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

EXTENSIONS

We will see the following extensions:

- BYM2 model
- Point processes
- Leaflet (we will see in the practicals)
- pc priors (we will see in the practicals)

EXTENSIONS – BYM2 MODEL

- In the classical model of Besag, York and Mollie, BYM (Besag et *al.*, 1991), spatially structured variation is not independent of unstructured variation (so-called non-identifiability problem).
- As a consequence, part of the spatial dependence (structured variation) could actually be heterogeneity (unstructured variation) and vice versa.

EXTENSIONS – BYM2 MODEL

- There are alternative formulations to the BYM model, such as the Leroux (Leroux et *al.*, 2000) and Dean (Dean et *al.*, 2001) models, which ensure that structured spatial variation is independent of unstructured variation.
- However, neither model 'scales' spatial variation.
- As a consequence, the hyperparameters depend on the spatial structure of the problem and cannot be interpreted correctly.

EXTENSIONS – BYM2 MODEL

- On the other hand, in the Bayesian context, the choice of the a priori distributions of the hyperparameters (priors) can have a considerable impact on the results.
- In the Leroux and Dean models, standard priors are used which lead to overfitting (collinearity).
- The main consequence of overfitting is that the variance estimators are larger than the actual ones and, therefore, the credibility intervals will be much wider than expected, which implies that we could end up not rejecting the null hypothesis (that the coefficients are equal to zero) when in fact we should have rejected it.

EXTENSIONS – BYM2 MODEL

- Simpson et *al.* (2017) proposed a modification of the BYM model, called BYM2, which solves these problems, as it scales spatially structured variation and uses priors that penalize complexity (called PC priors).
- These priors are robust, in the sense that they have no impact on the results and, in addition, they have an epidemiological interpretation.

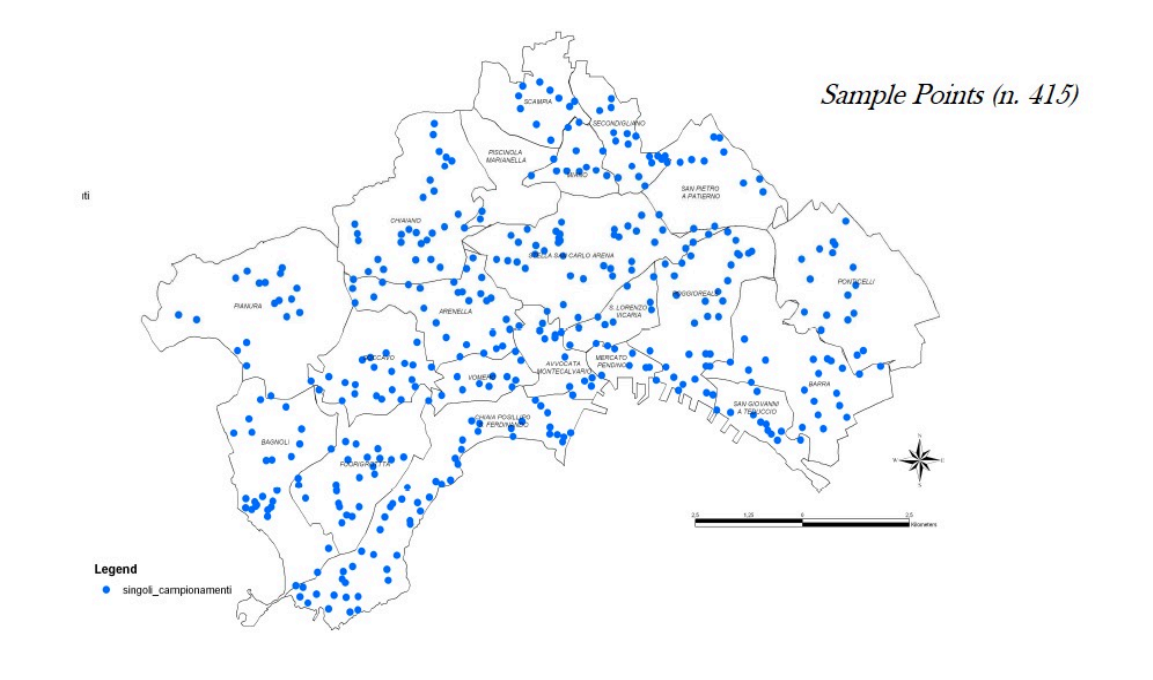
EXTENSIONS – POINT PROCESSES

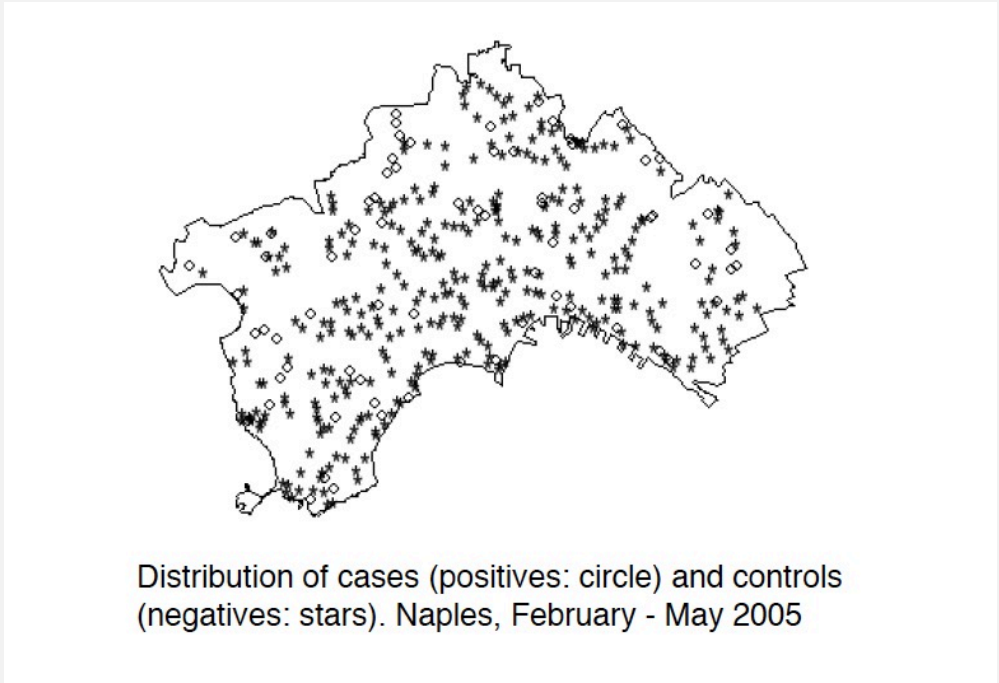
- ***Point processes*** correspond to Bernouilli random variables.
- In point processes, ***the exact location of the case is known and this is random***.

EXTENSIONS – POINT PROCESSES

- In this case, we may be interested in:
  - Assess whether the events follow a particular spatial pattern (aggregation, regular shape, etc.).
  - To study whether an observed pattern is associated with some variable (exposure to some environmental variable, such as atmospheric pollution; proximity to polluting sources; socioeconomic context, etc.).

EXTENSIONS – POINT PROCESSES





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

EXTENSIONS – POINT PROCESSES – LOG-GAUSSIAN COX PROCESS (LGCP)

- The LGCP model is the analog to the Gaussian linear model used for geostatistical data when data is modeled in the form of point processes.
- However, Diggle et *al.* (2013) propose to use this model to approximate the fit of spatial data of any type (that is, areal data, geostatistical data, and point processes data).

EXTENSIONS – POINT PROCESSES – LOG-GAUSSIAN COX PROCESS (LGCP)

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

- In 2012, Diggle proposed a paradigm shift and redefined spatial statistics as **'a set of statistical models and methods that aim to help scientists understand spatial phenomena, which cannot be observed directly, but only indirectly, with incomplete information'**, in the form of grid data, point processes and geostatistical data.
- The statistical model that he proposed, as the only core model, is the **log Cox model**.

EXTENSIONS – POINT PROCESSES – LOG-GAUSSIAN COX PROCESS (LGCP)

- That is, Diggle unified spatial statistics, just as McCullagh and Nelder (1989) unified generalized linear models.

EXTENSIONS – POINT PROCESSES – LOG-GAUSSIAN COX PROCESS (LGCP)

- Conditional on the true risk at location x_i, the probability of a case ocurring, $P(x_i)$, $i = 1, \ldots, n$, at this location is distributed as a binomial.

$$Y_i|P(x_i) \sim Binomial\big(E_i, P(x_i)\big)$$

- The model is (the link function can be another …)

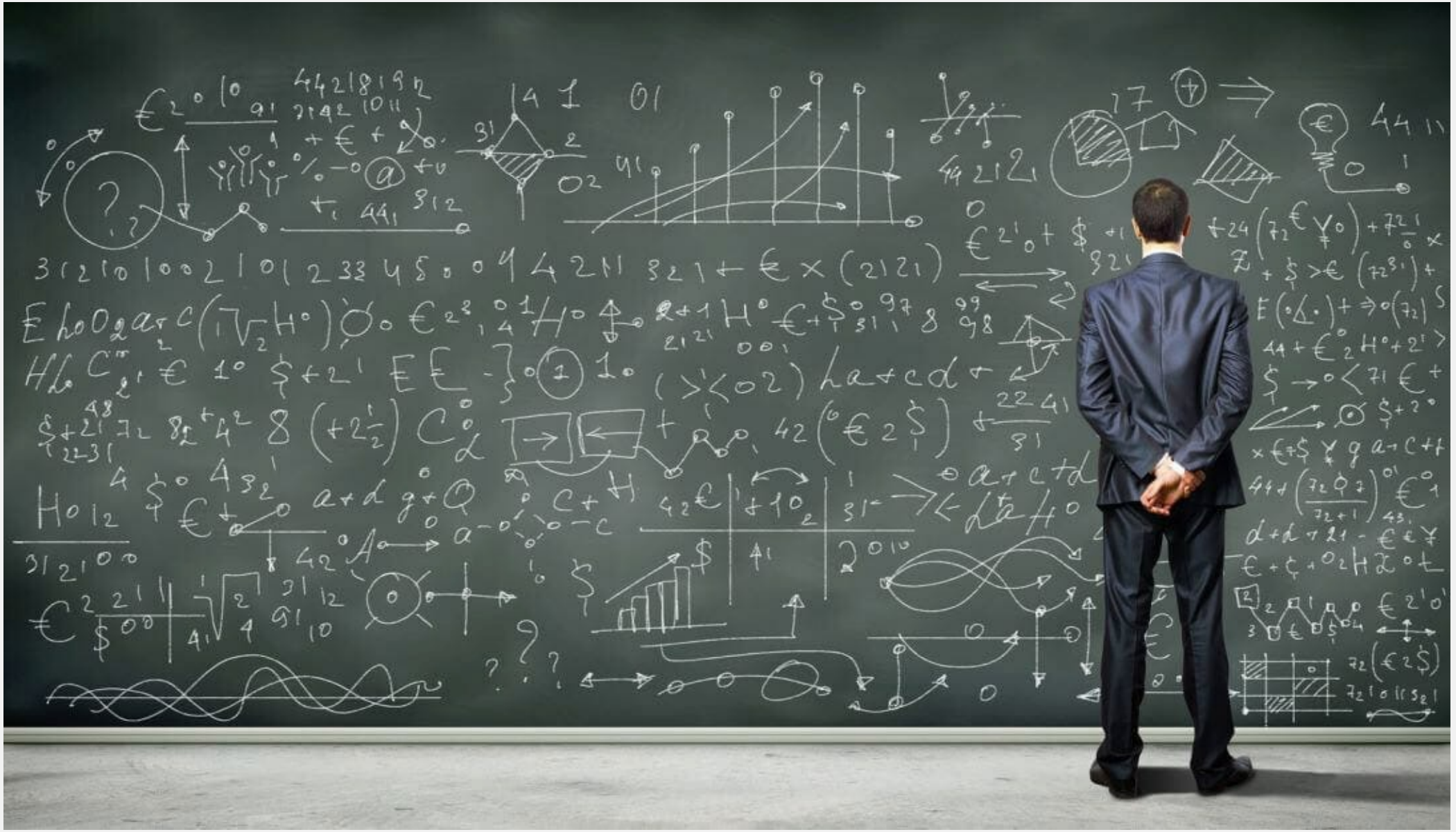
$$\log \mu = \beta_0 + \eta_{ji} + S\big(x_{ji}\big) + \tau_{ji} + offset(E_i)$$

$$Cov\Big(S(x_i), S(x_{i'})\Big) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\kappa \|x_i - x_{i'}\|\right)^{\nu} \mathrm{K}_{\nu} \left(\kappa \|x_i - x_{i'}\|\right)$$

GRECS
GRUP DE RECERCA EN ESTADÍSTICA ECONOMETRIA I SALUT

UdG

ciberesp
Centro de Investigación Biomédica en Red
Epidemiología y Salud Pública




I CAN... BUT I WON'T