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

GRECS
GRUP DE RECERCA EN ESTADÍSTICA ECONOMETRIA I SALUT

UdG

ciberesp
Centro de Investigación Biomédica en Red
Epidemiología y Salud Pública

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

GRECS
GRUP DE RECERCA EN ESTADÍSTICA ECONOMETRIA I SALUT

UdG

ciberesp
Centro de Investigación Biomédica en Red
Epidemiología y Salud Pública

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

There is currently a serious need for **spatial methods**, because:

- most of epidemiological studies have a spatial component
- we are interested in considering **the spatial component**
- very often, **the study area is small** and/or there is a large amount of **information at the individual level with geographic location information.**

EPIDEMIOLOGY AND SPATIAL EPIDEMIOLOGY

Why are we interested in considering the spatial component?

- Because we are explicitly interested in the spatial pattern of the risk factor: **disease maps**
- Because it contains a large part of the non-observed confounding: **spatial regression**
- Because we observe agglomerations in the space: **cluster detection**
- Because we are interested in the effects of a pollutant source on the health of the residents in the surrounding area: **source identification**

DISEASE MAPPING

Disease maps provide an initial overview of the spatial distribution of the disease, health event or its risk factors.

They are a visual summary of the geographical risk.

DISEASE MAPPING

Disease maps are used for :

- ***descriptive purposes***: to summarise the spatial and the spatial-temporal variation of the disease risk.
- ***to generate aetiological hypotheses***: exposure maps allow for informal examination (formal screening by spatial regression)
- **surveillance**: to highlight apparently high-risk areas
- to assist policy formulation and resource allocation.

DISEASE MAPPING

The representation of the spatial distribution of the disease on a map can be carried out at different scales:

- on a ***international scale***: comparisons between countries (WHO)
- on a ***national scale***: comparisons between autonomous communities, regions, ABS, etc.
- on a ***local scale***: studies in small areas

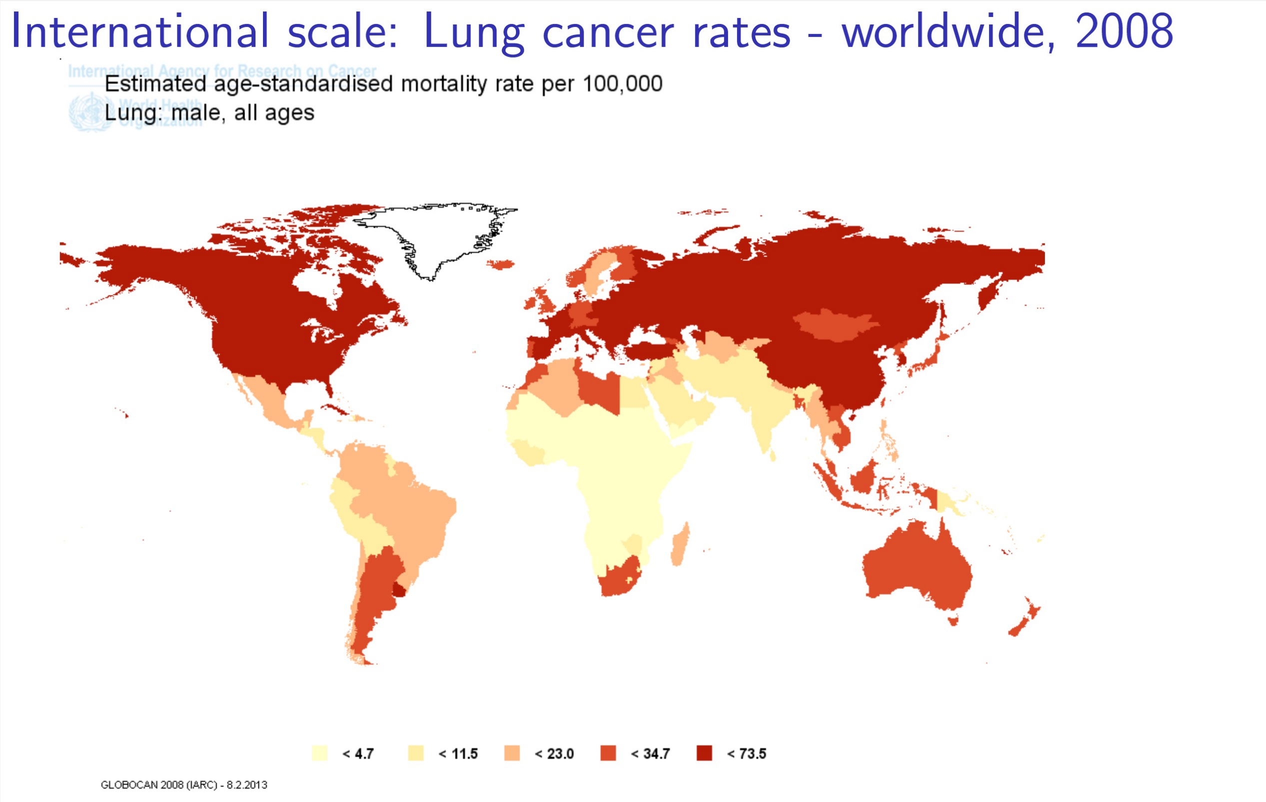
DISEASE MAPPING

Representation of the spatial distribution of the disease on a map

- on a ***international scale***: comparisons between countries (OMS)
  - large international differences in lung cancer mortality rates, potentially explained by differences in smoking prevalences
  - high rates of liver cancer in Africa and South-East Asia, related to hepatitis B

DISEASE MAPPING

- on a ***international scale***: comparisons between countries (OMS)



DISEASE MAPPING

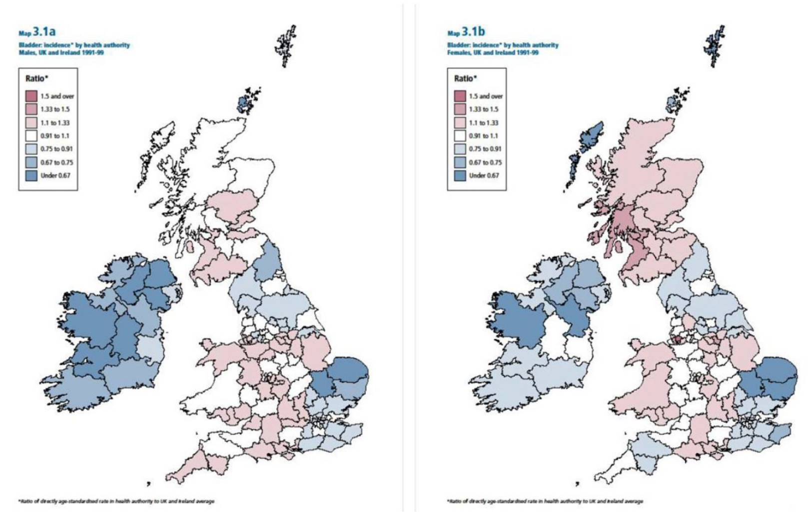
Representation of the spatial distribution of the disease on a map

- on a ***national scale***: comparisons between autonomous communities, regions, ABS, …
  - most of the published disease atlases fall into this category

DISEASE MAPPING

- on a ***national scale***: comparisons between autonomous communities, regions, ABS, …

National scale: Bladder cancer incidence in the UK and Ireland - Cancer Atlas of the United Kingdom and Ireland, 1991 - 2000



source: http://www.ons.gov.uk/ons/rel/cancer-unit/cancer-atlas-of-the-united-kingdom-and-ireland/1991—2000/index.html

DISEASE MAPPING

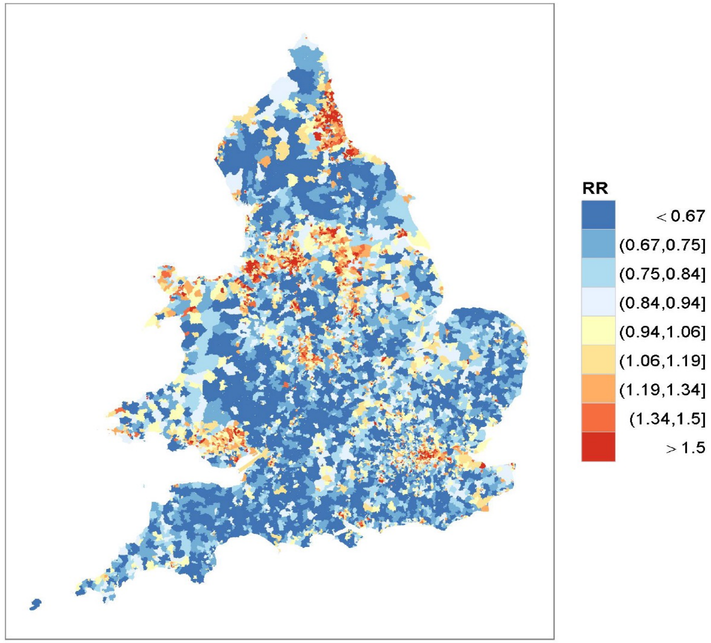
Representation of the spatial distribution of the disease on a map

- on a ***local scale***: studies in small areas
  - sub-national scale, for example, municipalities, neighbourhoods, census tracts, ...
  - increasingly frequent as data and methods improve.

DISEASE MAPPING

- on a ***local scale***: studies in small areas

Lung cancer incidence in males, 1985-2009, England and Wales



DISEASE MAPPING

At present, these maps are meaningless if they do not refer to small areas.

DESIGN AND ANALYSIS OF EPIDEMIOLOGICAL STUDIES IN SMALL AREAS

DISEASE MAPPING

Why are we interested in mapping these measures of disease occurrence in small areas?

DISEASE MAPPING

- Because there is an interest in mapping geographic variations in health outcomes on a small scale in order:
  - to highlight sources of heterogeneity and spatial patterns
  - to suggest public health determinants or etiologic clues
- Because on a small scale (at the level of municipality, neighbourhood, census tract, ...)
  - these measures are less susceptible to ecological bias (aggregation)
  - there is a greater capacity to detect highly localized effects

DISEASE MAPPING

But which occurrence indicators of the disease should we represent on these maps?

DISEASE MAPPING

Some of the most well-known measures of disease occurrence are:

- absolute numbers (counts)
- prevalence
- incidence
- mortality

DISEASE MAPPING

- **Morbidity**
  - ***prevalence***
  - ***incidence***

DISEASE MAPPING

- **Morbidity**
  - ***prevalence***
    - ✓ registers
    - ✓ hospital admissions (data quality problems due to administrative issues)

DISEASE MAPPING

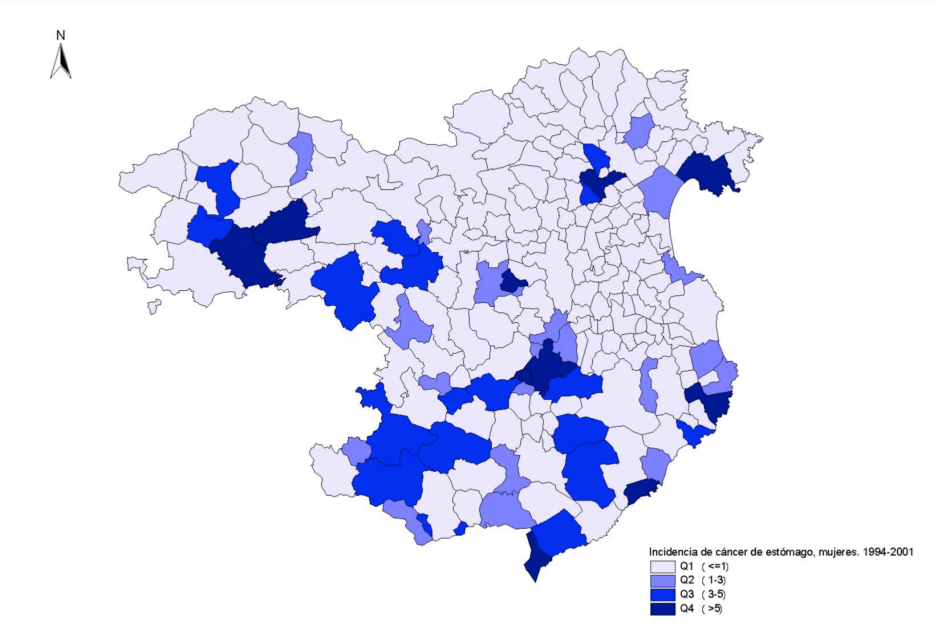
- **Morbidity**
  - ***incidence***
    - ✓ data usually available only for different types of cancers (registries)
    - ✓ may be more sensitive to the effects of exposure
    - ✓ shorter time lapse between exposure and event as compared to mortality

DISEASE MAPPING

- **Mortality**
  - is the most readily available source of data for all diseases
  - should be complete and relatively accurate

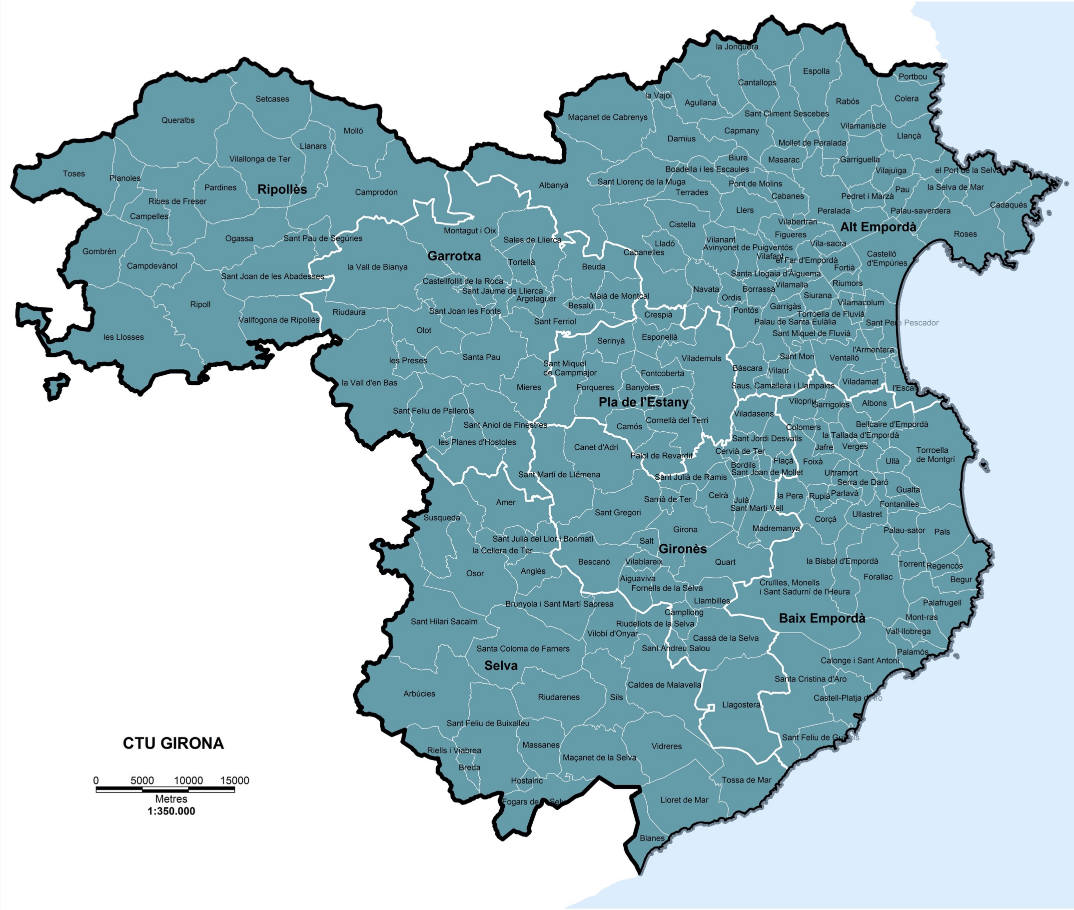
STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- **Observed cases vs. rates**



Incidence of stomach cancer among women, 1994-2001. ***Number of cases***

MAP OF MUNICIPALITIES OF THE HEALTH REGION OF GIRONA



STANDARDISATION OF INCIDENCE AND MORTALITY RATES

➢ **Crude rates vs. standardised rates**

Let us assume two municipalities of 10,000 inhabitantes, A and B.

In A, 125 cases of cancer were observed and in B, 182.

Crude rates: 12.5 (per 1,000 inhabitants) in **A** and **18.2** (per 1,000 inhabitants) in **B**.

Is there a higher risk of dying from cancer in municipality B?

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- **Crude rates vs. standardised rates**

Let us assume two municipalities of 10,000 inhabitantes, A and B.

In A, 125 cases of cancer were observed and in B, 182.

Crude rates: 12.5 (per 1,000 inhabitants) in **A** and **18.2** (per 1,000 inhabitants) in **B**.

Is there a higher risk of dying from cancer in municipality B?

KEY QUESTION: Is the age composition in the two municipalities the same?

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

➢ Crude rates vs. standardised rates

Before answering, we will calculate the **standardized rates** using the indirect method. To do this, we will look at the **age composition** in each municipality.

Grupos edad	Población		Composición poblacional (%)		Número de casos		Tasa específica (por 10^3)	
	A	B	A	B	A	B	A	B
0-4	1.000	4.500	10	45	63	90	63	20
5-14	3.500	3.500	35	35	50	84	14,29	24
15 o més	5.500	2.000	55	20	12	8	2,18	4
Total	10.000	10.000	100%		125	182	12,5	18.2

The composition by age is different in A (older age) than in B (younger).

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

➢ **Crude rates vs. standardised rates**

Now we will standardise by age using the indirect method.

	Población A			Población B			Población estándard		
Grupos edad	**Población**	**Casos**	**Tasa (por 10^3)**	**Población**	**Casos**	**Tasa (por 10^3)**	**Población**	**Casos**	**Tasa (por 10^3)**
0-4	1.000	63	63	4.500	90	20	5.500	153	27,82
5-14	3.500	50	14,286	3.500	84	24	7.000	134	19,14
15 o más	5.500	12	2,182	2.000	8	4	7.500	20	2,67
Total	10.000	125	12,5	10.000	182	18,2	20.000	307	15,35

Crude rate pobulation A

Crude rate pobulation B

Crude rate standard population

GRECS
GRUP DE RECERCA EN ESTADÍSTICA ECONOMETRIA I SALUT

UdG

ciberesp
Centro de Investigación Biomédica en Red
Epidemiología y Salud Pública

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- **Standardised rates by age using the indirect method**

$$Standardised\ rate_A = \frac{125}{27{,}82\times 1.000+19{,}14\times 3.500+2{,}67\times 5.500}\times 15{,}35 = 0{,}00114\times 15{,}35 = 0{,}01752$$

	Población A			Población B			Población estándard		
Grupos edad	**Población**	**Casos**	**Tasa (por 10^3)**	**Población**	**Casos**	**Tasa (por 10^3)**	**Población**	**Casos**	**Tasa (por 10^3)**
0-4	1.000	63	63	4.500	90	20	5.500	153	27,82
5-14	3.500	50	14,286	3.500	84	24	7.000	134	19,14
15 o més	5.500	12	2,182	2.000	8	4	7.500	20	2,67
Total	10.000	125	12,5	10.000	182	18,2	20.000	307	15,35

Crude rate population A

Crude rate population B

Crude rate standard population

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

➢ **Standardised rates by age using the indirect method**

$$Standardised\ rate_B = \frac{182}{27{,}82\times4.500+19{,}14\times3.500+2{,}67\times2.200}\times15{,}35 = 0{,}00092\times15{,}35 = 0{,}01414$$

	Población A			Población B			Población estándard		
Grupos edad	**Población**	**Casos**	**Tasa (por 10^3)**	**Población**	**Casos**	**Tasa (por 10^3)**	**Población**	**Casos**	**Tasa (por 10^3)**
0-4	1.000	63	63	4.500	90	20	5.500	153	27,82
5-14	3.500	50	14,286	3.500	84	24	7.000	134	19,14
15 o més	5.500	12	2,182	2.000	8	4	7.500	20	2,67
Total	10.000	125	12,5	10.000	182	18,2	20.000	307	15,35

Crude rate population A

Crude rate population B

Crude rate standard population

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- **Crude rates vs. standardised rates**

Let us assume two municipalities of 10,000 inhabitantes, A and B.

In A, 125 cases of cancer were observed and in B, 182.

Crude rates: **12.5** (per 1,000 population) in **A** and **18.2** (per 1,000 population) in **B**.

Standardized rates (by age) indirect method: **17.5** (per 1,000 population) in **A** and **14.1** (per 1,000 population) in **B**.

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- **Crude rates vs. standardised rates**

Another way of looking at this.

(1) Grupos edat	(2) Población estándard	(3) Tasas específicas
	A+B	
0-4	5.500	(63+90)/5.500=0,02782
5-14	7.000	(50+84)/7.000=0,01914
15 o més	7.500	(12+8)/7.500=0,00267
Total	20.000	

Grupos edad	Población		Tasas específicas		Casos esperados	
	A	**B**			**A**	**B**
0-4	1.000	4.500	0,02782	0.02782x1000 0.02782x4500	27,82	125,19
5-14	3.500	3.500	0,01914		66,99	66,99
15 o més	5.500	2.000	0,00267		14,685	5,34
Total	10.000	10.000	100%	**Esperados totales**	109,495	197,52
Observados totales	125	182				
Estandarizados	**125/109,495 =1,1416**	**182/197,52 =0,9214**				

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- **Crude rates vs. standardised rates**

Let us assume two municipalities of 10,000 inhabitantes, A and B.

Crude rates: **12.5** (per 1,000 population) in **A** and **18.2** (per 1,000 population) in **B**.

Standardized rates (by age) indirect method: **17.5** (per 1,000 population) in **A** and **14.1** (per 1,000 population) in **B**.

Standardized rates (by age) direct method: **23.14** (per 1,000 population) in **A** and **15.4** (per 1,000 population) in **B**.

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- Crude rates vs. standardised rates

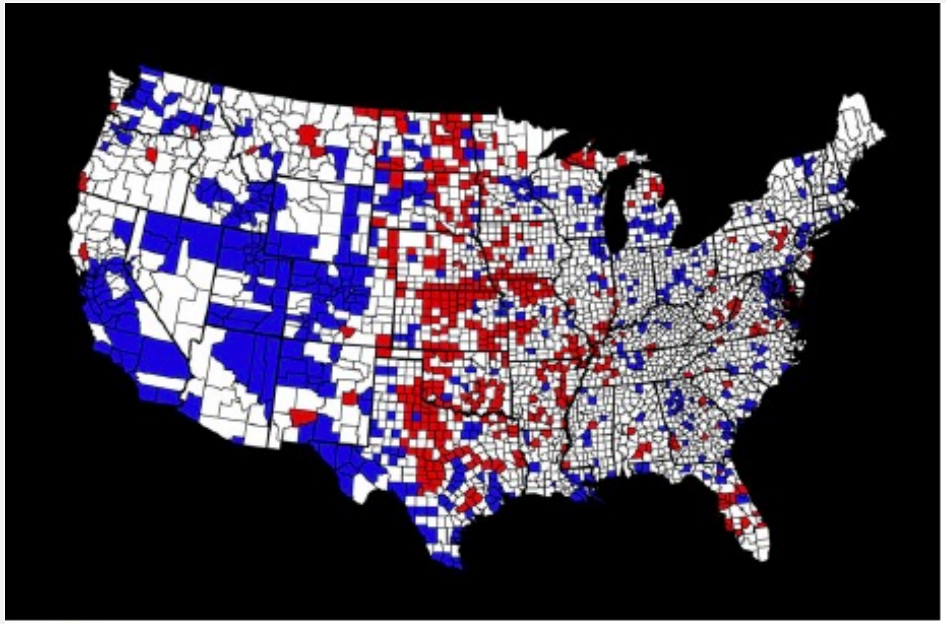


Figure 2
Unadjusted Mortality Rates 1993–1997 Red = High Mortality White = Normal Mortality Blue = Low Mortality

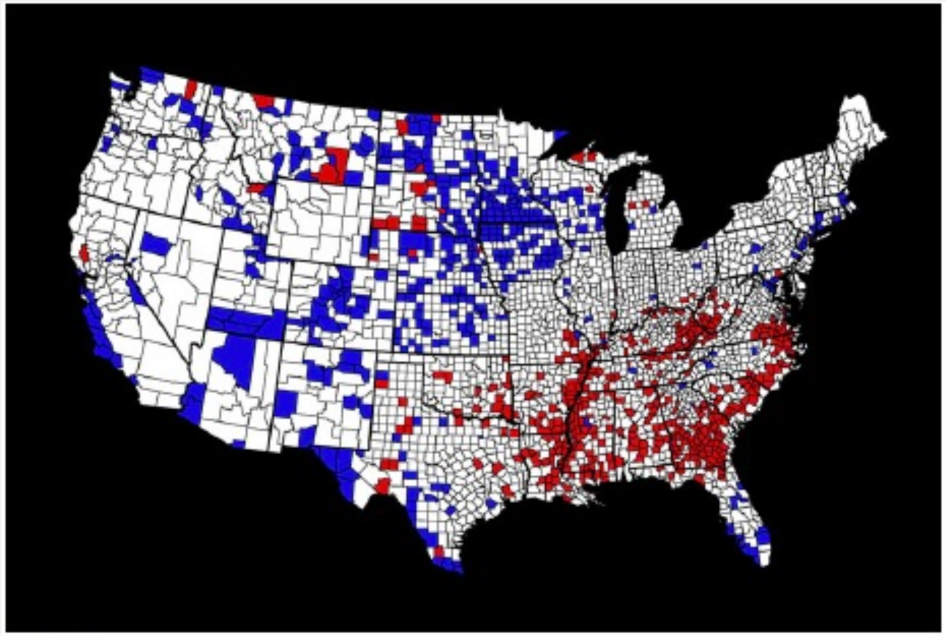


Figure 3
Age Adjusted Mortality Rates 1993–1997 Red = High Mortality White = Normal Mortality Blue = Low Mortality

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- **Crude rates vs. standardised rates**

Figura 3. Tasas crudas por muertes súbitas cardíacas. República mexicana, 2010.

Figura 4. Tasas estandarizadas por muertes súbitas cardíacas. República mexicana, 2010.

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

In conclusion, **rates must be standardized or adjusted**.

We have seen one type of standardization (indirect standardization), but **there are two methods of standardization**:

- **Direct standardisation**: the expected cases are calculated from the age and sex structure of the reference population.
- **Indirect standardization**: the expected cases are calculated from the specific disease rates of the reference population.

GRECS
GRUP DE RECERCA EN ESTADÍSTICA ECONOMETRIA I SALUT

UdG

ciberesp
Centro de Investigación Biomédica en Red
Epidemiología y Salud Pública

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

➢ **Standardised rates by age using the direct method**

$$Standardised\ rate_A = \frac{63\times 2.500 + 14{,}286\times 7.000 + 2{,}182\times 7.500}{20.000} = \frac{462.867}{20.000} = 23{,}14$$

	Población A			Población B			Población estándard		
Grupos edad	**Población**	**Casos**	**Tasa (por 10^3)**	**Población**	**Casos**	**Tasa (por 10^3)**	**Población**	**Casos**	**Tasa (por 10^3)**
0-4	1.000	63	63	4.500	90	20	5.500	153	27,82
5-14	3.500	50	14,286	3.500	84	24	7.000	134	19,14
15 o más	5.500	12	2,182	2.000	8	4	7.500	20	2,67
Total	10.000	125	12,5	10.000	182	18,2	20.000	307	15,35

Crude rate population A

Crude rate population B

Crude rate standard population

GRECS
GRUP DE RECERCA EN ESTADÍSTICA ECONOMETRIA I SALUT

UdG

ciberesp
Centro de Investigación Biomédica en Red
Epidemiología y Salud Pública

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- **Standardised rates by age using the direct method**

$$Standardised\ rate_B = \frac{20\times 5.500 + 24\times 7.000 + 4\times 7.500}{20.000} = \frac{308.867}{20.000} = 15{,}4$$

	Población A			Población B			Población estàndard		
Grupos edad	**Población**	**Casos**	**Tasa (por 10^3)**	**Población**	**Casos**	**Tasa (por 10^3)**	**Población**	**Casos**	**Tasa (por 10^3)**
0-4	1.000	63	63	4.500	90	20	5.500	153	27,82
5-14	3.500	50	14,286	3.500	84	24	7.000	134	19,14
15 o más	5.500	12	2,182	2.000	8	4	7.500	20	2,67
Total	10.000	125	12,5	10.000	182	18,2	20.000	307	15,35

Crude rate population A

Crude rate population B

Crude rate standard population

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

Data: we usually have data for a region of interest/reference area, at a geographical level and for a given time period. For example, Catalonia, census tract level, period 2011-2020.

- O_i: number of observed cases in the area i
- E_i: number of expected cases in the area i, based on the size of the population, adjusted by age, sex, other strata, etc.
- n_i : población a riesgo en el área i

Parameter of interest: Relative risk λ_i in each area compared with the chosen reference area

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- We calculate the expected number of cases assuming that the population has the same stratum-specific mortality/incidence rate as in the reference area.
- We adjust (stratify) by age, sex, etc.
- Indirect standardisation:

$$E_i = \sum_j n_{ij} r_j$$

where r_{ij} is the disease rate for stratum j in the reference population and n_{ij} is the population at risk in the area i, estratum j (if compared internally: $\sum_{i=1}^{N} O_i = \sum_{i=1}^{N} E_i$)

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

Another example: Incidence of lung cancer in men, for all ages, using the rates for England and Wales as a reference, for the period 1985-2009

Strata	Reference area=EW			Ward A		
Age group	Population n_j	Observed O_j	Age-specific rate per 100,000 males $r_j = \frac{O_j}{n_j}$	Population n_{ij}	Observed O_{ij}	Expected $E_{ij} = \frac{n_{ij} * r_j}{100000}$
0−4	41,400,692	15	0.04	11,438	0	0.00
5−9	41,143,722	6	0.01	9,697	0	0.00
10−14	41,469,696	9	0.02	9,026	0	0.00
15−19	43,087,823	39	0.09	8,650	0	0.01
20−24	45,441,353	79	0.17	12,409	0	0.02
25−29	46,873,725	172	0.37	16,963	0	0.06
30−34	46,927,658	518	1.10	17,303	0	0.19
35−39	46,936,367	1,465	3.12	13,847	0	0.43
40−44	45,304,711	4,136	9.13	11,843	1	1.08
45−49	41,657,557	9,835	23.61	9,457	5	2.23
50−54	38,451,416	20,929	54.43	8,561	3	4.66
55−59	35,842,426	40,427	112.79	7,613	8	8.59
60−64	32,480,032	68,230	210.07	6,968	5	14.64
65−69	28,231,499	95,794	339.32	6,290	15	21.34
70−74	23,315,240	110,371	473.39	5,098	27	24.13
75−79	17,297,264	102,038	589.91	4,049	22	23.89
80−84	10,498,214	68,273	650.33	2,616	20	17.01
85+	6,289,452	38,748	616.08	1,312	12	8.08
TOTAL	632,648,846	561,084		163,140	118	126.38

$$SIR_A = \frac{118}{126.38} = 0.93$$

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

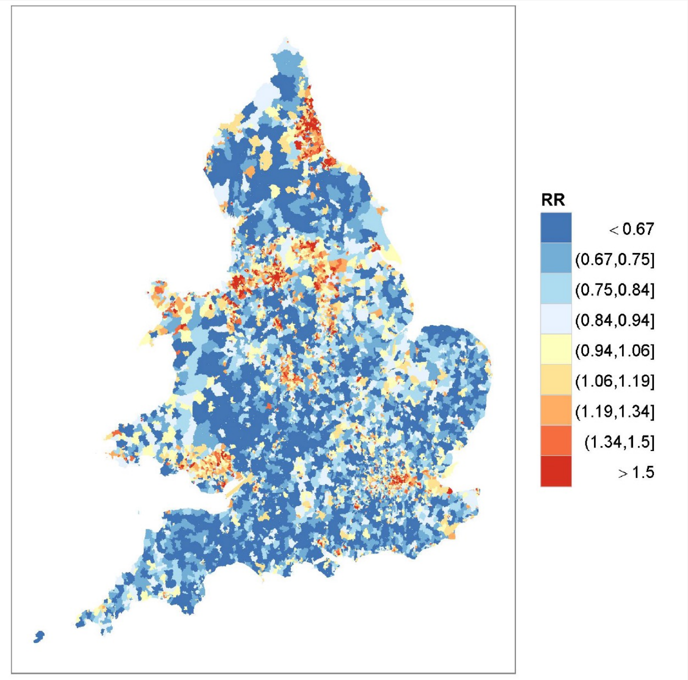
Interpretation:

$$SIR_A = \frac{118}{126.38} = 0.93$$

- We have fewer incident cases of lung cancer in men in health area A than would be expected (7% fewer) in this area once adjusted by age groups.

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- Standardised incidence ratio (SIR=O/E) of lung cancer in men in England and Wales (1985-2009) at county level



STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- The common practice is ***to plot*** the ***SMR (o SIR)*** on a map:
  - ***in the case of rare diseases and/or small areas, the SMRs (or SIRs) are very inaccurate*** because the variance is proportional to the square of the denominator (that is, it is proportional to the number of expected cases squared).
  - the variance associated with areas that have small expected cases will be very high

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- ***SMRs (or SIRs) are estimated independently*** in each area.
- SMRs (or SIRs) do not use the risk estimates in other areas of the map, even though they are likely to be similar. That is, ***they do not take into account the very likely spatial dependence***.

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- Standardized rates (**SMR or SIR**) are very unstable.

Observed = 100	Observed = 2
Expected = 80	Expected = 1.6
SIR = 1.25	SIR = 1.25
Observed = 101	Observed = 3
Expected = 80	Expected = 1.6
SIR = 1.26 (1% more)	SIR = 1.88 (50% more)

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- Let us remember that our ***parameter of interest*** is the relative risk λ_i in each area compared to the chosen reference area.

- Instead of calculating the parameter of interest (as we have done above), what is usually done is ***to estimate it by regression***.

- ***To estimate this parameter***, ***a Poisson regression*** tends to be used where the response variable is the numerator of the relative risk, in other words, the observed cases (discrete or count variable) and the expected cases are entered as an offset.

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- ***Standard statistical model*** if we have ***rare diseases and/or small areas***:

$$O_i \sim Poisson(\lambda_i\, E_i)$$

where E_i is the number of expected cases in the area i

- ***Parameter of interest:*** the relative risk λ_i, which is usually estimated using standardised mortality rates (SIR) and/or standardised incidence rates (SIR):

$$\hat{\lambda}_i = SMR_i \text{ or } SIR_i = \frac{O_i}{E_i} \qquad \text{and} \qquad \mathrm{Var}(\hat{\lambda}_i) = \frac{\lambda_i}{E_i} \rightarrow \widehat{Var}(\hat{\lambda}_i) = \frac{O_i}{E_i^2}$$

Remember that: $X \sim Poisson(\mu) \leftrightarrow E(X) = Var(X) = \mu$

GRECS
GRUP DE RECERCA EN ESTADÍSTICA ECONOMETRIA I SALUT

STANDARDISATION OF INCIDENCE AND MORTALITY RATES

- In the case of more ***common diseases, the Binomial model*** is usually used:

$$O_i \sim Binomial(p_i, N_i)$$

where N_i= population at riks and p_i= probability of the disease

- $logit(p_i) = \alpha + V_i$

- ***Parameter of interest:***

$$\text{odds ratio} = OR_i = exp(\alpha + V_i)$$

GRECS
GRUP DE RECERCA EN ESTADÍSTICA ECONOMETRIA I SALUT

UdG

ciberesp
Centro de Investigación Biomédica en Red
Epidemiología y Salud Pública

"Data don't make any sense, we will have to resort to statistics."