Lecture 1

Introduction to spatial analysis

MSc in Epidemiology @ Imperial College London
February 18, 2019

Module detail

Module goals

Through this module we will explore the main concepts and statistical methods used in spatial analysis and we will familiarize with a range of theoretical concepts and practical tools to **analyse**, **visualize**, **model**, and **interpret** spatially-related data with focus (but not only!) on epidemiological research.

We will handle model complexity within a Bayesian framework, with statistical inference carried out using Markov Chain Monte Carlo methods.

Module detail

From handbook, the major topics are:

- Week 1: Introduction to spatial analysis and areal unit data
- Week 2: Disease mapping (CAR model) and ecological regression models
- Week 3: Introduction to temporal processes and spatial-temporal models with focus on areal unit data
- Week 4: Point referenced data analysis, including variograms, kriging, spatial regression.
- Week 5: Point pattern data analysis, including Poisson process + QA session

Software

- Manipulation and visualisation of the data
  - R (https://www.r-project.org/) via RStudio environment
  - A suite of R packages for spatial analysis and visualization
- Modelling
  - OpenBUGS (http://www.openbugs.net/w/FrontPage) software via the R2OpenBUGS package in R
  - Post-processing (convergence checks and extraction of results for maps etc.) via `mcmcplots`, `coda` packages in R

Module detail

Assessment

- The examination will include both theoretical questions and the analysis of a real datasets using R and OpenBUGS
- Three hours in-class examination (open book)
- Date: May 7, 2019

People you will work with

- Monica Pirani, monica.pirani@imperial.ac.uk
- Areti Boulieri, a.boulieri@imperial.ac.uk
- Chiara Forlani, c.forlani16@imperial.ac.uk
- Niloofar Shoari, n.shoari@imperial.ac.uk

Lecture outline

1. Background
2. Spatial data
3. Areal unit data
4. Moving to modelling approaches: areal-referenced studies in epidemiology
    - General framework
5. Introduction to Bayesian disease mapping
    - Aims
    - Hierarchical framework
    - Interpretation of the parameters
6. Disease mapping in practice
    - Protocol
    - Introducing Practical 1

Objectives of spatial analysis

- To quantify phenomena referenced in space
- To examine the spatial variability of the phenomenon through explicative variables associated to the location as well as variables measured on different scale
- To identify and explain clusters in the observations
- To estimate and predict the phenomenon in areas/points where it is not sampled
- To study methods and models to describe and explain the process that operates in space based on (a sample of) observations taken at particular locations

Tobler (1970) first law of geography

> *Everything is related to everything else, but near things are more related than distant things*

- Data that are close together in space (time) are often more similar than data that are far apart
  (except when they aren't!)
- Can we quantify that for our data?
  Yes! Using statistics!

Spatial data - Outline

1. Background
2. Spatial data
3. Areal unit data
4. Moving to modelling approaches: areal-referenced studies in epidemiology
    - General framework
5. Introduction to Bayesian disease mapping
    - Aims
    - Hierarchical framework
    - Interpretation of the parameters
6. Disease mapping in practice
    - Protocol
    - Introducing Practical 1

Terminology

The formal framework to be used throughout the module is given by the so-called spatial stochastic process (also often referred to as spatial random field):

$$\{Z(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$$

where

- $Z(\boldsymbol{s})$ is a random variable at location $\boldsymbol{s}$
- $\boldsymbol{s}$ is the location of the measurement (e.g. lat, long)
- $\mathcal{D}$ is the domain, and it is called the index set = possible locations

Type of spatial data

Following Cressie (1993), we distinguish three types of spatial data, based on the nature of the spatial domain $\mathcal{D}$:

- Areal (or lattice) data: the spatial domain $\mathcal{D}$ (regular or irregular) is fixed and is partitioned into a finite number of areal units with well defined boundaries.

- Point-referenced (or geostatistical) data: $Z(s)$ is a random vector at a selected location $\boldsymbol{s}$, where $\boldsymbol{s}$ varies continuously over $\mathcal{D}$, a fixed subset of $\mathcal{R}^d$.

- Point pattern data: the set of locations in $\mathcal{D}$ is itself random. Its index set gives the locations of random *events* that are the spatial point pattern.

Example of areal unit data

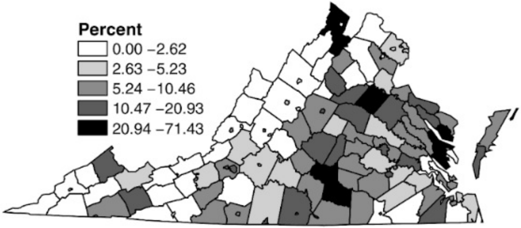


Figure: Percent of children under the age of 72 months with elevated blood lead levels in Virginia in 2000

- This figure is an example of a *choropleth map*, which uses grey scale (or shades of color) to classify values into a few broad classes
- From the choropleth map we know which regions are adjacent which other regions
- The "sites" $\boldsymbol{s} \in \mathcal{D}$ in this case are actually the regions (or blocks) themselves, which can be denoted not by $\boldsymbol{s}_i$ but by $B_i, i = 1, \ldots, N$ to avoid confusion between points $\boldsymbol{s}_i$ and regions B_i.

Source: Schabenberger and Gotway (2004)

Example of point referenced data

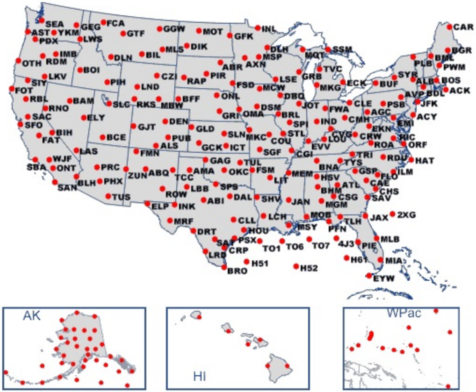


Figure: Map of wind and temperature stations in US (2019)

Source: https://www.aviationweather.gov/windtemp

Examples of point pattern data

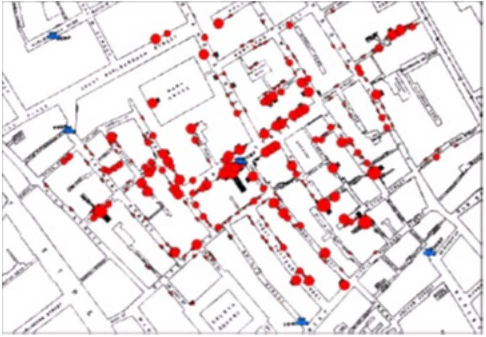


Figure: Map of cholera outbreaks by Dr John Snow (1854)

Areal unit data

- Data referenced at an aggregate level.
- Areal "units" are generally irregular geographic areas and in spatial analysis we have a collection of areal units (the methods may also apply to regular grids).
- Areal data are often defined by administrative boundaries (states, counties, districts etc.).
- We care about how areal units connect to each other and we use neighbour information to define spatial relationships.

Is there a pattern?

- Every real map presents sub-regions with some relatively higher values clustered in some parts of the map. Other parts present relatively lower valued areas.

- We also meet much "noise": sub-regions with high and low values mixed randomly without any spatially structured pattern.

- It is no easy just visually determine whether values are spatially clustered.

- Independent measurements will have no pattern, and would look completely random, but there may actually be an underlying pattern.

Maps can be misleading...

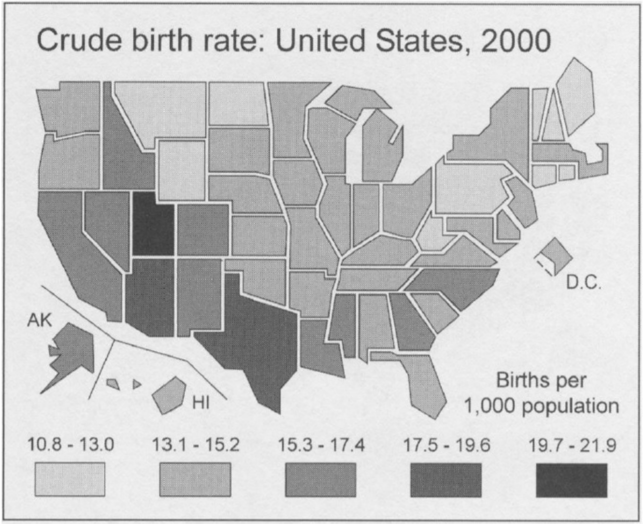

Crude birth rate: United States, 2000
AK
HI
D.C.
Births per 1,000 population
10.8 - 13.0
13.1 - 15.2
15.3 - 17.4
17.5 - 19.6
19.7 - 21.9

Figure: Crude birth rates by state based on equal-interval cut points

Maps can be misleading...

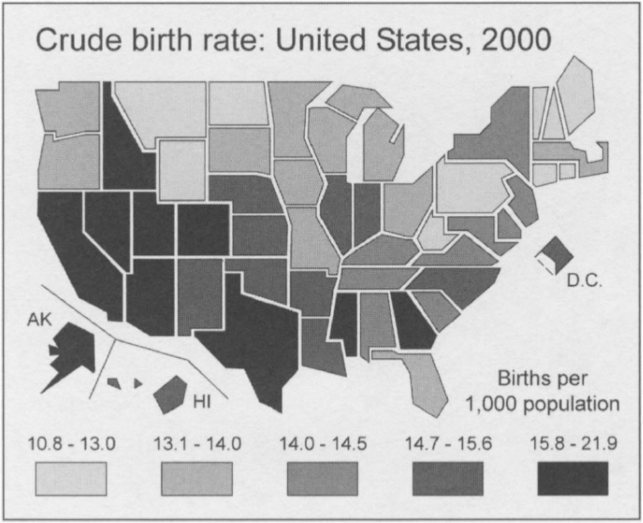


Figure: Crude birth rates by state based on quantile cut points

Spatial autocorrelation

Therefore, if there is a spatial pattern, how strong is it?

- The detection of spatial autocorrelation is useful in spatial analysis for identifying underlying data structures, the degree of spatial randomness, or clustering in the data.
- For a given variable, spatial correlation measures the level, the nature and the strength of interdependencies among the data points (or observational units) within the variable both in terms of space and the attribute under consideration (Oyana and Morgai 2015).
- Let Z_i be the variable of interest measured in areal unit B_i. The analysis may proceed as follows:
    - Representation of spatial proximity in the areal data (i.e. each B_i is supplemented with neighbourhood information: spatial weights)
    - Testing for spatial pattern using global measures, such as Moran's I statistic
    - Testing for spatial pattern using local measures (i.e. local testing such as local Moran's I)
    - Model (smooth) the data

Measuring spatial autocorrelation

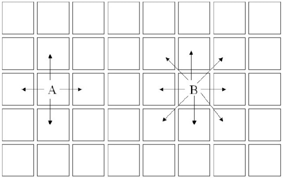
- A number of approaches have been suggested for measuring spatial autocorrelation
- Global measures of spatial autocorrelation share a common structure: calculate the similarity of values at locations i and j then weight the similarity by the proximity of locations i and j
- Form of the measure of global clustering

$$T = c\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} \times \text{similarity}_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}}$$

  - c constant
  - N number of areas
  - w_{ij} weight reflecting the proximity between areas i and j
  - similarity_{ij} measure of similarity between data values in areas i and j
- Null hypothesis: independent observations (no clustering)

Measures of proximity

- Various ways of measuring the *closeness* between areas
    - $w_{ij} = 1$ if areas i and j are adjacent (common boundary), 0 otherwise
    - $w_{ij} = 1$ if the centroids of areas i and j are within a certain distance of each other
    - $w_{ij} = d_{ij}^{-1}$ where d_{ij}^{-1} is the inverse of the distance between the centroids of areas i and j
- Operationally, we can distinguish between a rook (A) and a queen (B) criterion of contiguity



The rook criterion defines neighbours by the existence of a common edge between two spatial units.
The queen criterion defines neighbours as spatial units sharing a common edge or a common vertex.

- The choice of weights is arbitrary and should be made taking into account the specific problem under analysis

Moran's I test for spatial autocorrelation

Moran's I test can be applied to the data directly, or to the residuals from a regression model (as shown in Practical 1).
Let $\{Z_i : i, \ldots, N\}$ represent spatially referenced data (or residuals) for N spatial locations.
The Moran'I statistics is:

$$I = \frac{N \sum_i \sum_j w_{ij}(Y_i - \overline{Z})(Z_j - \overline{Z})}{\left(\sum_i \sum_j w_{ij}\right) \sum_{k=1}^{N}(Z_k - \overline{Z})^2}$$

- If there is no spatial dependence, I will be close to 0 (i.e. spatial pattern is random)
- If there is clustering then areas close together (as defined by w_{ij}) will tend to have similar responses (i.e. SMR), so I will be positive
- If $I < 0$, clustering of dissimilar values
- "Significance" will be done using Monte Carlo (MC) approach (i.e. the data are repeatedly randomly assigned to different areas, and the statistic calculated under each permutation, yielding a comparison distribution)

Types of studies with a spatial component

Disease mapping

- spatial and spatio-temporal variation in disease risk
- exploit spatial dependence in order to smooth rates and provide better predictions

Ecological regressions

- Aim is to examine how geographical variations in health outcomes relate to geographic variations in exposure of interest (e.g. air pollution)
- Association between risk and exposures at the area level
- Can address aetiological and/or public health questions

Disease clustering and cluster detection

- Disease clustering: specific patterns of heterogeneity in space
  → tendency of disease risk
- Cluster detection: identification of *unusual* aggregations of cases
  → *reveal* hotspots

General framework - I

- **Data** for a region of interest/reference area, geographical level and a specific period England, ward level, 2009-2012
    - O_i: Observed number of cases in area i
        - ⋆ Lung cancer deaths in males aged 45+
        - ⋆ Congenital anomalies
    - n_i: Population at risk in area i
        - ⋆ Male population aged 45+
        - ⋆ Live births and stillbirths
- **Parameter of interest** Relative risk λ in each area compared with the chosen reference area

General framework - II

- Standard statistical model if rare disease and/or small areas

$$O_i \sim \text{Poisson}(\lambda_i E_i)$$

where E_i = expected number of cases in area i

- λ estimated by Standardised Mortality/Incidence Ratio (maximum likelihood estimator)

$$\hat{\lambda}_i = \text{SMR}_i \text{ or } \text{SIR}_i = \frac{O_i}{E_i} \quad \text{and} \quad \text{Var}(\hat{\lambda}_i) = \frac{\lambda_i}{E_i} \rightarrow \hat{\text{Var}}(\hat{\lambda}) = \frac{O_i}{E_i^2}$$

So that areas with small E_i have hight associated variance

Recall: $X \sim \text{Poisson}(\mu) \Leftrightarrow \text{E}(X) = \text{Var}(X) = \mu$

Expected numbers of cases - definition

- Expected nb of cases if the population had the same stratum-specific mortality/incidence rates as in a reference area
- Adjustments (strata): age, gender ...

Indirect standardisation

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

with

r_j: disease rate for stratum j in the reference population

n_{ij}: population at risk in area i, stratum j

If internal comparison: $\sum_{i=1}^{N} O_i = \sum_{i=1}^{N} E_i$

Expected numbers of cases - calculation

Lung cancer incidence in males, all ages, using the rates in England and Wales as reference, for the period 1985-2009

Strata	Reference area=EW			Ward A		
Age group	Population	Observed	Age-specific rate per 100,000 males	Population	Observed	Expected
	n_j	O_j	$r_j = \frac{O_j}{n_j}$	n_{ij}	O_{ij}	$E_{ij} = \frac{n_{ij} * r_j}{100000}$
$0-4$	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

$\text{SIR}_A = \frac{118}{126.38} = 0.93$ → Fewer incident cases of lung cancer for males in ward A than expected in EW after adjusting for differences in age

Class exercise

Using the previous setting, calculate the SIR in ward A for men above 65 years old

Disease mapping - Outline

1. Background

2. Spatial data

3. Areal unit data

4. Moving to modelling approaches: areal-referenced studies in epidemiology
   - General framework

5. Introduction to Bayesian disease mapping
   - Aims
   - Hierarchical framework
   - Interpretation of the parameters

6. Disease mapping in practice
   - Protocol
   - Introducing Practical 1

Aims of disease mapping

Disease maps used variously for

- Descriptive purposes: to summarise spatial and spatio-temporal variation in disease risk → a visual summary of geographical risk
- To generating aetiological hypotheses: informal examination of maps with exposure maps (formal examination via spatial regression)
- For surveillance, to highlight areas at apparently high risk
- To aid policy formation and resource allocation

What to map

- **Mortality**
  - most readily available source of data for all diseases (England: Office for national statistics)
  - should be complete and relatively accurate
- **Morbidity**
  - **Incidence**
    - ★ incidence data usually only routinely available for cancers (registries)
    - ★ may be more sensitive to effects of exposure
    - ★ shorter time lag between exposure and event compared to mortality
  - **Prevalence**
    - ★ registries
    - ★ hospital admissions (many issues of data quality because admin purpose)

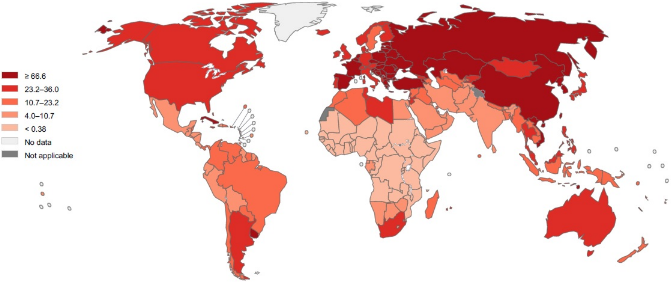
Geographical scales

Disease mapping may be carried out at a variety of scales:

- **International**: Comparisons between countries (WHO)
  - Large international differences in mortality rates of lung cancer, potentially explained by differences in the prevalence of smoking
  - High rates of liver cancer in Africa and SE Asia, related to Hepatitis B infection
- **National** Comparisons between e.g. regions or states
  - Most published disease atlases fall into this category
- **Small area studies**
  - Sub-national scale, e.g. wards or districts (UK)
  - Becoming increasingly common as data and methods improve

International scale: Lung cancer mortality rates - worldwide, 2012

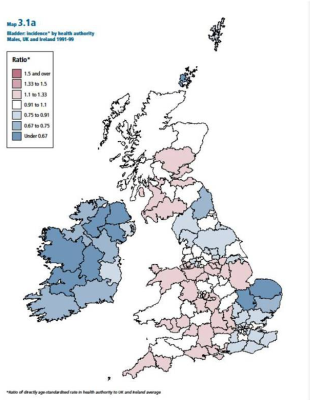
Estimated age-standardized rates (World) per 100,000 (males, all ages).



- Large international differences in mortality rates of lung cancer, potentially explained by differences in the prevalence of smoking.

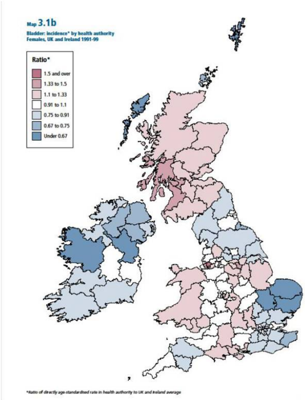
source: http://gco.iarc.fr/today/data/pdf/fact-sheets/cancers/cancer-fact-sheets-11.pdf

National scale: Bladder cancer incidence in the UK and Ireland - Cancer Atlas of the United Kingdom and Ireland, 1991 - 1999


Map 3.1a

Bladder: incidence* by health authority
Males, UK and Ireland 1991-99

*Ratio of directly age-standardised rate in health authority to UK and Ireland average


Map 3.1b

Bladder: incidence* by health authority
Females, UK and Ireland 1991-99

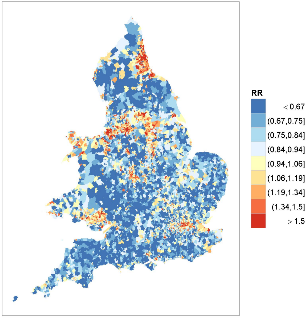
*Ratio of directly age-standardised rate in health authority to UK and Ireland average

source: https://www.ons.gov.uk/ons/rel/cancer-unit/cancer-atlas-of-the-united-kingdom-and-ireland/1991—2000/chapter-3–bladder.pdf

- Wide variability among the countries with higher rates in Scotland and Wales
- Similar rates in Northern Ireland and Ireland, 20-30% lower than the average

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

Standardized incidence rates (SIRs) at ward level



	Min	Q1	Median	Q3	Max
O	0	26	47	84	456
E	3.25	32.14	53.60	82.47	390.49
SMR	0	0.70	0.89	1.13	2.63

- Is the variability real or simply reflecting unequal E_is?
- Have the highlighted areas truly a raised relative risk?

Problems with mapping SMRs

- Common practice is to map SMRs
  - → SMR_i very imprecise for rare diseases and/or areas with small populations For the model $O_i \sim \mathrm{Poisson}(\lambda_i E_i)$, the MLE is $\mathrm{SMR}_i = \frac{O_i}{E_i}$, and its estimated variance is $\hat{\mathrm{Var}}(\hat{\lambda}) = \frac{O_i}{E_i^2}$
    - areas with small E_i have high associated variance

- SMR in each area is estimated independently
  - → makes no use of risk estimates in other areas of the map, even though these are likely to be similar

⇒ Highlights extreme risk estimates based on small numbers

⇒ Ignores possible spatial correlation between disease risk in nearby areas due to possible dependence on spatially varying risk factors

Problems addressed using Bayesian ‘smoothing’ estimators in a hierarchical formulation:

- Poisson-logNormal model: non spatial smoothing
- Poisson-logNormal-spatial model: spatial and non spatial smoothing

Hierarchical modeling for disease mapping

Poisson-logNormal model

$$
\begin{aligned}
O_i &\sim \text{Poisson}(\lambda_i E_i) \\
\log \lambda_i &= \alpha + V_i \\
V_i &\sim \text{Normal}(0, \sigma_v^2)
\end{aligned}
$$

Priors (vague, non informative):

- between-area variance σ_v^2:
  $\sigma_v^2 \sim$ Inverse Gamma$(0.5, 0.0005) \Leftrightarrow \tau_v \sim$ Gamma$(0.5, 0.0005)$
- mean log relative risk: $\alpha \sim N(0, 10000)$

where

- O_i, E_i: observed and expected nb of cases in area i
- $\lambda_i = \exp(\alpha + V_i)$: RR in area i compared with expected risk based on age and sex of population
- Parameters V_i: **area-specific random effects**
- residual RR $= \exp(V_i)$

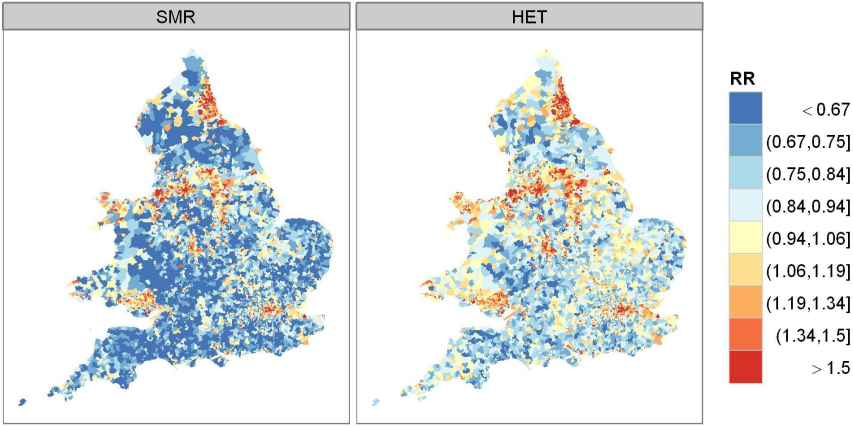
Recall $Y \sim$ Inverse Gamma$(a, b) \Leftrightarrow X = \frac{1}{Y} \sim$ Gamma(a, b), E$(X) = \frac{a}{b}$, Var$(X) = \frac{a}{b^2}$

Interpretation of the random effects

- Parameters V_i: **area-specific random effects** to take into account overdispersion
  - excess variation in the observed counts due to possible errors in numerator and denominator data
  - latent variable which captures the effects of unknown or unmeasured area level covariates
  - variance of the random effects (σ^2) reflects the amount of extra-Poisson variation in the data
  - σ^2 is the parameter controlling the spread of the random effects
- A useful summary of variability between units in a hierarchical model is to rank the random effects and calculate the ratio between two units at opposite extremes
  - $\lambda_{5\%}$: RR for the area ranked at the 5th percentile
  - $\lambda_{95\%}$: RR for the area ranked at the 95th percentile
  - $\mathrm{QR}_{90} = \frac{\lambda_{95\%}}{\lambda_{5\%}}$: ratio of RR between the top and bottom 5% of areas

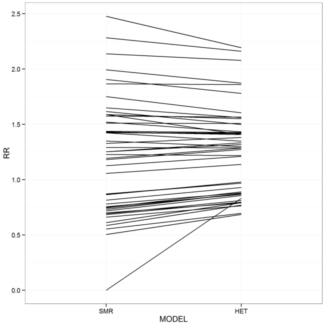
Lung cancer incidence in males, 1985-2009, England and Wales (I)

RR estimates using 2 methods



SMRs and non-spatially smoothed RRs

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



SMRs versus non-spatially smoothed RRs in selected areas

σ_v^2 = between-area variance of log relative risk of lung cancer
Posterior mean and 95% CI = 0.094 (0.091-0.098)

BUGS code for Poisson-logNormal model

```
model {
for(i in 1:N){
O[i] ~ dpois(mu[i])                  #mu[i]=lambda[i]*E[i]
log(mu[i]) <- log(E[i])+alpha+V[i]   #log(lambda[i])=alpha+V[i]
V[i] ~ dnorm(0, tau.v)               #area-specific RE
RR[i] <- exp(alpha + V[i])           #area-specific RR
resRR[i] <- exp(V[i])                #area-specific residual RR
e[i] <- (O[i]-mu[i])/sqrt(mu[i])     #residuals
 }
# Priors:
alpha  ~  dnorm(0.0,0.0001)          #overall log RR
overallRR <- exp(alpha)              #overall RR
tau.v ~ dgamma(0.5, 0.0005)          #prior on precision of RE
sigma2.v <- 1/tau.v                  #between-area variance of RR
QR90 <- ranked(resRR[],8360)/ranked(resRR[],440)  #90% relative risk ratio
 }
```

- tau.v=precision=1/variance, so if small precision ⇒ large variance
- Priors on alpha and tau.v
- `ranked(resRR[],8360)`: the 8360$^{\text{th}}$ smallest RR = 95% percentile of the distribution of RRs (8800*0.95=8360)

Data and initial values for Poisson-logNormal model

When using WinBUGS/OpenBUGS:

- Data:

  Number of areas (N), Observed (O) and Expected (E) numbers of cases in each area

  ```
  list(N=8800,
  O=c(7,35,11,...),
  E=c(17.49,39.78,12.20,...)  #same order as O
  )
  ```

- Initial values for the unknown parameters for each chain

  ```
  list(alpha=0.01, tau.v=10,
  V=c(0.01,0.01,0.01,...) #same length as O and E
  )
  ```

Binomial framework

- For more common diseases, Binomial model may be preferred

$$O_i \sim \text{Binomial}(p_i, n_i)$$

  where
  - n_i = population at risk
  - p_i = probability of disease
- $\text{logit}(p_i) = \alpha + V_i$
- Parameter of interest: Odds ratio $\text{OR}_i = \exp(\alpha + V_i)$

Disease mapping - Conclusion

- Use of routinely collected data (registries, surveys, censuses,...), easy to obtain, greatly improved in quality
- Non smoothed estimates of RR (i.e. SMR) can be unstable
- Non spatially-smoothed estimates of RR: more robust estimates of RR
- Useful visual summary of the geographical disease risk
- Extension to spatial smoothing
- Easy computation with OpenBUGS/WinBUGS and R

Carrying out a disease mapping study - protocol

Study health outcome, region of interest, period

- ICD codes, incidence, mortality
- Adjustments (strata): age, sex, socio-economic status
- Comparison area
- Map of the region of interest (shapefiles)

Data descriptive summary

- Population and observed cases per stratum and area
- Calculation of the expected numbers of cases, adjusted for the covariates, in each area
- Calculation of the SMRs

Non-spatially smoothed RRs using Poisson-logNormal model

- Write the model
- Format the data
- Run WinBUGS/OpenBUGS
- Sensitive analyses (priors on variance parameters)

Presentation of the results

- Quantitative summary (variance, quantile, ratio)
- Visual summary (histograms, maps)

Introducing Practical 1

Disease mapping study of incidence of lung cancer (all ages and both genders), in the 628 wards of Greater London, 2001-2005

- Lungcancer_strata_GL.csv: file containing all the data to carry out the analysis
- Shapefile of Greater London with Thames river
- model_HET.txt: file containing the model code
- Practical1.pdf and Practical1.Rmd: tutorial containing the text, questions and the R code

New script - beginning

- The beginning of a script is always the same:
    - Load packages you will need (can be done anytime). If not installed by default, install the packages using `install.packages()` function
    - Set a new working directory, i.e. the path where the files are

```
#load the packages
library(rgdal)       # Geospatial Data Abstraction Library,
                     # projection/transformation operations
library(maptools)    # Functions for manipulating and reading geographical data
library(sp)          # Classes and methods for spatail analysis
library(spdep)       # Functions and tests for evaluating spatial patterns and autoco
library(SpatialEpi) # Methods and Data for Spatial Epidemiology
library(R2OpenBUGS) # A Package for Running OpenBUGS from R
library(mcmcplots)  # A Package for plotting and viewing of MCMC output
library(coda)        # A Package for summarizing and plotting of MCMC output
                     # and diagnostic tests
#set a new working directory, i.e. the path where the files are
setwd("C:/SpatialAnalysis2019/Practicals/Practical1")
```

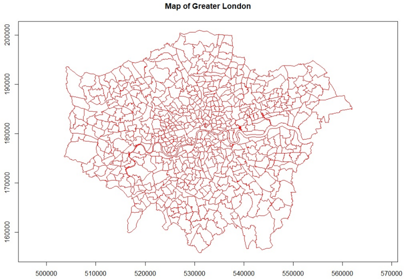
Map of Greater London

- Shapefile: list of spatial objects (list of points, lines, or polygons), possibility to attach data
- Collection of files with the same stem and different extensions
  - GreaterLondon_ward_river.shp it contains the geometry data
  - GreaterLondon_ward_river.shx it contains the spatial index
  - GreaterLondon_ward_river.dbf it contains the attribute data
  - GreaterLondon_ward_river.prj it contains information related to a coordinate system
- ID of the polygons in the .dbf file

Map of Greater London

```
# Read the shapefile of Greater London with the Thames river
shpriver <- readOGR(dsn = ".", layer = "GreaterLondon_ward_river")

# Map of Greater London
plot(shpriver, border = "red", lwd = 1, axes = TRUE)
title(main = "Map of Greater London")
```

- shpriver: Object of class SpatialPolygonsDataFrame
- List of 628 objects
- Each object of the list contains info for the area (ID, coordinates,...)



Lung cancer data

```
# Import the health data
lung <- read.csv("Lungcancer_strata_GL.csv")

# Print only the first rows of the data
head(lung)

# Names of the variables
names(lung)

# Levels for the variable SEX
levels(lung$SEX)
```

- Data stored in a data frame (object with different types of variables)

```
STwardcode POLY_ID STWardName SEX AGE_GROUP POPULATION CASES
 00AAFA       1    Aldersgate   M     50_54     568.01     0
 00AAFA       1    Aldersgate   M     55_59     485.00     0
 00AAFA       1    Aldersgate   M     60_64     303.97     0
```

  ⇒ `names`, `dim`, `head`

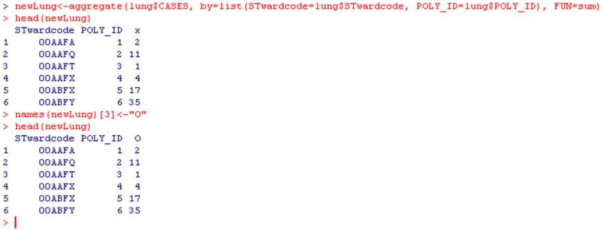
- Manipulating the data
  - subset: `lung[lung$SEX=="F" & lung$AGE_GROUP=="5_9",]`
  - functions: `summary`, `sum`, `aggregate`

Calculation of the observed counts of lung cancer

aggregate(x, by, FUN) from stats package

- x: R object
- by: a list of grouping elements, each as long as the variables in x
- FUN: function to compute the summary statistics which can be applied to all data subsets (mean, sum, median,...)

```
# Create a new data frame with the observed numbers of cases per ward
newLung <- aggregate(lung$CASES, by=list(STwardcode=lung$STwardcode,
                   POLY_ID=lung$POLY_ID), FUN=sum)

# Change the names
names(newLung)[3]<-"O"
```



```
> newLung<-aggregate(lung$CASES, by=list(STwardcode=lung$STwardcode, POLY_ID=lung$POLY_ID), FUN=sum)
> head(newLung)
  STwardcode POLY_ID  x
1     00AAFA       1  2
2     00AAFQ       2 11
3     00AAFT       3  1
4     00AAFX       4  4
5     00ABFX       5 17
6     00ABFY       6 35
> names(newLung)[3]<-"O"
> head(newLung)
  STwardcode POLY_ID  O
1     00AAFA       1  2
2     00AAFQ       2 11
3     00AAFT       3  1
4     00AAFX       4  4
5     00ABFX       5 17
6     00ABFY       6 35
> |
```

Calculation of the expected counts of lung cancer and SMR

`expected(population,cases,n.strata)` from `SpatialEpi` package

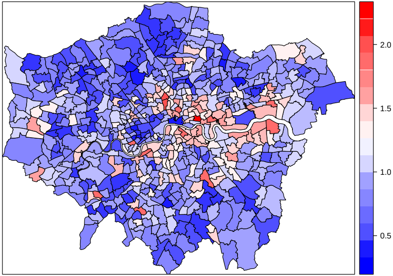
- `population`: vector of pop counts for each stratum in each area
- `cases`: vector of the observed number of cases in each area
- `n.strata`: number of strata considered

Caution: all counts are sorted by area first and then within each area the counts for all strata are listed (even if 0 count) in the same order

```
# Order the data set such as the strata are the same in each ward
lung <- lung[order(lung$POLY_ID, lung$SEX, lung$AGE_GROUP),]

# Expected numbers: n.strata = 2 genders * 22 age groups =44
newLung$E <- expected(population=lung$POPULATION, cases=lung$CASES,
                    n.strata=44)

# Check sum(O) = sum(E)
sum(newLung$E)
sum(newLung$O)

# Compute the SMRs
newLung$SMR <- newLung$O/newLung$E
```

Map of the SMRs

```
# Transform the spatial polygon file into dataframe
data.London <- attr(shpriver, "data")
# Merge the shapefile with the health data
attr(shpriver, "data") <- merge(data.London, newLung, by="STwardcode")

# Define the colors and plot of the SMRs
color.palette <-  colorRampPalette(c("blue", "white", "red"))
spplot(obj=shpriver, zcol= "SMR", col.regions=color.palette(20), asp=1)
```



Smoothing the RRs of lung cancer

- Check the model in a text file called model_HET.txt
- With R, format the data and the initial values in a list

```
# Just to be sure, order the data according to POLY_ID
newLung<-newLung[order(newLung$POLY_ID),]

# Format the data
data<-list(N=628,        # nb of areas
           O=newLung$O,  # observed nb of cases
           E=newLung$E)  # expected nb of cases

# Initialise the unknown parameters, 2 chains
inits <- list(
  list(alpha=0.01, prec.v=10, V=rep(0.01,times=628)), # chain 1
  list(alpha=0.5, prec.v=1, V=rep(0.05,times=628)))   # chain 2
```

bugs(): setting it

- R2OpenBUGS works through the main function bugs()
- You will need to set the MCMC and specify the list of parameters to be monitored

```
# MCMC setting

ni <- 3000  # nb iterations
nt <- 1     # thinning interval
nb <- 1000  # nb iterations as burn-in
nc <- 2     # nb chains


# Parameters monitored
parameters <- c("sigma2.v", "overallRR", "QR90", "RR", "resRR", "e")
```

bugs(): running it

- Now you can run the MCMC through `bugs()`

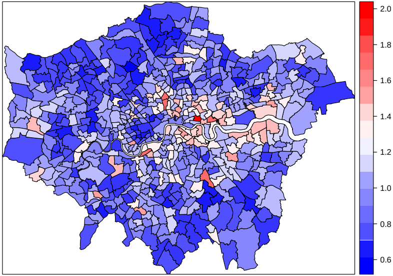
```
modelHET.sim <- bugs(data = data, parameters = parameters,
 inits = inits, model.file = "model_HET.txt",
 n.chains = nc, n.iter = ni, n.burnin = nb,
 n.thin = nt, debug = FALSE,
 working.directory = getwd(),
 codaPkg = FALSE, summary.only = FALSE,
 bugs.seed = 9)
```

- `codaPkg = FALSE` means that the programme will not save the coda for the parameters monitored.

Map of the smoothed RRs

- After convergence is reached (more on this in the practical)

```
# Attach bugs object
attach.bugs(modelHET.sim)
# Obtain posterior mean
RR_HET <- as.data.frame(apply(resRR,2,mean))
# Create an ID for each area
RR_HET$POLY_ID <- 1:628
colnames(RR_HET) <- c("HET","POLY_ID")
# Merge the ResRR with the original data
newLung<-merge(newLung, RR_HET, by="POLY_ID")

# Produce the map of resRR
attr(shpriver, "data") <- merge(data.London, newLung,
 by="STwardcode")
spplot(obj=shpriver, zcol= "HET",
   col.regions=color.palette(20),
       asp=1)
```



Useful Books

Banerjee, S., Carlin, B. P. and Gelfand, A. E. Hierarchical Modeling and Analysis for Spatial Data, 2nd ed. CRC press, 2014
(Chapters 1 and 6; for the latter, paragraph 6.4)

Lawson, A. Bayesian Disease Mapping, 3rd ed. CRC press, 2018
(Chapter 5)

R computing environment

Brunsdon, C. and Comber, L. An introduction to R for Spatial Data Analysis and Mapping, 2nd ed. Sage, 2018

Bayesian statistics

Lambert, B. A Student's Guide to Bayesian Statistics, Sage, 2018

Papers/books cited in this lecture

Cressie, N. (1993), *Statistics for Spatial Data*, John Wiley & Sons, Inc.

Monomier, N. (2005), "Lying with Maps," *Statistical Science*, 20, 215–222.

Oyana, T. J. and Morgai, F. (2015), *Spatial Analysis: Statistics, Visualization, and Computational Methods*, CRC press.

Schabenberger, O. and Gotway, C. A. (2004), *Statistical Methods for Spatial Data Analysis*, Chapman and Hall/CRC.

Tobler, W. R. (1970), "A computer movie simulating urban growth in the Detroit Region," *Economic Geography*, 46, 234–240.

Recap on distributions: Binomial distribution

- Convenient model for binary variables

- Distribution followed by a random variable (X, say) counting the number of successes out of n independent binary (yes / no) trials, where each trial has probability of success $= p$, $p \in [0, 1]$

If $X \sim \text{Binomial}(n, p), \quad p \in [0, 1]$

- $\Pr(x \text{ successes}) = \Pr(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$ for $x \in \{0, 1, \ldots, n\}$
- Mean of this distribution is $\mathrm{E}(X) = np$
- Variance is $\mathrm{Var}(X) = np(1-p)$

Recap on distributions: Poisson distribution

- Convenient model for counts (number of occurrences of an event across time or over an area)

- Assumes events occur independently and at constant rate

- Characterised by a single parameter, μ, representing the mean number of events in an interval of time or space ($\mu > 0$)

If $X \sim \text{Poisson}(\mu), \quad \mu > 0$

- $\Pr(x \text{ events in the interval}) = \Pr(X = x) = \frac{e^{-\mu}\mu^{x}}{x!}$ for $x \in \{0, 1, \ldots\}$
- Mean of this distribution is $\mathrm{E}(X) = \mu$
- Variance is $\mathrm{Var}(X) = \mu$

Recap on distributions: Normal or Gaussian distribution

- Convenient model for continuous variables

- Has symmetric ‘bell’ shape

- Characterised by two parameters: mean μ and variance σ^2

$X \sim \text{Normal}(\mu, \sigma^2)$

- $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x-\mu)^2}{2\sigma^2}$ for $x \in (-\infty, +\infty)$
- Mean of this distribution is $\mathrm{E}(X) = \mu$
- Variance is $\mathrm{Var}(X) = \sigma^2$

Recap on distributions: Gamma distribution

- Flexible distribution for positive quantities

$X \sim \text{Gamma}(a, b), \quad a > 0, b > 0$

- $p(x) = \frac{b}{\Gamma(a)} x^{a-1} e^{-bx}$ for $x > 0$
- Mean of this distribution is $\text{E}(X) = \frac{a}{b}$
- Variance is $\text{Var}(X) = \frac{a}{b^2}$

- If $Y \sim \text{Inverse Gamma}(a, b) \Leftrightarrow X = \frac{1}{Y} \sim \text{Gamma}(a, b)$