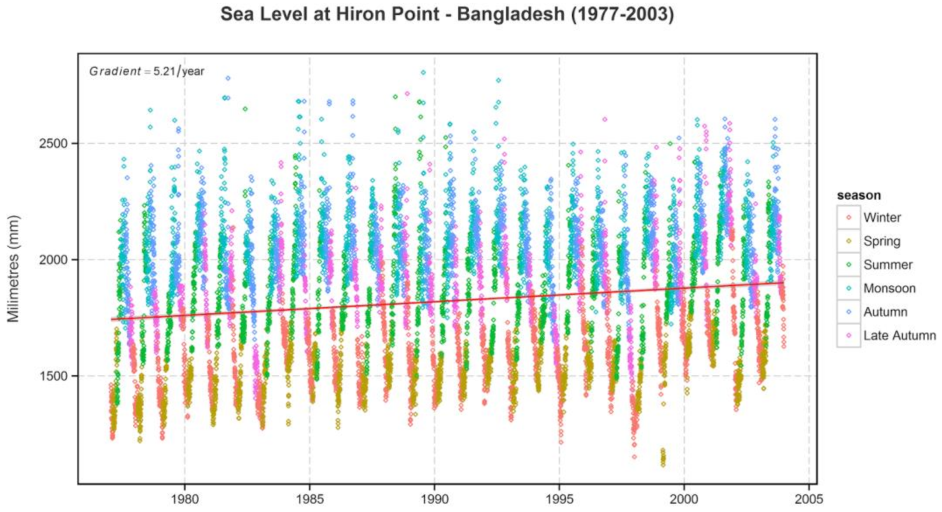
Data analysis
Spatio-temporal data and hierarchical models

Introduction (Ch1)

Djemel Ziou

Motivations

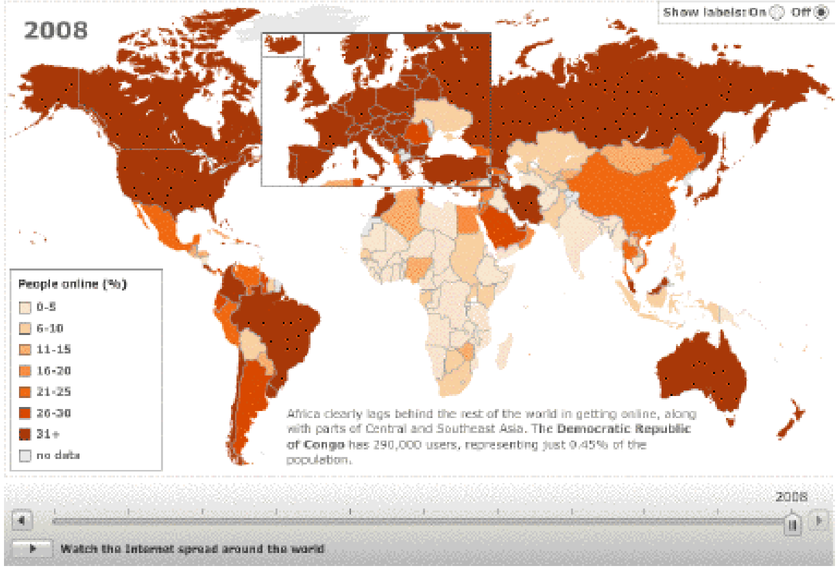
- Water level of the river



Sea Level at Hiron Point - Bangladesh (1977-2003)

Source: University of Hawaii Sea Level Centre / Bangladesh Inland Water Transport Authority (BIWTA) - 2014

- Evolution of the Internet population in the world



Africa clearly lags behind the rest of the world in getting online, along with parts of Central and Southeast Asia. The **Democratic Republic of Congo** has 290,000 users, representing just 0.45% of the population.

Measurement

- Many phenomena (physics, chemistry, biology, social, economic ...) change over time and/or space.
- The study of phenomena can lead to perform measurements (data acquisition) of their properties, to a certain scale.
- Some phenomena are measurable and others not (e.g. Minimum uncertainty principal).
- Data acquisition
    - sensors, survey ...
    - experimental data: Data acquisition should be performed according to the principle block-randomization-replication.
    - observational data: Direct measurement of the phenomena without control. The uncertainty is higher.

Measurement

- Acquired data
  - subject to errors caused by instruments, handling, sampling, archiving ...
  - used to produce knowledge (abstraction, prediction, inference ...).
- Many data are produced every day. The data are important for the economy, health, culture, security ...
- Data increasingly complex.

Data analysis

- Objectives
  - knowledge production (eg. abstract) from observed data, non-measurable phenomena (hidden variables) and knowledge (eg. law of physics ...).

Ex. You manufacture a metal plate for aircraft engines. You have some data about its thickness. The thickness of the plates measured in your manufactory is equals to the mean of the data. But, your client denies the plate because of the thickness. Why?

Thickness is sampled from a some function of several variables (material feature, temperature, original thickness)

Temperature is hidden phenomena

Physical law : Δthikness= material feature x original thickness x Δ temperature

  - It requires the understanding the problem and phenomena, data acquisition, archiving, data cleaning, specification models, estimation models, model validation, prediction, interpretation of results ...

Data analysis

- The approach for data analysis
  - The errors can be ignored, leading to deterministic approaches (e.g. partial differential equations, convolution, variational calculus...).
  - Taking errors into consideration leads to the approaches of uncertainty analysis (i.e. descriptive statistics, probability, random processes, stochastic differential equations, heuristics...).
    - when formal reasoning is used we talk about data analysis/statistics/probability.
    - when heuristics are used we talk about data mining.
    - when data analysis is embedded into the management, we talk about business intelligence.
    - what about the big data?

- The probabilistic approach will be used in this course to move from measurement to knowledge.

Model

- The data can be described by a model (equations, graph, tree ...)

- Model
  - must capture the properties to be studied (e.g. the spatiotemporal dynamics of the phenomenon of interest)
  - must have a level of complexity that can answer the question
  - it is another source of uncertainty
  - must be evaluated (theoretical and / or experimental)
  - it can be a probability density, a mixture of probability densities, time series, Kriging, regression, or a combination of these concepts.

Model

- Example (Social network)
  - let us consider a source of information and a ratio of people (e.g. number of people per unit area) located at a distance x which are influenced by this information at a specific time t.
  - the propagation of the information can be seen as a diffusion process.
  - the observed data (ratio of people) can be seen as sampled from a diffusion process.
  - let I(x, t) the ratio of people and Let f (I, x, t) the variation of this ratio (people add / leave at time t and the position x).

$$\frac{\partial I(x,t)}{\partial t} = a\frac{\partial^2 I(x,t)}{\partial x^2} + f(I,x,t)$$

- Example (Time series)
  - given the budgets of an institution since 1954. How will be the budget in 2016?

$$b_{16} = a_1 b_{15} + a_2 b_{14} + \cdots + a_5 b_{11} + f(2015) + \varepsilon_{16}$$

Model

- Models will be studied in this course
  - approaching spatiotemporal data with probability density functions (pdf).
  - considering the phenomena (process), it leads to hierarchical models/analysis (data, process 1, process 2 ..., parameters).
  - solve the underlying problems concerning the specification of the model (choice of hierarchy and pdfs) and the estimation or selection of different parameters.
  - use the probability for the data, process, and sometimes parameters.
- The use of probabilities can compensate for the lack of data, poor data quality... and can lead to a more accurate analysis. However, the resulting algorithms may have a higher computational complexity.

Model

- Change of support
  - the scale (global, regional, local), the unit, hidden factors can lead to absurd conclusions.
  - example of the stock exchange : the daily prediction from biannual measurements can lead to financial ruin.
  - example of management: imagine international laws, laws by countries, laws by province ...
  - example of treatment of kidney stones.

Surgery	Ultrasound
78% (273/350)	**83%** (289/350)

Treatment with ultrasound is better?

Change of support

Not sure

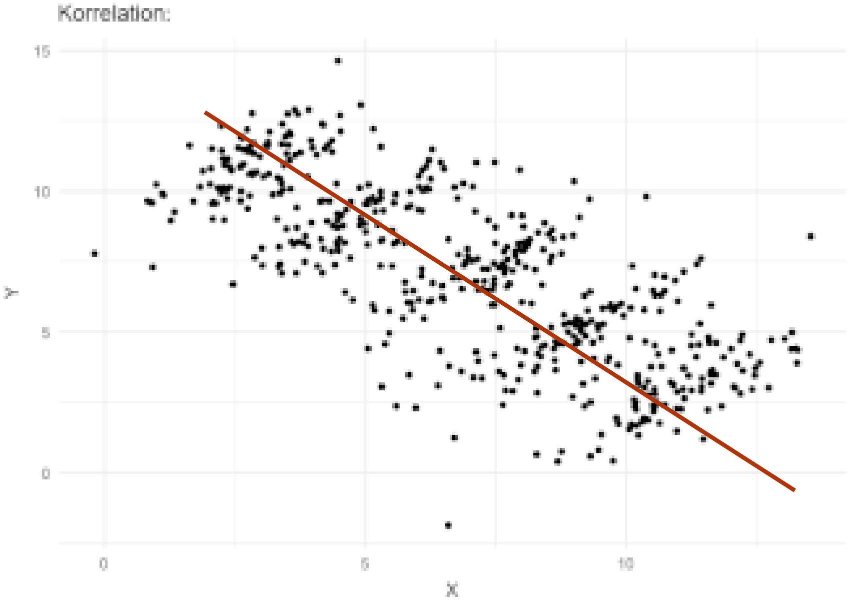
	Surgery	Ultrasound
Small stone (<2 cm)	**93%** (81/87)	87% (234/270)
Large stone (≥ 2cm)	**73%** (192/263)	69% (55/80)

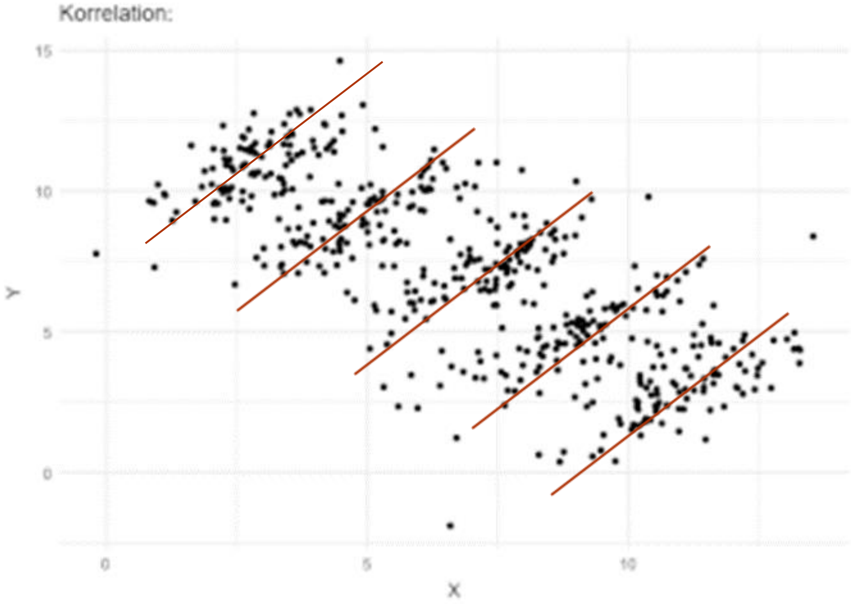
Treatment with surgery is better

- Simpson paradox: Contradiction between groups and all.
- the sizes of the kidney stones and groups (block) influence the scores.

Change of support

Another example (from Wikipedia)





Course

- According to some, this century will be devoted to the amount of data, large dimensions, the hierarchy of models and spatiotemporal phenomena.
- Content
  - models of marginal and conditional probability of spatial, temporal and spatiotemporal
  - hierarchy of these models
  - estimators of these models.
- Illustrations with examples of social networks, geography, images ...
- Basic concepts will be presented in order to be able to follow the course.

Homework

Hw1: You have some scalar data and you have no idea about their distribution:

- How can you tell if the distribution is peaky or not?
- How can you tell if the distribution is symmetric or not?
- How can you tell if the distribution is unimodal or multimodal?
- How can you estimated the mean and the variance of these data? You need to provide all the assumptions you used.

Hw2: The goal is to study the experimentation design in statistics and data analysis. For this purpose, you need to find and read documents that allow to answer the following questions:

- Summarize the experimentation design.
- Illustrate the method by studying the kindly stones case. You need to explain the blocks and the randomization.
- We would like to dress conclusions about the kindly stones in all province of china during the last 40 years. Identify the change of support issues that you will be faced with.

Homework

Important

- You need to provide a personal work. The cut and past from existing documents is not accepted.
- You can help people but make sure the people is capable of doing the homework independently.
- You can do the work by a group of three persons.
- Deadline: April, 1st, before starting the class.

Bibliography

- N. Cressie and C.K. Wikle. Statistics for Spatio-Temporal Data. Wiley, 2011.

- H. Wang et al. Modeling Information Diffusion in Online Social Networks with Partial Differential Equations. ICIAM 2015.