Hierarchical models in spatial Statistics

David V. Conesa Guillén

VAlencia BAyesian Research group

Universitat de València

1 Introduction to Spatial Statistical Models

Variability is all around

- Data obtained when performing any experiment show variability.
- Statistics is a discipline that has been developed in response to the experimenters whose data exhibit variability.
- A mere description of the data is useful, but proper science needs to know why we have observed those data and to do so we need to MODEL the situation.
- Once we have data collected, we use models with a structure like:
  - Response Variable to be explained.
  - A systematic component that contains the "general" information of the system under study and it is expressed as a combination of explanatory variables in the form of a parametric equation.
  - A random component reflecting the inherent variability in each particular situation.

Statistical models

- Depending on the type of variable, the explanatory are classified as:
  - Qualitative $\longrightarrow$ Factors (with their corresponding "levels")
    - ⋆ Fixed effects (if levels are previously fixed: Sex)
    - ⋆ Random effects (if the levels are a random sample of the possible levels: ship)
  - Quantitative $\longrightarrow$ Covariates: bathymetry, temperature, etc.
- Often the systematic component is expressed as a linear combination (but can also be non-linear).

Examples of Statistical models

- Linear models, as when we want to describe the age of a fish (RV) depending on its lenght and weight (covariates), that is,

$$\text{Age}_i \sim N(\mu_i, \sigma^2)\,;$$

$$\mu_i = \beta_0 + \beta_1 \text{lenght}_i + \beta_2 \text{weight}_i\,.$$

- Generalized linear models, when we want to describe the ocurrence (presence/absence) of a species or the number of species or the proportion of discards depending on some covariates.
- Generalized linear mixed models, when we also include a random effect in the model.
- Generalized additive models, when the relationship with the covariate is not linear.

Independence and autocorrelation

- All of previous models assume independence of the observations.
- But, as it happens in time series, sometimes some or all outcome measures exhibit autocorrelation.
- When the relative outcomes of two points are related to their distance, we found spatial autocorrelation ...
- Nearby observations tend to be more alike than those far apart.
- Tobler (1970) called this the first law of geography: everything depends on everything else, but closer things more.

Spatial variability

- We can find how to get rid of autocorrelation or ... much better ... to use models (no just mere descriptions of the spatial data) and methods that recognize the presence and importance of that spatial information.
- At the end of the day we want to answer the really important questions:
  - where things are,
  - in which amount can we find them, and also,
  - which is the probability of finding something in a particular location.

Not a new thing: John Snow mapped out the spread of a cholera outbreak in London back in 1854 by relating with the location of pumps

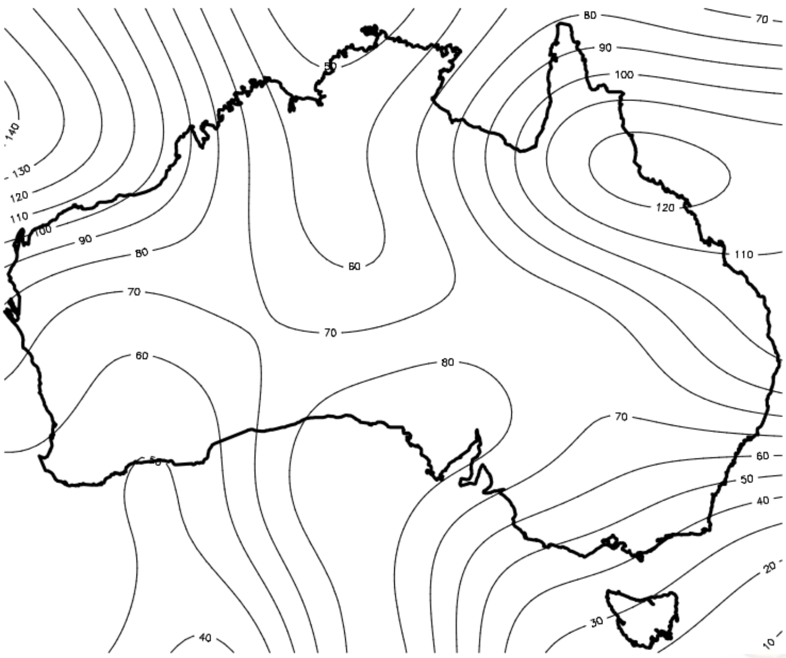
Types of spatial data

- Three important prototypes:
  1. Geostatistical data
  2. Lattice/areal/regional data
  3. Spatial point patterns
- Determining the prototype is not always an easy task.

2 Types of spatial data

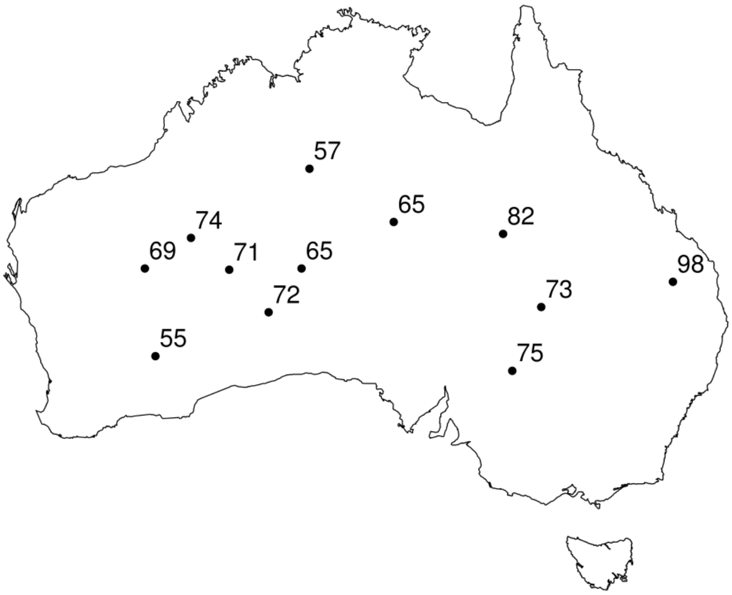
Geoestatistical data

The quantity of interest has a value at any location, ...



Geoestatistical data (2)

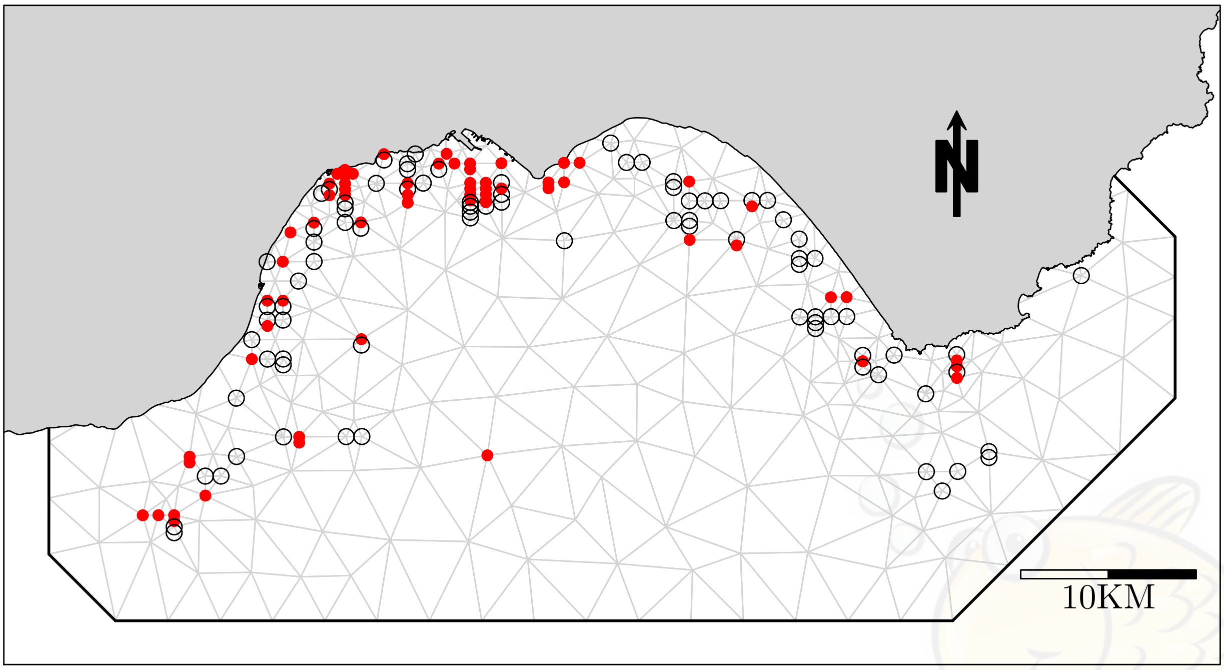
... but we only measure the quantity at certain sites. These values are our data.



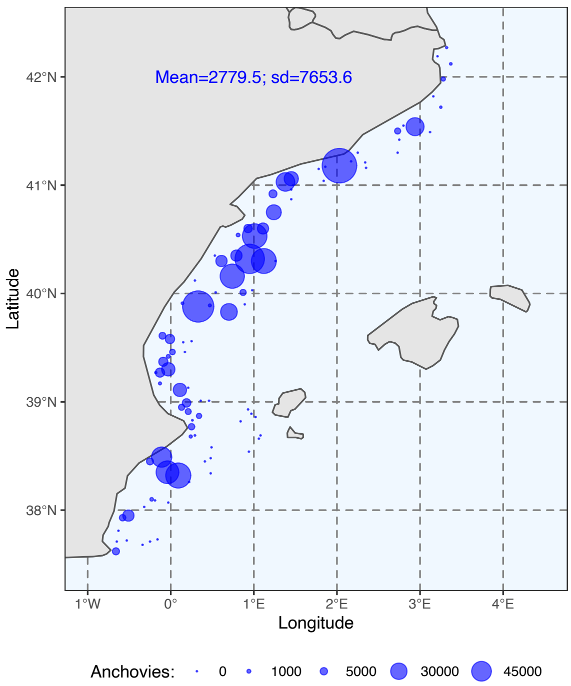
Geoestatistical data (3)

- In other words, point observations of a continuously varying quantity over a region.
- Final aim is to obtain (predict) the value of interest at any location.
- Examples:
  - A famous problem was to predict the ore grade in a mining block.
  - Weekly concentrations of ozone in Europe; annual acid rain deposition in Europe.
  - Abundance of a species; Prevalence of a disease.

Ocurrence of Mackerel in Gulf of Almería

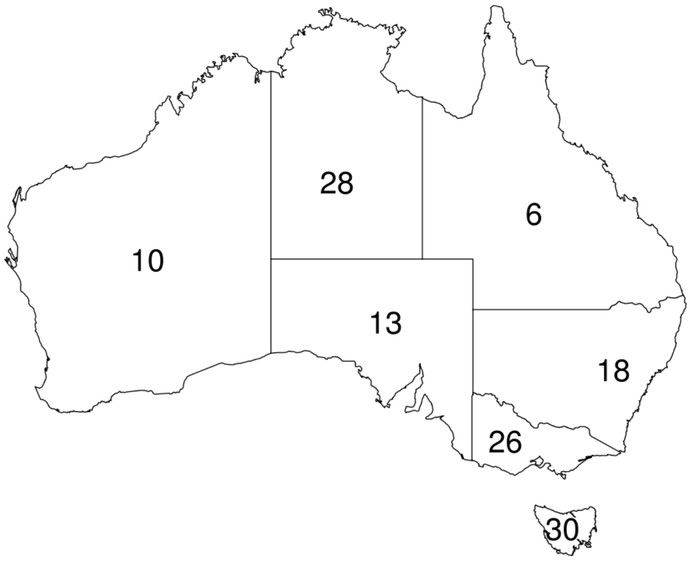


Fish species: captures of anchovies in the Mediterranean



Areal data

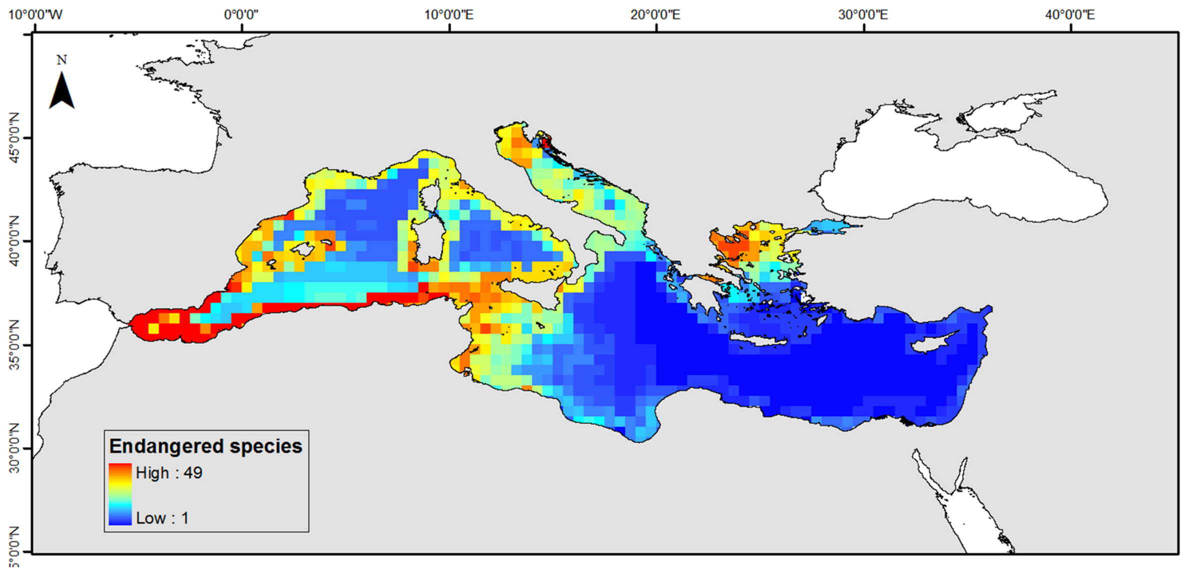
The quantity of interest is only defined for regions/municipalities/census districts .... It is measured/reported for certain fixed regions.



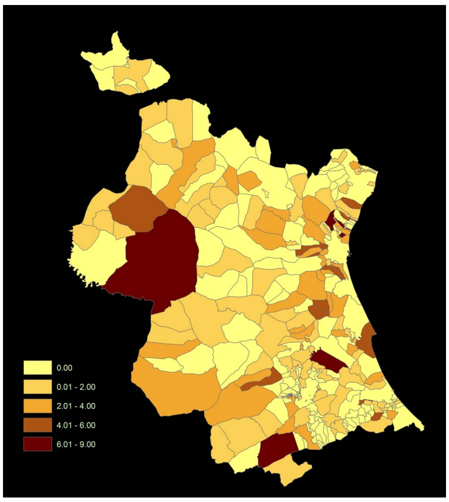
Regional data (2)

- Data are usually counts, averages or proportions of a quantity on subregions that make up a larger region.
- A lattice of locations evokes an idea of regularly spaced points. These will be referred to as regular lattices, allowing for the possibility of irregular lattices, whose relative displacements do not follow a predictable pattern.
- Statistical models for lattice data need to express the fact that observations nearby (in time or space) tend to be alike.
- Examples:
  - Presence or absence of a plant species in square quadrats over a study area
  - Number of deaths due to a disease in the provinces of a country
  - Number of different species in each quadrat of a regular grid

Species Richness in the Mediterranean Sea



Disease mapping



Prostate cancer mortality (aggregated from 1975-1980). Colours indicate intensity of the characteristic studied.

Point Pattern data

The main interest is in the locations of all occurrences of some event (e.g. tree deaths, meteorite impacts, observation of a whale). Exact locations are recorded.

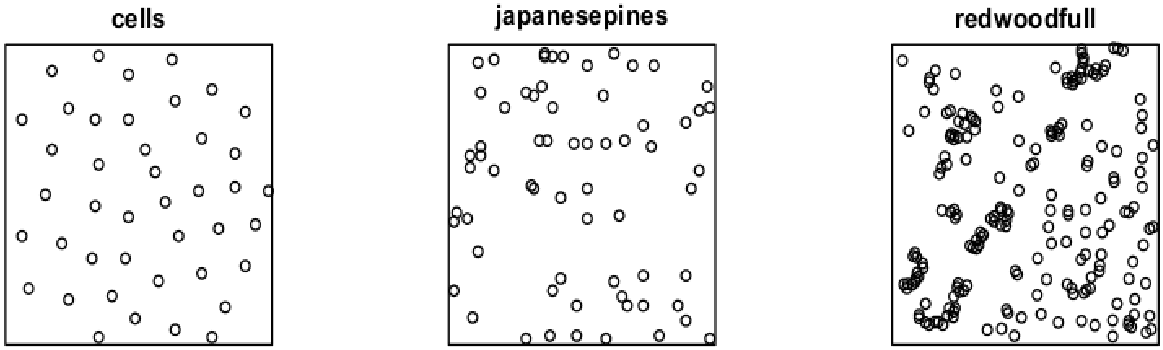
Examples:

- Location of bird's nests in a suitable habitat – evidence of territoriality?
- Location of longleaf pines in a natural forest – evidence of clustering?

Cholera in London

Point Pattern data (2)

The question of interest is whether the pattern is exhibiting complete spatial randomness, clustering, or regularity.
These maps show differences in the behaviour (different patterns)

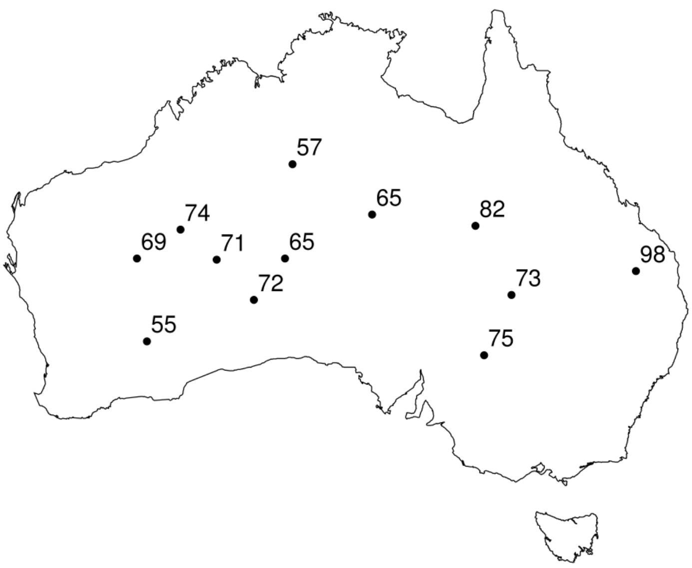


Are we able to guess which pattern?

Left: Regularity. Middle: complete spatial randomness. Right: clustering.

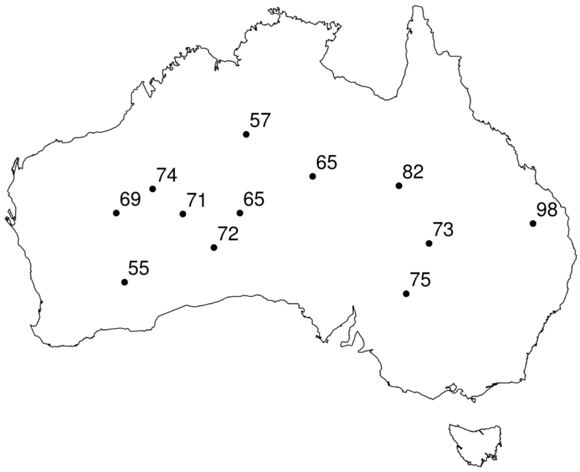
Points with marks

Points may also carry data (e.g. tree heights, meteorite composition, amount of the robberie)



Point pattern or geostatistical data?

Graphs in pages 13 and 26 are (artificially) the same, ...



... but do they show the same type of data?

Point pattern or geostatistical data?

- Response Variable is the quantity that we want to be predicted or explained
- Explanatory variable is the quantity that can be used to predict or explain the response
- Geostatistics treats the spatial locations as explanatory variables and the values attached to them as response variables.
  "Temperature is increasing as we move from South to North"
- Spatial point pattern statistics treats the spatial locations, and the values attached to them, as the response.
  "Trees become less abundant as we move from South to North"

Task

For the following situations find the type of spatial data involved. It would help you:

- Identify the variable of interest. Is it the spatial location?
- If so, do we have marks in each location?
- Identify the explanatory variables.
- Is the response explained by the spatial location?
- Does the quantity of interest vary continously along the space?
- Does the quantity of interest vary in certain fixed regions or subregions?

Task

1. A study of the noise level in Barcelona shows observations in different places of the city, with the aim of predicting the noise level in other parts of the city.
2. The spur-thighed tortoise is a protected species that lives mainly in the provinces of Almería and Murcia. A study carried out in the Sierra de Almenara has provided information of the sighting of specimens and the location of each one of they.
3. Numerous medical studies have confirmed that most malignant lung tumors are caused by tobacco. What's more, other factors may condition the risk of developing this Cancer. To know the geographic pattern of these other factors of risk, the number of cases detected in each municipality of the Valencian Community since 2000 are observed.
4. After the floods in Valencia, an ecologist group sought to analyze the lead content in the waters of the Albufera. For this, water samples were taken at 200 points distributed throughout the extension following a regular grid. That is, locations were separated every 100 meters in both coordinates. At each location, a bottle of water was collected and analyzed to assess the level of lead present.
5. A group of people went to sightseeing around Canary Islands and they were able to watch whales. They have provide us the exact location of the encounters.

3 Hierarchical Bayesian Modelling of Spatial data

How can we analyse spatial data?

- First of all: understanding the model behind each prototype of data.
- But the three prototypes can be seen under the same hierarchical structure.
- Hierarchical Statistical Modeling
  - Data model: distribution of the data given a hidden process
  - Process model: uncertainty in the true (but hidden) process through a probability distribution of the phenomenon of interest
  - Characterized by conditional probability distributions
- Once the model is set you need to find your way to make inference on the parameters involved: Bayesian.

Bayesian inference

- Inference and prediction about the unknown parameters usually done with the frequentist (a.k.a. classical) approach.
- Another option is to estimate and predict using Bayesian statistics:
  - Based on the fact that information and uncertainty about all the unknown can be better (and easily) expressed in terms of probability distributions.
  - Probability is then used (and understood) as a "subjective" measure of the uncertainty.
  - Consider as unknown elements of the problem not only data but also the parameters of the distribution that govern their behavior.
  - Data are used to update knowledge about the parameters.
  - Inference is made in terms of (posterior) probability distributions:

$\boxed{\text{Prior Info}} + \boxed{\text{Data}} \longrightarrow \boxed{\text{Posterior}}$

Why Bayesian is so popular now?

- The Bayesian point of view is not a technique in the field of Statistics.
- It is another way of understanding and performing Statistics.
- And so, when our data bring us to a any model ...
- ... we can solve it using both Bayesian and Frequentist methods.
- Bayesian statistical analysis has benefited from the explosion of cheap and powerful desktop computing over the last two decades: MCMC, INLA.
- Bayesian techniques can now be applied to complex modeling problems where they could not have been applied previously. SO FAMOUS IN MANY APPLIED FIELDS!!
- Bayesian perspective will probably continue to challenge, and perhaps sub-plant, traditional frequentist statistical methods which have dominated many disciplines of science for so long.

The Bayesian learning process

- Construction of the joint posterior distribution of unknowns:
  $l(\theta; x) = p(\mathrm{x}|\boldsymbol{\theta})$ is the likelihood,
  $p(\boldsymbol{\theta})$ is the prior distribution,

$$p(\mathrm{x}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}) p(\mathrm{x}|\boldsymbol{\theta}) \,.$$

- Obtaining the posterior distribution via Bayes theorem:

$$p(\boldsymbol{\theta}|\mathrm{x}) = \frac{p(\mathrm{x}, \boldsymbol{\theta})}{p(\mathrm{x})} = \frac{p(\boldsymbol{\theta}) p(\mathrm{x}|\boldsymbol{\theta})}{p(\mathrm{x})} = \frac{p(\boldsymbol{\theta}) p(\mathrm{x}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}) p(\mathrm{x}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}$$

  As $p(\mathrm{x})$ does not depend of $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\mathrm{x}) \propto p(\boldsymbol{\theta}) \times p(\mathrm{x}|\boldsymbol{\theta}) \,.$$

Example: Interest of estimating π, the proportion of COVID

- Our model: $Y \sim \mathrm{Ber}(\pi)$
- Data $Y_i \sim \mathrm{Ber}(\pi), i = 1, \ldots, n$
- Observed data $\mathrm{y} = \{y_i; i = 1, \ldots, n\}$
- Likelihood: $p(\mathrm{x}|\pi) = \ell(\pi) = \binom{n}{r} \pi^r (1-\pi)^{n-r}$
- A Beta Conjugate Prior distribution with parameters a and b:

$$p(\pi) \propto \theta^{a-1} (1-\theta)^{b-1}$$

- Posterior distribution:

$$\begin{aligned} p(\pi|\mathrm{y}) &\propto p(\mathrm{y}|\pi) \times p(\pi) \propto \pi^{r+a-1} (1-\pi)^{n-r+b-1} \\ \pi|\mathrm{y} &\sim \mathrm{Beta}(r+a, n-r+b) \end{aligned}$$

- The posterior predictive distribution that predicts (knowing what has happened before) the number of persons with COVID, Z, out of m is:

$$\begin{aligned} p(Z|\mathrm{y}) &= \int \mathrm{Binomial}(m, \pi) \mathrm{Beta}(r+a, n-r+b) \mathrm{d}\pi \\ &= \text{Beta-Binomial}(m, r+a, n-r+b) \end{aligned}$$

Estimating the proportion of COVID depending on age

- Our previous simple model only had one parameter. Life is usually more complex.
- Let's now include a dependence with a covariate (AGE): Logistic regression (GLM).
- Model in two pieces (being X_i, the age of person i):

$$
\begin{aligned}
Y_i &\sim \text{Ber}(\pi_i), \quad \forall i = 1, \ldots, n \\
\text{logit}(\pi_i) &= \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i\,, \quad \forall i = 1, \ldots, n \\
\pi_i &= \text{logit}^{-1}(\beta_0 + \beta_1 X_i) = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}\,, \quad \forall i = 1, \ldots, n
\end{aligned}
$$

- Likelihood:

$$
\begin{aligned}
\ell(\beta_0, \beta_1) = p(\mathrm{x}, \mathrm{y}|\beta_0, \beta_1) &= \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{y_i} \left(1 - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{1-y_i}.
\end{aligned}
$$

- Classical estimation not easy: but so easy to use `glm` function in `R`. Do we really know what happens behind?

Estimating the proportion of COVID depending on age (2)

- To make inference in a GLM within the Bayesian framework, we first assign priors to the parameters:
  - independent Gaussian prior for the coefficients of the regressors (centered at zero with large variance), e.g. N(0,10000), that is, $p(\beta_0, \beta_1) = N(\beta_0|0, 10000) \times N(\beta_1|0, 10000)$, o
  - Improper flat prior $p(\beta_0, \beta_1) \propto 1 \times 1 \propto 1$.
- The posterior distribution does not have an analytical expression, neither the predictives:

$$
\begin{aligned}
p(\beta_0, \beta_1|\mathrm{y}, \mathrm{x}) &\propto p(\beta_0, \beta_1) \times \ell(\beta_0, \beta_1) \\
&\propto 1 \times \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{r} \left(1 - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{n-r}.
\end{aligned}
$$

- Numerical tools are needed.
- But life is even more complicated, as we would like to have more covariates and our friends the random effects.

Estimating proportion of COVID depending on age and city

- Imagine now that we have data from a random sample of c cities in which we still analyze whether having COVID is related to age (GLMM).
- Model in three pieces (being X_i, the age of person i):

$$\begin{aligned} Y_i &\sim \mathrm{Ber}(\pi_i), \quad \forall i = 1, \ldots, n \\ \mathrm{logit}(\pi_i) &= \beta_0 + \beta_1 X_i + \mathrm{City}_{j(i)}, \quad \forall i = 1, \ldots, n; \forall j = 1, \ldots, c \\ \mathrm{City}_{j(i)} &\sim N(0, \sigma_C^2) \end{aligned}$$

- To make inference in a GLMM within the Bayesian framework, we first assign priors to the parameters and hyperparameters:

$$p(\beta_0, \beta_1, \sigma_C^2 | \mathrm{y}, \mathrm{x}) \propto p(\beta_0, \beta_1, \sigma_C^2) \times \ell(\beta_0, \beta_1, \sigma_C^2)$$

- The posterior distribution does not have an analytical expression, neither the predictives, and again Numerical tools are needed.

Hierarchical Bayesian modelling

- In many areas such as epidemiology, climatology, real estate marketing, and ecology, data are:
  - highly multivariate, with many important predictors and response variables,
  - geographically referenced, and often presented as maps.
- A good way to to analyse complex (and large spatial) data sets is by means of Hierarchical modelling.
- Though analysis of hierarchical modelling can be attempted through nonBayesian approaches, the Bayesian paradigm can be very convenient to tackle uncertainty assessment within the given spatial specification.

Numerical approaches for Bayesian statistics

- Bayesian statistics do not come for free.
- Most of the complications that appear in the Bayesian methodology come from the resolution of integrals that appear when applying the learning process:
  - the normalization constant of the posterior distribution,
  - moments and quantiles of the posterior,
  - credible regions, probabilities in the contrasts, etc.
- Most two popular solutions:
  - Monte Carlo Markov Chain (MCMC) methods;
  - Integrated Nested Laplace Approximation (INLA).

Integrated Nested Laplace Approximation: INLA.

- Although in theory MCMC is always possible to implement, this approximation has some backwards:
  - in terms of both convergence and computational time and
  - the implementation itself (that can be very tricky for those non experts in programming).
- INLA appears as an option to perform a numerical approximation much faster for a particular kind of models, latent Gaussian models (LGM).
- By its own construction (hierarchical structure in three stages), latent Gaussian models are a class of hierarchical Bayesian models.
- We only need then to work with latent Gaussian models and the good thing is that all spatial models are.

Model comparison within INLA.

Model comparison in terms of best fit (DIC and WAIC) and best one predicting (LCPO):

- DIC: *Deviance Information Criterion* for comparing hierarchical models. Penalizes models that are very complex (Spiegelhalter et al., 2002).
- WAIC: *Watanabe Information Criterion* for comparing hierarchical models. It also penalizes complexity (Watanabe, 2010).
- LCPO: Logarithm of the *Conditional Predictive Ordinate* as a predictive measure of the models (Gneiting and Raftery, 2007).

4 Hierarchical Modelling of Spatial data

Spatial models more detailed (Cressie, 2015)

- Spatially continuous (Geostatistical or point-referenced) Data
  - R is a fixed subset of the plane of positive area (2-D) or volume (3-D).
  - $Y(s)$ is a random variable at each of the infinite continuous locations $s \in R$.
- Area (Lattice) Data
  - $R = \{s_1, s_2, \ldots, s_n\}$ is a fixed regular or irregular lattice on the plane.
  - $Y(s_i)$ is a random variable at each of location s_i, $i = 1, \ldots, n$.
- Spatial Point Process Data
  - $R = \{s_1, s_2, \ldots, s_n\}$ is a random collection of points on the plane
  - $Y(s)$ is not specified or it is a random variable at a location $s \in R$ (marked point process).

Hierarchical Modelling of Geostatistical data

- Data Model:

$$Y(s_i) \sim Distribution(\theta_i)$$

  where the distribution of response variable can be Normal, Poisson, Binomial, etc.
- Process model:

$$g(\mu_i) = \mathrm{X}_i\boldsymbol{\beta} + u_i\,,\ i = 1, \ldots, n$$

  - g is the link function
  - $\mathrm{X}(s_i)$ known covariates
  - $\boldsymbol{\beta}$ fixed effects parameters
  - U is a latent Gaussian Field approximated by a Gaussian Markov random field zero-centered with covariance Matérn structure that depends on precision τ_w and the range κ: $W \sim N(0, \mathbf{Q}(\kappa, \tau_w))$.
- The classical way of analyzing this kind of data is via Kriging that provides an estimator of the expected value at several locations based on the estimated semivariogram.

Hierarchical modeling of Areal data

- Spatial Markovian assumption: value of the random variable at a given site only depends on the values at a specified set of neighboring sites.
- Data Model

$$Y(s_i) \sim Distribution(\theta_i)$$

where the distribution of response variable can be Normal, Poisson, Binomial, etc.
- Process model

$$g(\theta_i) = \mathrm{X}(s_i)\boldsymbol{\beta} + u_i$$

with:
  - g the link function
  - $\mathrm{X}(s_i)$ known covariates
  - $\boldsymbol{\beta}$ fixed effects parameters
  - u has an ICAR (Intrinsic Conditional Auto-Regressive) distribution

$$[u_i|u_j, i \neq j, \tau_u] \sim N\left(\bar{u}_i = \frac{1}{\sum_j w_{ij}} \sum_j u_j w_{ij}, \tau_i = \frac{1}{\tau_u \times \sum_j w_{ij}}\right)$$

    where $w_{ij} = 1$ if adjacent (0 otherwise).
  - Predictor can incorporate smooth nonlinear effects of covariates, time trends, seasonal effects, random intercept and slopes and temporal random effects.

Hierarchical modelling of Point Patterns

- Data Model: data Y_i are a realization of an inhomogeneous Poisson point process with intensity λ_i.
- Process model:

$$\log(\lambda_i) = \mathrm{X}_i\boldsymbol{\beta} + u_i, i = 1, \ldots, n$$

with:
  - X_i known covariates at each location i,
  - $\boldsymbol{\beta}$ fixed effects parameters
  - U is a stationary and isotropic Gaussian Latent Field with a Matérn covariance with parameters (κ, τ).

5 References

References

1. S. Banerjee, B.P. Carlin, and A.E. Gelfand (2014). Hierarchical modeling of analysis for spatial data, CRC Press. 2nd edition.
2. R.S. Bivand, E.J. Pebesma, and V. Gómez-Rubio (2013). Applied Spatial Data Analysis with R, Springer. 2nd Edition.
3. N. Cressie (2015). Statistics for Spatial Data (revised edition).
4. T. Gneiting and A.E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* 102 (477), 359–378.
5. F. Lindgren, H. Rue and J. Lindstrom, (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the SPDE approach (with discussion). *Journal of the Royal Statistical Society, Series B*, 73: 423-498.
6. H. Rue, S. Martino and N. Chopin, 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, 71(2): 319-392.
7. D.J. Spiegelhalter, N.G. Best, B.P. Carlin and A. van der Linde (2002). Bayesian measures of model complesity and fit. *Journal of the Royal Statistical Society: Series B*, 64: 583-616.
8. S. Watanabe (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, 11, 3571–3594.