UNIVERSIDADE DO MINHO

Spatial Data Analysis for a Spatial World

Raquel Menezes

Centre of Mathematics, University of Minho, Portugal
rmenezes@math.uminho.pt

July 5, 2024

Where can I find these presentation slides?

Raquel Menezes's GitHub – http://github.com/rmenezesmotaleite/SummerSchoolUV

Outline

1. Introduction to Spatial Statistics
   - ▷ Data Analysis and Visualization
2. How do **Areal**, **Point Pattern** and **Geostatistical** Data differ ?
   - ▷ Try a QUIZ
3. Additional Information on Geostatistics
   - ▷ What's Spatial Correlation?
4. Applications to Fisheries and Marine Sciences
   - ▷ How Can Spatial Statistics Help?
   - ▷ Try DYNAMIC MAPS

The Beginning of Spatial Statistics

John Snow and the 1855 London Cholera Epidemic

Background

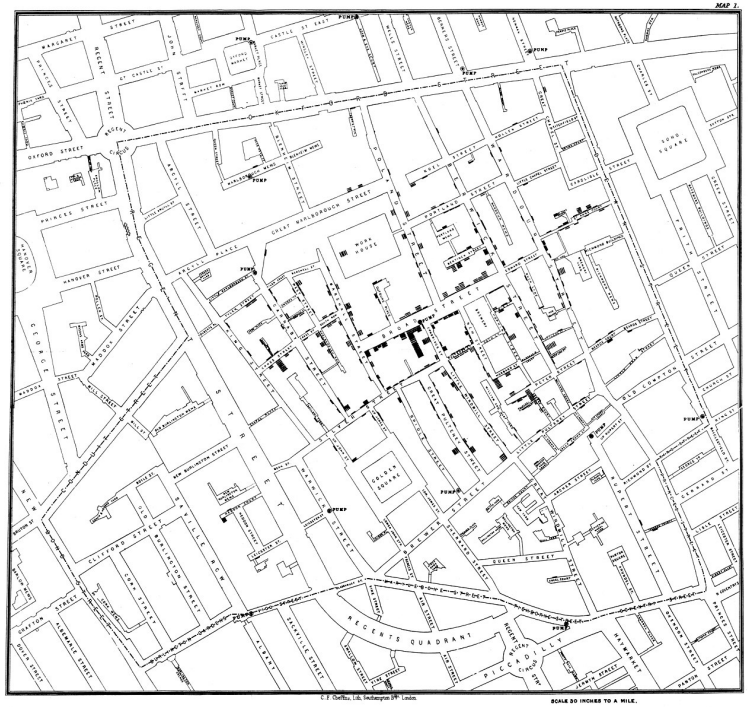
- **John Snow (1813-1858):** A pioneering British physician known for his work in epidemiology.

The Cholera Epidemic

- **London, 1854:** A severe cholera outbreak hits Soho, causing widespread panic and numerous deaths.
- **Common Belief:** Cholera was believed to be spread by "miasma" or bad air.

Snow's Insight

- **Hypothesis:** Snow theorized that cholera was waterborne, not airborne.
- **Investigation:** He meticulously mapped cholera cases in Soho.



The Breakthrough

- **Pump Analysis:** Snow identified a pattern centering around Broad Street water pump.
- **Action Taken:** Persuaded authorities to remove the pump handle, leading to a dramatic decline in cases.

Impact on Spatial Statistics

- **Mapping Cases:** Snow's use of maps to identify the outbreak's source is considered a foundational moment in spatial statistics.
- **Legacy:** Demonstrated the power of spatial data in understanding and solving public health issues.

Exploratory Data Analysis

- Exploratory data analysis (EDA) is an integral part of Statistics in the context of applications, which **must precede any modeling approach**. Spatial Statistics is no exception.

- In our case, EDA is naturally oriented towards the preliminary investigation of the **spatial aspects** of the data that are relevant to our modeling assumptions.

- However, **non-spatial aspects** can and **should also be investigated**.

EDA and Its Main Objectives

We must consider:

- **Numerical summaries**, known as descriptive statistics (means, medians, quantiles, variance, . . .)
- **Graphical summaries** related to preliminary analysis, through data visualization

Know your data!

- distributions (symmetric, normal, asymmetric)?
- data quality issues?
- outliers or extreme values?
- correlations and interrelations?
- subsets of interest?
- suggestions of functional relationships?

Sometimes, EDA or data visualization can be the main objective!

Why Punctual Statistics Are Not Sufficient?

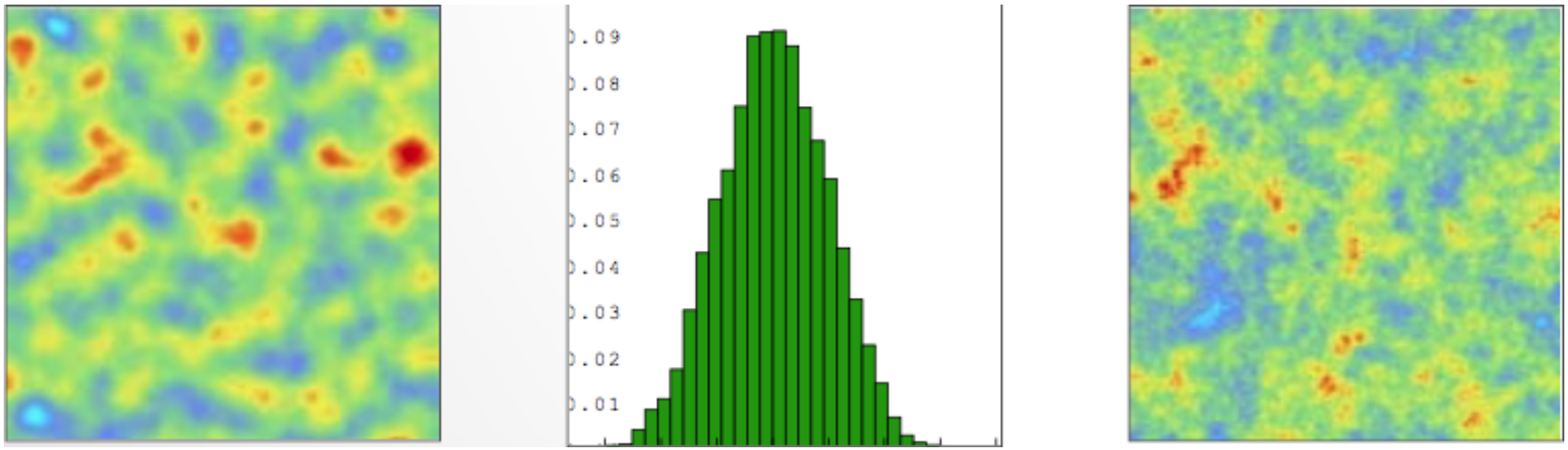


Figure: Two random fields (highest values and lowest values) with the same histogram.

Data Visualization

Human beings are the best pattern identifiers, so the analysis of graphical summaries can be quite productive.

The **visual methods** to be chosen will depend on whether we have

- one, two, or more variables?
- categorical or quantitative variables?
- geographic reference or temporal reference?

The methods may include

- boxplots or histograms
- scatter plots or bar charts
- **mapping observed spatial data in a region**
- time series plot over the observed period

An Example - Meuse River Data (*gstat* R package)

Top soil **heavy metal concentrations (ppm)**, collected in a flood plain of river Meuse, near village Stein:

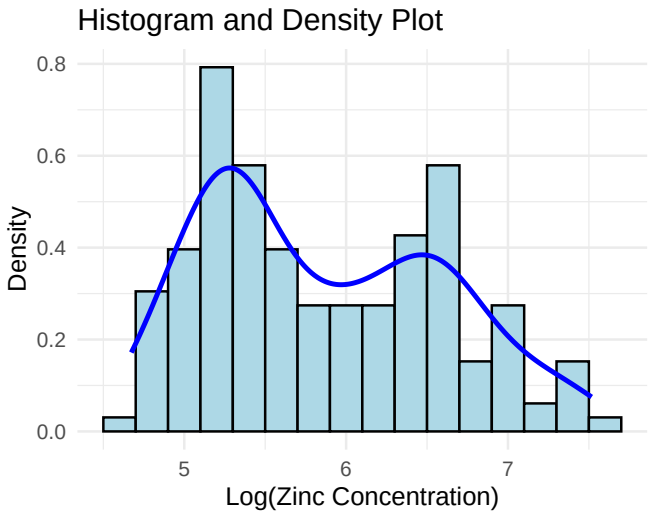
- cadmium (Ca)
- copper (Cu)
- lead (Pb)
- zinc (Zn)

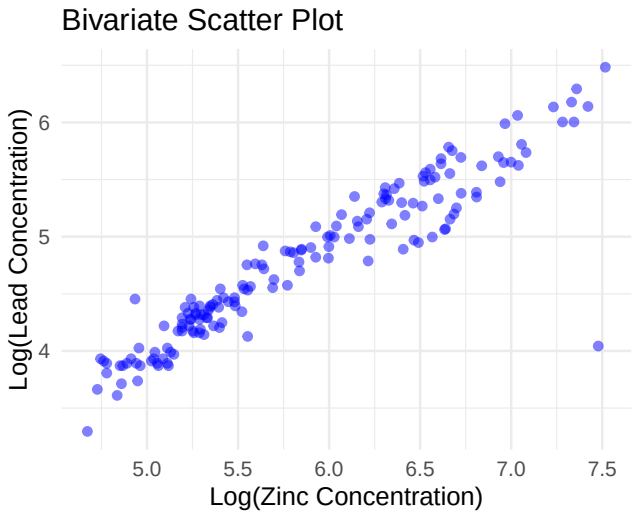
Data	Ca	Cu	Pb	Zn
Original				
mean	3.11	39.42	148.55	464.60
median	1.9	29.5	116.0	307.5
st.dev	3.49	23.40	110.18	376.57
Logarit.				
mean	0.52	3.53	4.77	5.86
median	0.64	3.38	4.75	5.73
st.dev	1.21	0.51	0.67	0.73

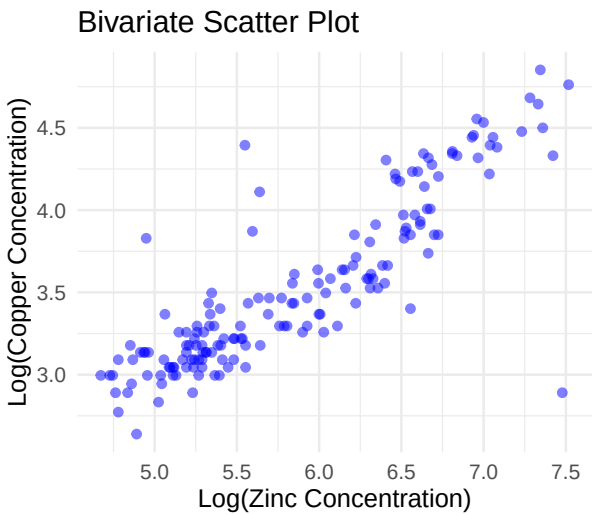
Table: Example of numerical summaries.

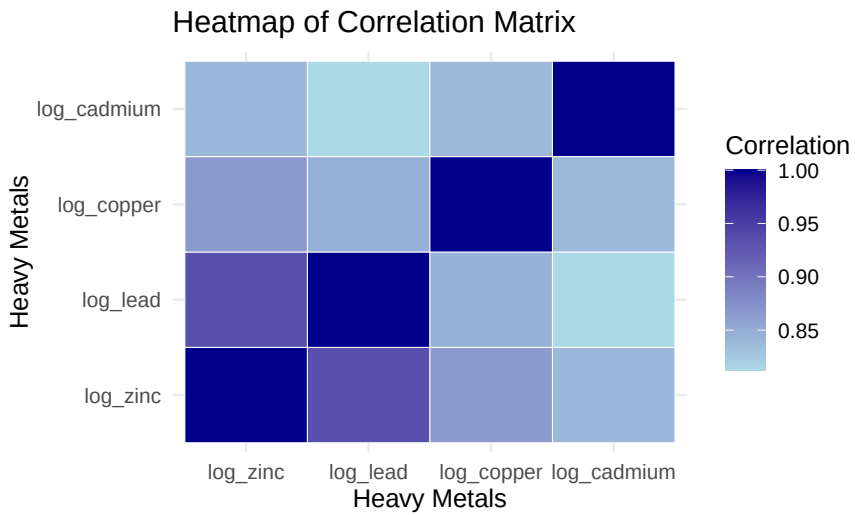
Examples of Graphical Summaries

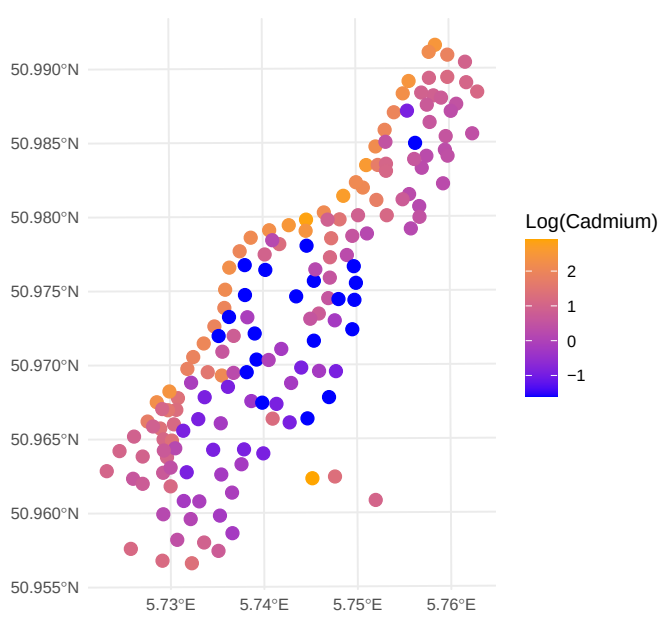
Univariate, Bivariate and Multivariate

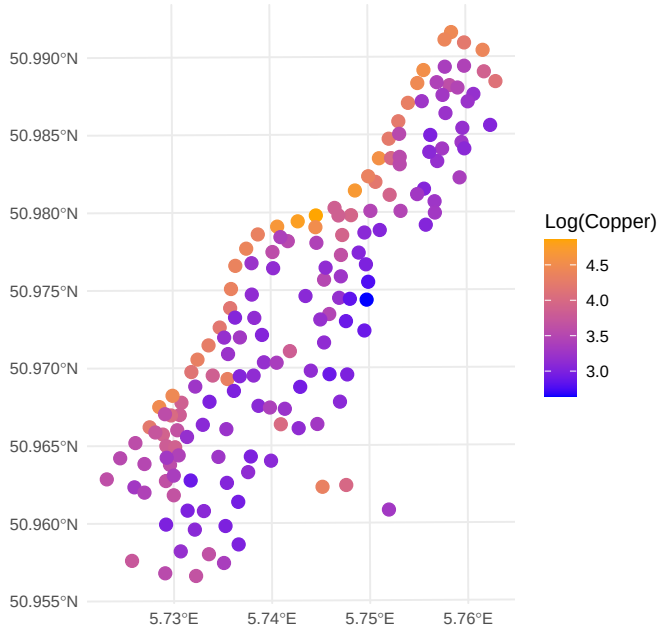


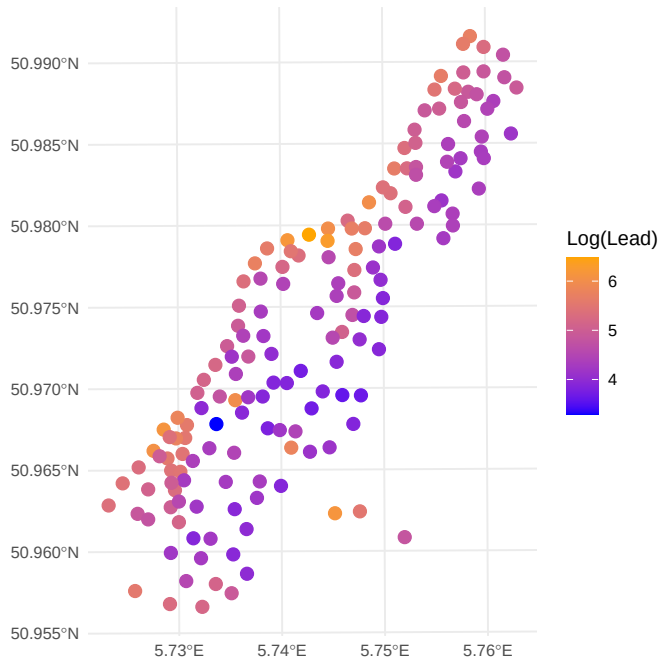


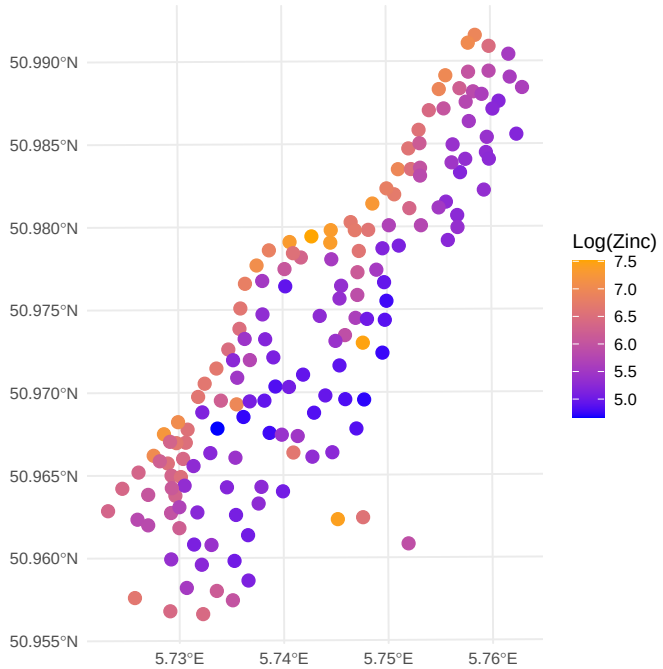





50.990°N
50.985°N
50.980°N
50.975°N
50.970°N
50.965°N
50.960°N
50.955°N
5.73°E
5.74°E
5.75°E
5.76°E
Log(Cadmium)
2
1
0
−1


50.990°N
50.985°N
50.980°N
50.975°N
50.970°N
50.965°N
50.960°N
50.955°N
5.73°E
5.74°E
5.75°E
5.76°E
Log(Copper)
4.5
4.0
3.5
3.0


50.990°N
50.985°N
50.980°N
50.975°N
50.970°N
50.965°N
50.960°N
50.955°N
5.73°E
5.74°E
5.75°E
5.76°E
Log(Lead)
6
5
4


50.990°N
50.985°N
50.980°N
50.975°N
50.970°N
50.965°N
50.960°N
50.955°N
5.73°E
5.74°E
5.75°E
5.76°E
Log(Zinc)
7.5
7.0
6.5
6.0
5.5
5.0

Summarizing remarks

Exploratory data analysis and data visualization:

- It helps to get a general sense of the data.
- You should always look at every variable - you will learn something!
- It's a data-driven (model-free) approach.

So it is always well worth looking at your data, before moving on to any modelling approach!

Outline

1. Introduction to Spatial Statistics
   - ▷ Data Analysis and Visualization
2. How do **Areal**, **Point Pattern** and **Geostatistical** Data differ?
   - ▷ Try a QUIZ
3. Additional Information on Geostatistics
   - ▷ What's Spatial Correlation?
4. Applications to Fisheries and Marine Sciences
   - ▷ How Can Spatial Statistics Help?
   - ▷ Try DYNAMIC MAPS

Spatial Data

In spatial statistics, our process of interest $Y(x)$ is defined over a spatial region $x \in A$, and there are observations at specific locations $x_1, \ldots, x_n$.

Depending on the **nature of the data** and the **spatial aggregation** we give them, we can differentiate three types of spatial data:

1. Areal data, also known as lattice data
2. Data referring to point patterns
3. Point-referenced data, known as geostatistical data

1. Lattice Data

- Data referring to areas, or simply lattice data, represent an **aggregation of observations** $Y(x)$ over a predefined *unit of area*[1], such as
  - Number of cases of a specific disease in each area
  - Percentage of unemployment in each area

- The region of study is divided into a **finite collection of area units** with well-defined boundaries, such as
  - Autonomous communities of Spain
  - Districts or municipalities in Portugal
  - Countries in the European Union

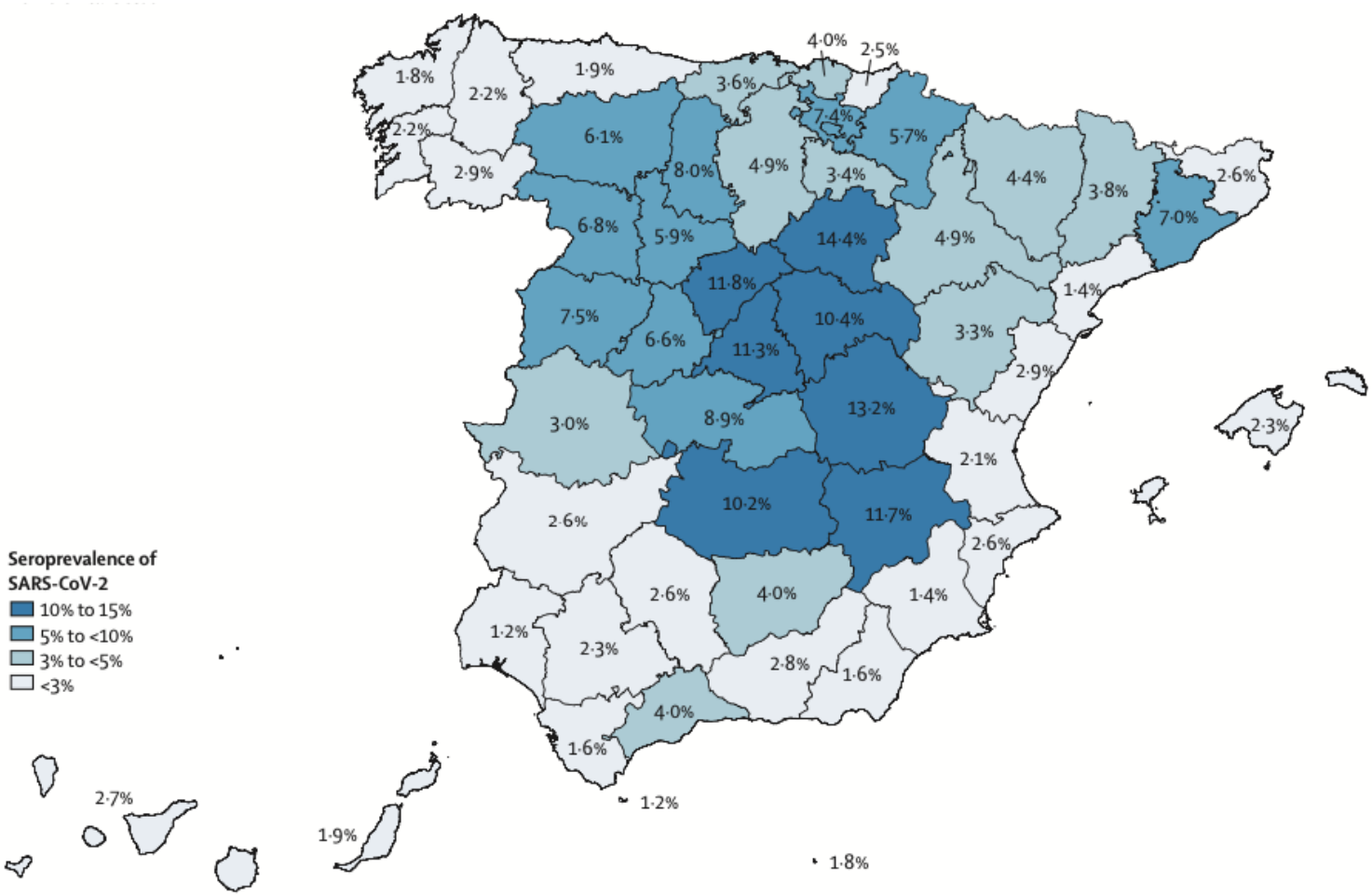
In the modeling of lattice data, one must consider **whether adjacent regions have similarities**, in the sense that it is expected that nearby areas have more in common than distant areas.

[1] Point x might represent the centroid of that unit area.

Modeling Aggregated Data

- Modeling area data involves obtaining information from adjacent regions.
- One common model structure in these cases is the **conditional autoregressive model** (Besag et al., 1991), better known as the **CAR** or **BYM** model after the authors' initials.
- These models consider spatial autoregressive correlation through an adjacency structure of the area units.

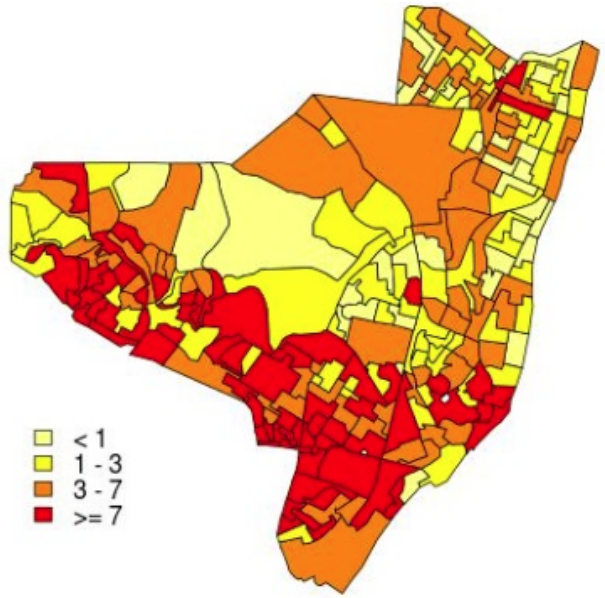
Example of Aggregated Data



Prevalence of SARS-CoV-2 in Spain, serological survey from April 27 to May 11, 2020, involving 61,075 participants (Pollán et al., 2020)

Example of Aggregated Data

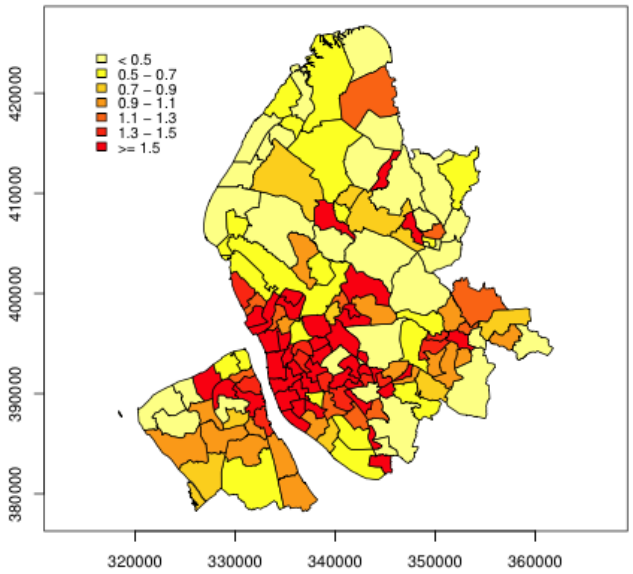
Leprosy Surveillance, Olinda NE Brazil (Bailey, 2008)



- Olinda is a municipality in the state of Pernambuco, in northeastern Brazil, composed of **241 sectors** with approximately 350,000 inhabitants.
- Available data on the incidence of new leprosy cases by sector in the period 1991-1995 (total of 1135 cases), along with population estimates for the mid-period (1993) in these sectors.
- A simple deprivation indicator is also available - proportion of household heads with a monthly income below the legal minimum wage (US$80).

Example of Aggregated Data

Laryngeal Cancer, Northwest England (Bailey, 2008)



- Data on 876 cases of laryngeal cancer diagnosed in 1982-1991, in **144 electoral wards** of Mersey and West Lancashire.
- Research allowed the definition of a smoking prevalence indicator in each district ('low', 'medium', or 'high').
- Available annual air pollution measure based on traffic flow.

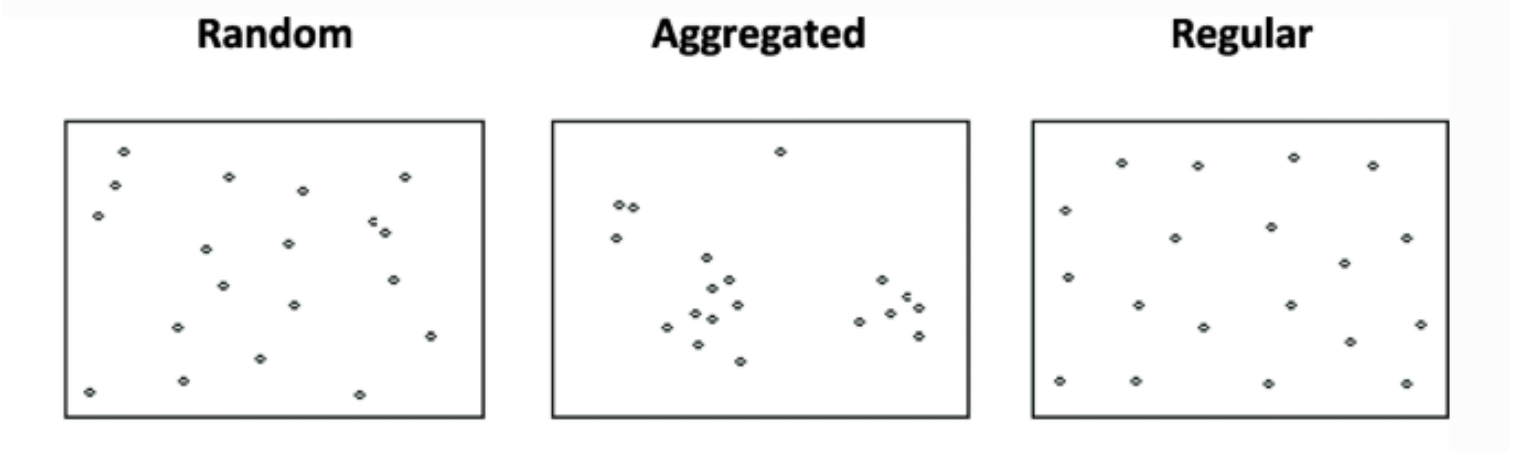
2. Data referring to Point Patterns

- The theory of point patterns emerged from the need to model the location of events of interest as random.
- The first applications were in the fields of ecology (e.g., locations **where a rare bird is sighted**) and forest sciences (e.g., **location of a type of trees**).
- However, their range of applications is very broad. We can consider the distribution of vessels at sea, or the **distribution of known cases of a particular contagious disease**.

In these various contexts, the goal is to study the spatial arrangement of the event of interest in space, to identify important areas (e.g., favorable areas for a given phenomenon).

Modeling Point Patterns

- One starts with the hypothesis of complete spatial randomness, i.e., the spatial pattern shows no apparent structure (left-panel).
    - It is assumed that the number of events in a region follows a **Poisson distribution** with a mean proportional to the area of the region and the average number of events per unit area – the intensity of the process.



- Alternatively, we may have a pattern corresponding to strongly clustered events (central-panel), or a regular pattern (regular-panel) if a minimum distance between events is imposed.

Modeling Point Patterns

- The response $Y(x)$ is fixed (1=presence) and the locations x are generated randomly from the spatial random field Λ.

- We need to study the underlying spatial structure, using the topological, geometric, or geographic properties of the observed locations.

Note: We have a **marked point process** if some information is associated with the point x (for example, if the point process is defined by the occurrence of a disease, we can associate the gender of the individual to each point).

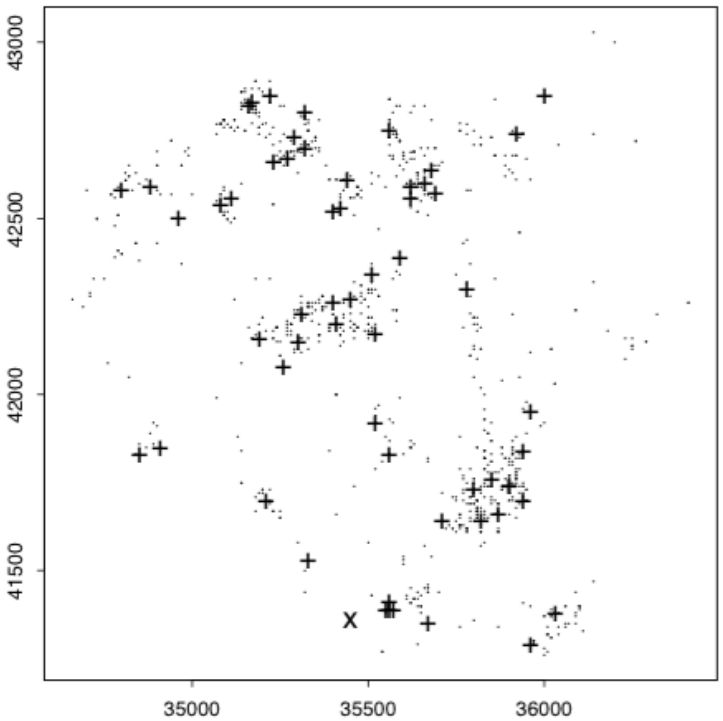
Possible Scientific Questions

Some of the most relevant questions for point patterns data are typically related to **clustering events**.

- Is the spatial distribution of the observed points homogeneous in space?
- Or do we have a clustering process?
- If there is clustering, what are the reasons that could justify it?

Example of Point Patterns

Lung and Larynx Cancers (Diggle, Gatrell, and Lovett, 1990)



The map shows:

- cases of lung cancer (dots)
- cases of larynx cancer (crosses)
- a deactivated industrial incinerator (x symbol)

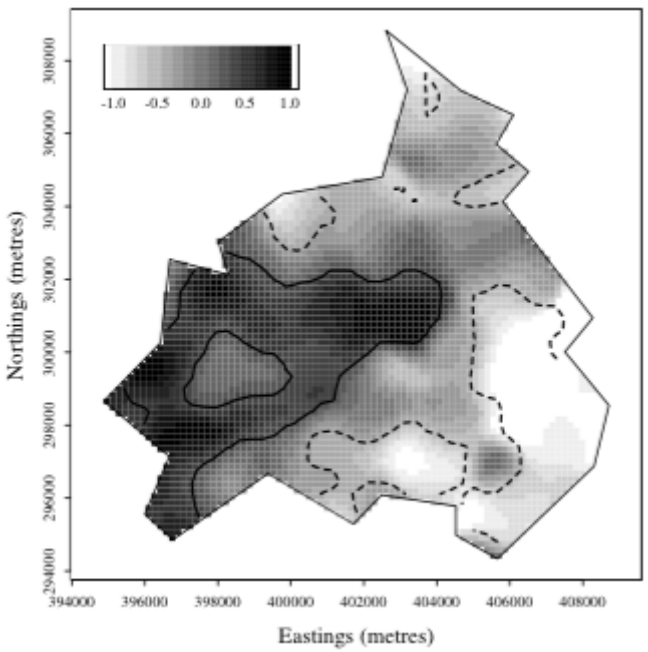
Important Questions:

- Do the cases show a *surprising* tendency to cluster?
- Does the disease risk vary spatially?
- Is the disease risk elevated near a specific location?

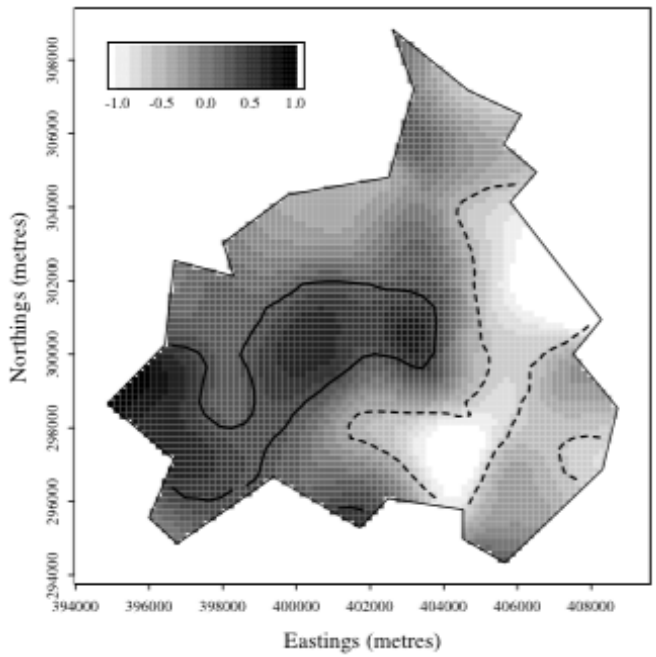
Example of Point Patterns

Estimated Risk Surfaces (based on kernel smoothing)

lung cancer



stomach cancer



Some comments:

- Similar variation pattern in both diseases
- Solid and dashed lines identify the boundaries of regions where the risk is significantly higher or lower, respectively, than the average

3. Geostatistical Data

- Geostatistical data (or point-referenced data) consist of a **stochastic process $Y(x)$ associated with a fixed set of locations x over a continuous spatial field $S(x)$**.

- The space is typically treated as **two-dimensional**, defined by its **longitude** and **latitude**, but it can also include altitude or depth to make it three-dimensional.

- Examples of geostatistical processes:
  - $S(x)$ represents the pollution surface over the city of Lisbon, and $Y(x)$ represents a pollution indicator measured at the monitoring station at location x.
  - In fishing, $S(x)$ may represent the abundance surface of a species and $Y(x)$ the catch of that fish at location x.

Geostatistical Data

- They can be referred to as **spatially continuous data**.
- The term **continuous** does not **mean** that the variable of interest is continuous, but rather that **it can be measured at any location in the study region**.
- Such continuously distributed variables were traditionally used in geosciences for the analysis of mineral concentrations, which explains the term Geostatistics.
- Nowadays, they are widely used in various contexts, as long as geographic location is explicitly used in the data analysis.

  Examples: **sea surface (or air) temperature**, or **salinity**, or some measure of **fish abundance**, such as egg concentration.

Modeling Geostatistical Data

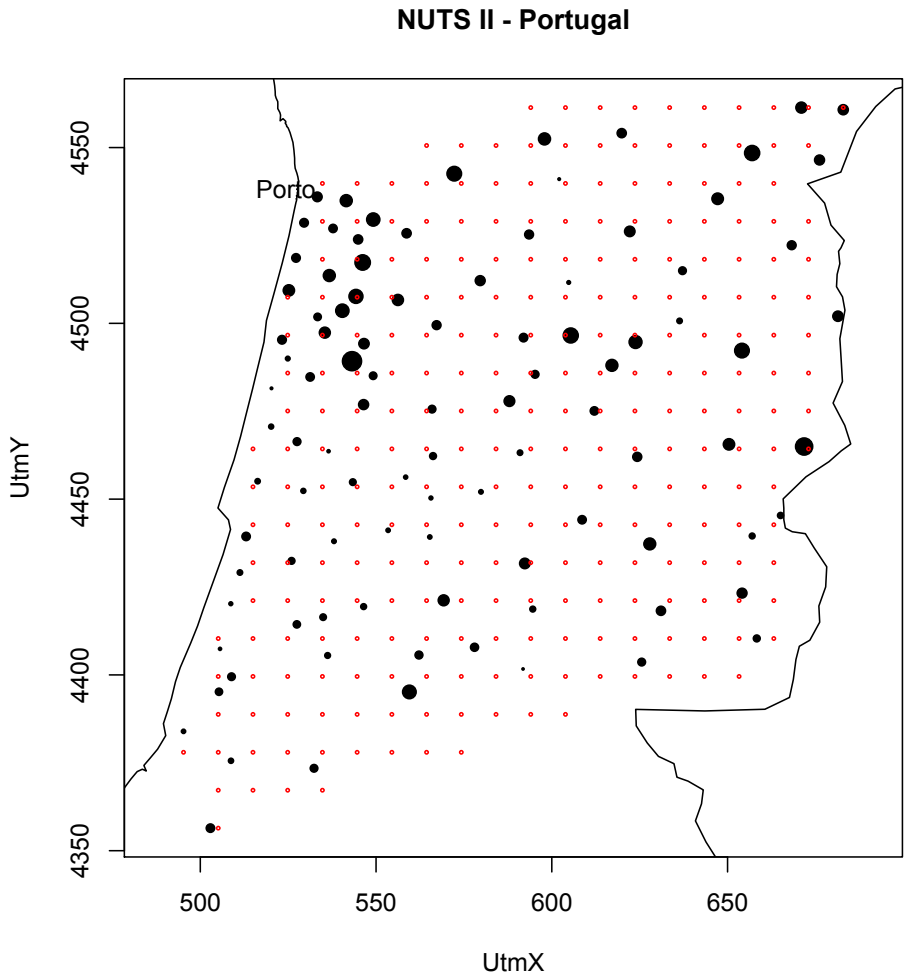
- When modeling our **variable of interest**, conditioned on **predictors**, it is expected that the respective **residuals are spatially correlated**.

- Our main objective is to **infer the spatial correlation** underlying our data to improve prediction, using (for example) **kriging** techniques at unsampled locations.

Obviously, in all three sub-areas of Spatial Statistics, the **spatial dimension can be extended to the spatiotemporal domain** by adding the correlation of the variable of interest between temporal events (e.g., studying fish abundance every hour, day, etc).

Example of Geostatistical Data

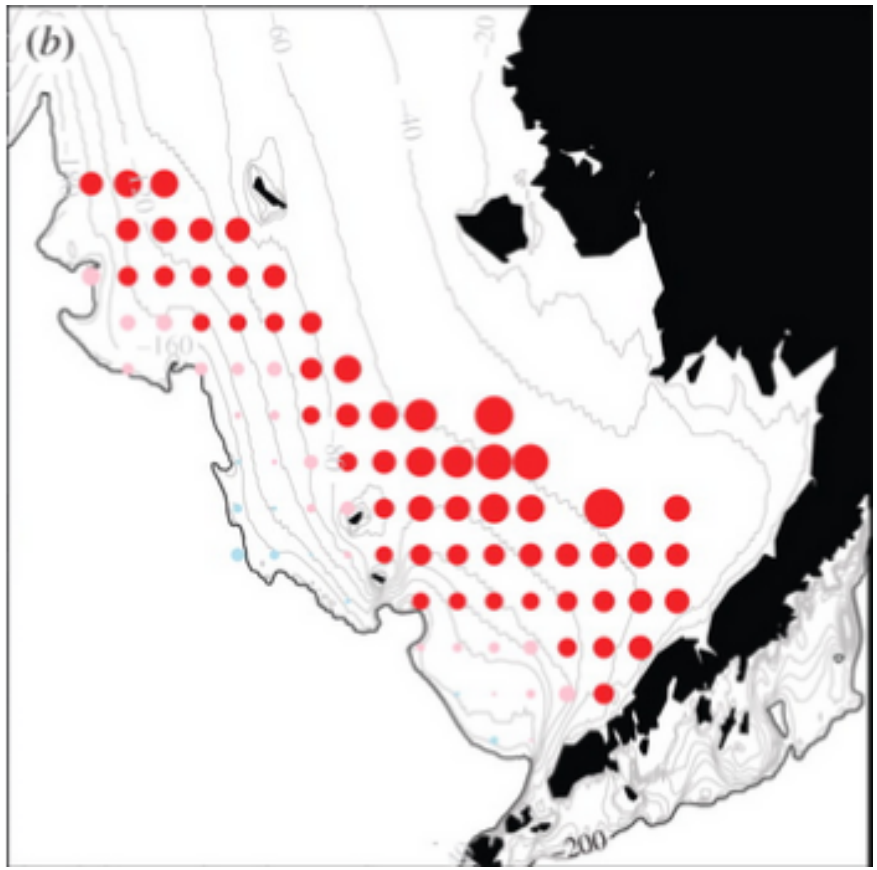
Arsenic pollution data $Y(\mathbf{x})$ with $\mathbf{x} \in A \subset IR^2$ (Garcia-Soidán and Menezes, 2017).



Note: Black circles identify the 98 measurements of *As*, with the circle diameter proportional to the observed value. Red points identify the grid points where *As* is to be estimated.

Example of Geostatistical Data

Ciannelli et al. (2012) modeled the **distribution of adult fish** in the Bering Sea. Measurements $Y(x)$ are made at discrete locations x of the continuous domain A.



Interactive Quiz – Identify the Area of Spatial Statistics

Areal data? Point patterns? Geostatistical data?

Instructions:

- Use your phone to scan the QR code
- Take the quiz to match each case study with the corresponding area of spatial statistics

SCAN ME

Outline

1. Introduction to Spatial Statistics
   - ▷ Data Analysis and Visualization
2. How do **Areal**, **Point Pattern** and **Geostatistical** Data differ ?
   - ▷ Try a QUIZ
3. Additional Information on Geostatistics
   - ▷ What's Spatial Correlation?
4. Applications to Fisheries and Marine Sciences
   - ▷ How Can Spatial Statistics Help?
   - ▷ Try DYNAMIC MAPS

Fundamental Concept of "Spatial Dependence"

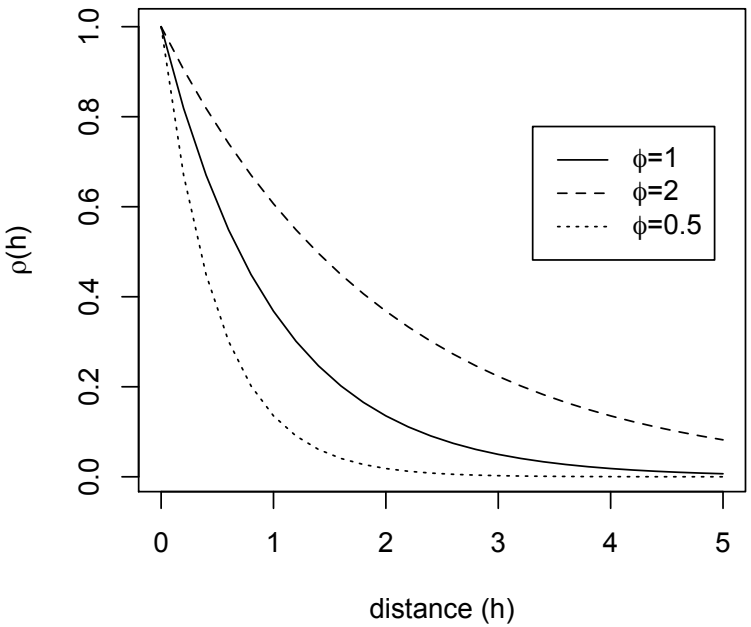
Everything is related to everything else, but near things are more related than distant things (first law of geography)

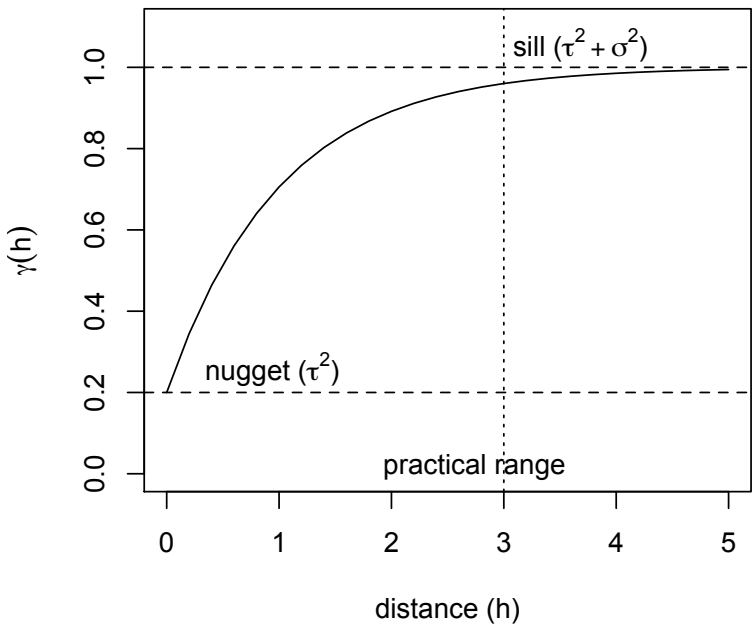
Waldo Tobler, 1970

Moreover . . . typically, beyond a certain distance between two observations, they become (spatially) uncorrelated/independent.

How to Control the Spatial Correlation?

Under certain stationarity (stability) conditions:

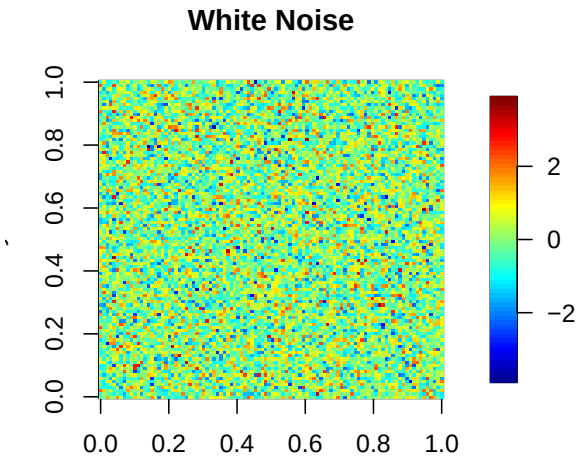


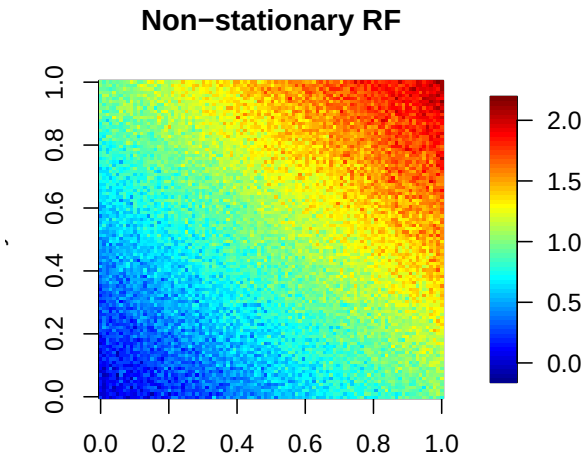


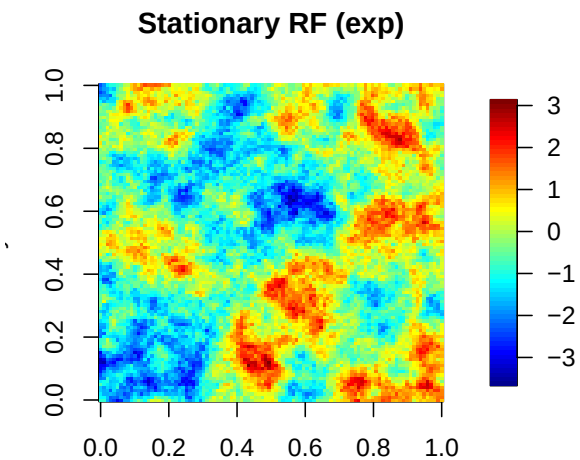
Left Panel: Exponential correlation function for different values of the radius of influence, also called *range*, ϕ.

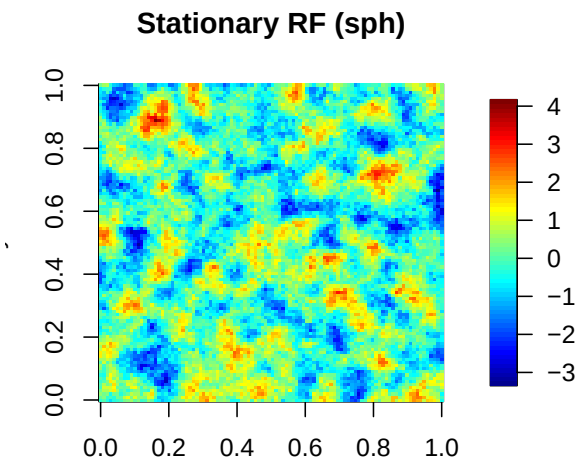
Right Panel: Schematic representation of a variogram with its structural parameters.

Simulations of Geostatistical Data (also named Random Fields, RF)









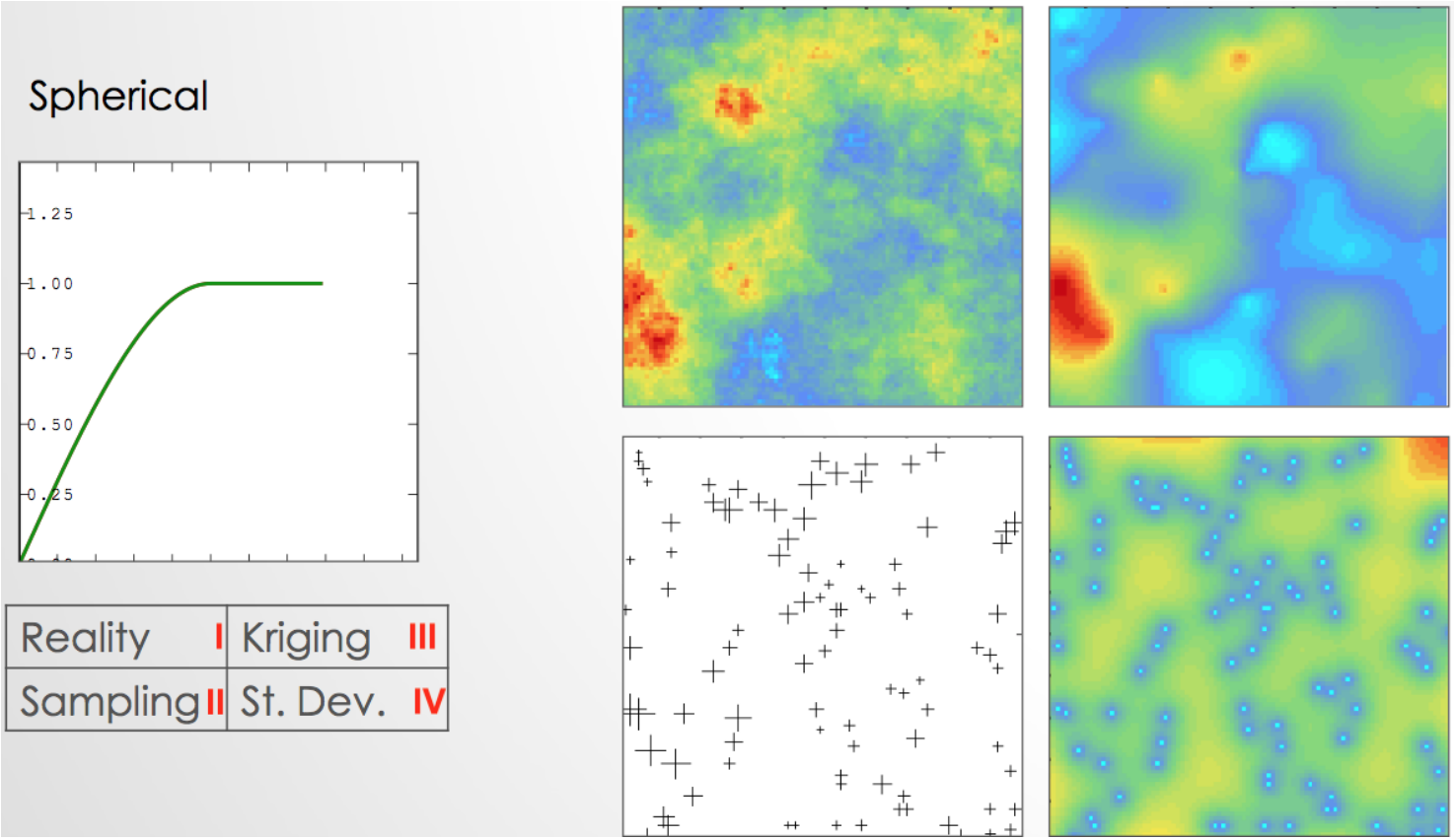
Simulation Example: True RF versus Predicted RF

I **Simulate a RF** given a spatial correlation

II **Extract a sample** with 100 observations

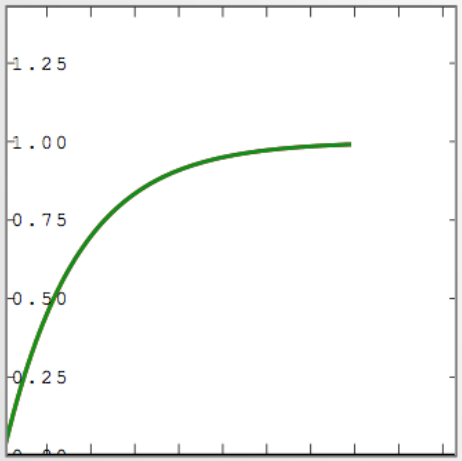
III **Predict the RF** given the sample (kriging method)

IV Estimate a prediction error

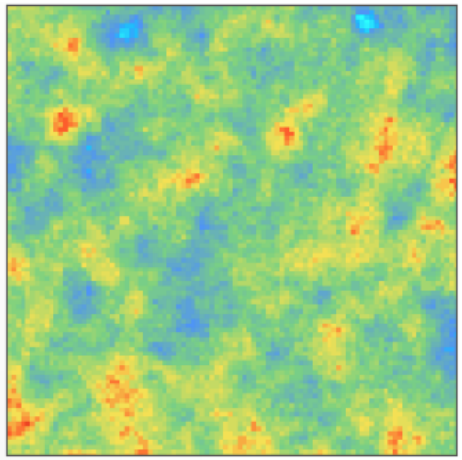


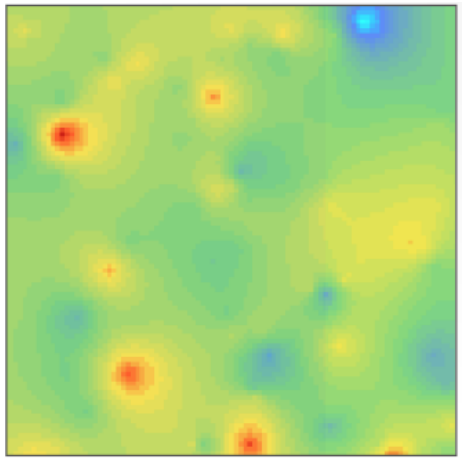
Simulation Example: True RF versus Predicted RF

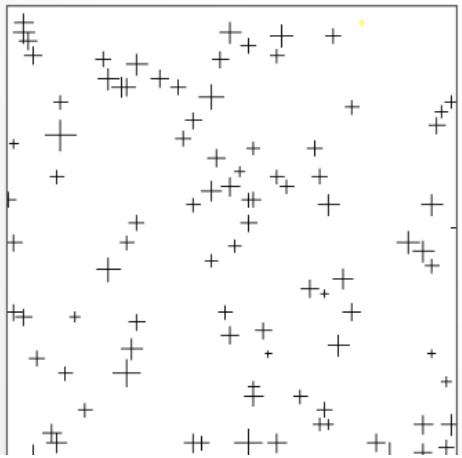
Exponential

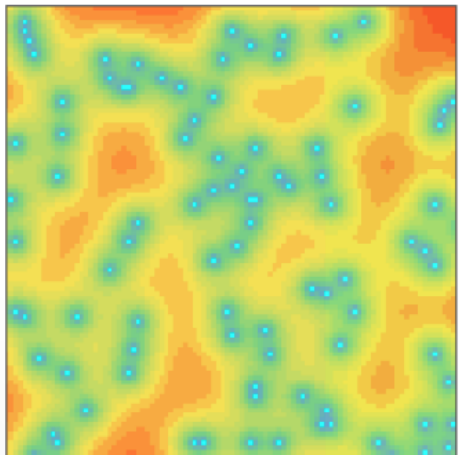


Reality	I	Kriging	III
Sampling	II	St. Dev.	IV

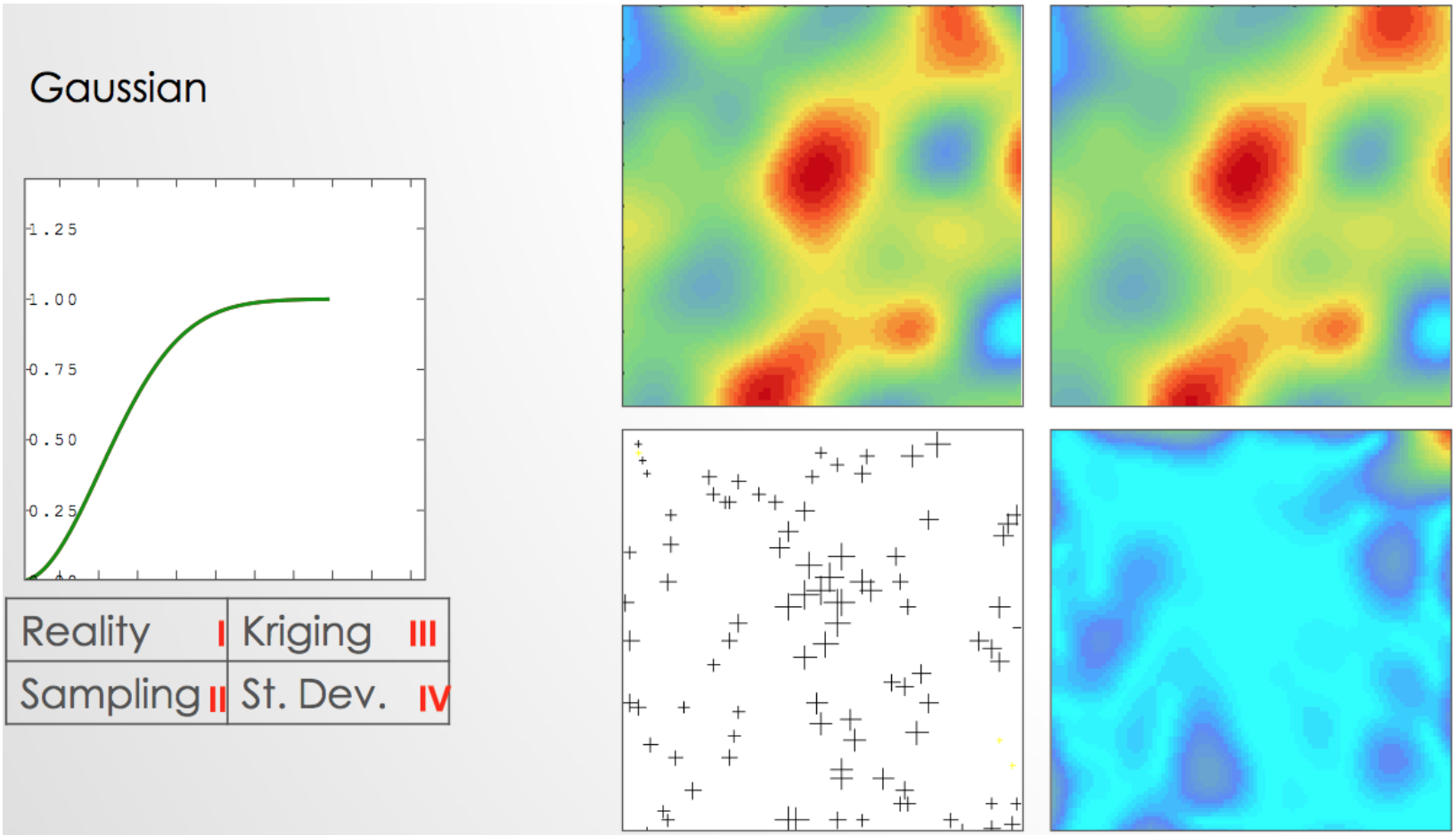








Simulation Example: True RF versus Predicted RF



Outline

1. Introduction to Spatial Statistics
   - ▷ Data Analysis and Visualization
2. How do **Areal**, **Point Pattern** and **Geostatistical** Data differ ?
   - ▷ Try a QUIZ
3. Additional Information on Geostatistics
   - ▷ What's Spatial Correlation?
4. Applications to Fisheries and Marine Sciences
   - ▷ How Can Spatial Statistics Help?
   - ▷ Try DYNAMIC MAPS

Marine Spatial Planning (MSP)

- MSP started in the 1980's, with involvement of ICES[2].
- Conflicts arise in **resource usage**, **environmental impact**, and **legal frameworks**.

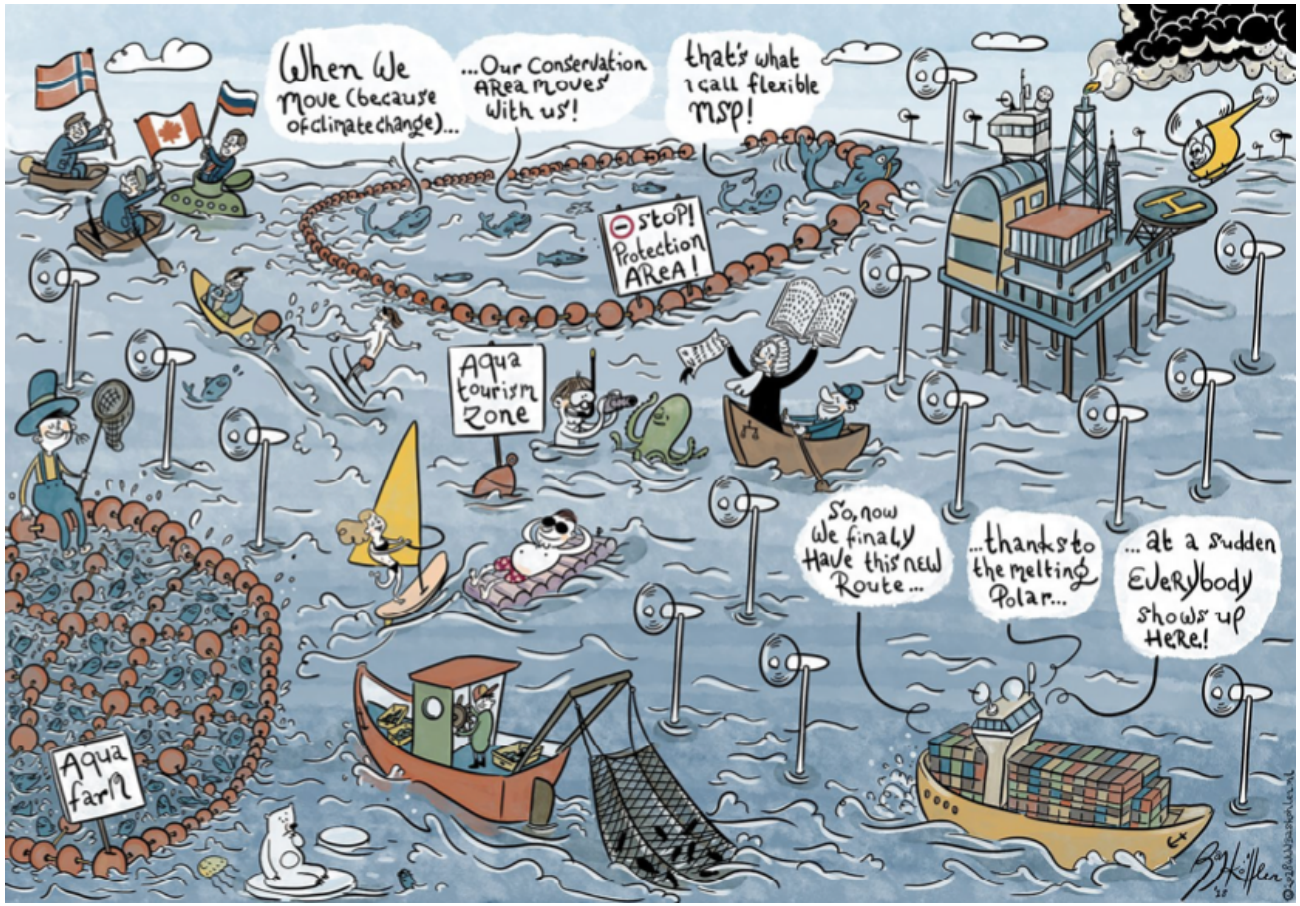


Figure: A crowded ocean under a changing climate. Cartoon created by Bas Kohler (Santos et al., 2020).

[2]International Council for the Exploration of the Sea (ICES)

The Portuguese Case – IPMA / ICES

The **Portuguese Institute for Sea and Atmosphere (IPMA)** is a public institution under the Ministry of Economy and Sea, supporting the **strategic plans in term of analysis and survey of Marine Resources**.

Key Objectives:

1. **Combating marine biodiversity loss**, whose main causes are
   - Climate changes
   - Illegal fishing activities
   - Difficult management of fishery resources
2. **Improving knowledge on the population dynamics** of marine species with commercial interest, such as sardines, anchovies, hake, etc.

recovery of marine biodiversity

$+$

sustainability of fisheries

Applications to Fisheries and Marine Sciences

Kickoff project PTDC/MAT-STA/28243/2017 – PREFERENTIAL

Improving spatial estimation . . . through preferential sampling in fishery and biological applications

UMinho - CMAT / CEAUL / CBMA

Raquel Menezes

Daniela Silva

Guido Moreira

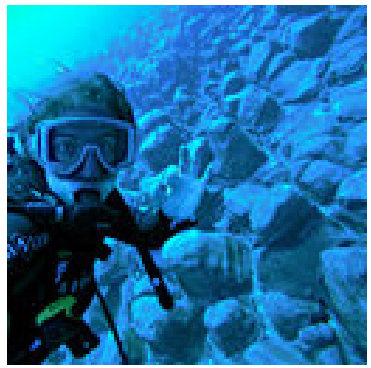
Fran Iziquierdo

IPMA - Portuguese Institute of Sea and Atmosphere

Susana Garrido

Alexandra Silva

Ana Machado



Barbara Pereira

How can Spatial Statistics help?

Creating mathematical/statistical models for ...

- **Space-time analysis** of abundance indicators. Examples:
  - What are the best places to fish for sardines? And at what time of the year?
- **Studying the influence** of environmental/biological variables. Examples:
  - Do sardines prefer salty water? And what temperature?
- **Construction of maps**. Examples:
  - What are the preferred locations for juvenile individuals?

Summarizing... **spatial statistics helps to assess the status of fish populations and make informed management decisions!**

There Are No *Omelets* Without *Eggs*!

Omelets = **Models/Maps** and Eggs = **DATA**

Sources of DATA:

1. **Scientific campaigns** (+ expensive)
   - Based on research surveys, often using standardized sampling techniques
   - For example, IPMA Spring/Autumn campaigns

2. **Commercial fisheries** (+ economical)
   - Information collected directly from fishing activities, with data typically collected through logbooks and geo-reference obtained through Automatic Identification System (AIS)

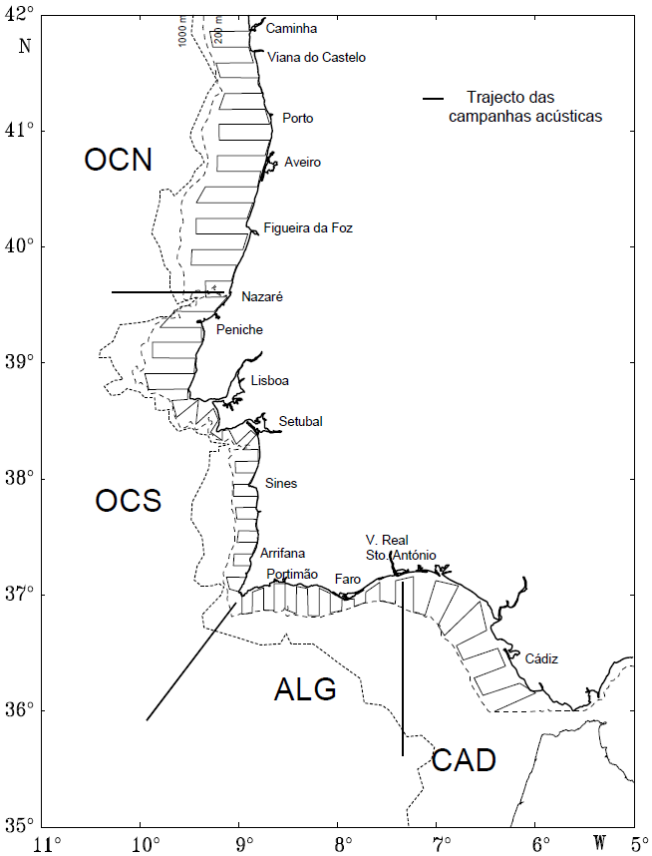
3. **Satellites data**
   - ▷ sea surface temperature (SST) in $^{\circ}C$
   - ▷ chlorophyll-a concentration (CHL) in mg/m^3
   - ▷ bathymetry in m
   - ▷ intensity and direction of surface ocean currents in m/s and degrees

Data sources

Scientific surveys

↓

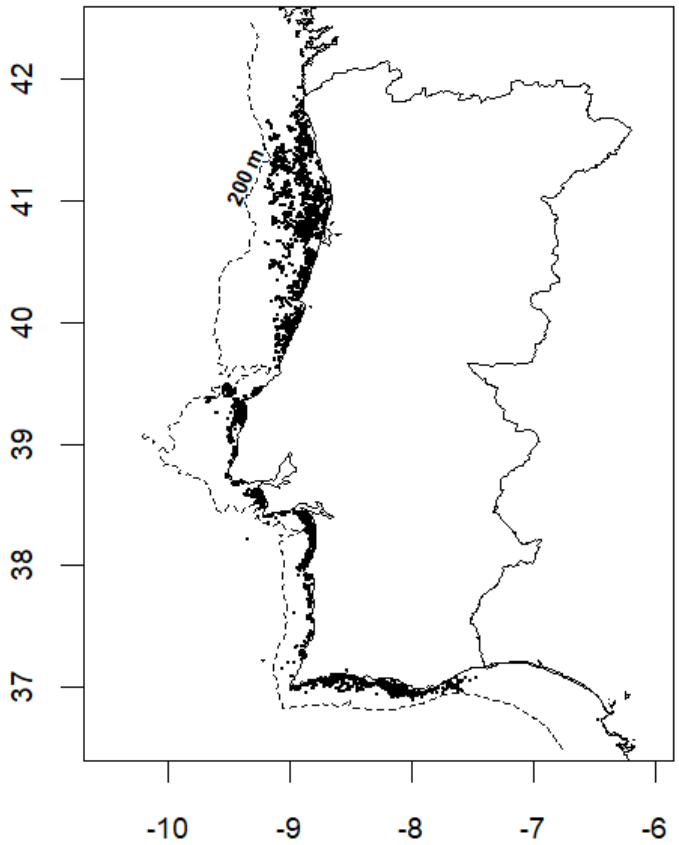
Fishery-independent data



Commercial fisheries

↓

Fishery-dependent data



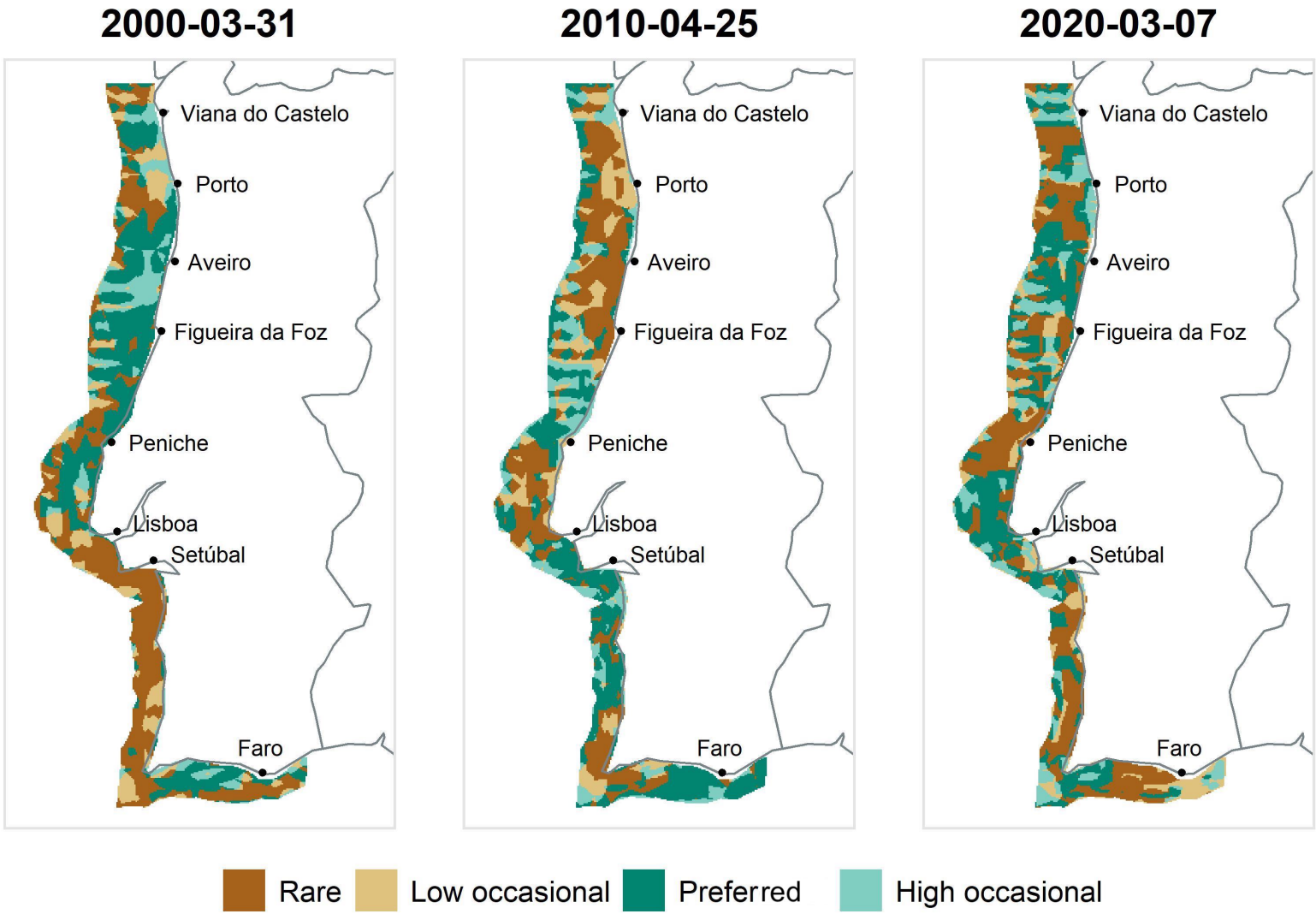
What are the (dis)advantages of each data source?

- Scientific Surveys:
  1. **Moderate-dimensional datasets**
  2. **Short time** (1 or 2 times a year) and **long spatial coverage**
  3. **Many zeros** (zero-inflated issues)
  4. **Smaller variability** (standardized data collection)

- Commercial Fisheries:
  1. **High-dimensional datasets**
  2. **Long time** (almost all year) and **short spatial coverage**
  3. **Preferential sampling** (a poor representation of the study region)
  4. **Larger variability** (data need to be standardized)

Spatial Statistics Supporting Decision Making

Areas of occupancy for sardine



Thank you for your time and interest!

Feel free to ask any questions while we explore dynamic maps together.

Instructions:

1. Use your phone to scan the QR code.
2. Open the HTML file found on Raquel's GitHub.
3. Explore the interactive boxplots and other graphics.
4. **Navigate and zoom in on the world map to find the coast of Portugal**... and data from the scientific campaigns.