

# Geographic Information

DELHI GIS-R School  
9-12<sup>th</sup> April 2019

**Hadrien Commenges & Paul Chapron**

[hadrien.commenges@univ-paris1.fr](mailto:hadrien.commenges@univ-paris1.fr)

[paul.chapron@ign.fr](mailto:paul.chapron@ign.fr)

# Definitions

## Geographical Information

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

## Geographical information types :

1. Geographical objects (volcanos, railways, forest, etc.)
2. Event occurrences (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. Dwellings point patterns (spatial object)
2. Dwellings sales (occurrences)
3. Dwellings prices (measure points)
4. Average price by district (« statistics »)

# (1) Geographical Objects

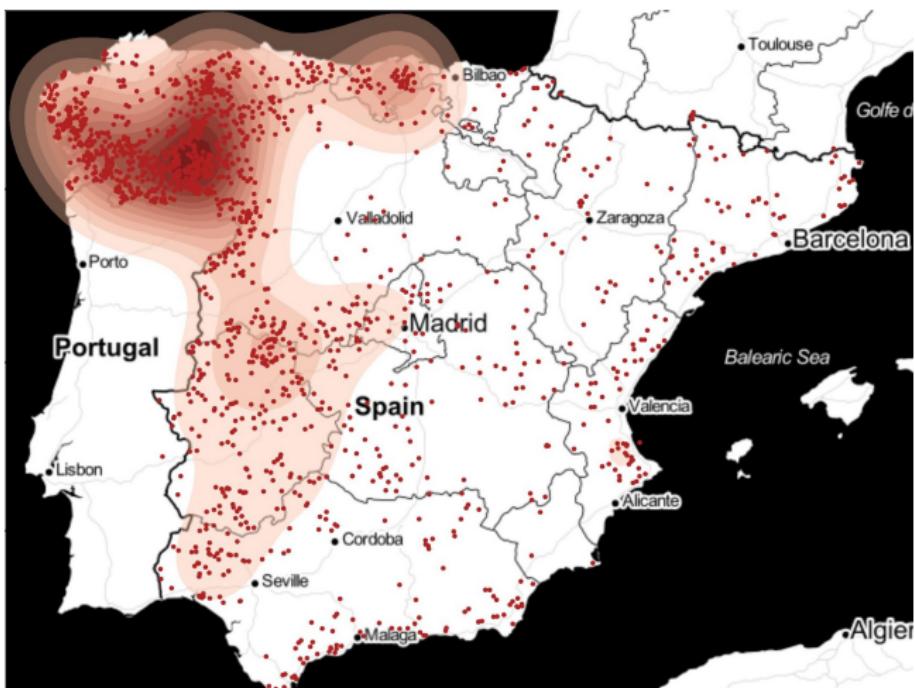
Geographical objects come in three types : **points**, **lines** and **areas**.

Geographical data analysis focus on their **geometry** (e.g. length, morphology) and their **topology** (e.g. neighborhood , distance).



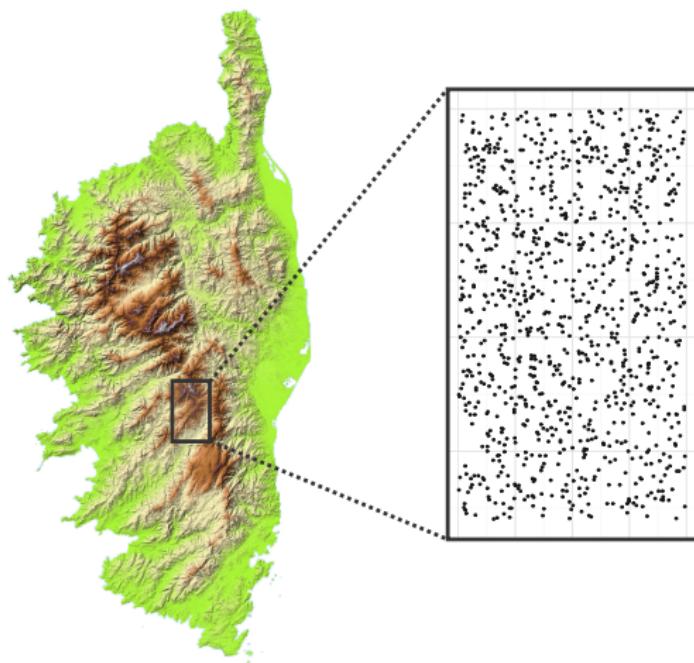
## (2) Event occurrence

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**.

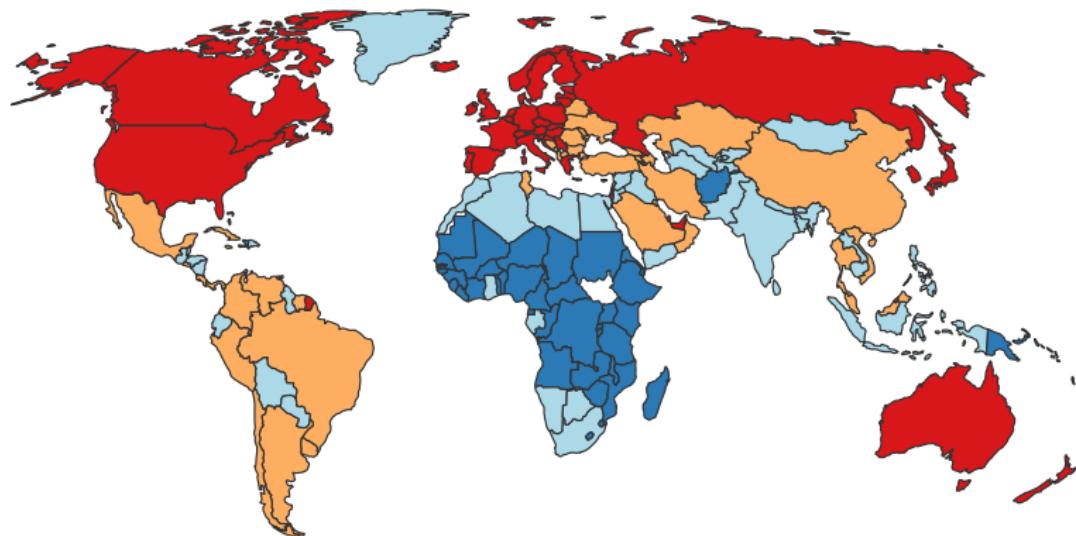


## (4) Statistics

«Statistics», from *statista*, «state man» in italian.

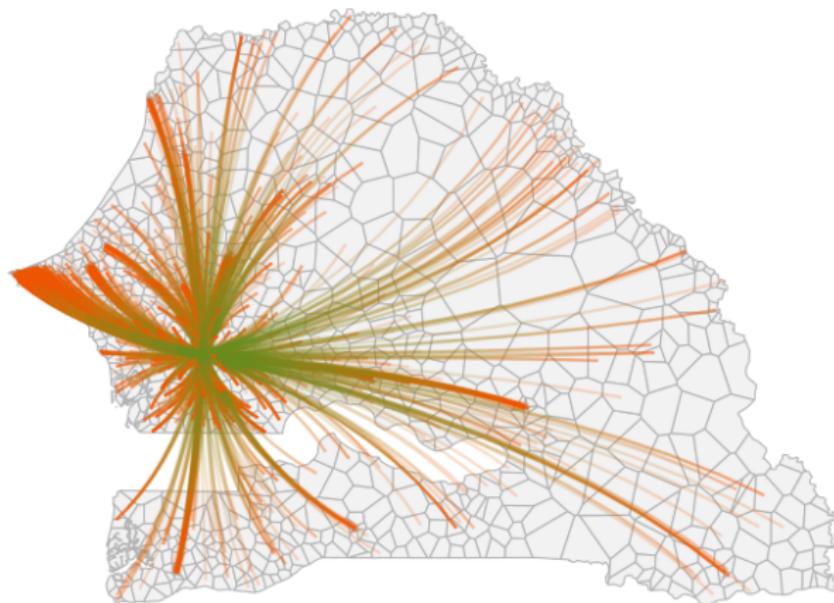
**Zonal extent variables** created from census - **sampled or extensive** - for territorial management.

Such variables are attributes measured or computed **within spatial units**.

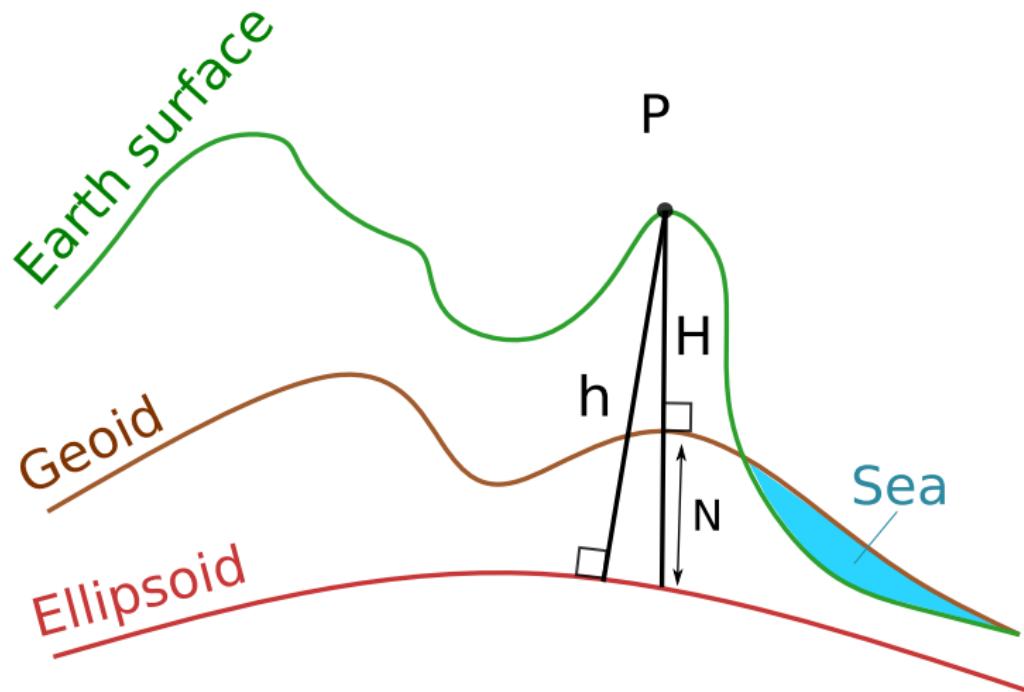


## (5) Interactions

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

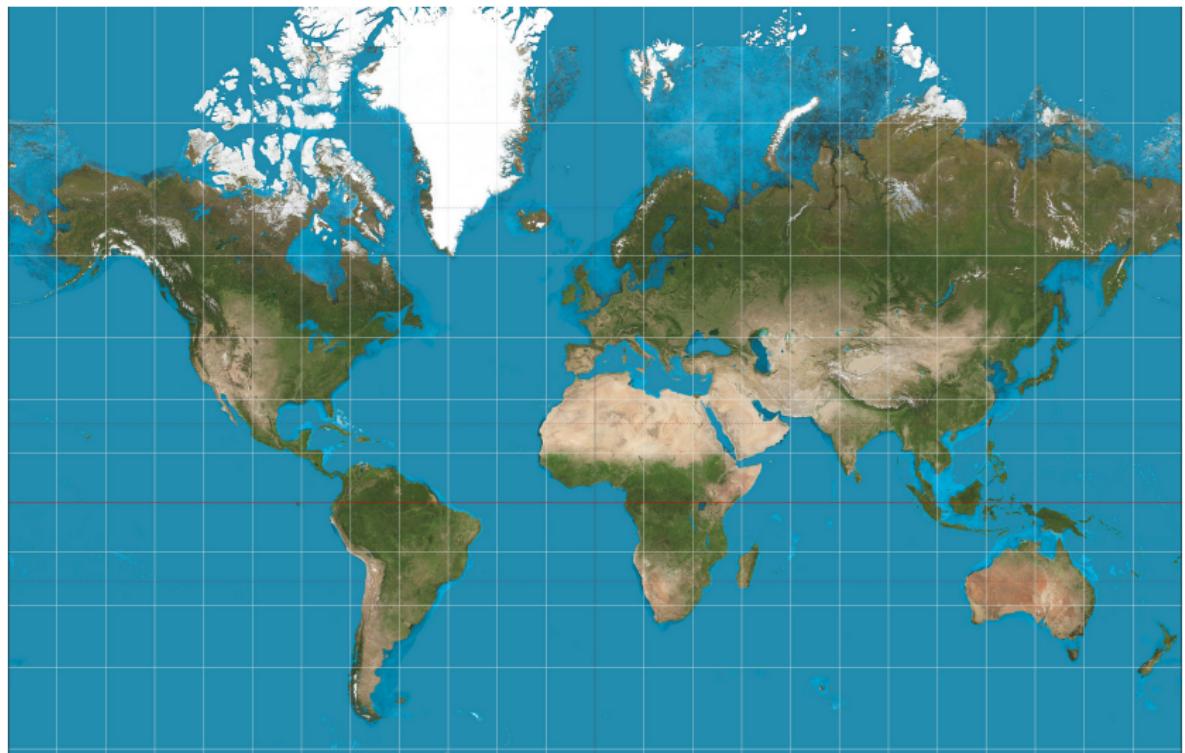
- ▶ **conformal** : conserve angles (shape)
- ▶ **equivalent** : conserve areas
- ▶ **equidistant** : conserve distances

Some others don't.

UTM (Universal Transverse Mercator, conformal) allows almost everywhere an acceptable projection.

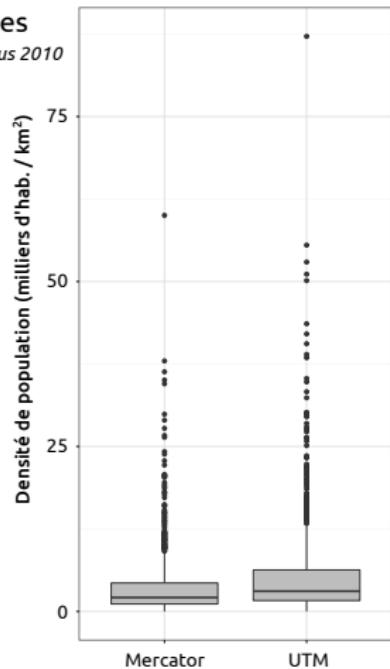
What is the recommendation for India ? Specific ? Kalianpur 5 zones ?

# Coordinates, areas, distances



Source : Wikimedia, Mercator projection

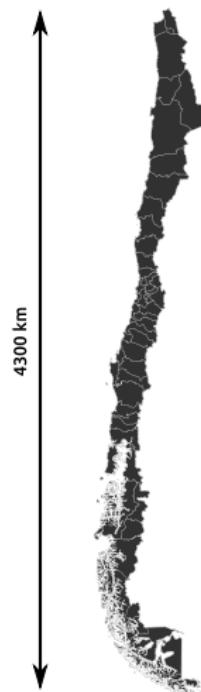
# Coordinates, areas, distances



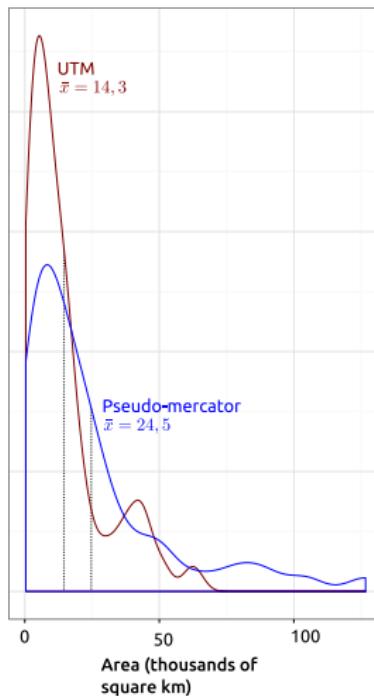
Med. 2111      Med. 3069  
Moy. 3764      Moy. 5483

Paris : 22 000 hab./km<sup>2</sup>  
Île-de-France : 1 000 hab./km<sup>2</sup>

# Coordinates, areas, distances



Chile's provinces area



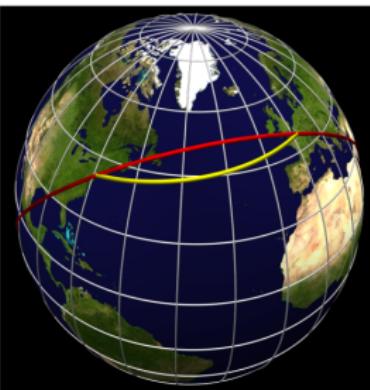
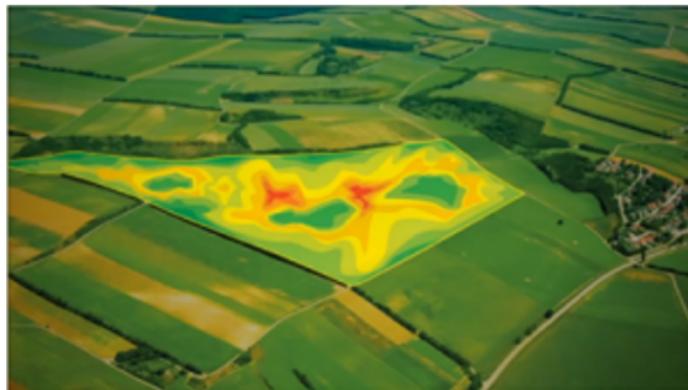
# Coordinates, areas, distances

Basic precautions regarding projection :

- ▶ Density → any areas alteration ?
- ▶ Distance → any length alteration ?

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

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



# Modeling and representation

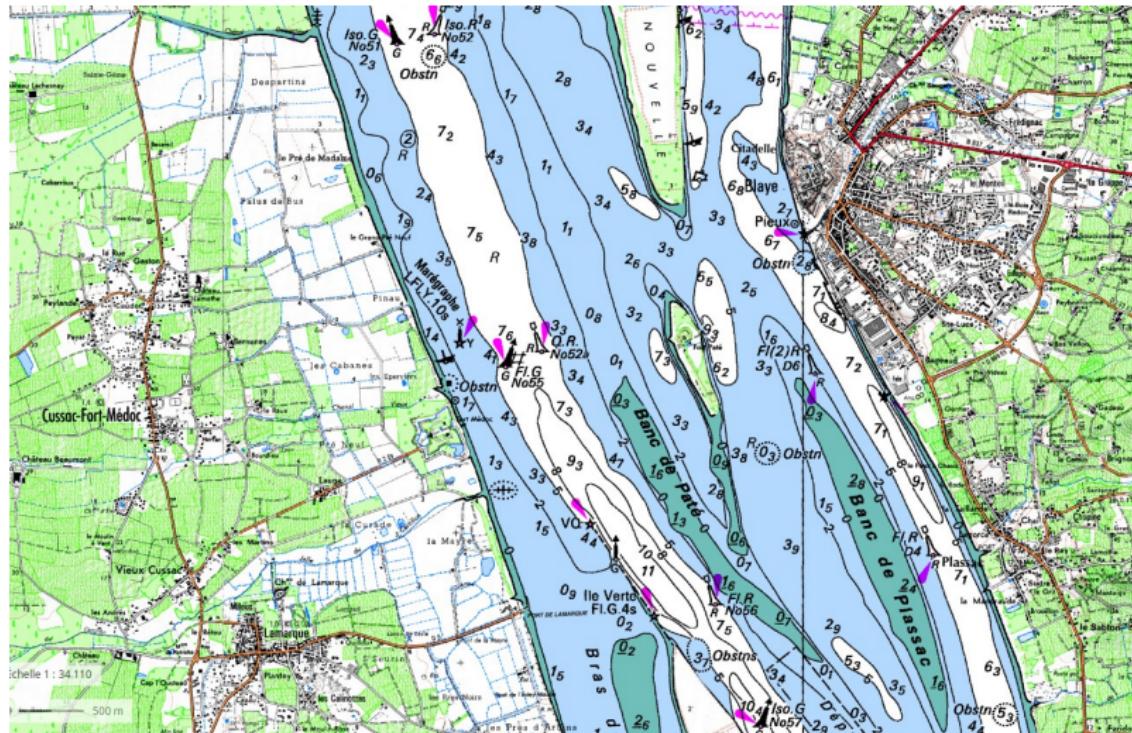
to model (here), is defining **categories** of objects **depicting** real-world objects (somehow linked to ontologies).

The raster view of the world	Happy Valley spatial entities	The vector view of the world
	 Points: hotels	
	 Lines: ski lifts	
	 Areas: forest	
	 Network: roads	
	 Surface: elevation	

Credit: Indiana University

# Modeling and representation

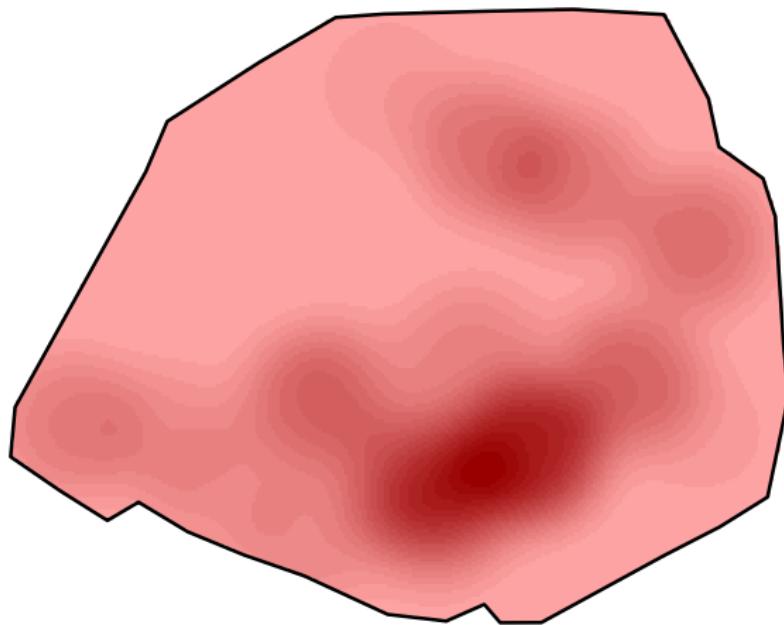
## Objects and Fields.



*Sources : IGN and SHOM*

# Fields

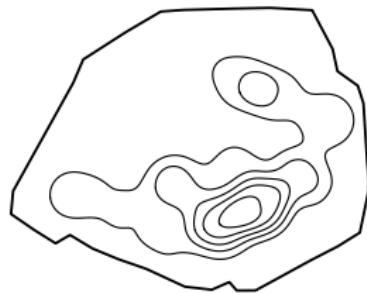
What does this field represent ?



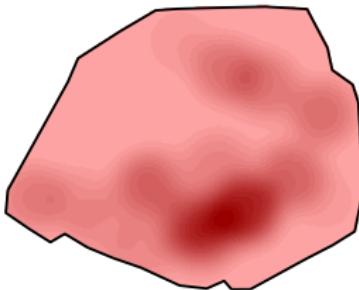
# Fields

## Representation modes

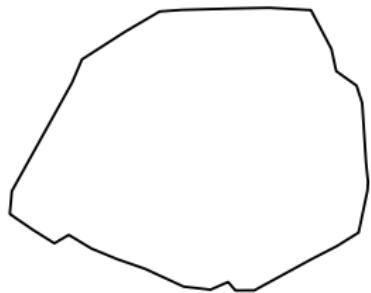
CONTOUR LINES



GRADIENT

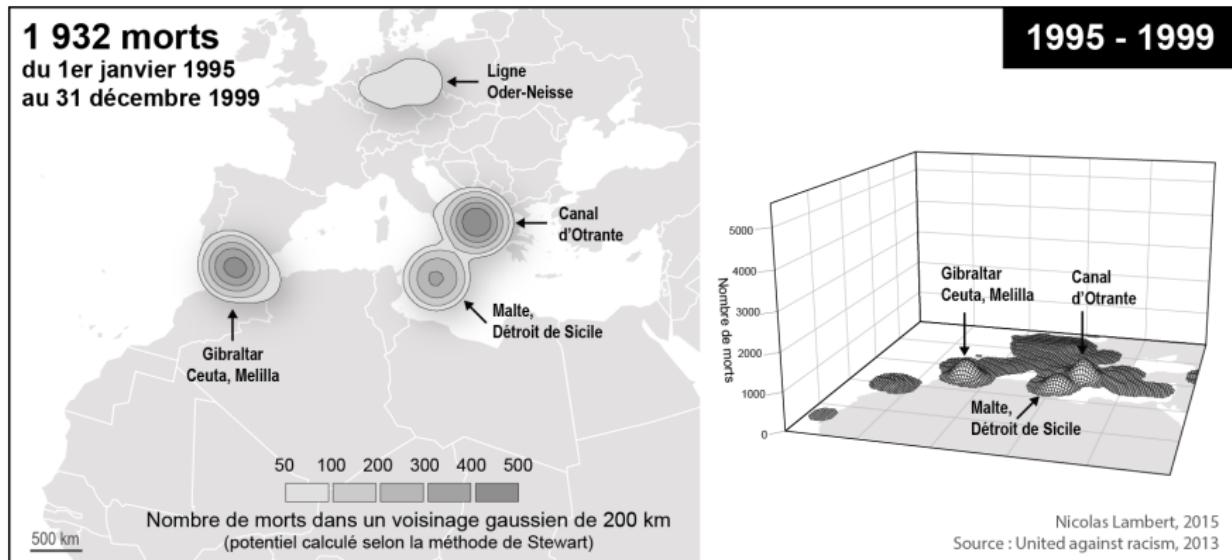


3D



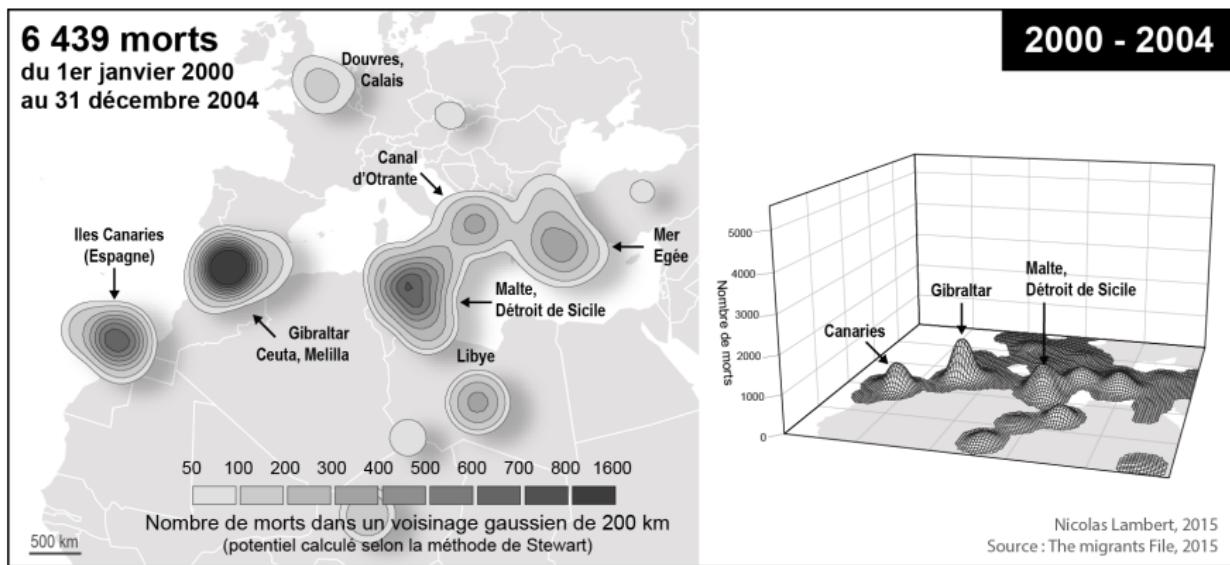
# Fields

## Representation modes examples



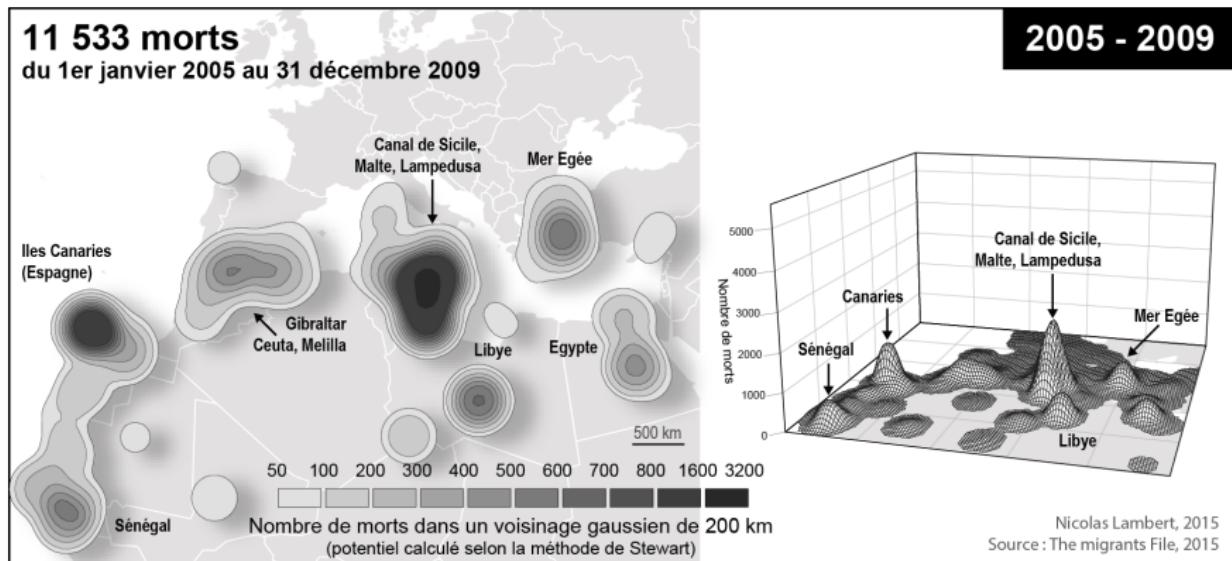
# Fields

## Representation modes examples



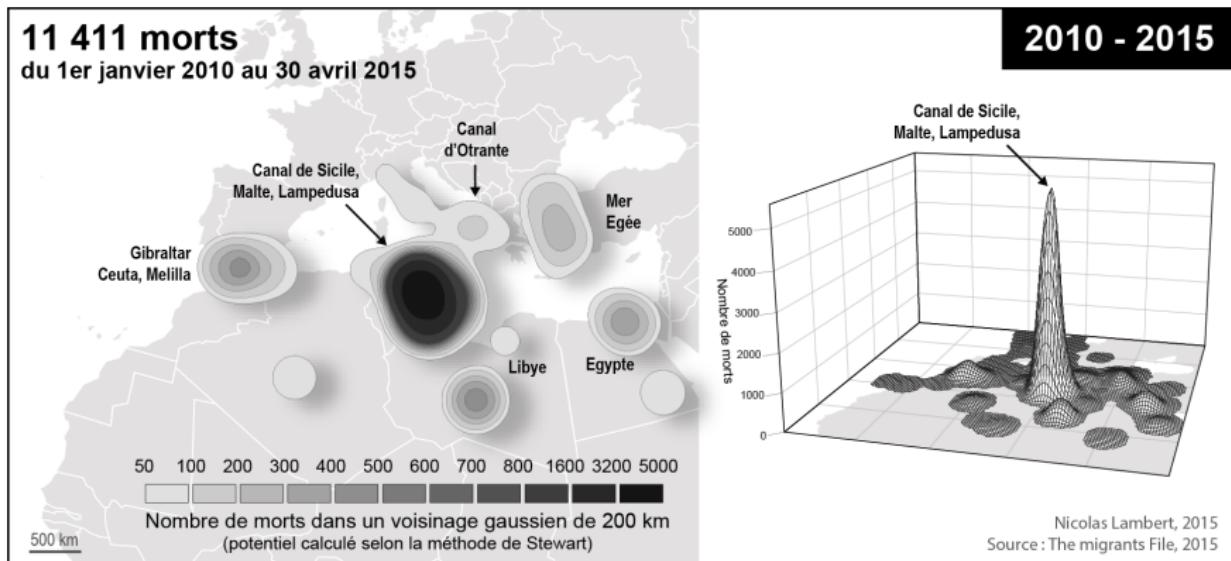
# Fields

## 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  
9-12<sup>th</sup> April 2019

**Hadrien Commenges & Paul Chapron**

[hadrien.commenges@univ-paris1.fr](mailto:hadrien.commenges@univ-paris1.fr)

[paul.chapron@ign.fr](mailto:paul.chapron@ign.fr)

# Stochastic simulation

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

- ▶ **Monte Carlo** : generic term referring 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** : exemplify the Law of Large Numbers

**Expected value  $\mu$  of a dice roll :**

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

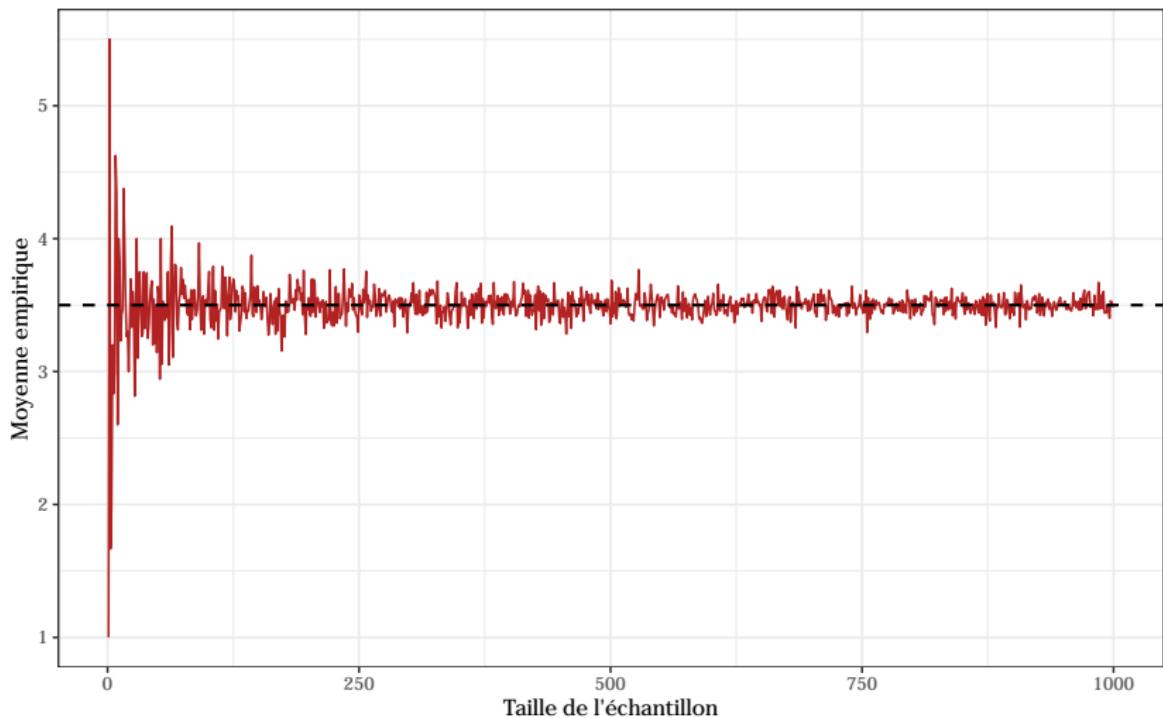
$$\mu = \frac{1+2+3+4+5+6}{6} = 3,5$$

**(weak) Law of Large Numbers :**

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

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

# Monte Carlo



# Bootstrap

## Classical inference :

- ▶ Goal : approach  $\mu$  and  $\sigma$  **parameters** of a distribution
- ▶  $\bar{X}$  and  $\sigma_X$  are computed on a **sample X**
- ▶ Sampling of X and  $\bar{X} / \sigma_X$  computation are repeated

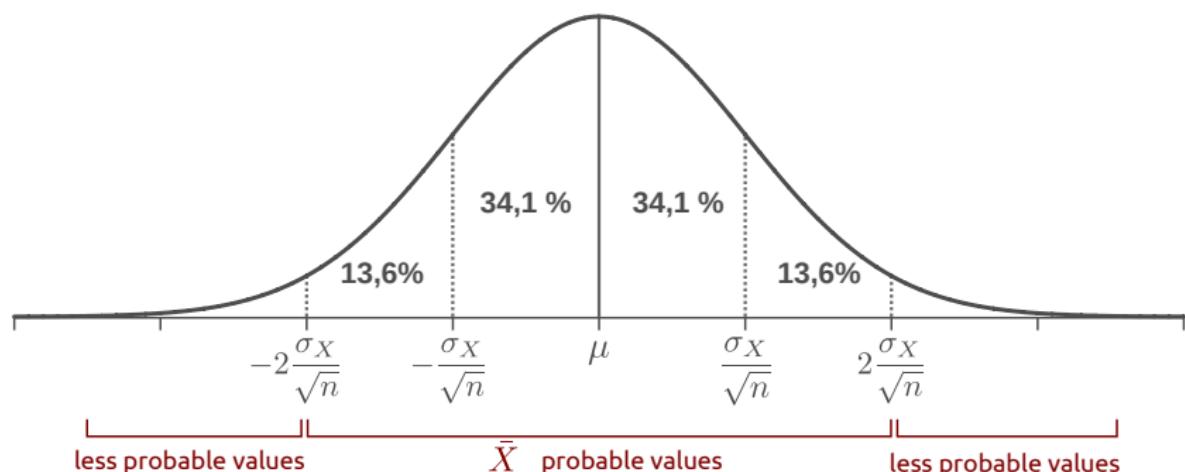
## Example :

- ▶ 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 .

→ these 500 average values form a *distribution* : the **sampling distribution of the mean**

# Bootstrap

Sampling distribution of the mean (500 mean values) :



where  $\mu$  et  $\sigma$  are the **parameters** – real mean and standard deviation of the population – and  $n$  is the **sample size**.

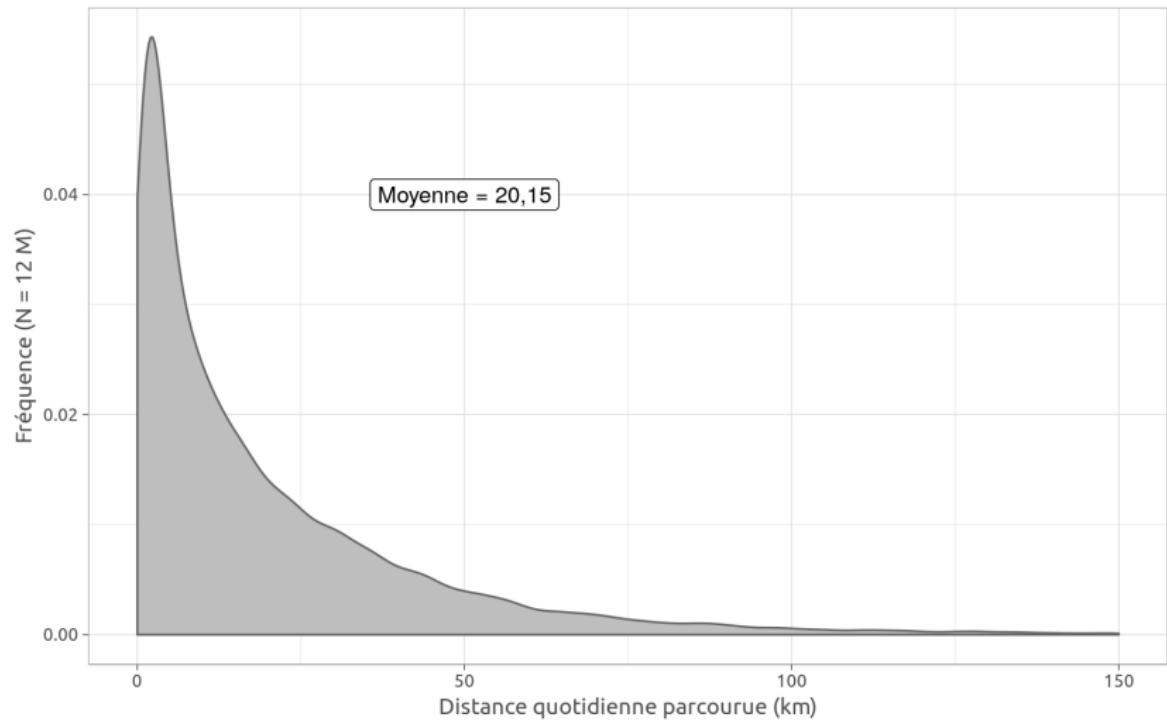
→ This is **central limit theorem (CLT)**.

# 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

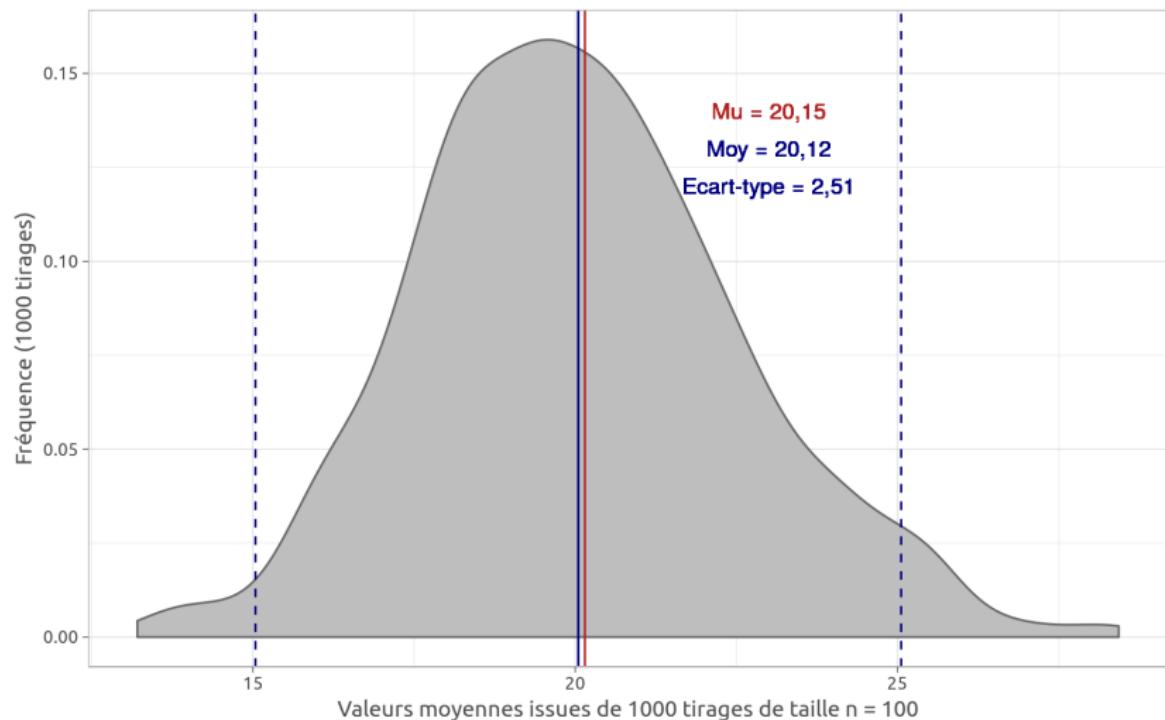
## Daily travel distance for Île-de-France people



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

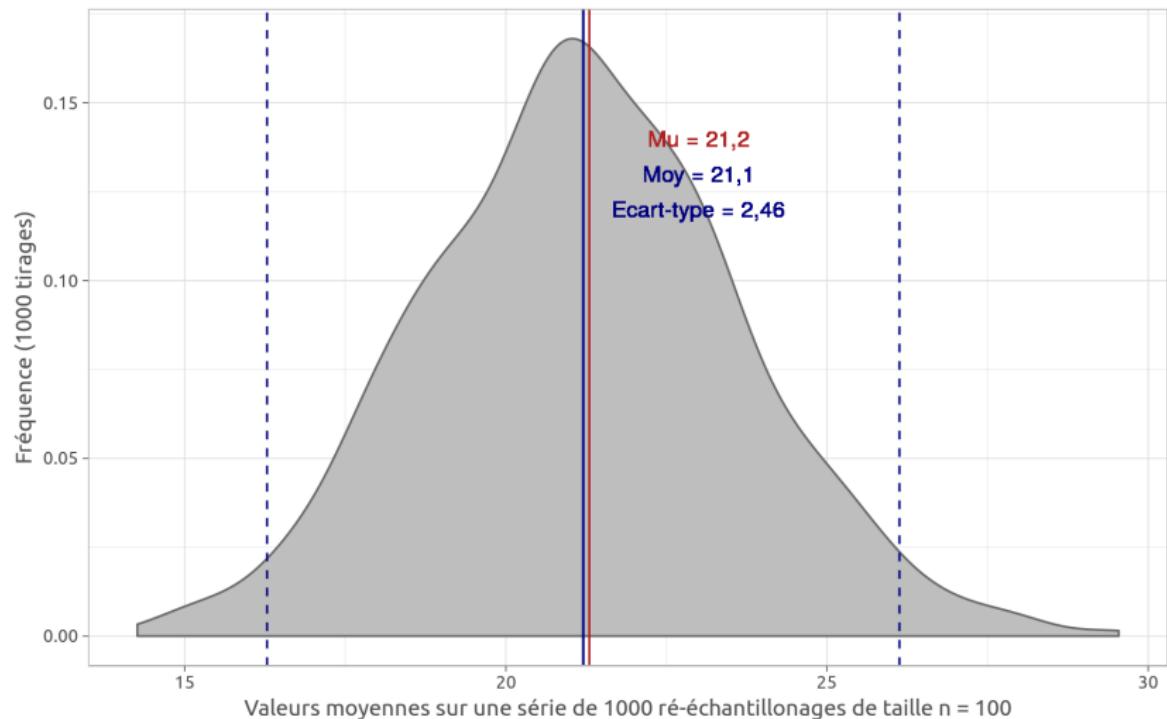
# Bootstrap

## 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  
9-12<sup>th</sup> April 2019

**Hadrien Commenges & Paul Chapron**

[hadrien.commenges@univ-paris1.fr](mailto:hadrien.commenges@univ-paris1.fr)

[paul.chapron@ign.fr](mailto: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 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** ).

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

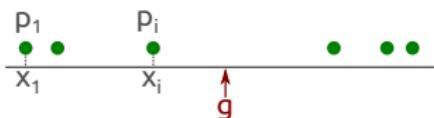
→ weights  $w_i$  might be constant or varying, depicting localized stocks variations.

# Spatial distribution parameters

2-Dimensions variance  $\approx$  inertia.

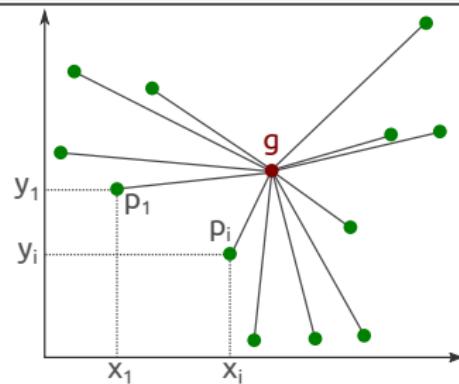
**General formula**  
-> squared distances mean

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



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

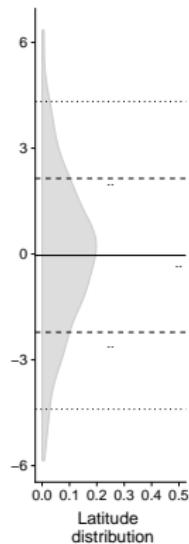
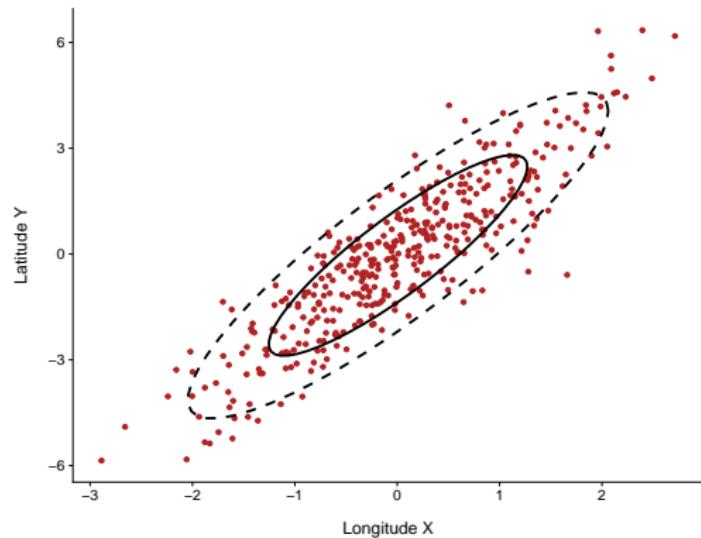
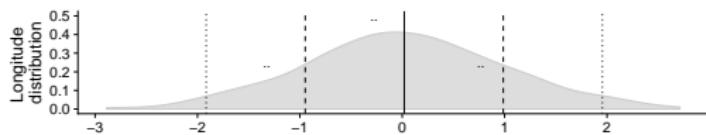
1D



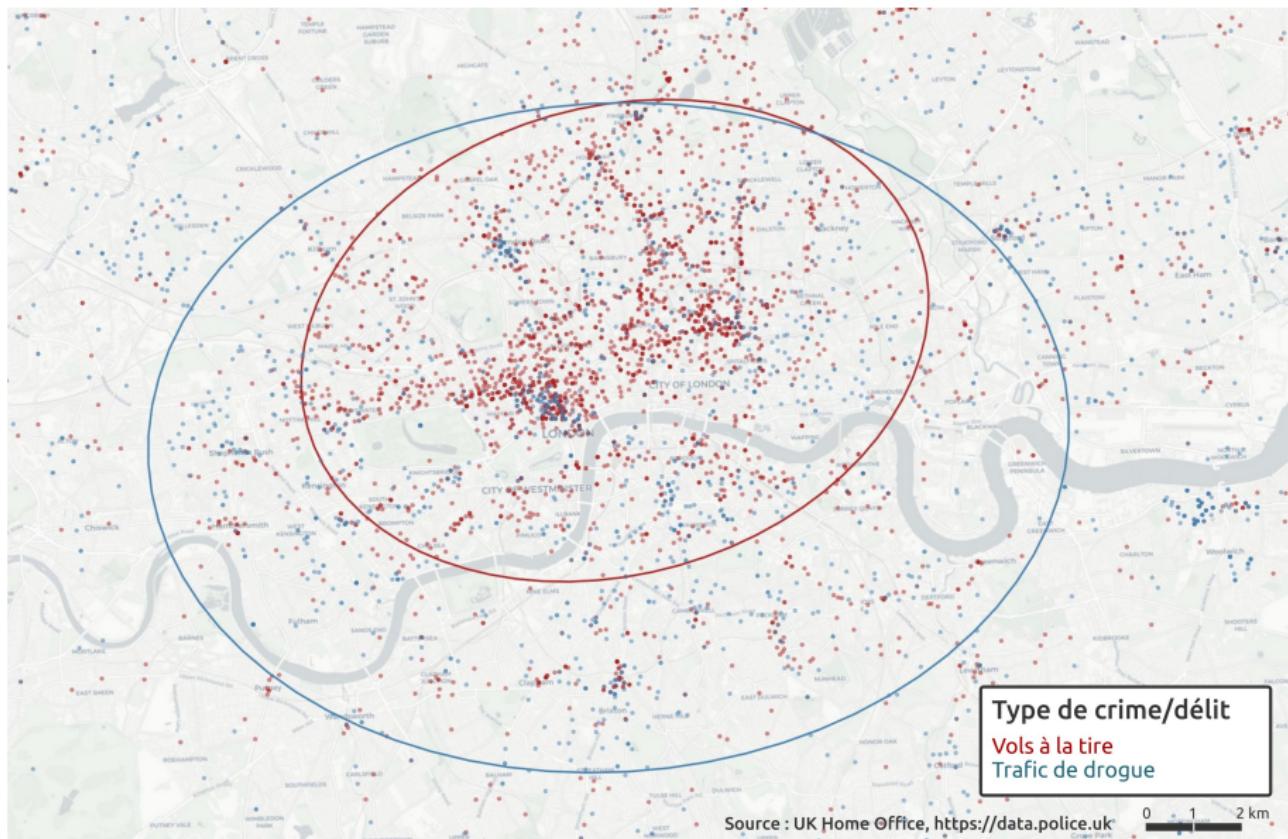
$$I = \frac{1}{n} \sum_{i=1}^n [(x_i - x_g)^2 + (y_i - y_g)^2]$$

2D

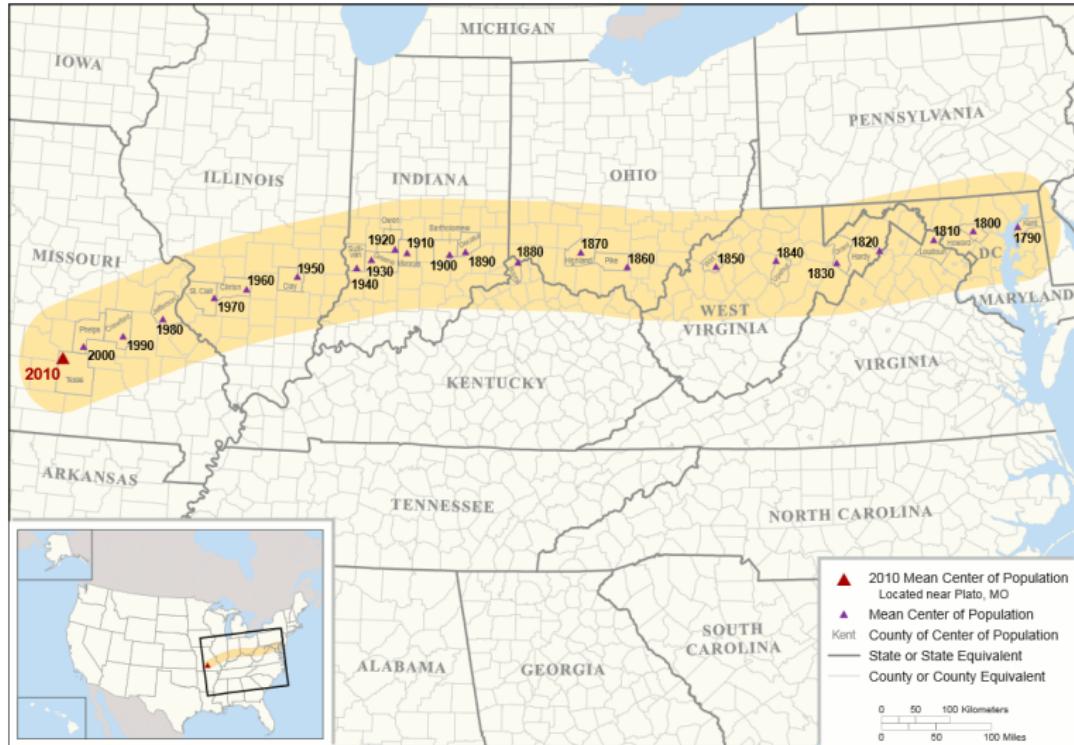
# Spatial distribution parameters



# Spatial distribution parameters



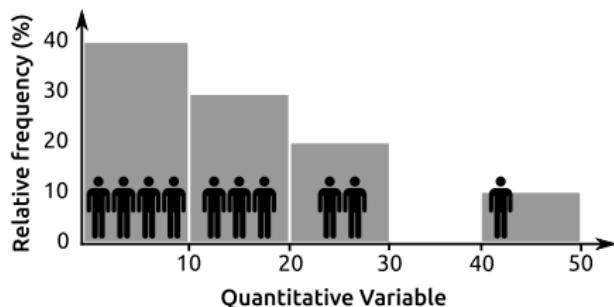
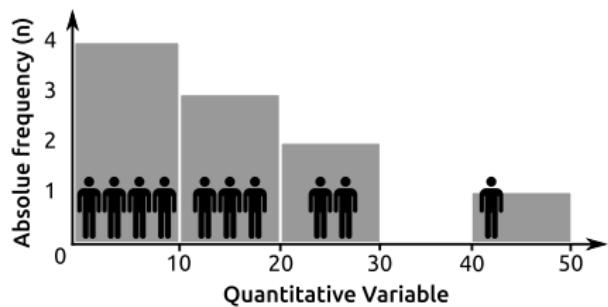
# Spatial distribution parameters



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

# 1D distribution graph (discrete)

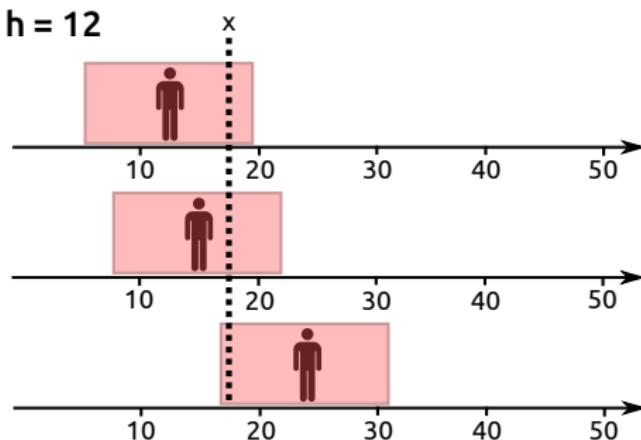
## Histogram



→ histogram estimated density is **discrete** by construction.

# Distribution graph (Parzen)

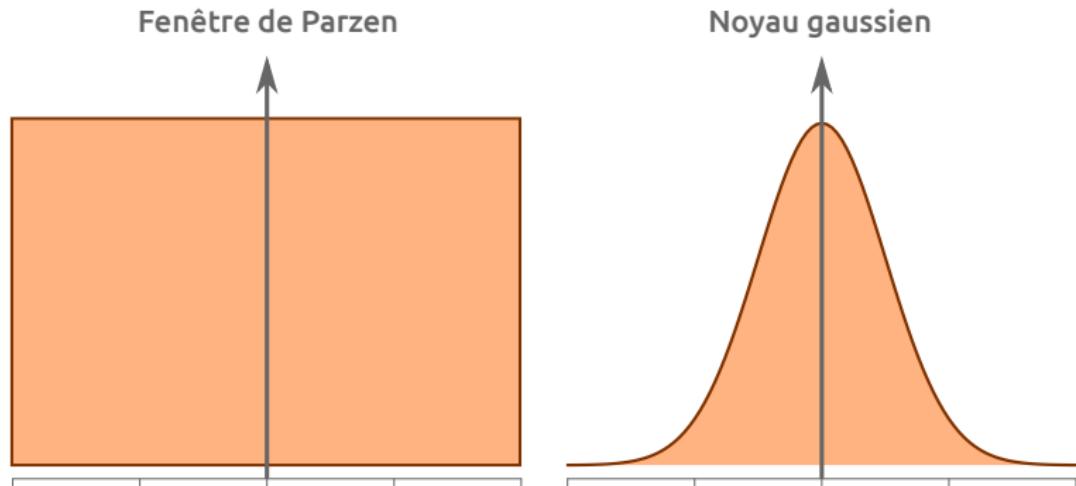
Histogram generalization : Parzen window



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

# Distribution graph (continuous)

Parzen Generalization : Gaussian Kernel



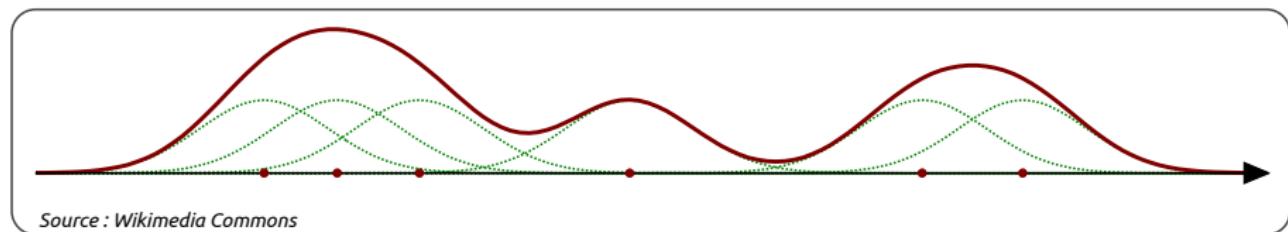
# 1D distribution graph (continuous)

## 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

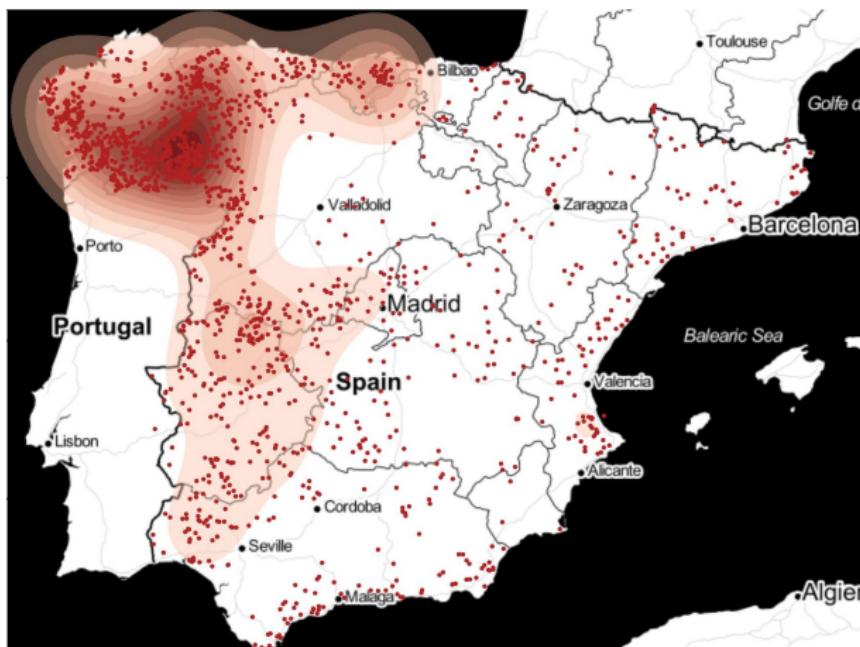


«Close observations contribute strongly, far ones contribute slightly »

*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

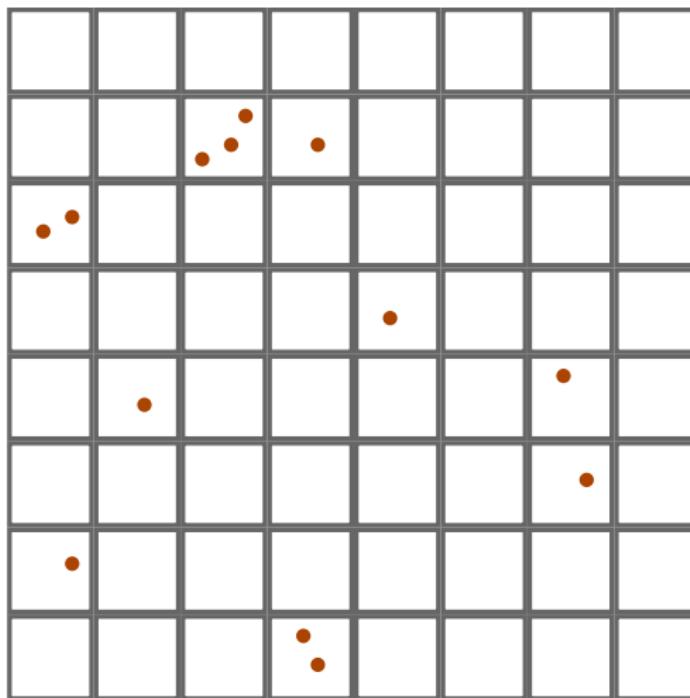
May this distribution have been generated by a stochastic process ?

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.

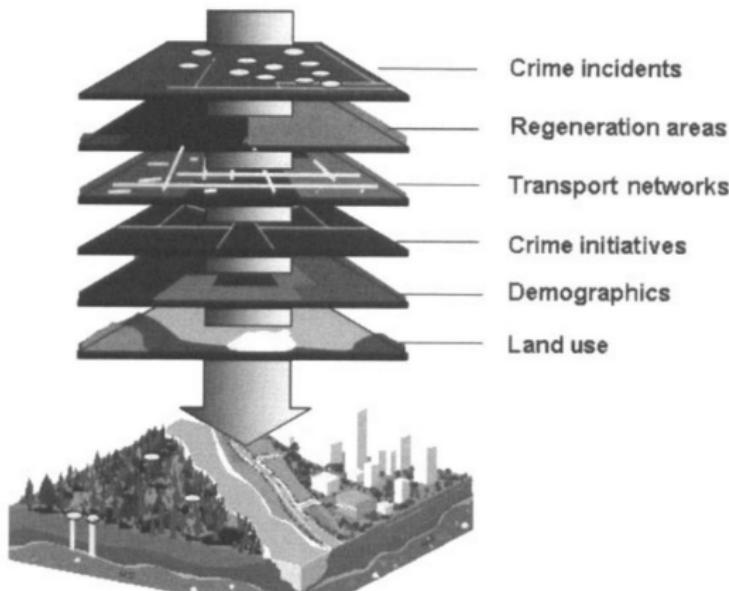
# Distribution testing

May this distribution have been generated by a stochastic process ?



# Distribution testing

May this distribution have been generated by a stochastic process ?

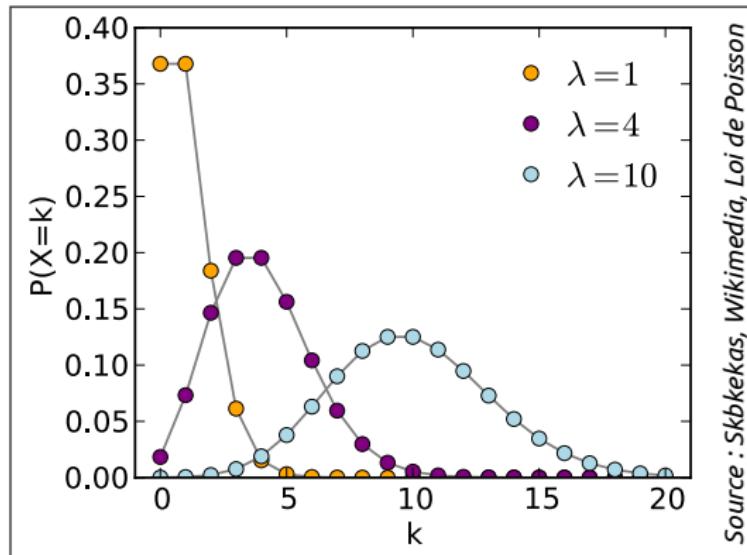


Source : Chainey, Ratcliffe (2005) *GIS and crime mapping*, Wiley.

# Poisson distribution

Poisson's distribution  $\lambda$  parameter is both the **mean** and the **variance** of the distribution.

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



Source : Skbekkas, Wikimedia, Lai de Poisson

# 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 occurrences in a zone  $z$  is modeled by a Poisson distribution whose mean is  $\lambda \times \text{area}(z)$ .

# Dispersion index

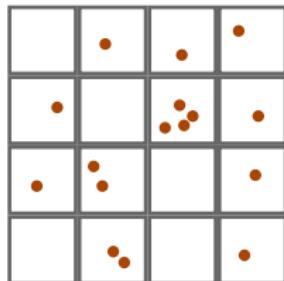
**VMR** *Variance-to-Mean Ratio* =  $\frac{\mu}{\sigma^2}$

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

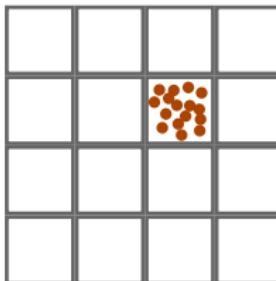
## VMR interpretation

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

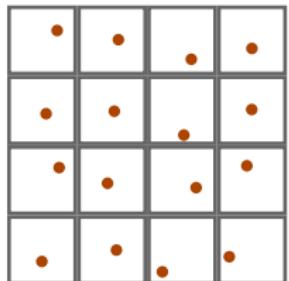
# Dispersion index



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



X (nbr. occurrences)  
[0 0 0 0 0 0 1 6 0 0 0 0 0 0 0 0]



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

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

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

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

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

# Quadrat methods

« A spatial  $\chi^2$  »

- ▶ Construct a regular grid (quadrats) : observed distribution
  - ▶ Theoretical distribution (null model) is given by a spatial Poisson process.
  - ▶ Contingency tables with observed and expected occurrences
- **test** : Observations differ significantly from expected values ? ( $\chi^2$  test)

# Interactions & Networks

DELHI GIS-R School  
9-12<sup>th</sup> April 2019

**Hadrien Commenges & Paul Chapron**

[hadrien.commenges@univ-paris1.fr](mailto:hadrien.commenges@univ-paris1.fr)

[paul.chapron@ign.fr](mailto: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

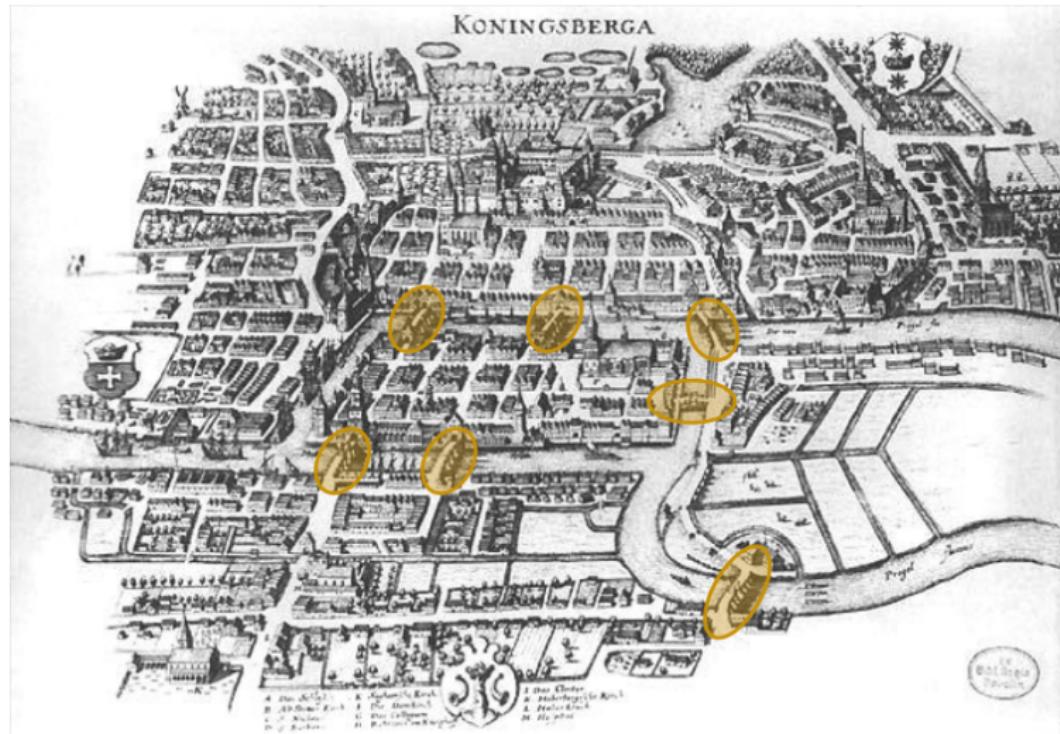
ori	des	poids
A	B	2
A	C	5
B	A	3
B	C	4
C	A	0
C	B	0

matrice d'interaction

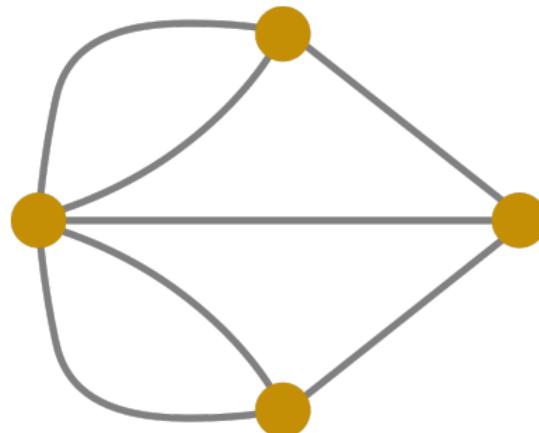
	A	B	C
A	0	2	5
B	3	0	4
C	0	0	0

# Network Analysis

Leonhard Euler and the Seven Bridges of Königsberg (Kalininograd).



# 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 :

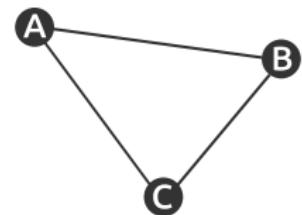
*edges list*

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".

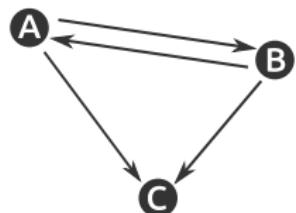
*edges list*

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 :

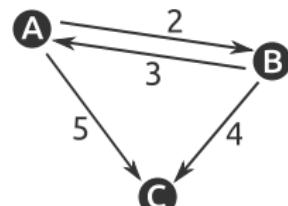
*edges list*

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

## → Maximum number of edges (no loops) :

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

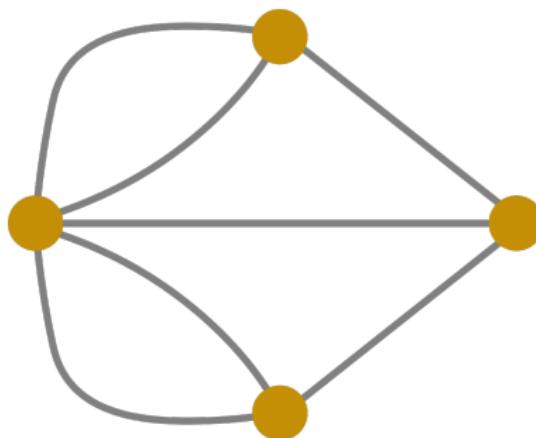
# Network analysis measures

## Nodes measures :

- ▶ Degree : number of edges (neighbors) of a node
- ▶ in-degree / out-degree : number of incoming arcs / outgoing arcs
- ▶ Weighted degree : sum of edges weights
- ▶ Closeness centrality : inverse of the sum of the shortest 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 are many (an infinity) paths connecting one node to every other node.



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

## Dijkstra Algorithm principles

Dijkstra (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 edges 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)

# Dijkstra's pseudo code

Init

$$P \leftarrow \emptyset$$

$$d(v) := 0, \forall v \in V \text{ (distances to } V_{orig})$$

$$d(V_{orig}) := 0$$

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>

# Accessibility

DELHI GIS-R School  
9-12<sup>th</sup> April 2019

**Hadrien Commenges & Paul Chapron**  
[hadrien.commenges@univ-paris1.fr](mailto:hadrien.commenges@univ-paris1.fr)  
[paul.chapron@ign.fr](mailto:paul.chapron@ign.fr)

# Use case

## Accessibility

Accessibility depicts the potential access of an agent to a resource.

two meanings :

1. **Access to a location** : required effort for agents to get to a location ( $\approx$  service area). Such a measure depends on the network (infrastructure + services) and the agent's capacity to travel.
2. **Access to a resource** : required effort for agents to get to a resource. Such a measure depends on the network, the agent's capacity to travel and the resource **spatial distribution**.

What kind of geographical information is concerned ?

- *TYPE 1 - Geographical Objects*
- *TYPE 2 - Occurrences*
- *TYPE 3 - Measure points*
- *TYPE 4 - Statistics*
- *TYPE 5 - Interaction Measures*

# Distance and interaction

- ▶ **friction** (*impedance*) is the key-concept, distance is a measure of friction.
- ▶ Spatial interaction is addressed via **interactions between locations** (flows) or **attraction/influence of a location on the others**.
- ▶ Accessibility models rely on the latter.