ANALYSIS OF SPATIAL DATA IN EPIDEMIOLOGY

Prof. Dr. Maria A Barceló and Prof. Dr. Marc Saez

September 8, 10, 14 and 16, 2021

Research Group on Statistics, Econometrics and Health (GRECS), University of Girona
CIBER of Epidemiology and Public Health (CIBERESP)

GRECS
GRUP DE RECERCA EN ESTADÍSTICA ECONOMETRIA I SALUT

UdG

ciberesp
Centro de Investigación Biomédica en Red
Epidemiología y Salud Pública

COURSE INTRODUCTION

1. Course introduction
2. **Introduction to epidemiology and spatial statistics**
3. Overview of mixed models
4. Overview of mixed models - Practicals
5. Introduction to INLA and R INLA
6. R INLA - Practicals

Wednesday 8

Friday 10

Marc Saez Maria Antònia Barceló

COURSE INTRODUCTION

7. Disease mapping. Standardisation of incidence and mortality rates
8. Disease mapping. Smoothing standardised incidence and mortality rates
9. Disease mapping – Practicals
10. Geographical association studies. Spatial ecological regression
11. Spatial ecological regression - Practicals

Tuesday 14

Marc Saez Maria Antònia Barceló

COURSE INTRODUCTION

12. Clustering
13. Extensions: BYM2, point processes, leaflet, pc priors
14. Extensions – Practicals

Thursday 16

Marc Saez Maria Antònia Barceló

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Epidemiology

- **Epidemiology** is dedicated to the study of the distribution, frequency, causes and control of factors related to health and disease in well-defined human populations and to the application of this study to protect and improve the health of the population.
- **Epidemiology** especially studies the relationship between exposure and disease.
- **Epidemiology** is considered the basic science for preventive medicine and a source of information to formulate public health policies.

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

- Diseases do not occur randomily. They have causes, many of them human in origin, which can be avoided.
- Epidemiologists have been pivoltal in identifying numerous etiological factors which, at the time, justified the formulation of health policies geared towards the prevention of diseases.

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

- Epidemiology emerged from the study of infectious disease epidemics.
- Nowadays, epidemiology is concerned with the demographic study of any disease with the help of statistics.

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Spatial epidemiology

"Spatial epidemiology is the description and analysis of geographic variations in disease with respect to demographic, environmental, behavioural, socioeconomic, genetic, and infectious risk factors".

Elliot P, Wartenberg D. Spatial epidemiology: current approaches and futures challenges**.** *Environ. Health Perspect.* 2004; 112(9):998-1006. doi: 10.1289/ehp.6735.

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

- **Spatial epidemiology** is related to describing diseases and the study of their causes and prevention using different perspectives of analysis where location of events is a basic component, given that it also studies the geographical variations of the diseases.
- **Spatial epidemiology** is a sub-field of epidemiology focused on the study of the spatial distribution of health outcomes.
- **Spatial epidemiology** is based on a concept of health where individuals are seen in their socio-cultural context.

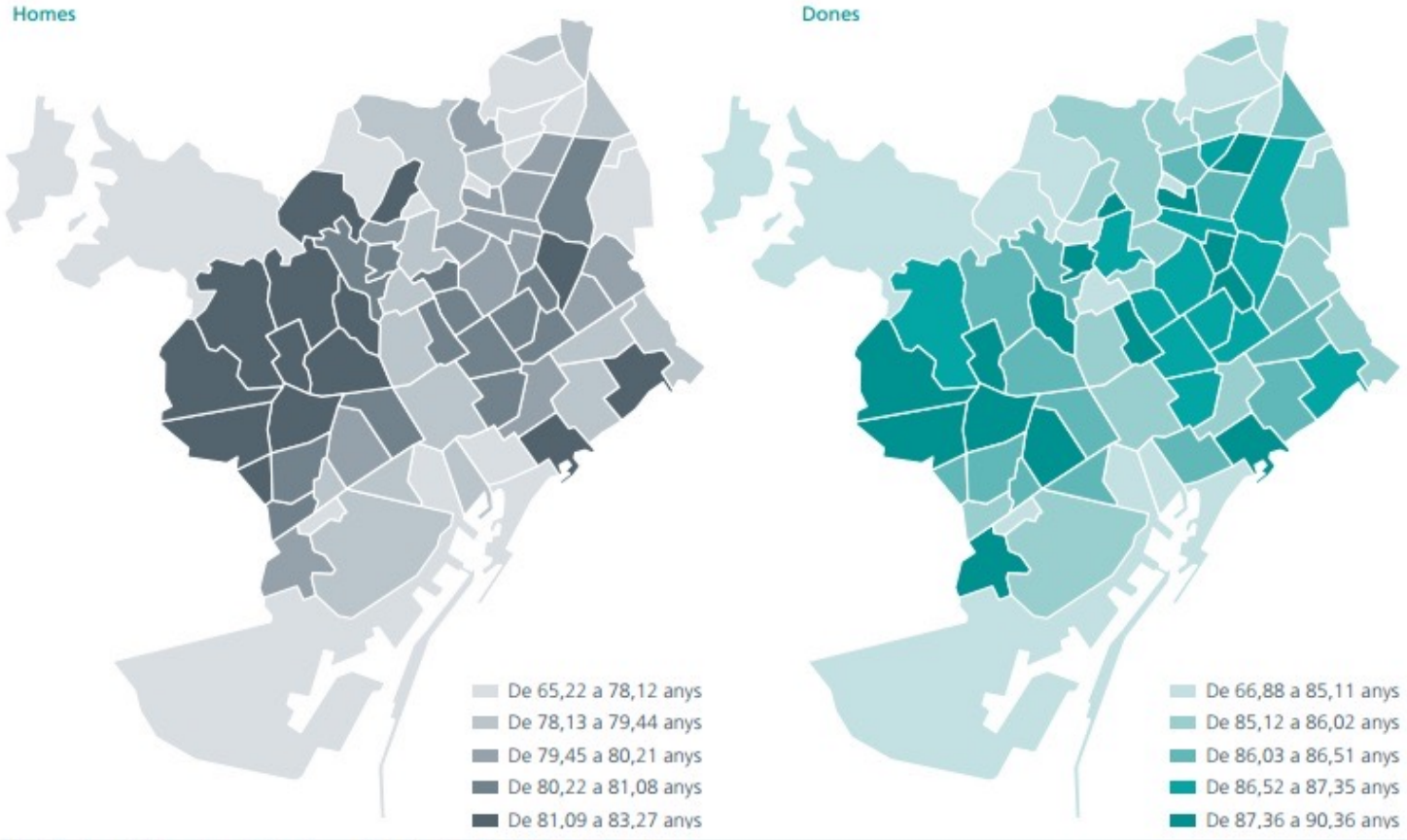
EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Main objective of spatial epidemiology:

- To show what part of the spatial variation of the distribution of the occurrence of a health event is not explained by either the spatial distribution of known explanatory factors, or by a random variation.
- In fact, very often we are interested in finding clues about an unknown risk factor of a certain disease.

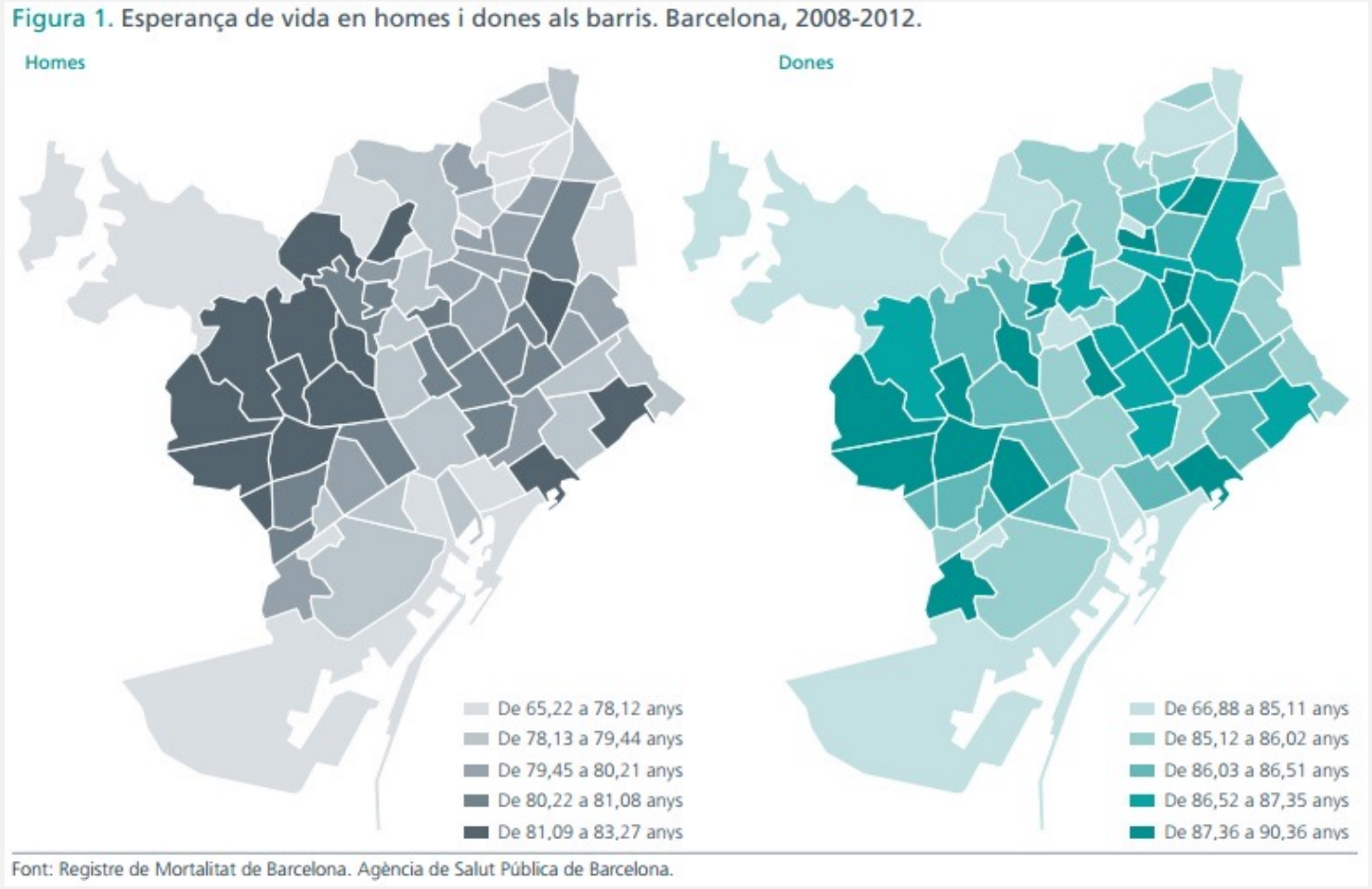
EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

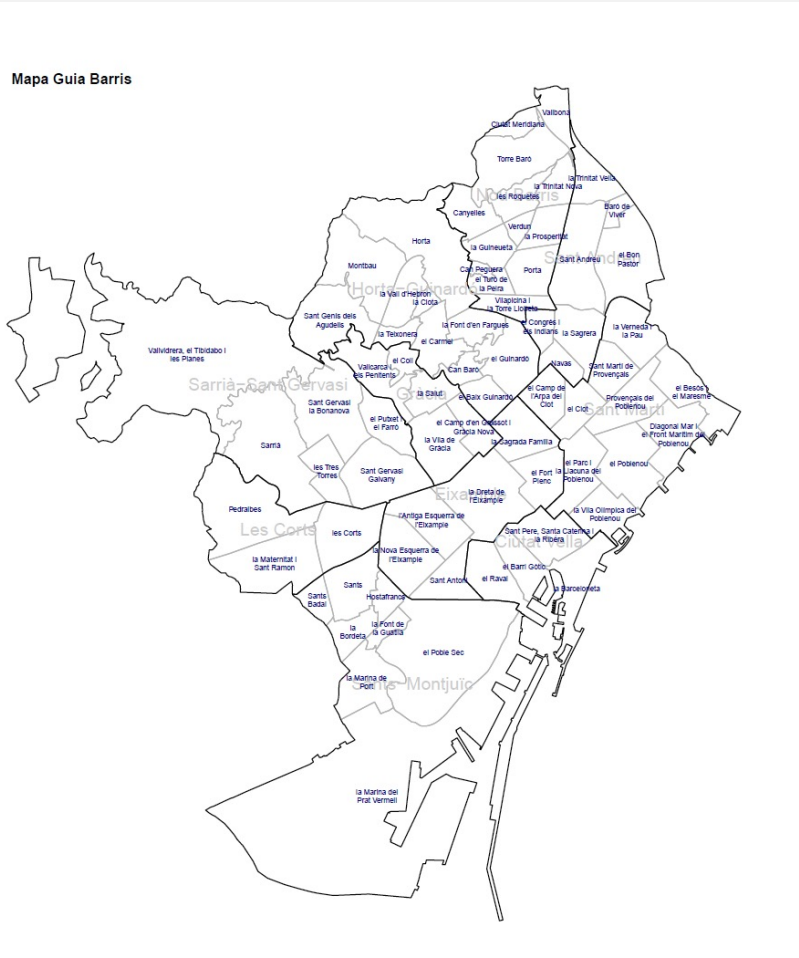
Figura 1. Esperança de vida en homes i dones als barris. Barcelona, 2008-2012.



Font: Registre de Mortalitat de Barcelona. Agència de Salut Pública de Barcelona.

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY





2. Introduction to epidemiology and spatial statistics

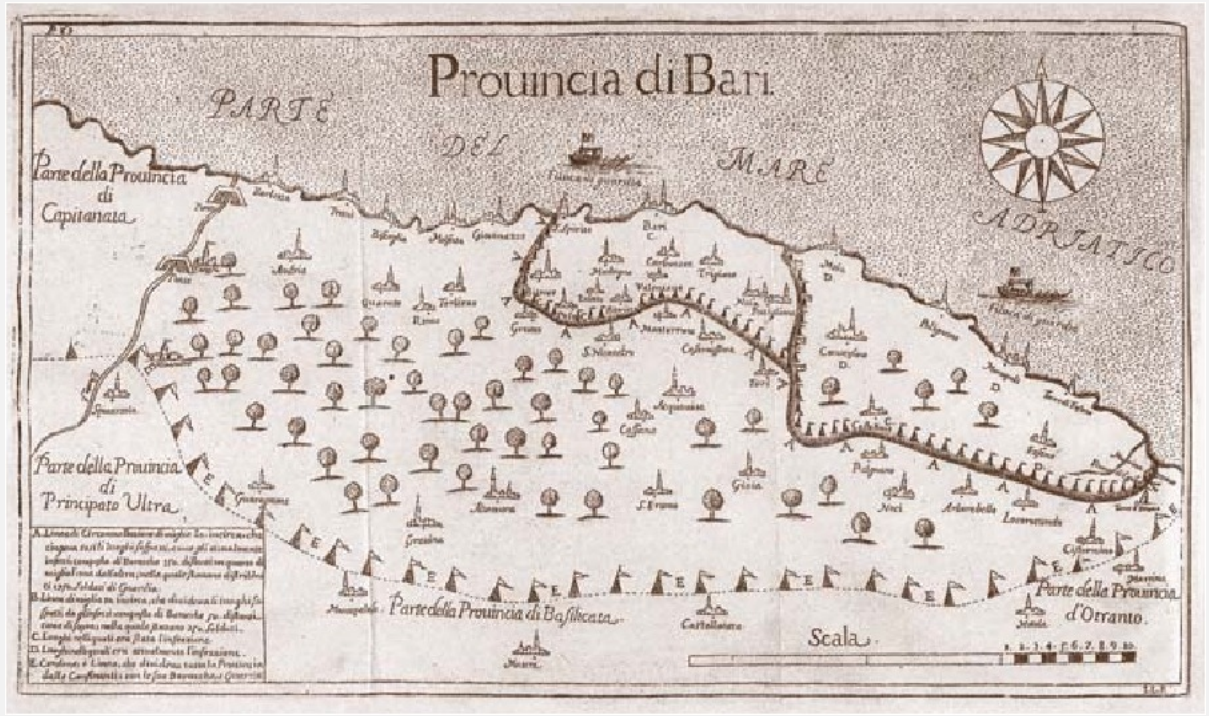
EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

- The study of the geographical distribution of health events has become enormously interesting for epidemiologists in recent decades despite, as we will see below, having a history of more than 200 years.
- The **first disease maps** were created to represent the location of cases of infectious disease:
  - Yellow fever in New York (Seaman, 1798)
  - Cholera in London (Snow, 1854)

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Map of Fillipo Arrieta

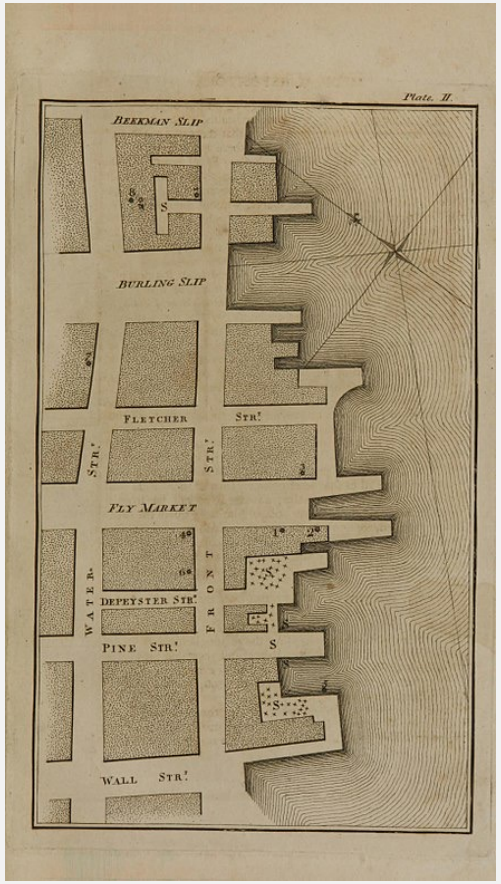
Visualisation of the strategy to contain the propagation of the black death in the region of Bari, Italy 1690-92.



EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Seaman's map of cases of yelllow fever

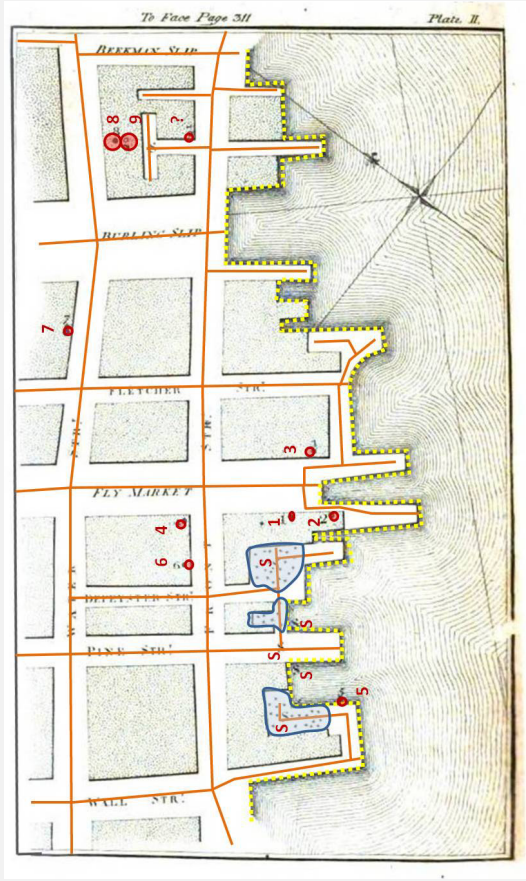
The points represent the deaths due to yellow fever and the "S" the dumping sites.



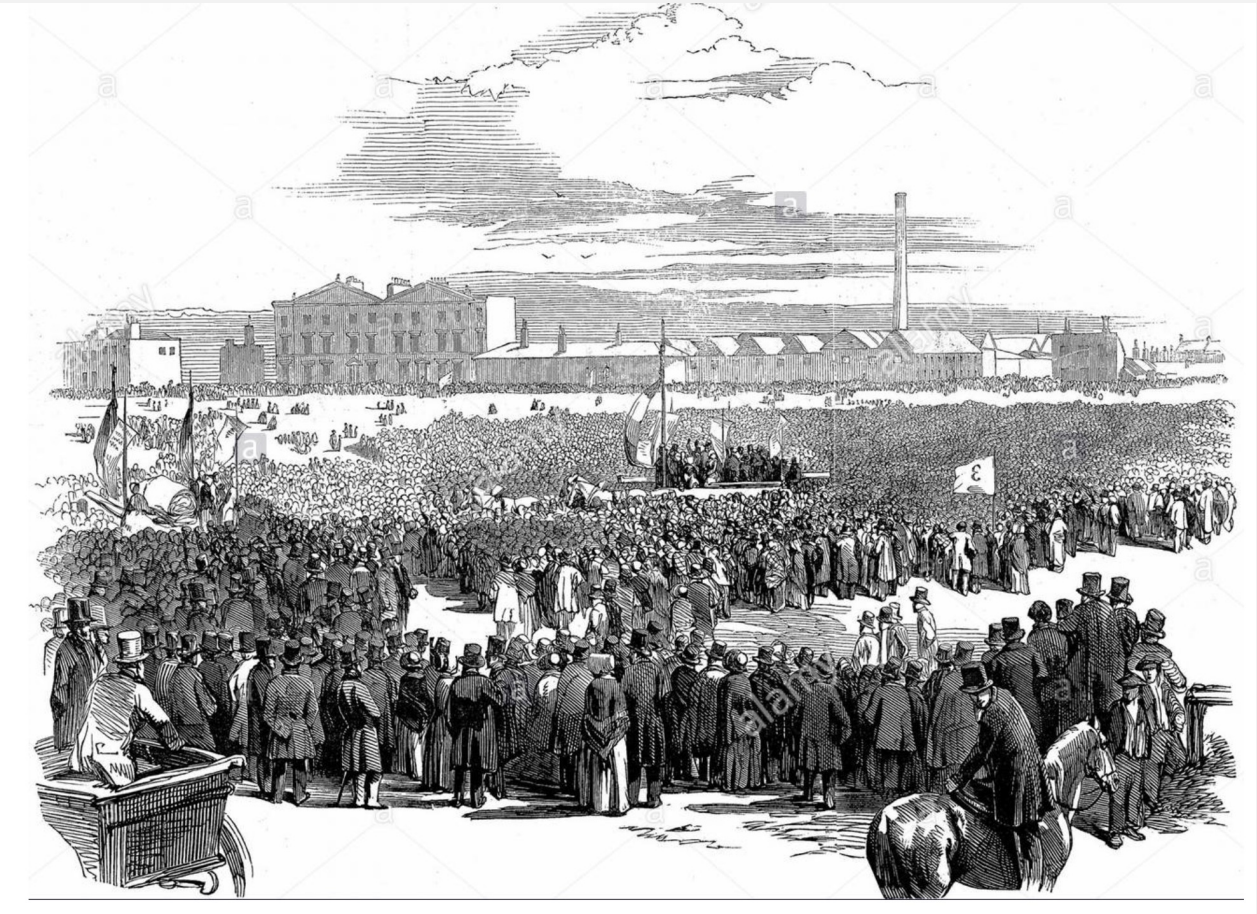
EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Seaman's map of cases of yellow fever

The points represent the deaths due to yellow fever and the crosses and the "S", the dumping sites.



EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY



2. Introduction to epidemiology and spatial statistics

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

- In the autumn of 1848, there was a second cholera epidemic in London, causing a huge number of deaths.
- Neither its aetiology nor the way the disease was transmitted was known for sure.
- Two current theories:
  - Contagion due to contact
  - Miasmas

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Snow observed that:

- **Miasmas could not be causing the disease:** patients must have presented respiratory symptoms due to inhaling the “miasmas” and not the acute diarrhoea symptoms present in cholera.
- **The deaths due to cholera** that occurred between 1848-49 (second epidemic) were **concentrated** in the districts of **South London**. The death rates observed in this area were way higher than in the rest of the city (8.0 and 2.4 deaths per 1,000 inhabitants, respectively).

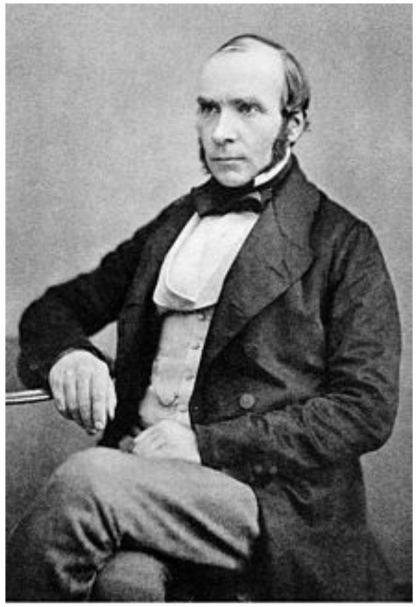
EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

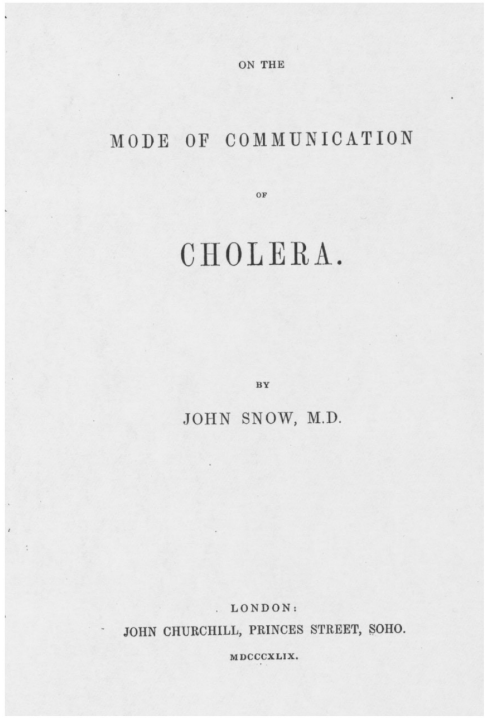
Furthermore, Snow observed that:

- The inhabitants of South London obtained their drinking water from waters below the Thames (highly contaminated waters).
- The other areas of London obtained their drinking water from less contaminated sections of the river (water from the Thames or its tributaries).

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

John Snow
1813-1858





ON THE

MODE OF COMMUNICATION

OF

CHOLERA.

BY

JOHN SNOW, M.D.

LONDON:
JOHN CHURCHILL, PRINCES STREET, SOHO.
MDCCCXLIX.

Deaths from Cholera in London, registered from September 23d, 1848, to August 25th, 1849.

Districts of London.	Population in 1841.	Deaths from Cholera.	Deaths to each 1,000 inhabitants.
West . .	300,711	533	1·77
North . .	375,971	415	1·10
Central . .	373,605	920	2·48
East . . .	392,444	1,597	4·06
South . .	502,548	4,001	7·95
Total . .	1,948,369	7,466	3·83

On the Mode of Communication of Cholera. John Snow. 1849

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

- In 1853-1854, there was the third wave of the cholera epidemic in London.
- The inhabitants of some districts in the south of the city took their water from the small tributaries of the river Thames or from the numerous public use water pumps.
- Human faeces were thrown into improvised sewers or directly into the river.

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

- There were two companies responsible for the water supply: Sothwark and Vauxhall Company and Lambeth Water Company.
- During the second cholera epidemic in 1848-49, both companies extracted their water from the contaminated sections of the Thames, with the districts supplied by the two companies recording similar number of deaths.

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

- In 1853:
  - Lambeth Water Company had moved their infrastructure upriver (non-contaminated water).
  - Southwark and Vauxhall Water Company continued to extract the lower waters.
- Snow observed that the mortality rate due to cholera in homes supplied by the second company were 8.5 times higher than that of the homes supplied by the first.

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

- At the beginning of September 1854:
  - In the “Golden Square” area of London (in Soho), there was an unusually intense epidemic outbreak of cholera (500 deaths in just 10 days).
  - Most of the section’s residents extracted their water from a public use pump located in Broad Street.

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

- **A map of the sector was created**, marking the points corresponding to the deaths due to cholera and the different existing drinking water pumps.
- It was surmised that the outbreak was due to the ingestion of contaminated waters from the pump on Broad Street.

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Snow identified the spatial aggregation of the cases of cholera in London in 1857.

Spot map of cholera cases (Snow, 1854)

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

- Snow took samples from the pump on Broad Street and 4 other nearby pumps (there were differences regarding the clearness of the water).
- The distance between the residence of each victim and the nearest water pump was calculated.
- It was observed that in 73 of the 83 cases, the pump in Broad Street was the nearest. This fact was subsequently communicated to the health authorities, who closed the pump.

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Evolution of the maps in spatial epidemiology

- Spot maps
- Choropleth maps
- Atlas of diseases, at the national and the international levels

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Evolution of the maps in spatial epidemiology

- Spot maps
  - Yellow fever in New York (Seaman, 1798)
  - Cholera in London (Snow, 1854)

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Spot map of cholera cases (Snow, 1854)

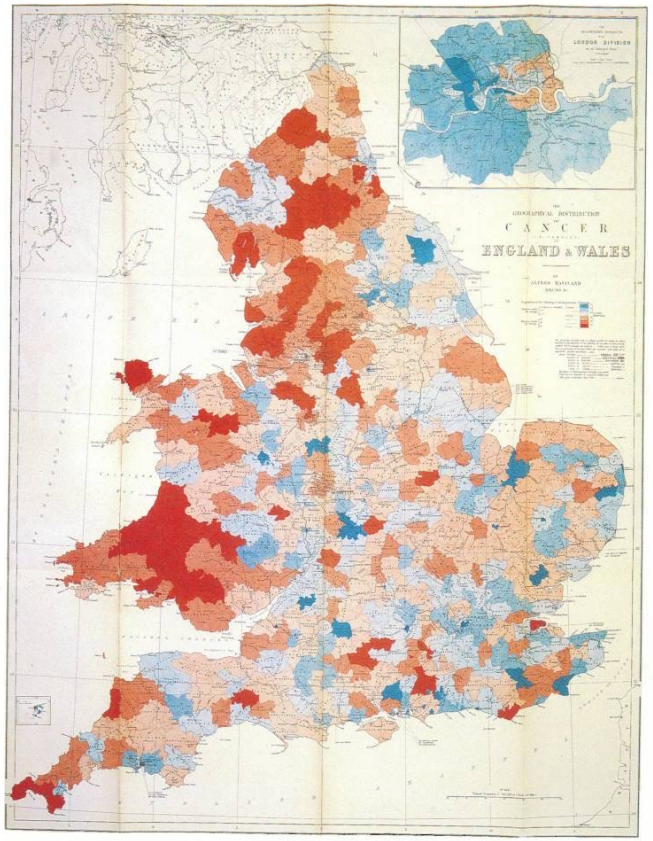
EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Evolution of the maps in spatial epidemiology

- Choropleta maps
  - Geographical distribution of deaths due to heart disease, cancer and tuberculosis in England and Wales (Haviland, 1878)
  - Mortality by county in England and Wales, adjusted by age and sex (Stocks, 1936,1937,1939)

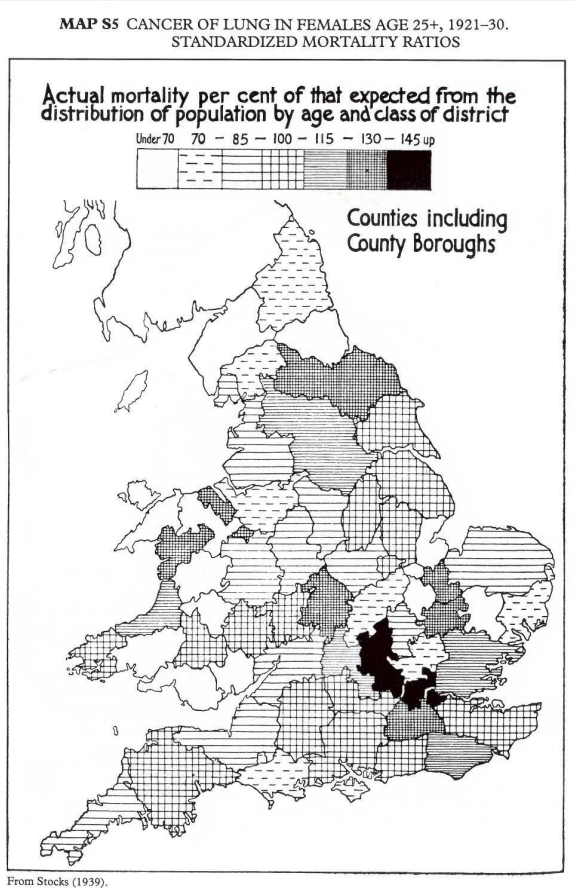
EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Cancer among females, 1851-1860 (Haviland, 1878)



The geographical distribution of cancer in females in England and Wales, 1851–60. From Haviland (1878).

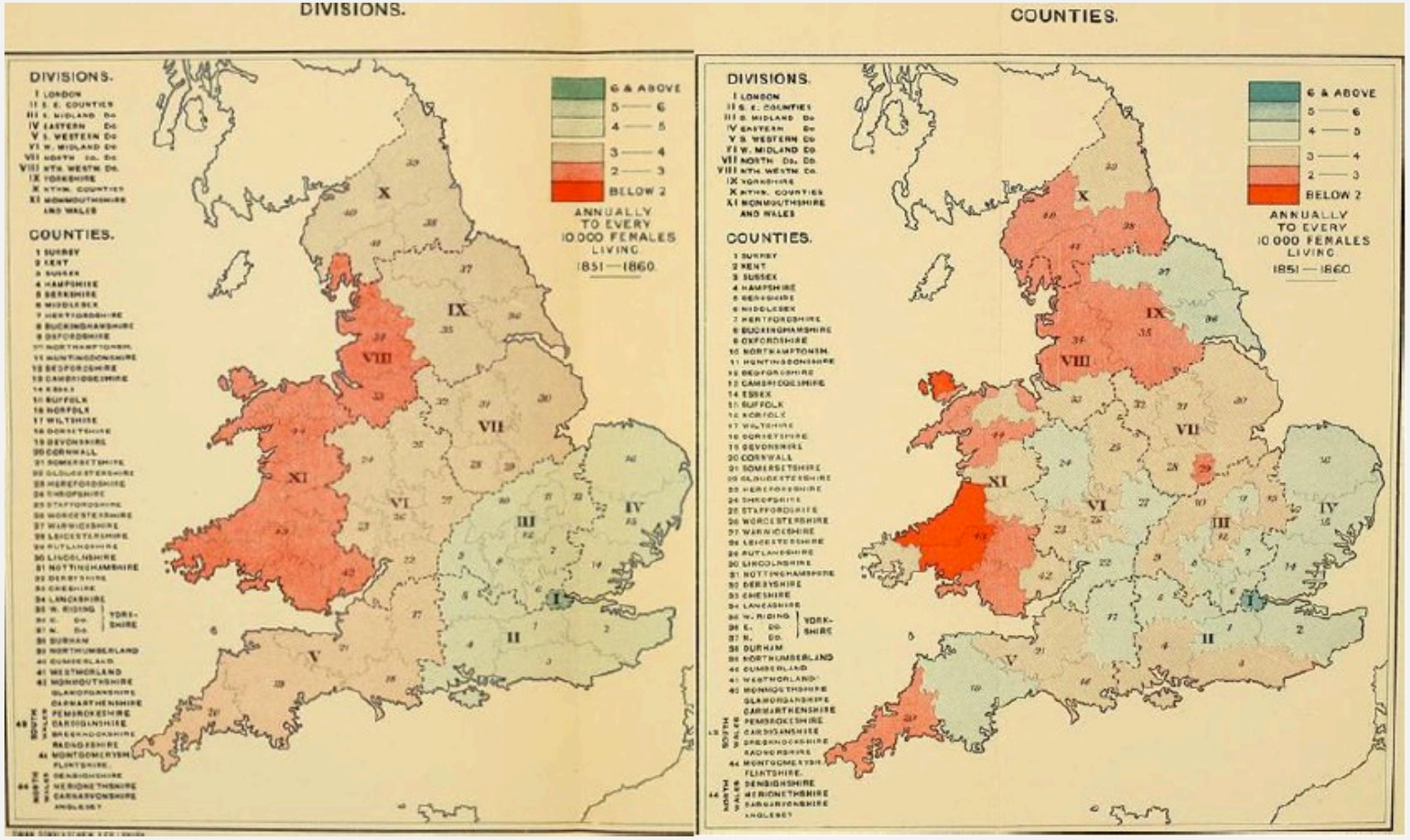
Lung cancer in females, 1921-1930 (Stocks, 1939)



From Stocks (1939).

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Cancer in females, 1851-1860
(Haviland, 1878)



EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

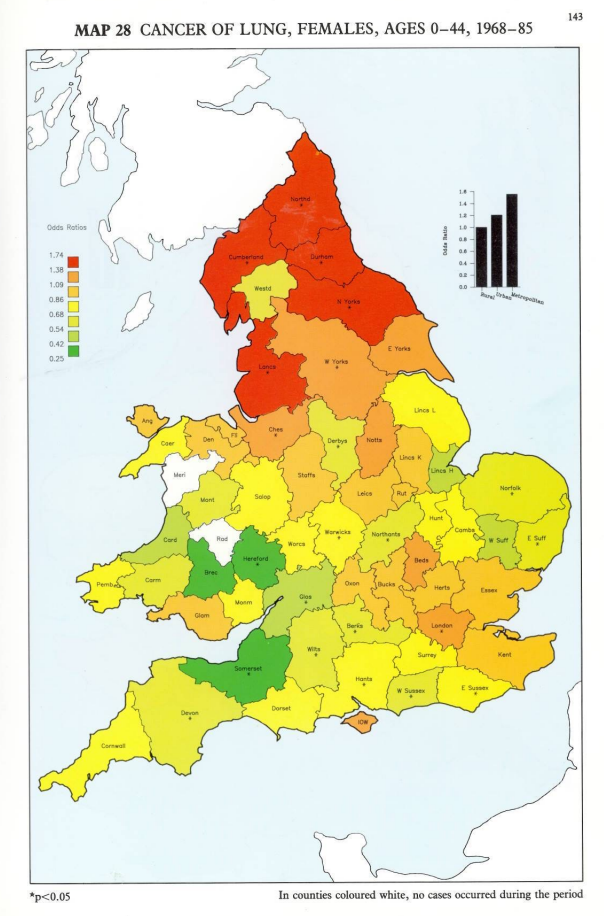
Evolution of the maps in spatial epidemiology

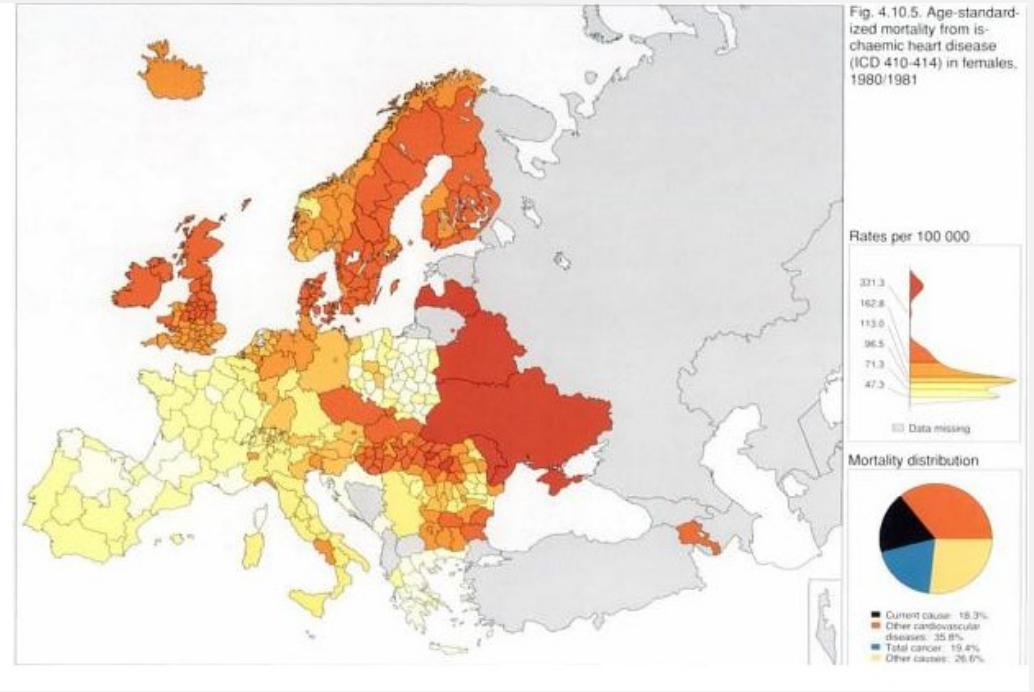
- Atlas of diseases, at the national and international levels
  - Atlas of the incidence of cancer in England and Wales 1968-85 (Swerdlow and dos Santos Silva, 1993)
  - Atlas of mortality in Europe 1980-81 and 1990-91 (OMS, 1997)

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Incidence of lung cancer in females 1968-1985

(Swerdlow and dos Santos Silva, 1993)





Deaths due to myocardial infarction standardised by age, 1980-81

(OMS)

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

- As we can see, spatial epidemiology has stemmed from the representation of the spatial distribution of health events with the dual objective of characterising their extension and establishing hypotheses concerning the possible causes.
- Spatial epidemiology has since grown enormously in terms of complexity (methods of analysis, small units) and use.
- The confluence of epidemiology, statistics and informatics, together with huge technological advances, have meant that spatial epidemiology has developed in leaps and bounds.

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Developments since 1996

- Geographical Information Systems (SIG)
- Software to represent maps and to analyse spatial data
- Greater availability of georeferenced data (GPS, etc.)

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Developments since 1996

- Development of specialised statistical methods
  - Sophisticated techniques to separate noise signal
  - Methods to control spatial and time dependency
  - Methods to detect clusters

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Problems in spatial epidemiology:

- Different scales (for example, autonomous regions, provinces, municipalities, districts, neighbourhoods, post codes, census tracts, etc.)
- Changes in the boundaries of some of these units
- There can be georeferencing errors (due to wrong or inexistent addresses, etc.)
- Misalignment

SPATIAL STATISTICS

- **Spatial statistics** is concerned with the exploration, description, visualisation and analysis of data, considering their distribution characteristics in the space, which are usually expressed using geographical coordinates.
- **Spatial statistics** is the branch of statistics that analyses georeferenced data or, in other words, data for which their spatial coordinates are available (spatial data).

SPATIAL STATISTICS

- **Spatial data** is understood to be the measures and observations made in specific locations or areas. In addition to the value of the measure, they incorporate the location/position of the observed values.

SPATIAL STATISTICS

Characteristics of spatial data:

1. **Spatial heterogeneity**
   - Observations are not homogenous in the space

2. **Spatial dependency**
   - The observations in a location depend on other observations in other locations (generally close by).

SPATIAL STATISTICS

Characteristics of spatial data:

1. **Spatial heterogeneity**
   - Observations are not homogenous in the space

SPATIAL STATISTICS

There are two types of **spatial heterogeneity**:

- ***Heteroscedasticity***:
  - *Structural cause:* caused by using data coming from arbitrary spatial units
  - *Sample cause:* existence of outliers, omission of relevant variables, errors of measurement and other errors of specification.

- ***Structural change-instability:*** individuals are not homogeneous in the space (for example, north-south; centre-periphery, etc.).

SPATIAL STATISTICS

Characteristics of spatial data:

2. **Spatial dependency**
   - The observations in a location depend on other observations in other locations (generally close by).

SPATIAL STATISTICS

There are two types of **spatial dependency**:

- ***Spatial interaction (substantive spatial dependency):*** this is a spillover effect of an individual on another individual.
- ***Autocorrelation in the residual values (residual spatial dependency):*** caused by spatially related errors of measurement.

SPATIAL STATISTICS

Types of spatial data:

Spatial data has traditionally been categorised in three large groups (Cressie, 1993):

1. **Lattice data** or **areal data**
2. **Point processes**
3. **Geostatistical data**

SPATIAL STATISTICS

1. **Lattice data** or **areal data**

- **Lattice data** or **areal data** corresponds to discrete random variables, or count data.
- In areal data, ***the exact location of the case is unknown***.
- ***The locations are areas*** with well-defined geographical limits, usually administrative units (cities, neighbourhoods, census tracts, etc.).

SPATIAL STATISTICS

1. **Lattice data** or **areal data**

- The ***response variable is the aggregated number of cases in this area*** in a determined period.
- This was the type of data most commonly used when spatial data was in its infancy.

SPATIAL STATISTICS

1. **Lattice data** or **areal data**

- Observations coming from a random process about a set of spatial regions distributed in a regular or an irregular manner.
- They are defined mathematically as a set of location indices with an associated set of neighbors.
- ***Neighbors:*** neighbouring areas (proximity, contiguity, etc.) of a specific area.

SPATIAL STATISTICS

1. **Lattice data** or **areal data**

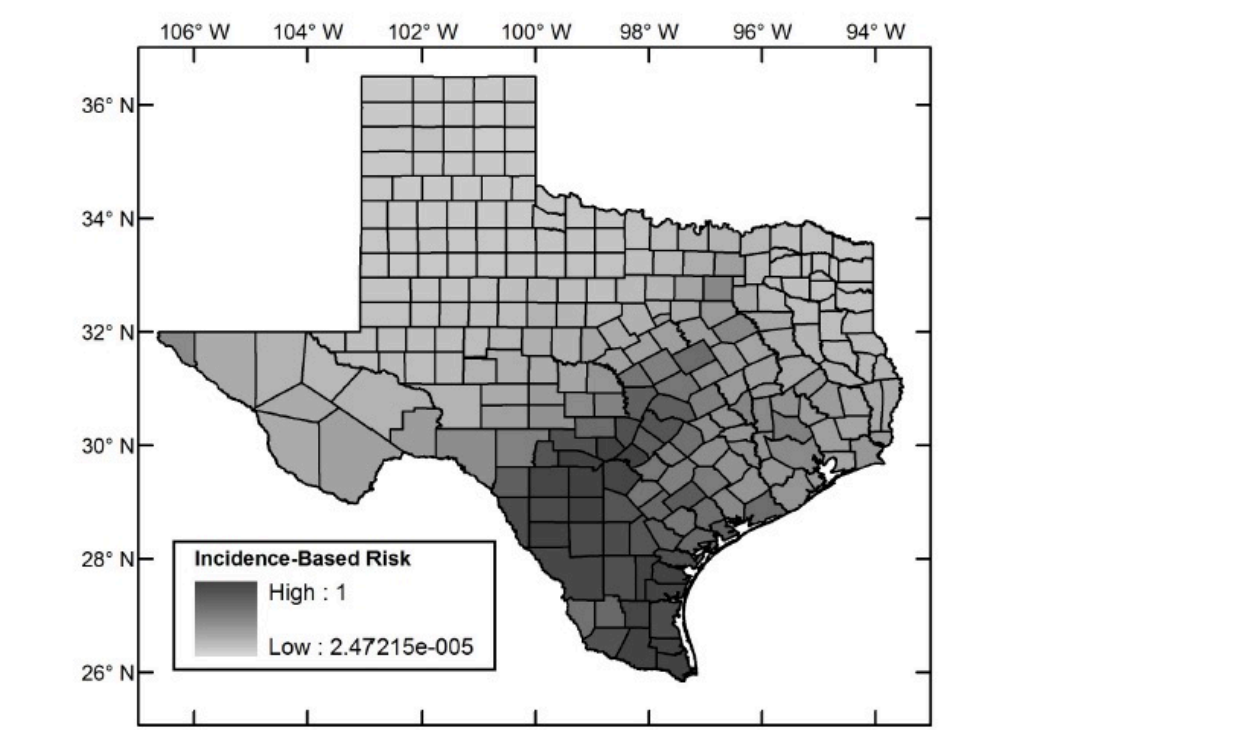


Figure 5. Incidence-based risk map for Chagas Disease in Texas.

SPATIAL STATISTICS

2. Point processes

- **Point processes** correspond to Bernouilli random variables.
- In point processes, ***the exact location of the case is known and is random***.
- The ***locations*** of the event of interest ***are observed in a determined region***. For example, the coordinates of the addresses of the cases of ALS in Catalonia.

SPATIAL STATISTICS

2. Point processes

- ***Information*** about these data ***is not public***. They are collected in studies of cases and controls, or in cohort studies.
- They can be aggregated by spatial units, creating areal data.

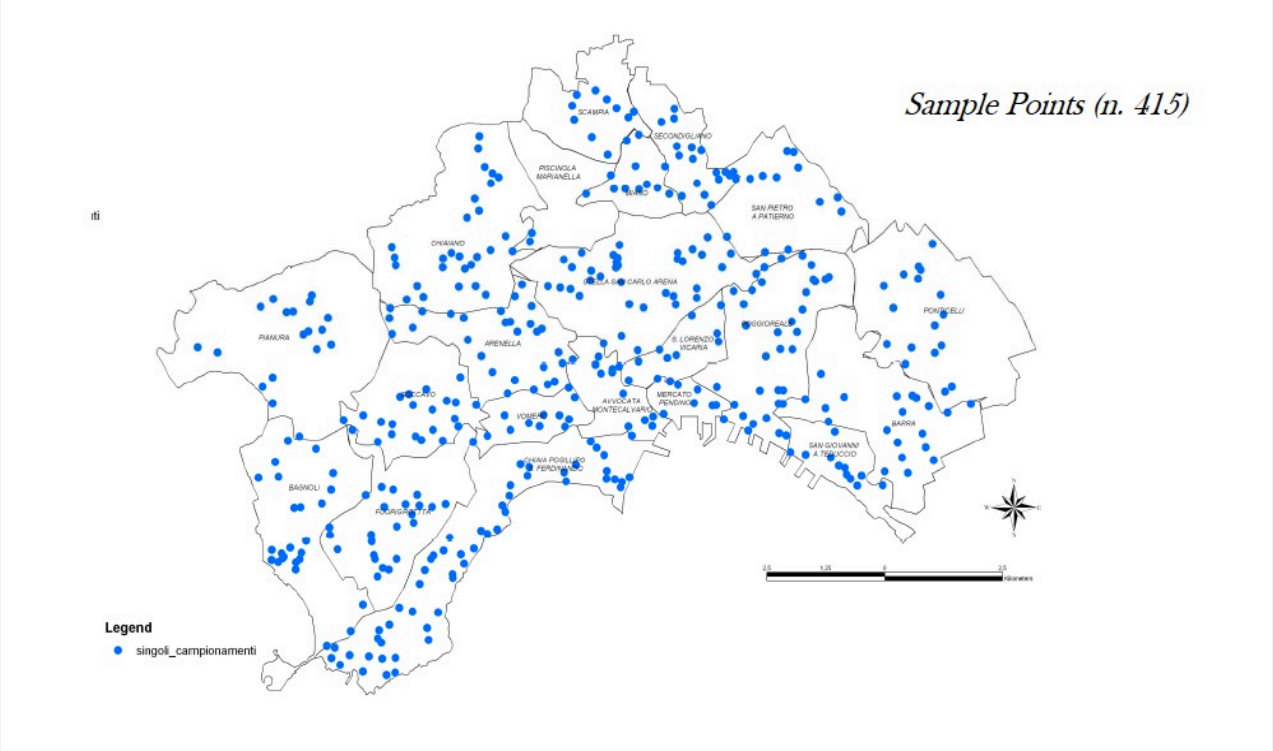
SPATIAL STATISTICS

2. Point processes

- In this case, it may be of interest:
  - To evaluate whether the events follow a determined spatial pattern (aggregation, regular shape, etc.)
  - To study whether an observed pattern is associated with a certain variable (exposure to an environmental variable such as atmospheric contamination; proximity to contaminating focal points; socioeconomic context, etc.)

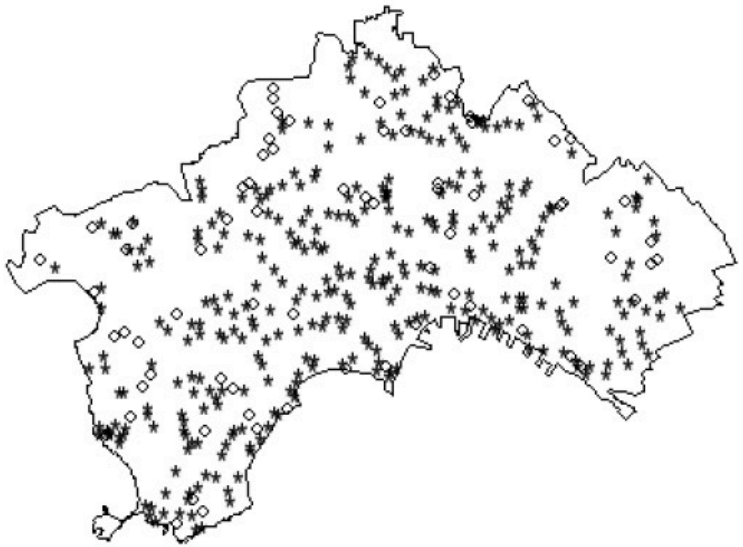
SPATIAL STATISTICS

2. Point processes



SPATIAL STATISTICS

2. Point processes



Distribution of cases (positives: circle) and controls (negatives: stars). Naples, February - May 2005

SPATIAL STATISTICS

3. Geostatistical data

- **Geostatistical data** correspond to continuous random variables.
- In geostatistical data ***the exact location of the cases is known and is fixed*** (atmospheric contamination monitoring centres, seams of a certain metal, etc.)
- ***The response variable is measured in each location*** (for example, measures of atmospheric contamination, measures of the chemical composition of the ground, temperature, etc.).

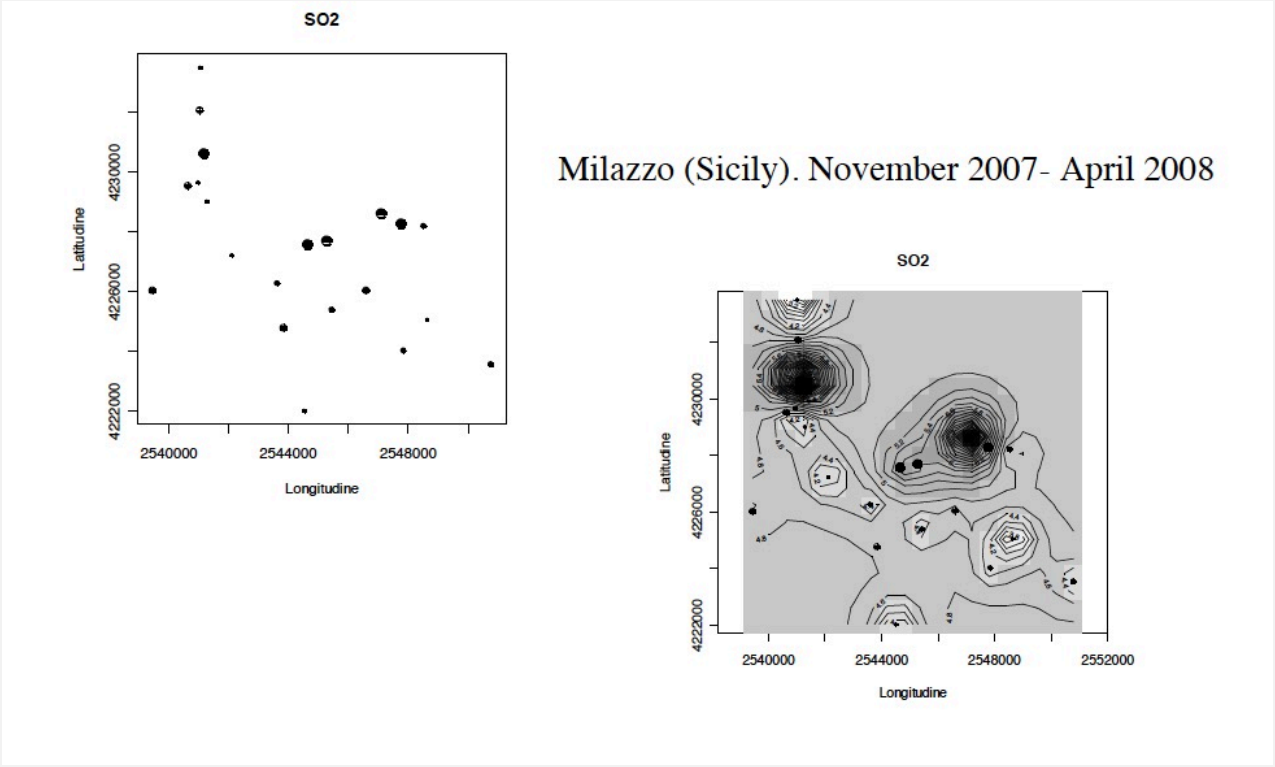
SPATIAL STATISTICS

3. Geostatistical data

- This means we have measures taken at fixed points, defined anywhere in the space, such that their ***localities define a spatially continuous surface area***.
- ***The spatial distribution*** of the values of an attribute are usually extended ***across the study region*** using mathematical models (for example, kriging).

SPATIAL STATISTICS

3. Geostatistical data



SPATIAL STATISTICS

But, what actually are the types of spatial design? Processes? Models? Methods?

- In 2012, Diggle proposed a change of paradigm and re-defined spatial statistics as ***'a set of statistical models and methods that aim to help scientists to understand spatial phenomena that cannot be observed directly, to observe them indirectly with incomplete information'***, in the form of lattice data, point processes and geostatistical data. The statistical model he proposed as the single base model is the **log-Cox model**.

SPATIAL STATISTICS

- That is to say, Diggle (2012) unified spatial statistics much in the same way as McCullagh and Nelder (1989) unified generalised linear models.


INTELLIGENT DESIGN APPLIED...
chase
"The theory is fine! We just didn't pray hard enough!"