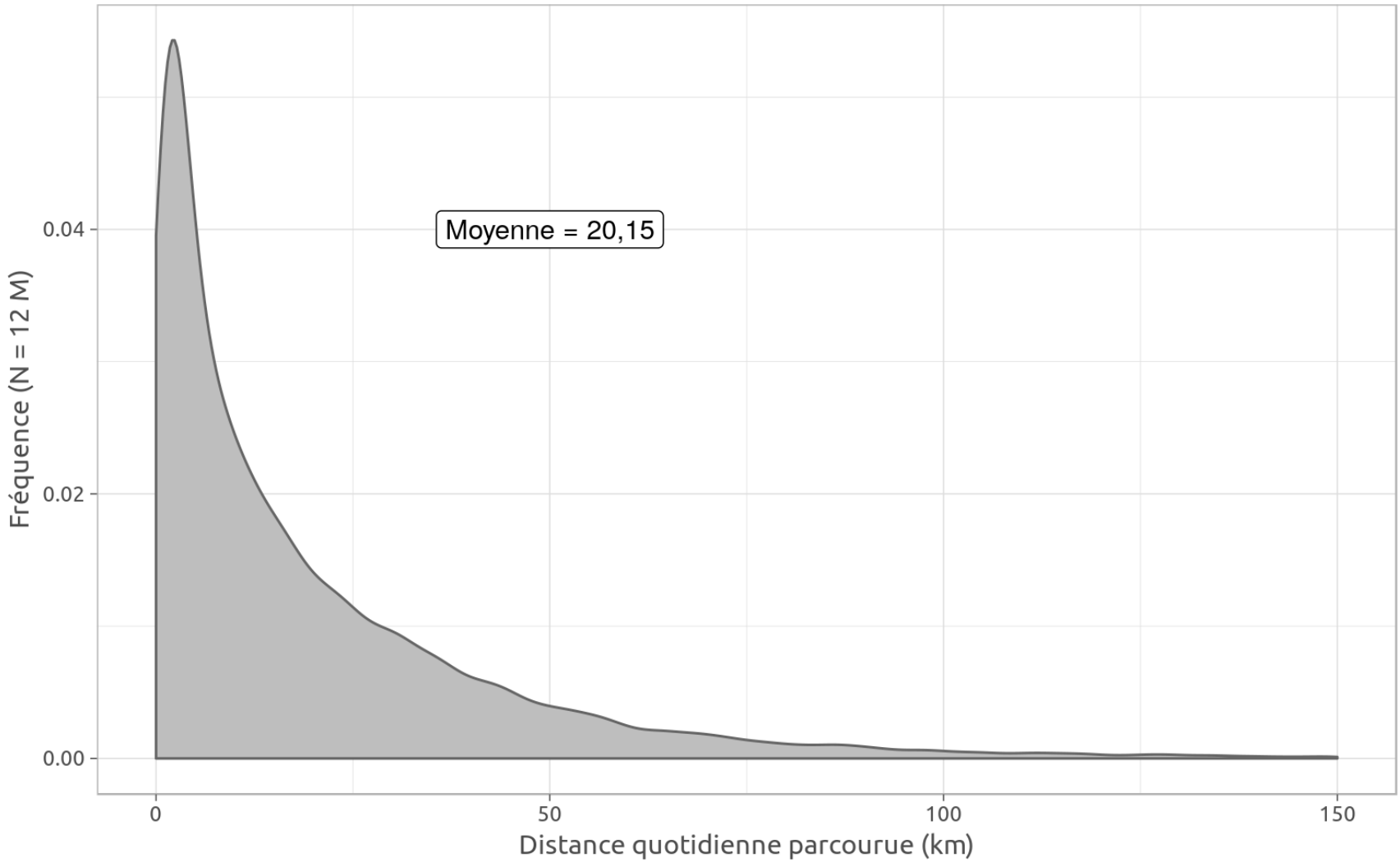
$\rightarrow$ This is **central limit theorem** (**CLT**).

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

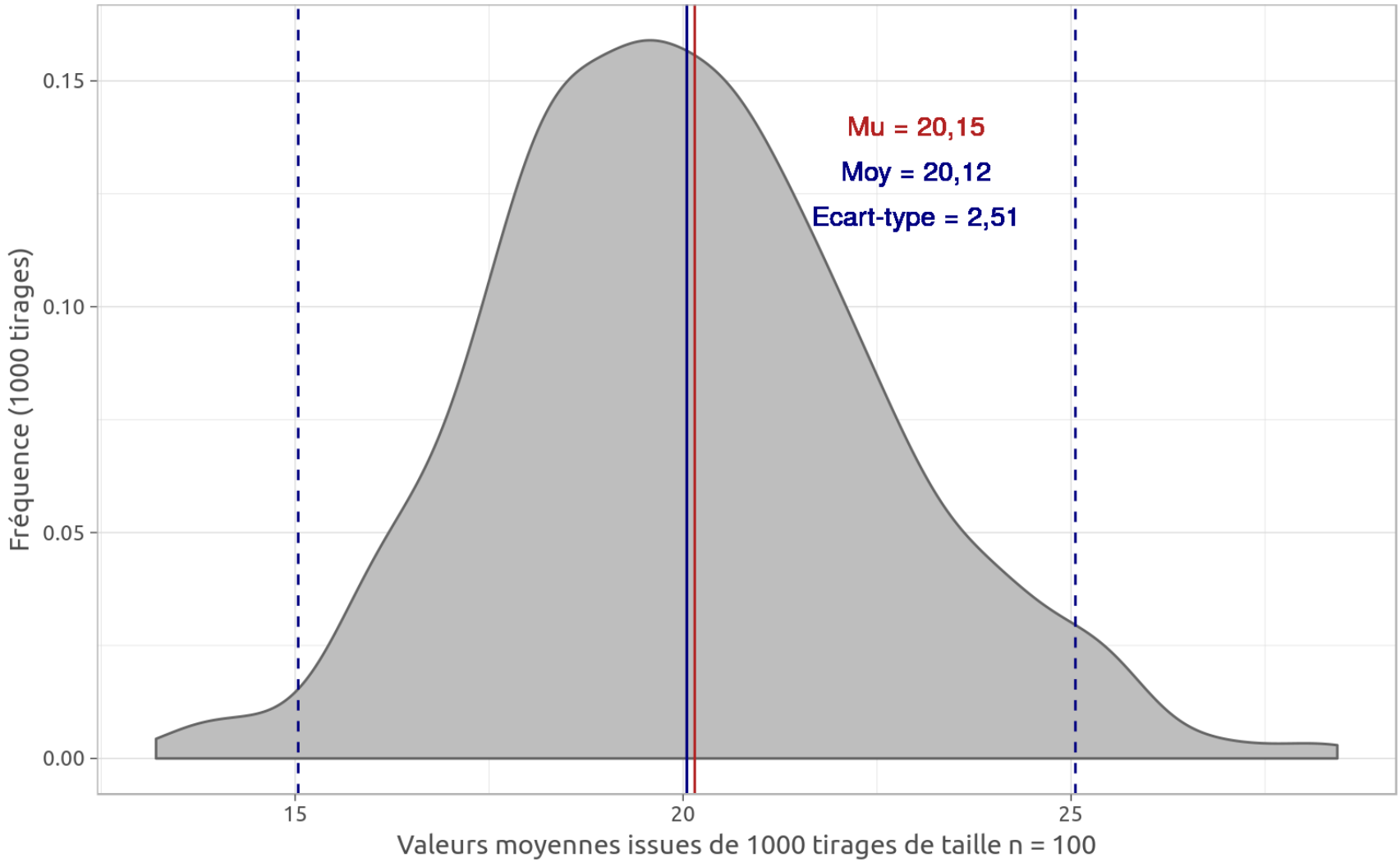
Daily travel distance for Île-de-France people



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

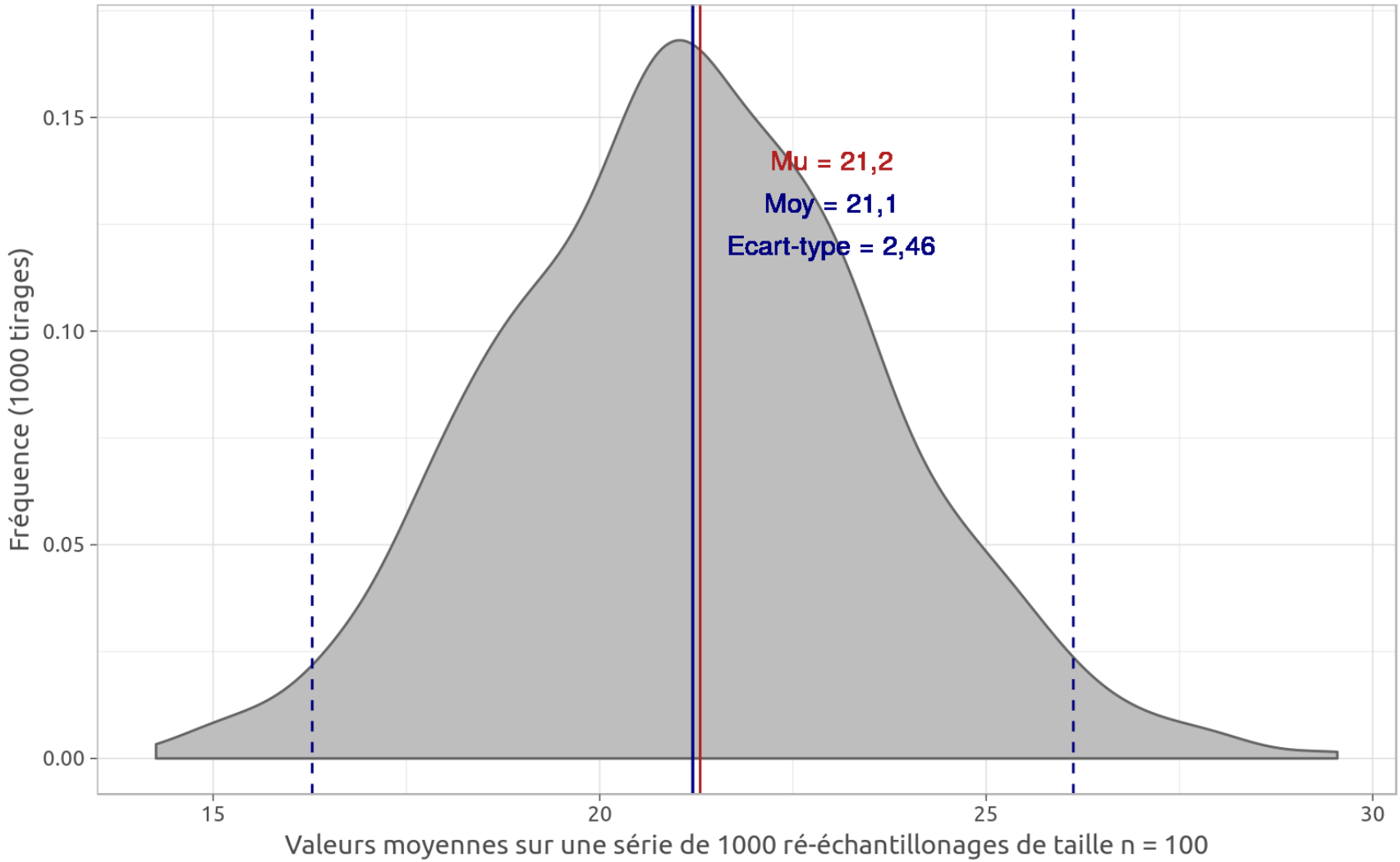
Bootstrap

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.

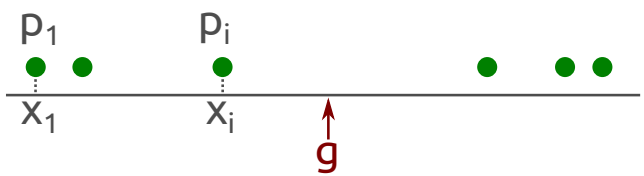
Spatial distribution parameters

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

General formula
-> squared distances mean

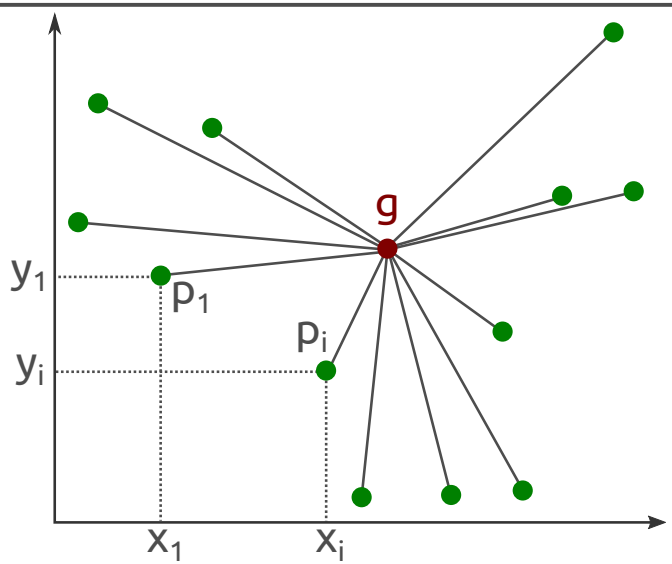
$$I = \frac{1}{n} \sum_{i=1}^{n} d^2(p_i, g)$$

1D



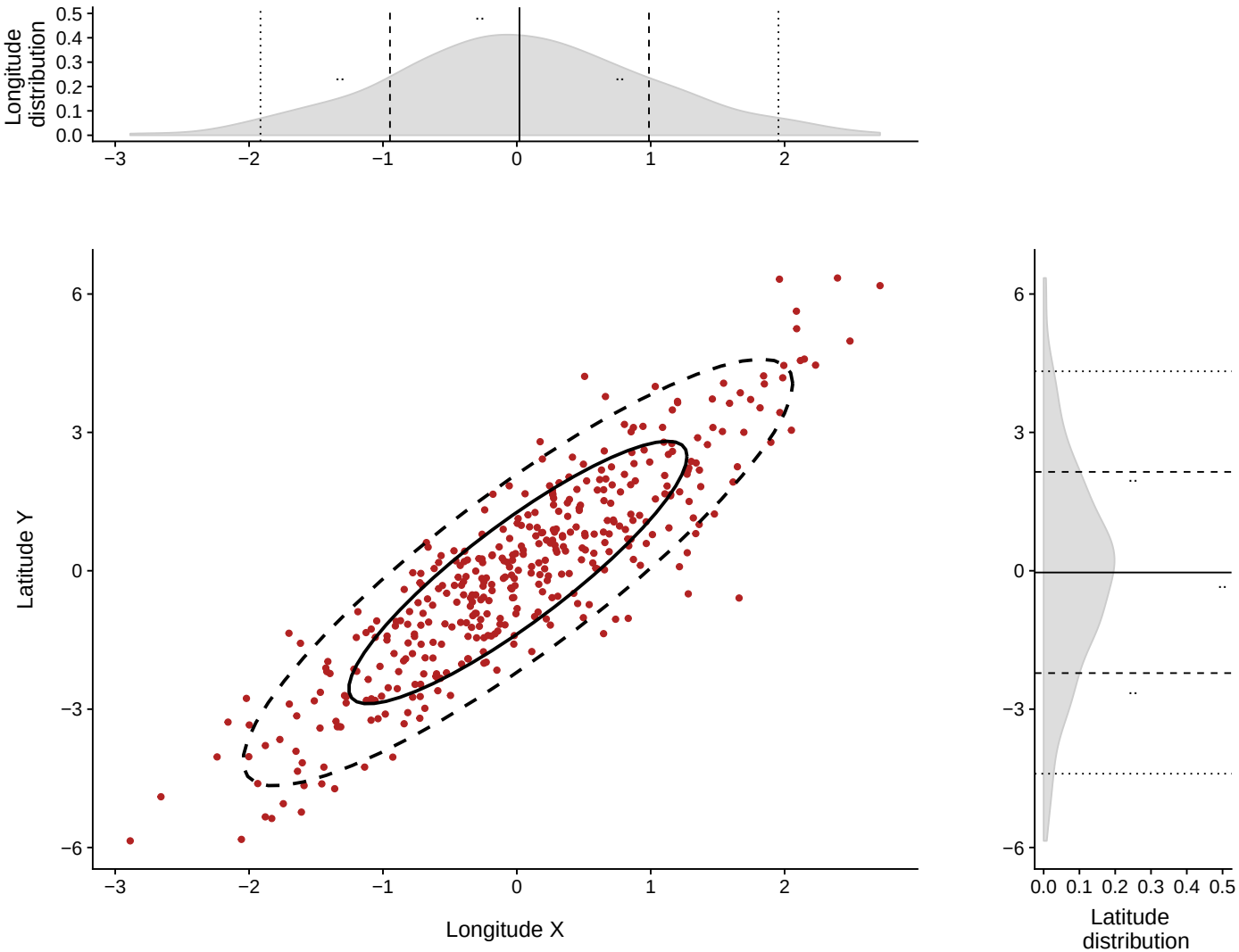
$$I = \frac{1}{n} \sum_{i=1}^{n} (x_i - x_g)^2$$

2D

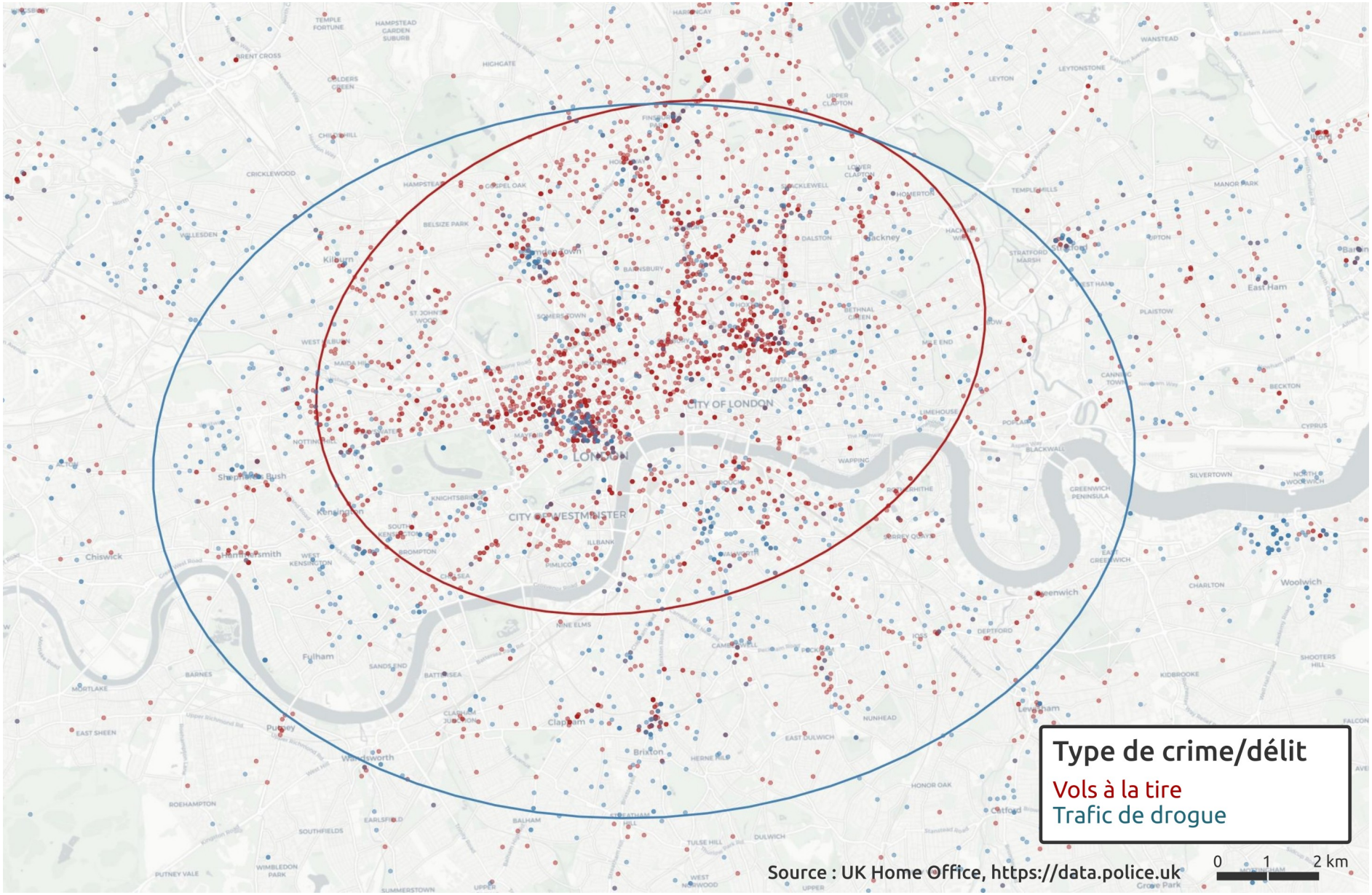


$$I = \frac{1}{n} \sum_{i=1}^{n} \left[(x_i - x_g)^2 + (y_i - y_g)^2 \right]$$

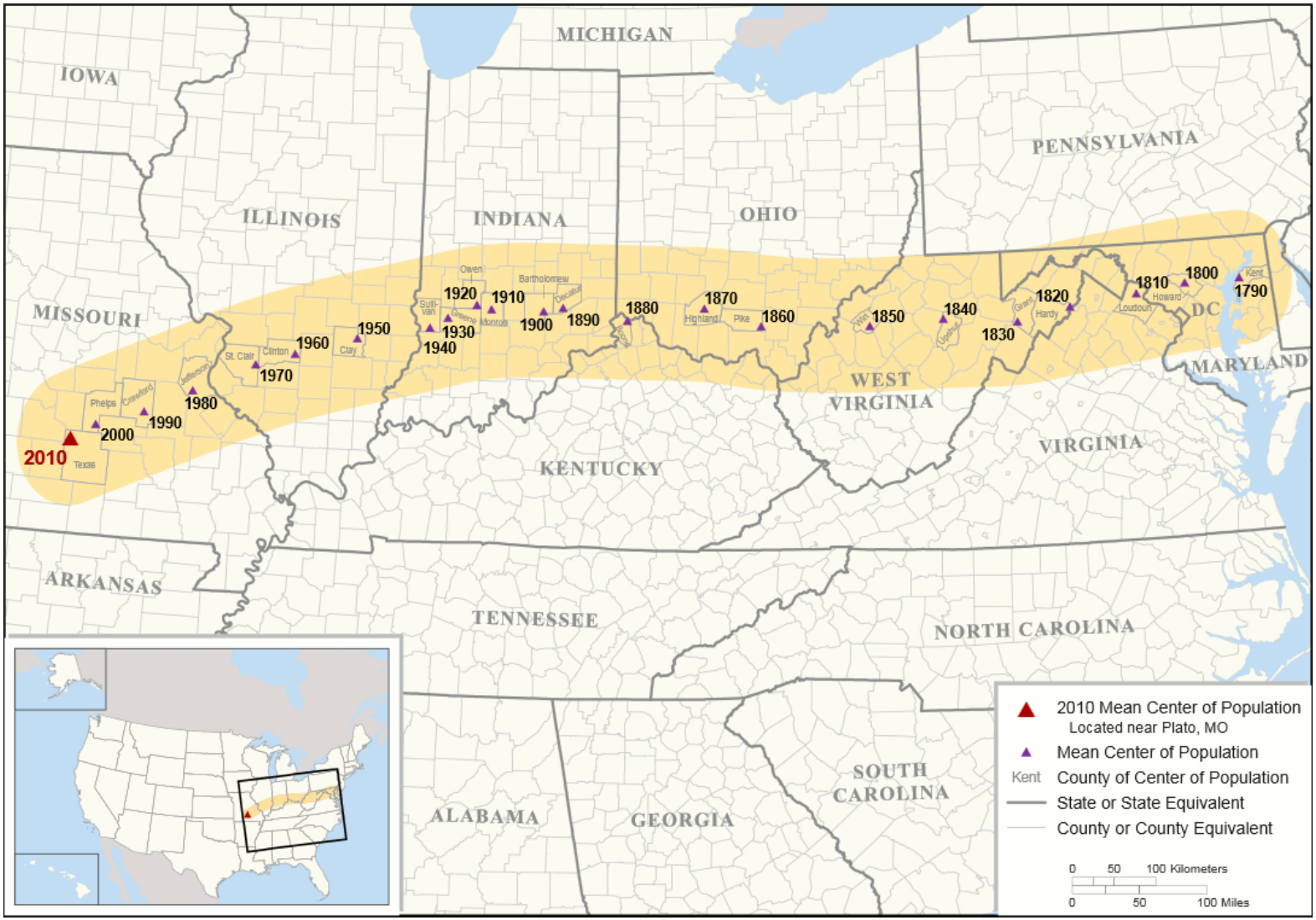
Spatial distribution parameters



Spatial distribution parameters



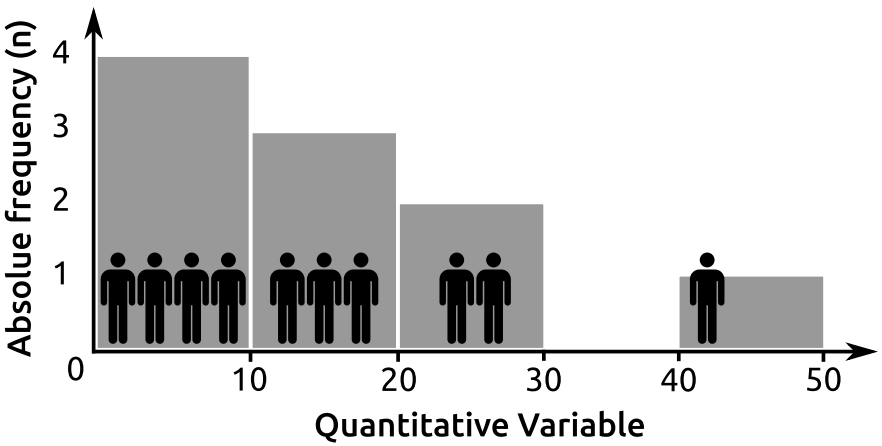
Spatial distribution parameters

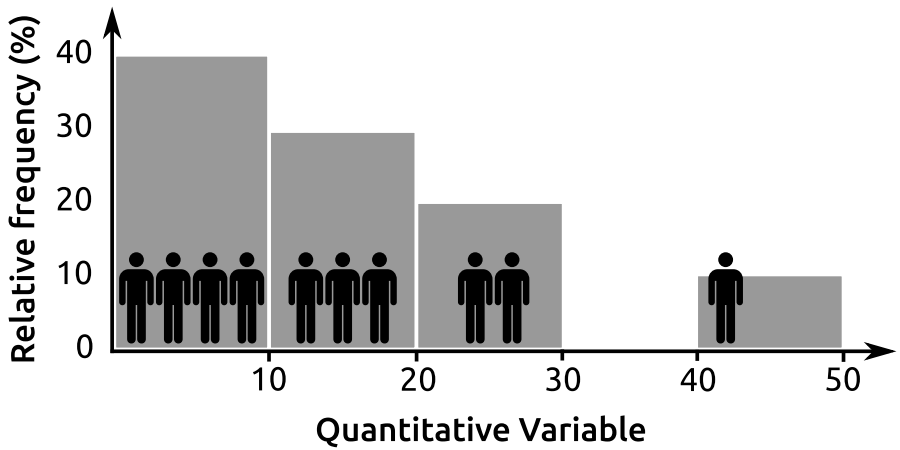


Source : US Census, `https://www.census.gov/geo/reference/centersofpop.html`

1D distribution graph (discrete)

Histogram

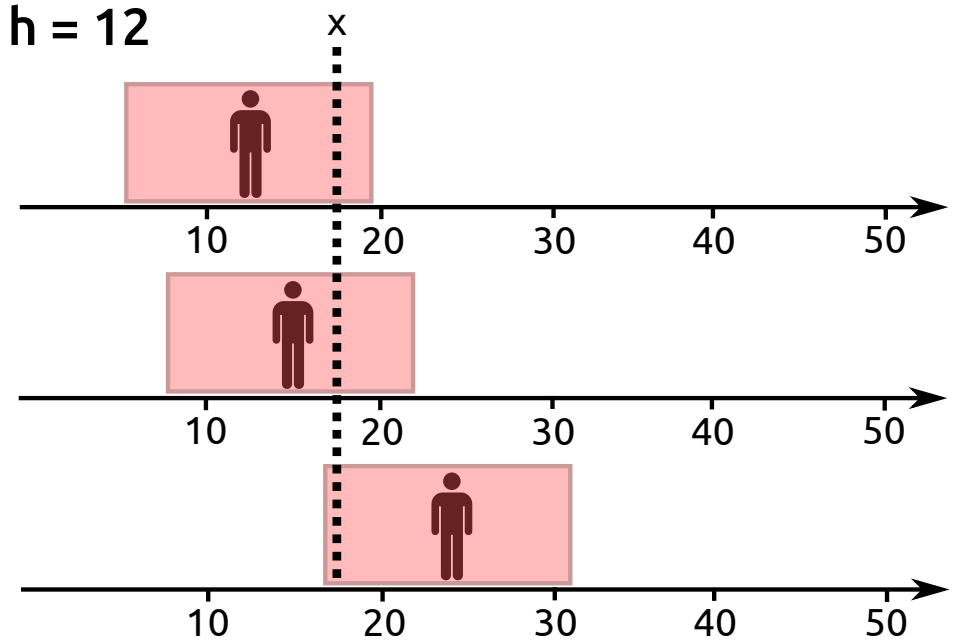




$\rightarrow$ histogram estimated density is **discrete** by construction.

Distribution graph (Parzen)

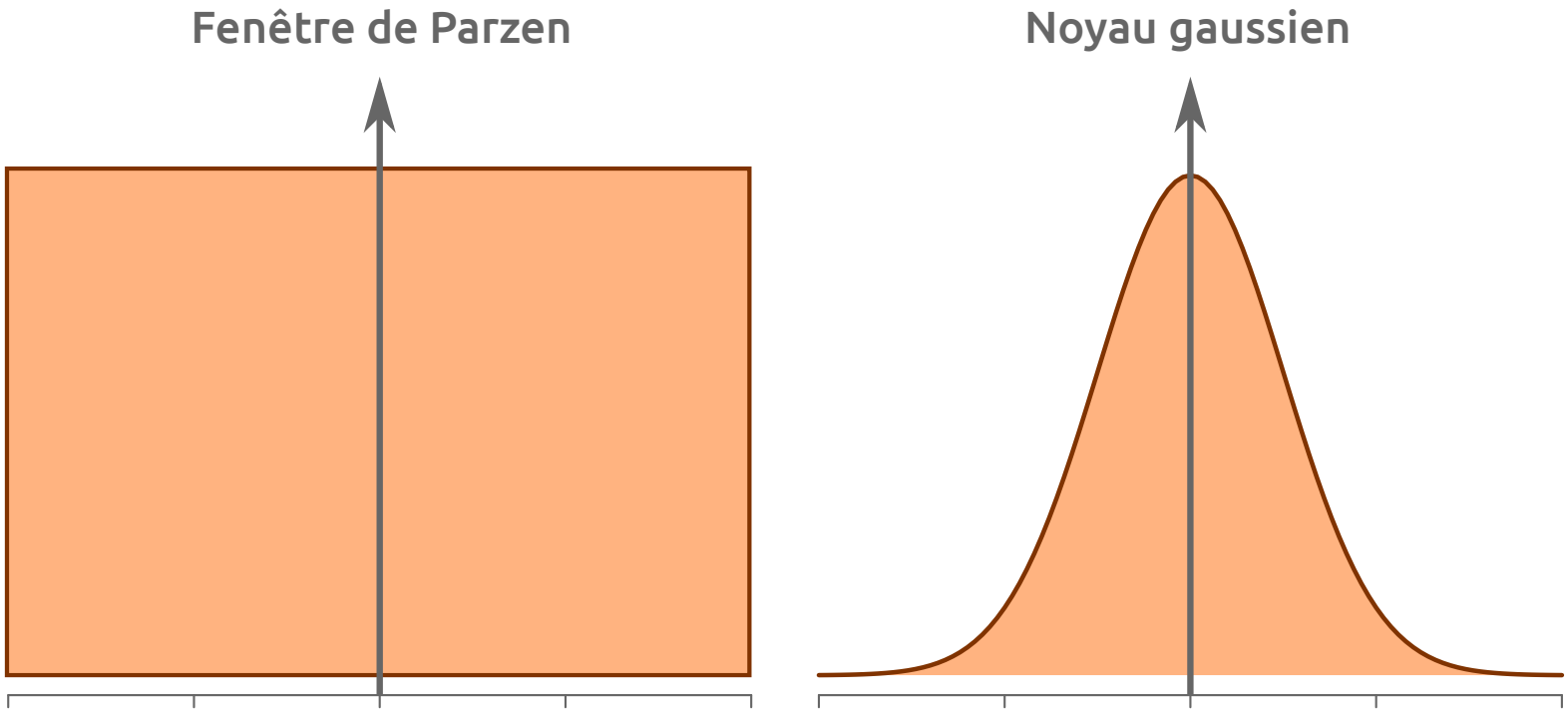
Histogram generalization : Parzen window



$$D(x) = \frac{1}{3 \times 12}(1 + 1 + 1) = \frac{1}{12}$$

Distribution graph (continuous)

Parzen Generalization : Gaussian Kernel



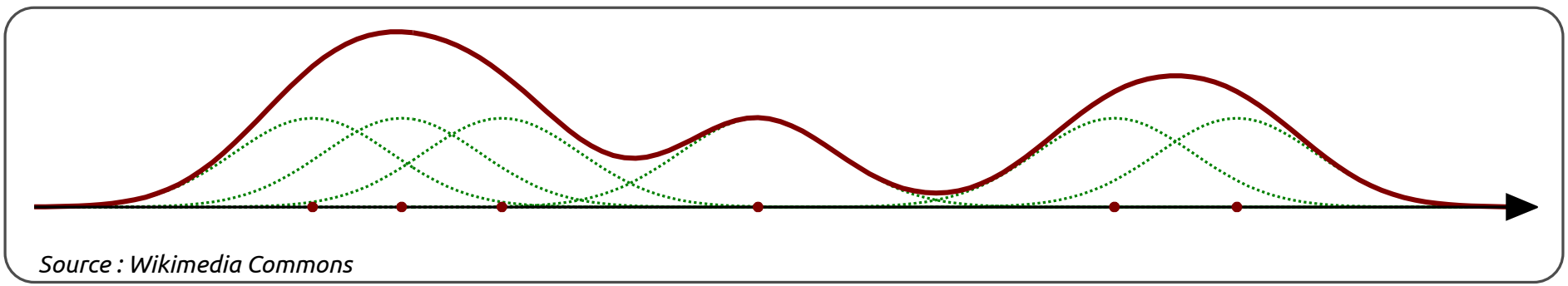
1D distribution graph (continuous)

Kernel Density Estimate

General idea : density for a value x is estimated by the proportion of observation *near* x

«Near» is defined by a certain window described by a **kernel function** (usually gaussian).

Contribution of each observation within the window is given by kernel function value taken for each x. *implies* smoothing



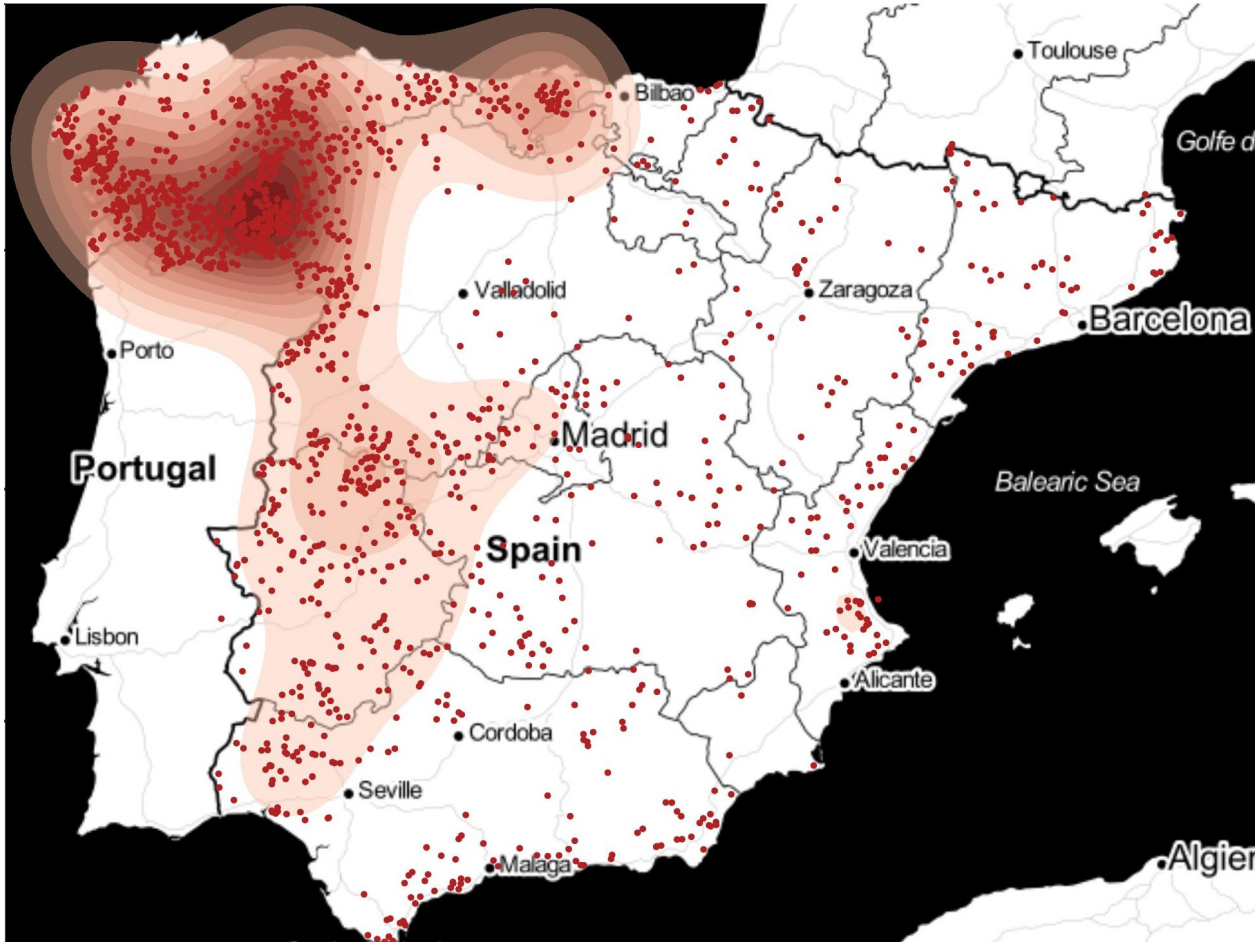
Source : Wikimedia Commons

«Close observations contribute strongly, far ones contribute slightly »

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

Distribution mapping

Density obtained by (KDE) in **2 Dimensions**.



Distribution testing

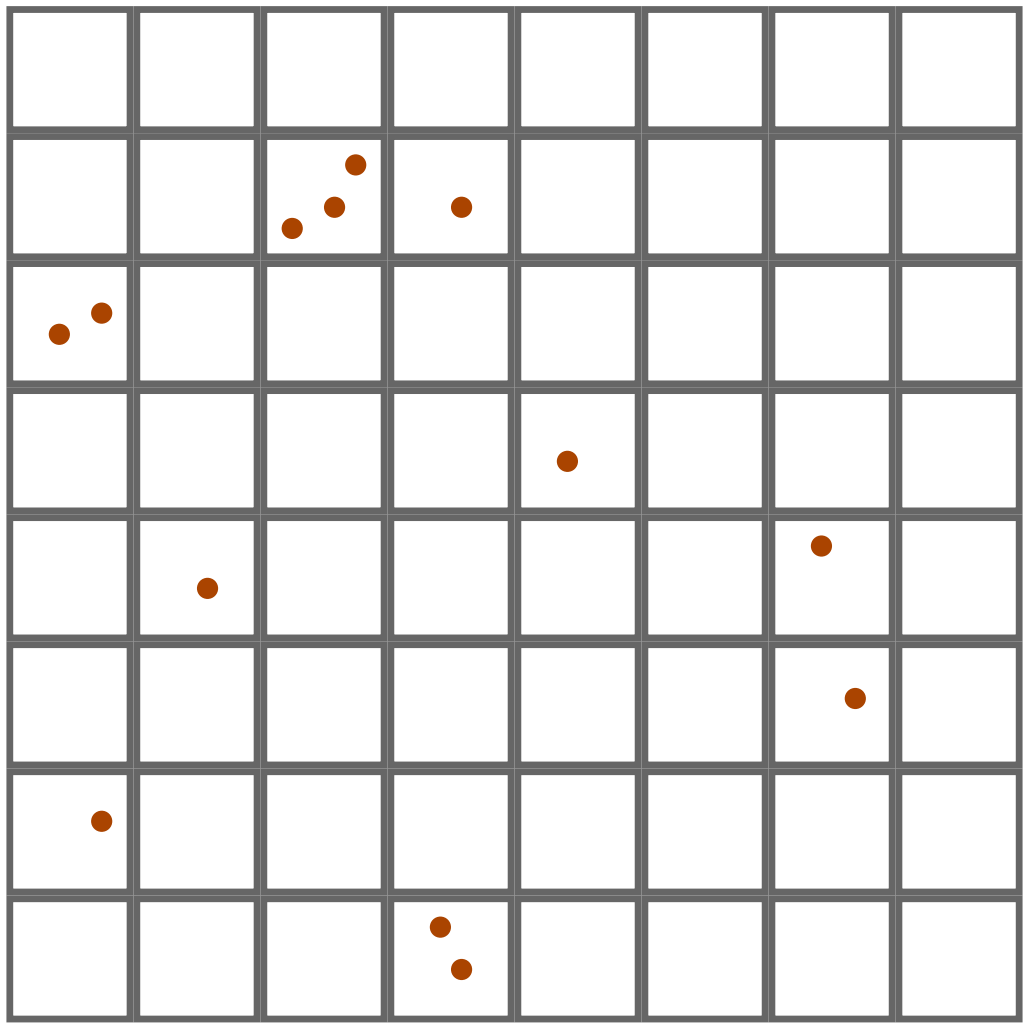
May this distribution have been generated by a stochastic process ?

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.

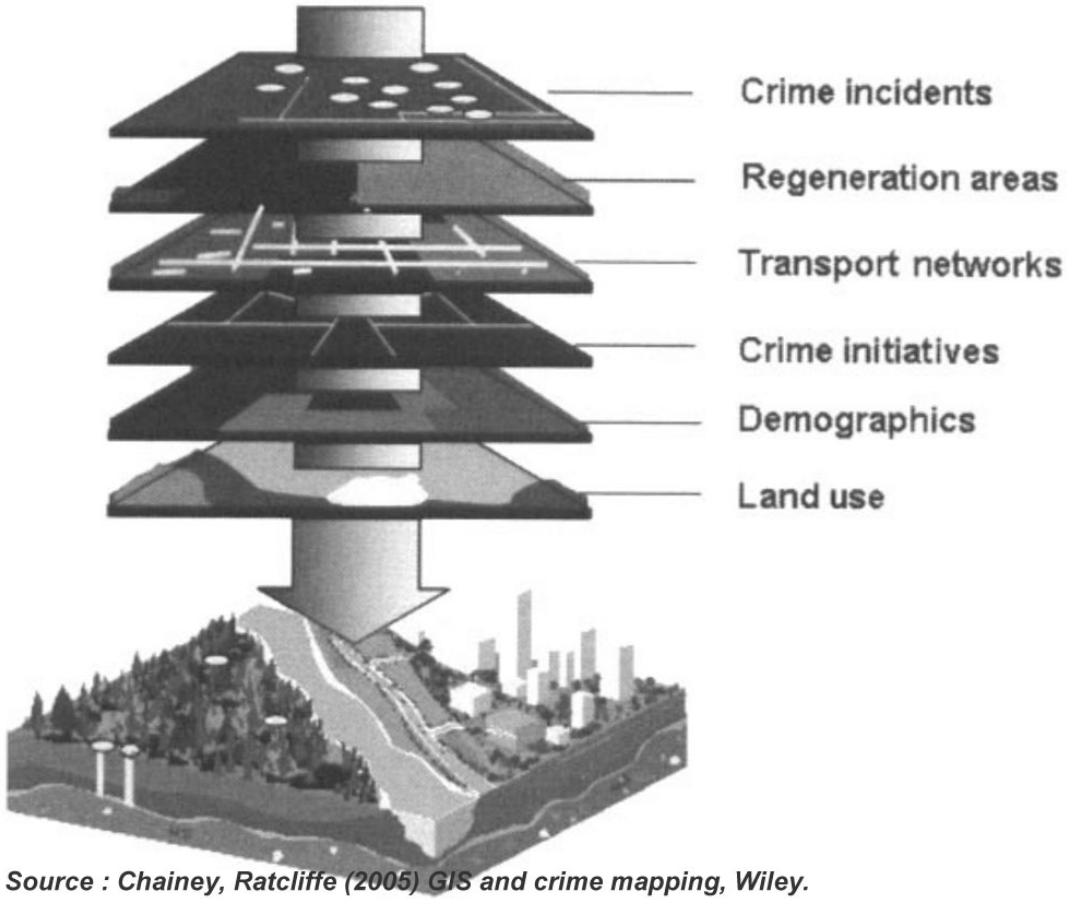
Distribution testing

May this distribution have been generated by a stochastic process ?



Distribution testing

May this distribution have been generated by a stochastic process ?

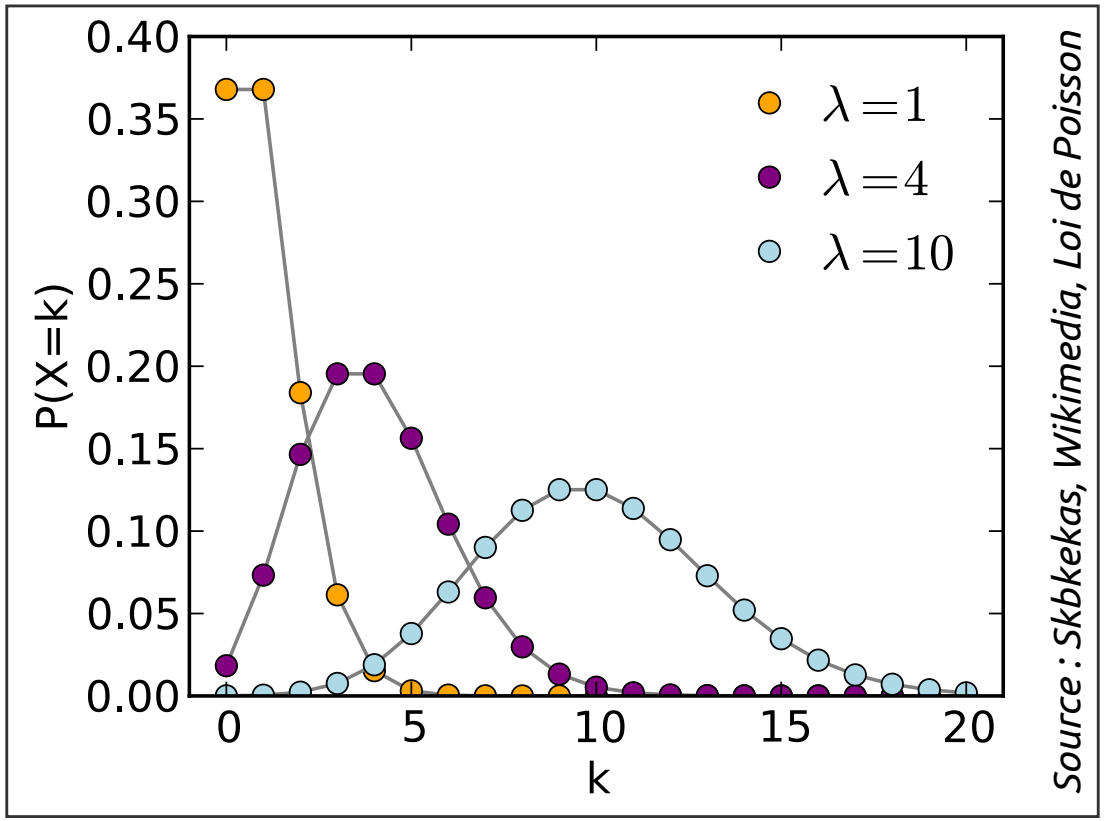


Source : Chainey, Ratcliffe (2005) GIS and crime mapping, Wiley.

Poisson distribution

Poisson's distribution λ parameter is both the **mean** and the **variance** of the distribution.

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



Poisson spatial distribution

A **Poisson spatial process** is a **spatial stochastic process**, sometimes labelled as :

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

Given a partitioned space Z, the probability of a given number of occurences in a zone z is modeled by a Poisson distribution whose mean is $\lambda \times area(z)$.

Dispersion index

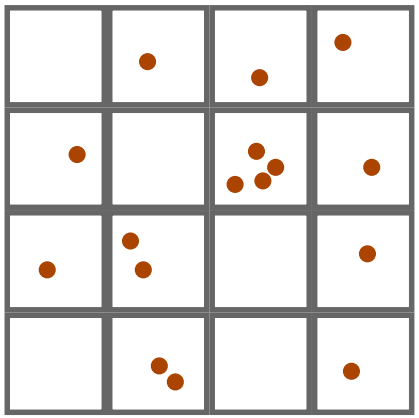
VMR *Variance-to-Mean Ratio* $= \frac{\mu}{\sigma^2}$

- Construct a regular grid
- Count occurrences
- Compute variance, mean and VMR

VMR intrepretation

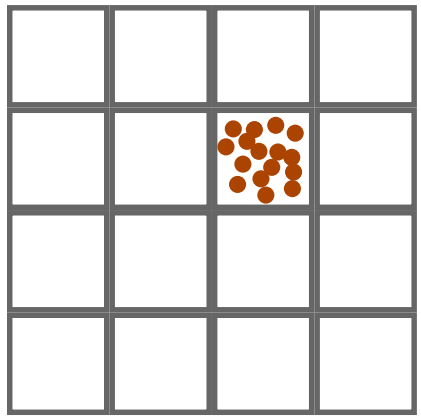
- VMR $= 0$: not dispersed
- VMR $= 1$: may have been obtained by a Poisson process
- VMR < 1 : uniform / periodic
- VMR > 1 : concentrated / clusters

Dispersion index



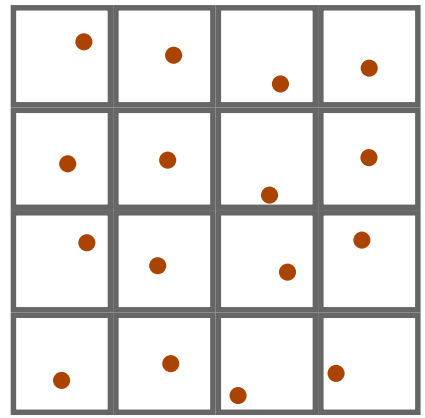
X (nbr. occurrences)
[0 1 1 1 1 0 4 1 1 2 0 1 0 2 0 1]

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

→ **test :** does VMR differs from 1 significantly ? (Student)

Quadrat methods

« A spatial χ^2»

- Construct a regulard grid (quadrats) : observed distribution
- Theoretical distribution (null model) is given by a spatial Poisson process.
- Contingency tables with observed and expected occurrences

$\rightarrow$ **test :** Observations differ significantly from expected values ? (χ^2 test)

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

GÉOGRAPHIE-CITÉS
UMR 8504

Interactions & Networks

DELHI GIS-R School

9-12th April 2019

Hadrien Commenges & Paul Chapron

hadrien.commenges@univ-paris1.fr

paul.chapron@ign.fr

Use case

Interaction

Relationship among objects. An interaction may be unidirectional, bi-directional or multi-directional . If the entities are spatial objects, interaction always integrate a spatial dimension.

Relationships come in many forms

- Environment : migratory birds, climate refugees, home-to-work commutes, etc.
- «Material realm» : commercial exchanges, percolation,
- «Immaterial realm» : twin-towns, Facebook friendship, co-authorship, etc.

What kind of geographical information is concerned ?

→ *TYPE 1 - Geographical Objects*

→ *TYPE 2 - Occurrences*

Relations Modeling

Interacting systems can be represented by the **relations** between constitutive **entities** , either as an **(adjacency) matrix** or as a **list of links**, weighted or not. (cf Graphs)

liste de liens

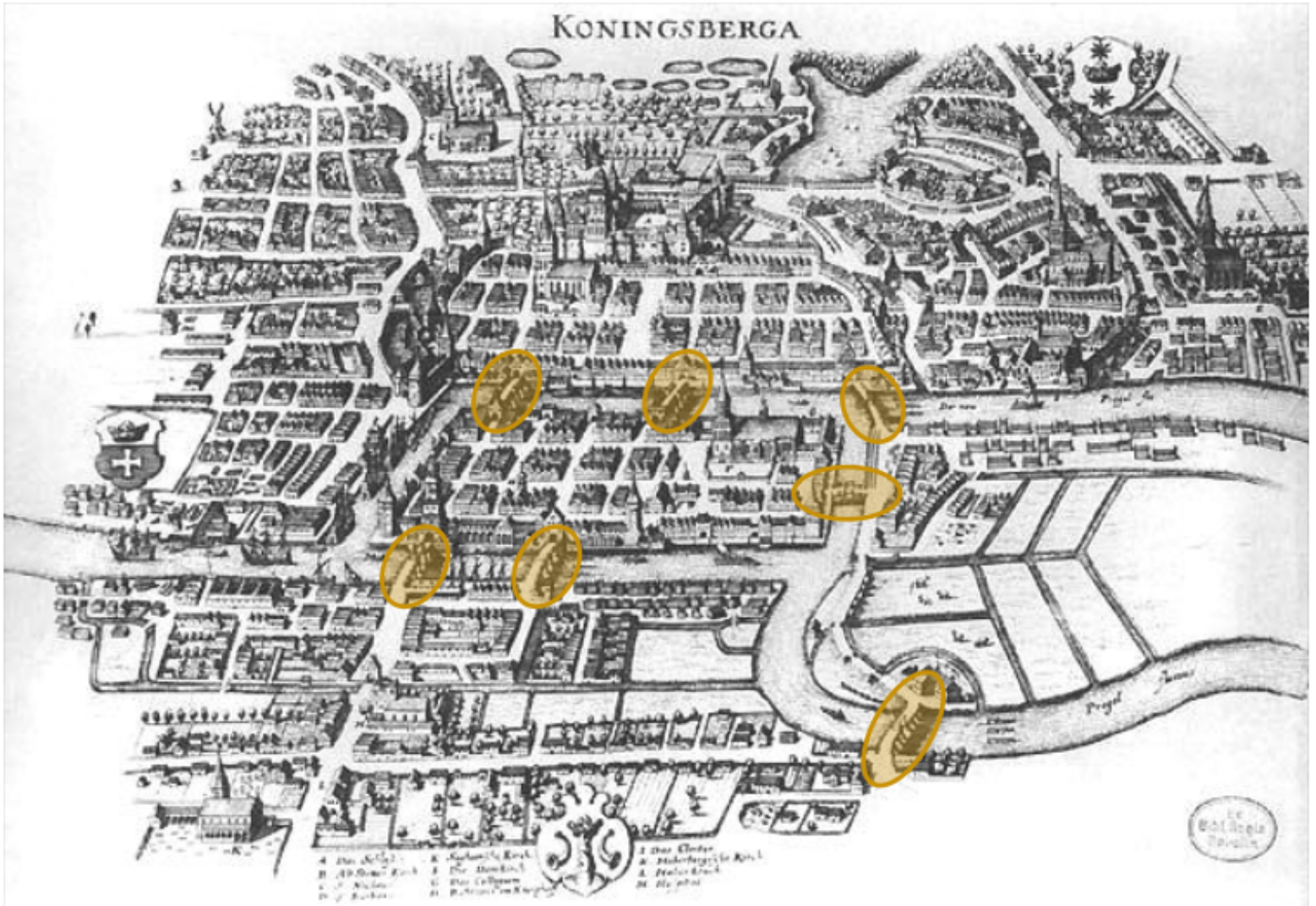
ori	des	poids
A	B	2
A	C	5
B	A	3
B	C	4
C	A	0
C	B	0

matrice d'interaction

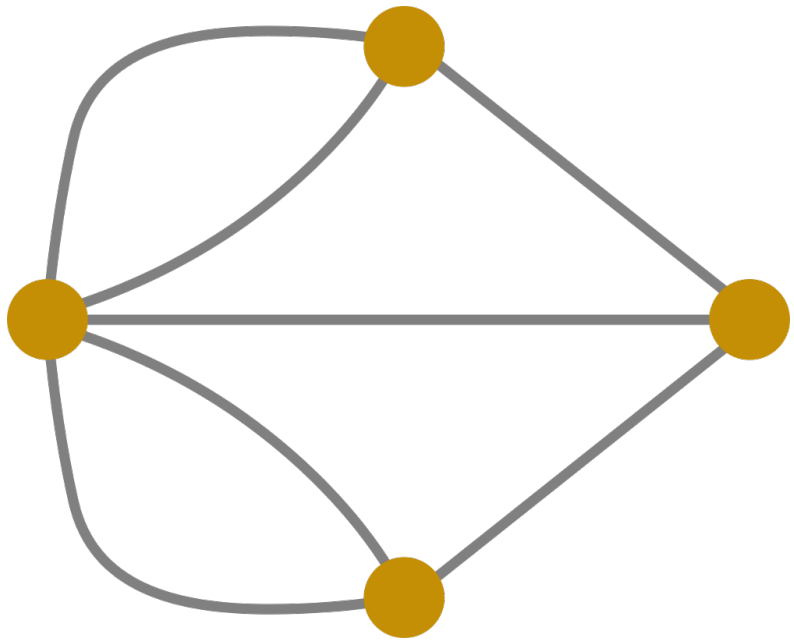
	A	B	C
A	0	2	5
B	3	0	4
C	0	0	0

Network Analysis

Leonhard Euler and the **Seven Bridges of Königsberg** (Kaliningrad).



Network analysis



- No **cycle eulérien** (parité des liens incidents)
- Pas de **chaîne eulérienne** (continuité du tracé)

Graphs

A graph is a set of *vertices*) connected by **edges**.
Examples :

- **Transportation networks :** e.g. Subway : nodes are stations, edges are lines.
- **Social networks :** individuals (nodes) , social interactions (edges)
- **Scientific collaboration networks :** Co-autorship (edges) among researchers (nodes)
- ...

Types of graphs

Undirected graph :

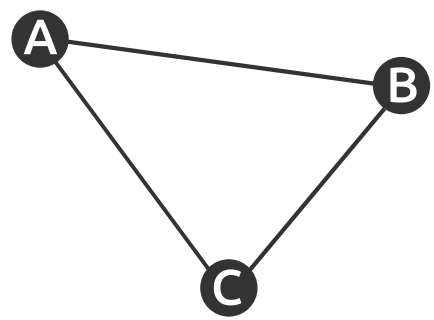
edges list

ori	des
A	B
A	C
B	C

adjacency matrix

	A	B	C
A	0	1	1
B	1	0	1
C	1	1	0

graph plot



Types of graphs

Directed graph :
arcs are employed instead of *edges* who are "undirected".

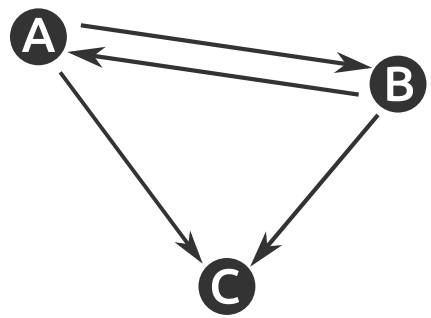
edges list

ori	des
A	B
A	C
B	A
B	C
C	A
C	B

adjacency matrix

	A	B	C
A	0	1	1
B	1	0	1
C	0	0	0

graph plot



Types of graphs

Weighted directed graph :

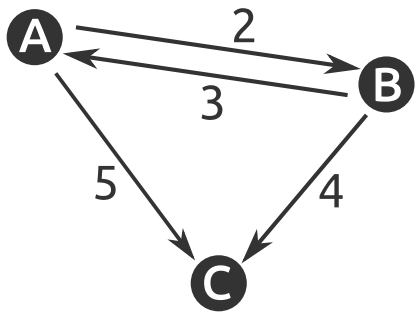
edges list

ori	des	flux
A	B	2
A	C	5
B	A	3
B	C	4
C	A	0
C	B	0

adjacency matrix

	A	B	C
A	0	2	5
B	3	0	4
C	0	0	0

graph plot



Also :

- Weighted undirected graphs
- Multi-graph ()
- Multi-partite graphs (several types of nodes)
- Hypergraphs (links between more than 2 nodes)
- ...

Network analysis : measures

Global measures :

- number of nodes
- number of edges
- (number of) connected components
- Density (ratio between nodes and edges)
- Diameter : max length among shortest paths
- Connectivity : ratio between number of edges and possible number of edges

$\rightarrow$ **Maximum number of edges (no loops) :**

- Planar graph : $3V - 6$
- Undirected non-planar graph : $\frac{V(V-1)}{2}$
- Directed non-planar graph : $V(V-1)$

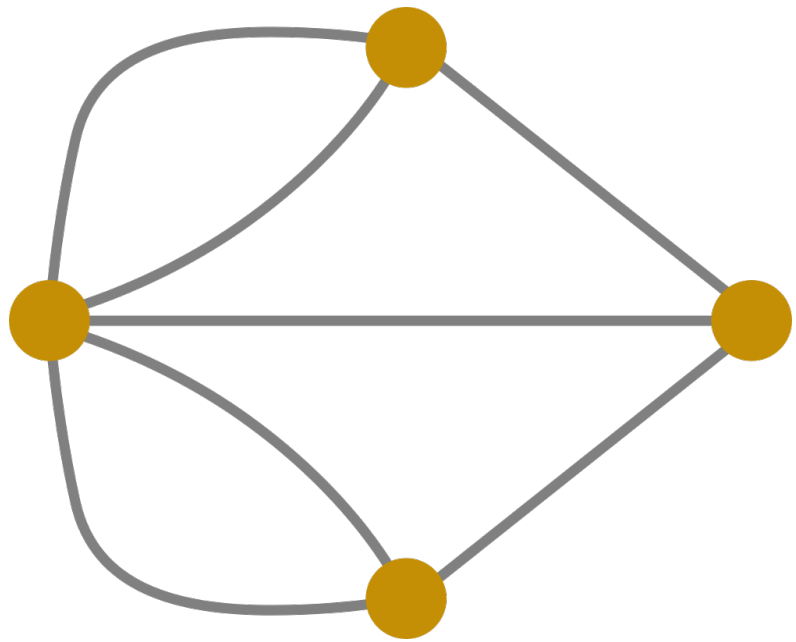
Network analysis measures

Nodes measures :

- Degree : number of edges (neighbors) of a node
- in-degree / out-degree : number of inomling arcs / outgoing arcs
- Weighted degree : sum of edges wights
- Closeness centrality : inverse of the sum of the shorstest paths length to any other node in the graph
- Betweenness centrality : number of shortest paths between any two vertices of the graph that contain the node.

Distances in Network

In a connected graph, there is many (an infinity) paths connecting one node to every other node.



The **shortest** of these paths is adopteed as a «**network distance**» indicator.

Djikstra Algorithm principles

Djikstra (1959) proposed a breadth-first algorithm for directed weighted graphs, to compute the shortest path between a node and the others

Given $G(V, E)$ a graph,
$V_{orig} \in V$ a node,
$w(a, b)$ a function giving the weight of the edeges linking vertices a and b (or $+\infty$ if none)

- The algorithm builds P , a *sub-graph* of G so that any distance from V_{orig} to a node $a \in P$ is known and minimum in G
- A Distance list from V_{orig} to any $v \in V$ are maintained. Distance of a given path is the sum of its edges weights
- P grows by adding an edge (a, b) of $P \times G$ if $d(V_{orig}, b)$ is minimum.
- Algorithm terminates when P is a *minimum spanning tree*, i.e. a tree visiting every node with a minimal sum of edges weights.

(Also works on undirected graphs, and for a given V_{dest} destination node)

Djikstra's pseudo code

Init
$P \leftarrow \emptyset$
$d(v) := 0, \forall v \in V$ (distances to V_{orig})
$d(V_{orig}) := 0$

While $\exists v \in V$ such that $v \notin P$:
    pick a node $a \notin P$ such that $d(a)$ is minimum
    add a to P
    For each $b \notin P$ such that $w(a, b) \neq \infty$
        $d(b) = min(d(b), d(a) + w(a, b))$
    End for each
End While

A good visualization is available at :
`https://www.cs.usfca.edu/~galles/visualization/Dijkstra.html`

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

GÉOGRAPHIE-CITÉS
UMR 8504

Concentration, Segregation, Autocorrelation

DELHI GIS-R School
9-12th April 2019

Hadrien Commenges & Paul Chapron

A good starting point

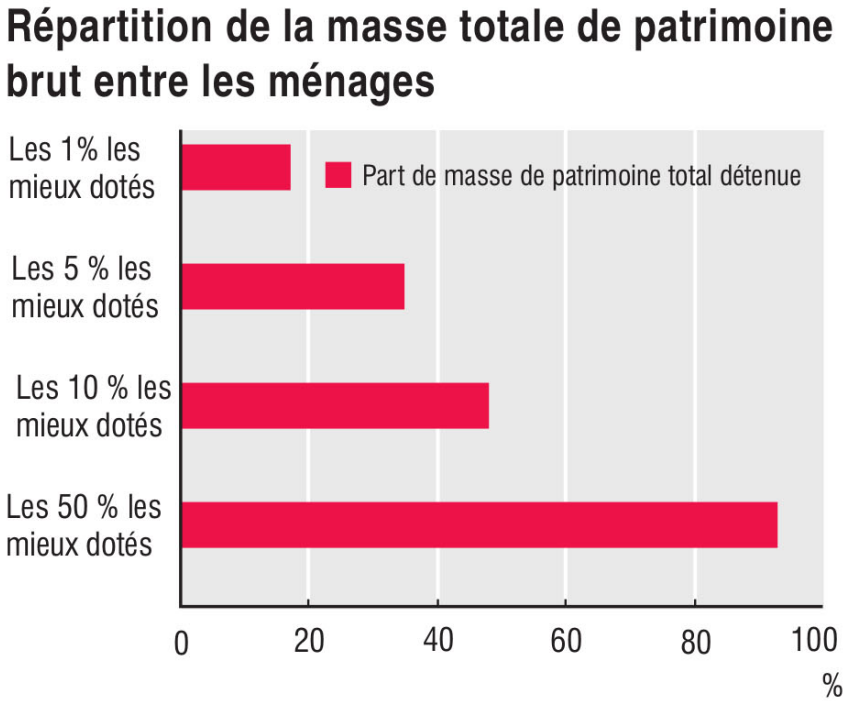
Information about tools and methods

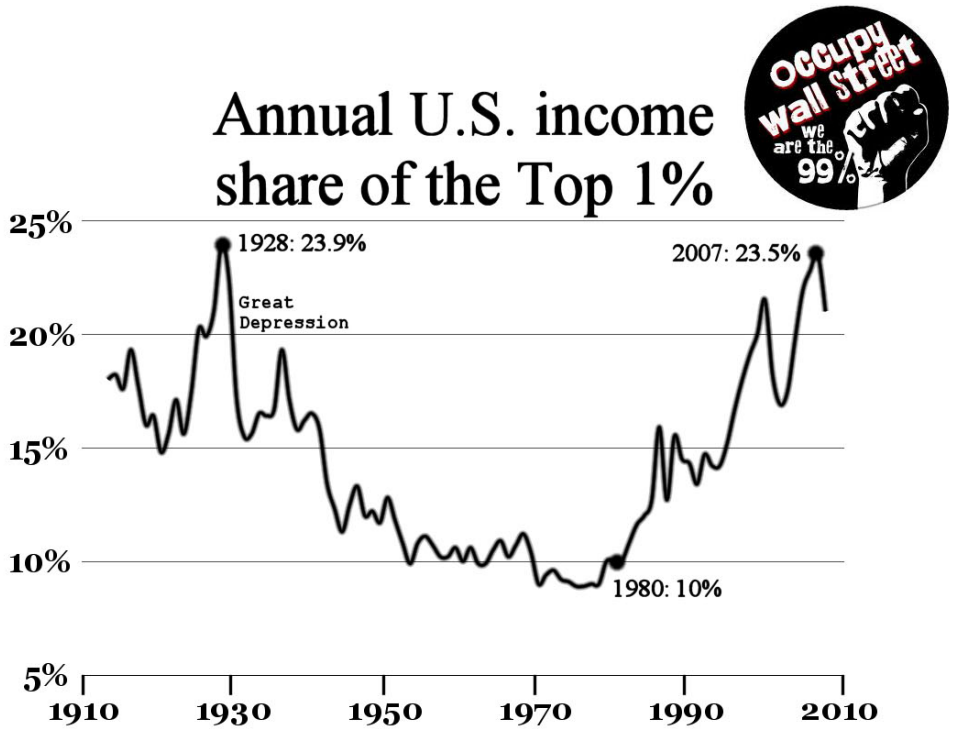
GeoDa Center for Geospatial Analysis and Computation (Luc Anselin)
`https://geodacenter.asu.edu`

→ logiciels *standalone* (GeoDa) + Python and R libraries
→ tutorials, references, online courses

Concentration

Classical methods from inequality economics





Concentration

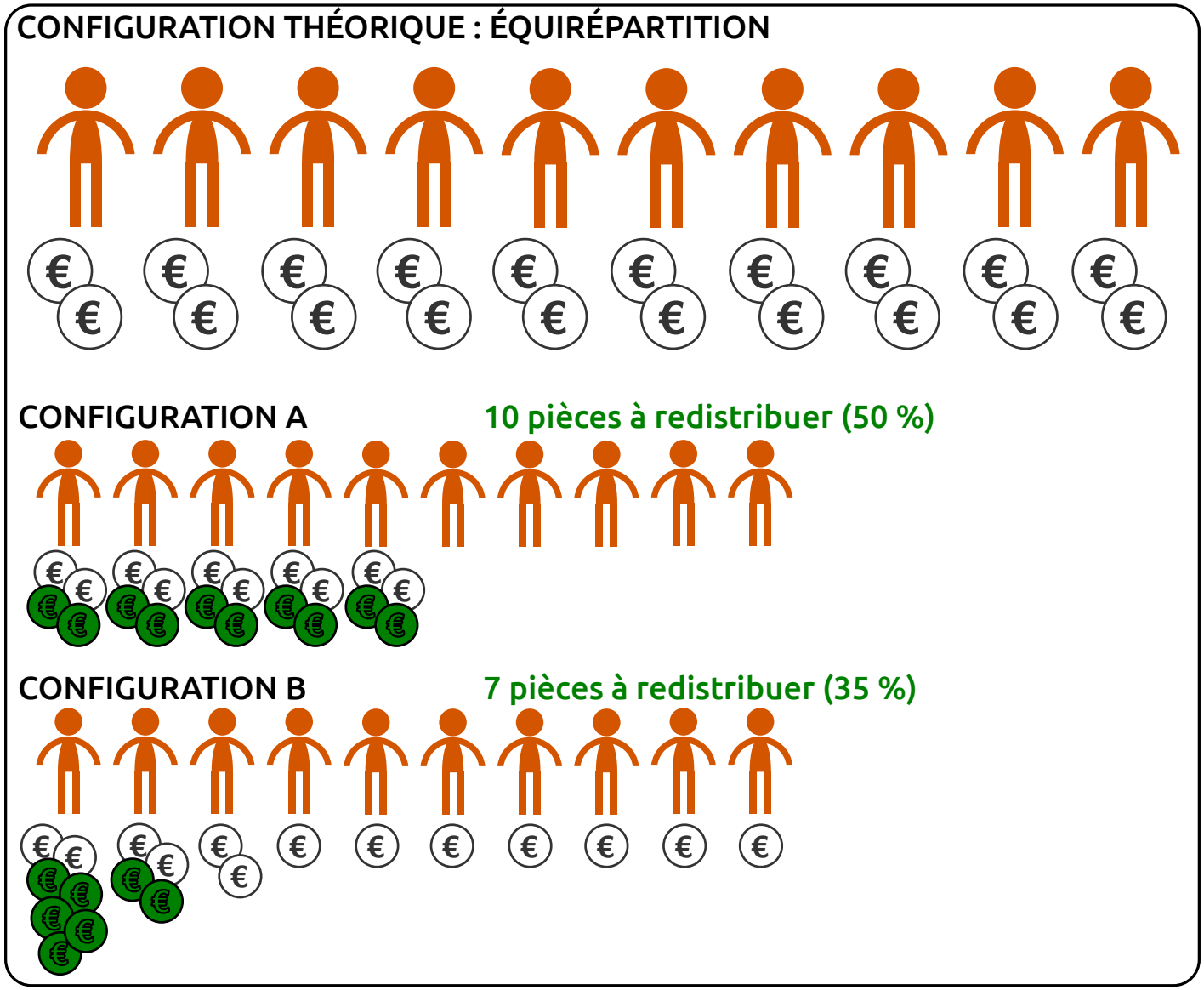
Hoover index
(Dissimilarity index or *Duncan's I)*

For two classes x and y of a population that sum up to 100% (e.g. male / female),

$$H = \frac{1}{2} \sum_{i=1}^{n} \left| \frac{x_i}{x_{tot}} - \frac{y_i}{y_{tot}} \right|$$

(also the longest vertical distance between the lorenz Curve and the 45 degrees line of perfect equality)

Concentration



Concentration

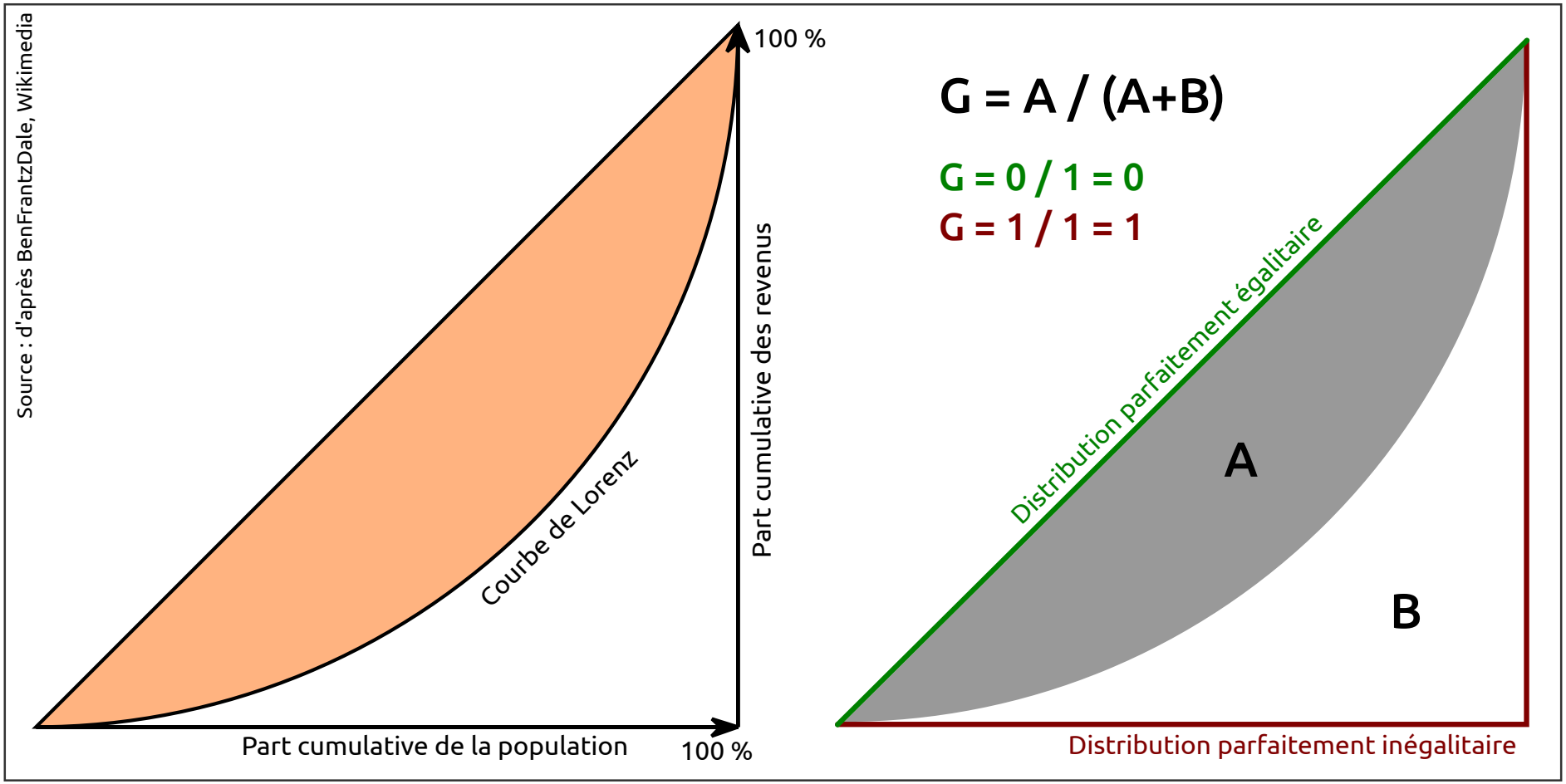
Gini index

relative mean of absolute differences between every pair of a stock

$$G = \frac{\sum_i \sum_j |x_i - x_j|}{2n \sum_i x_i}$$

ranges from 0 (perfect equality) to 1 (perfect inequality), but these theoretical values are never reached.

Concentration



Segregation

Entropy-based measures

These meassures comes from the field of information theory (Shannon)

Quantity of information I (in bits) of a message m : logarithm (base 2) of the ratio between the finite set of every possible message before a message is recieved (M, the universe) and the finite set of possible message after it is recieved (m) :

$$I = \log\left(\frac{M}{m}\right) = -\log(p(m))$$

This quantity is expressed in *bits* since the logarithm base is 2.

Segregation

Entropy-based measures

I : Information brought by the occcurence of a message (informative event) .
H (entropy) : information stored in a finite set of informative event (e.g. all the messages of a transmission)
- i.e. average quantity of information stored in a set of messages
- i.e. each message bring as much information as its quantity of information multiplied by its probability of apparition.

$$H(x) = -\sum_{i=1}^{n} p_i \log(p_i)$$

Relative entropy : ratio between an entropy and the maximum entropy H_{max}

$$H_{max} = \log(n)$$

Segregation

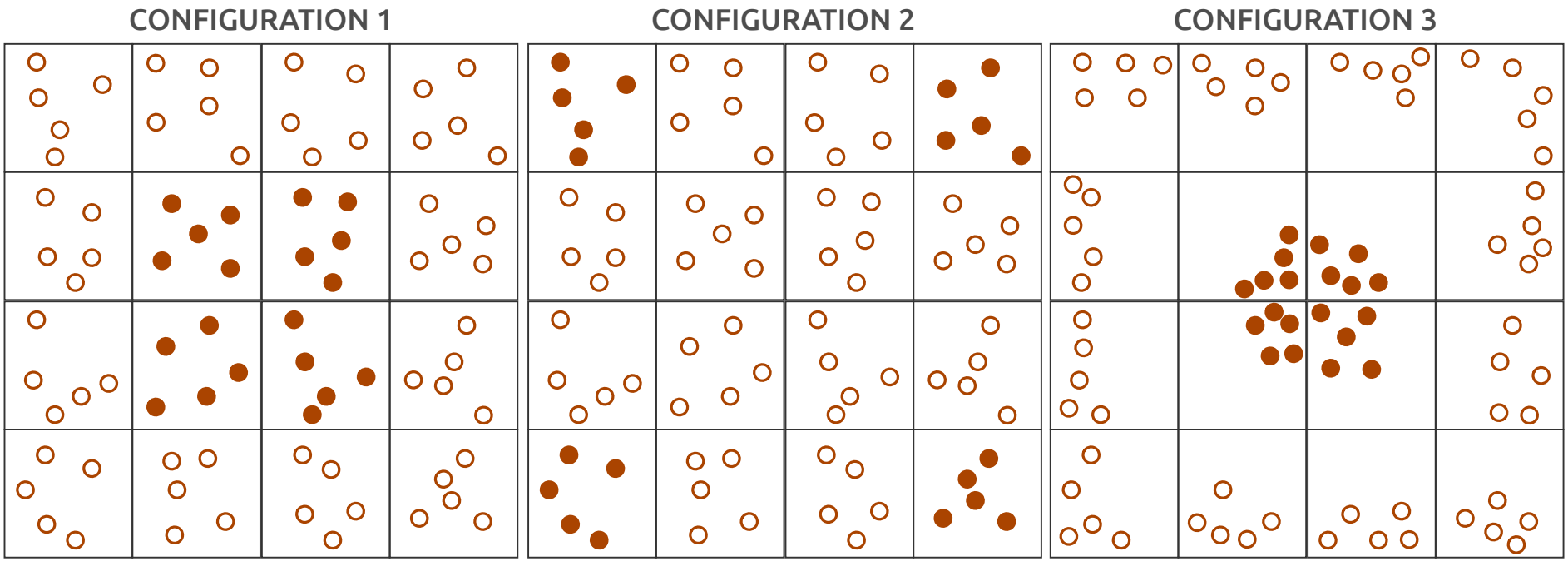
Entropy interpretation guide

Low (relative) entropy values tend to indicate segregated configurations

High (relative) entropy values tend to indicate more evenly distributed configurations

Autocorrelation

Hoover, Gini, Shannon (Theil) indexes can't distinguish between these configurations



Autocorrelation indexes (**Geary** and **Moran**) can distinguish between **1** et **2**.

Autocorrelation

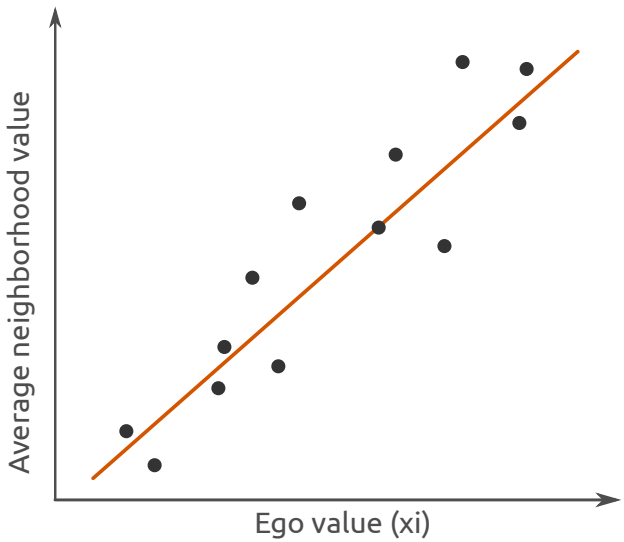
Moran's I (1950) :

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \times \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

n : number of spatial units
x_i et x_j : values of x in i and j
$\bar{x}$: average value of x
w_{ij} : weighting matrix corresponding to the neighborhood definition

Autocorrelation

Moran's I interpretation based on **Moran's plot**



Ordinary Least Square (OLS)

$$y = \beta_o + \beta_1 x + \epsilon \quad \text{with} \quad \beta_1 = \frac{COV(xy)}{VAR(x)}$$

In case of a simple contiguity matrix with constant weights:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \times \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

$$I = \beta_1$$

Autocorrelation

Local contribution to the global autocorrelation index
LISA - *local indicators of spatial autocorrelation*

$$I_i = z_i \sum_j w_{ij} z_j$$

z_i standardized value of x for spatial unit i
z_j standardized value of x for spatial unit j
w_{ij} weighting matrix

Spatial autoregressive model

The autocorrelation can be used in a **descriptive** approach or in a **modeling** approach.

Ex. of land values modeling (econometrics) : the value of a dwelling is function of :

- Intrinsic attributes : surface area, garden, etc.
- Contextual attributes : i.e. the value of the dwellings in the neighborhood

In this case, it may be useful to build a **spatial autoregressive regression** (SAR), i.e. to inject the **lagged values** (neighborhood average) as a regressor.