

# Approximate Bayesian inference for the analysis of population health data

Virgilio Gómez Rubio\*

Universidad de Castilla-La Mancha

\*this is joint work with quite a few people (including H. López-Gómez and G. García-Donato Layrón)



[Virgilio.Gomez@uclm.es](mailto:Virgilio.Gomez@uclm.es)



@precariobecario



<https://becarioprecario.github.io>



Castilla-La Mancha



GENERALITAT  
VALENCIANA

Una manera  
de hacer Europa

Fondo Europeo de  
Desarrollo Regional



Unión Europea



GOBIERNO  
DE ESPAÑA

MINISTERIO  
DE CIENCIA, INNOVACIÓN  
Y UNIVERSIDADES



AGENCIA  
ESTATAL DE  
INVESTIGACIÓN

# Introduction

- The integrated nested Laplace approximation (INLA).
- Population-based analysis of ischemic stroke in Poland.
- Limitations of INLA.
- “INLA con cosas”: using INLA to fit a wider range of models.
- Models for changepoint detection.
  - Example: Yearly coal mining disasters in the U.K.
- Analysis of health exposure using case-control point patterns.
  - Example: Exposure to pollution sources in Alcalá de Henares (Madrid).

# INLA (in one slide)

- INLA stands for “integrated nested Laplace approximation”.
- Numerical approach for Bayesian inference.
- INLA is faster than typical Markov chain Monte Carlo (MCMC).
- For a model with I vectors of latent effects  $x$  and hyperparameters  $\theta$ , INLA provides estimates of the *marginal* posterior distributions:
  - $\pi(\theta. | y)$ .
  - $\pi(x. | y)$ .
- INLA can provide estimates of other quantities of interest:
  - Marginal likelihood:  $\pi(y)$ .
  - Model selection criterion: DIC, WAIC, etc.

# Population-based analysis of ischemic stroke

- Dataset about 500,000 inhabitants published by Polish National Health Fund.
  - Demographic information.
  - Date of ischemic stroke.
  - Administrative region (379 *powiat*-level entities)
  - Drug prescriptions.
- Smetkowski et al. (2015) propose a deprivation index:
  - Income.
  - Employment.
  - Living conditions.
  - Education.
  - Access to goods and services.

# Models

- Logistic regression:
  - To assess prevalence of stroke.
  - Outcome variable: evento or no event (binary).
  - The logit function is the canonical link function for the Bernoulli distribution.
- Survival model:
  - To study time until a stroke occurs.
  - Outcome variable: time-to-evento.
  - The Weibull distribution is commonly used in Accelerated Failure Time models.
- Models fit with INLA on a Linux cluster (~15 minutes to fit).

# Logistic regression models

- Let  $p_{ij}$  be the probability that individual  $i$  living in region  $j$  will suffer an ischemic stroke:

$$\text{logit}(p_{ij}) = x'_{ij}\beta + \gamma_j$$

- $\beta$  is the regression coefficient vector.
- $x_{ij}$  is the vector of covariates.
- $\gamma_j$  are the random effects associated with region  $j$ .

# Survival models

- Let  $T_{ij}$  be the time that individual  $i$  living in region  $j$  suffers an ischemic stroke since entering the study:

$$\log(T_{ij}) = x'_{ij}\beta + \gamma_j + \sigma\varepsilon_{ij}$$

- $\beta$  is the regression coefficient vector.
- $x_{ij}$  is the vector of covariates.
- $\gamma_j$  are the random effects associated with region  $j$ .
- $\sigma$  is a scale parameter.
- $\varepsilon_{ij}$  are i.i.d. random variables with a standard Gumbel distribution.

# Fixed and random effects

- Covariates:
  - Gender.
  - Age.
  - Type of region (city indicator).
- Spatial random effects are using a Leroux specification, with the following precision matrix:

$$(1 - \phi)I + \phi Q$$

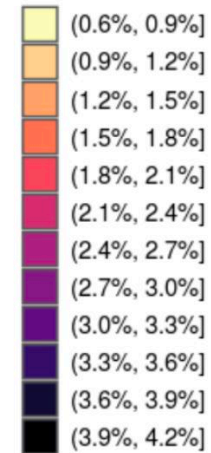
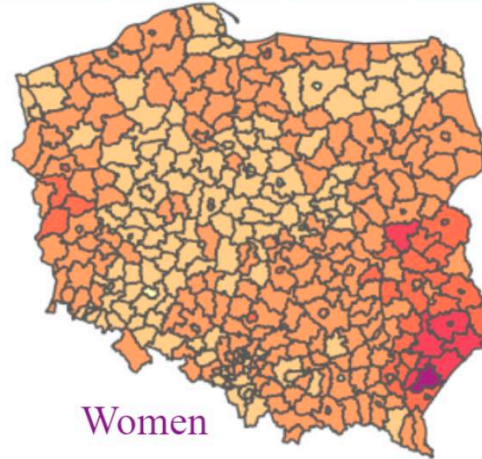
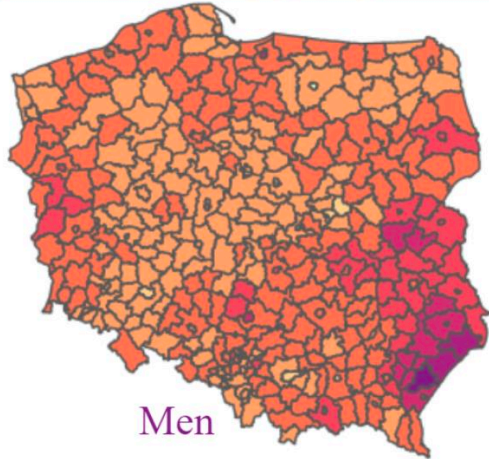
- $I$  is the identity matrix.
- $\phi \in [0, 1]$  is a weight parameter.
- $Q$  is the precision matrix of an ICAR specification.

# Results

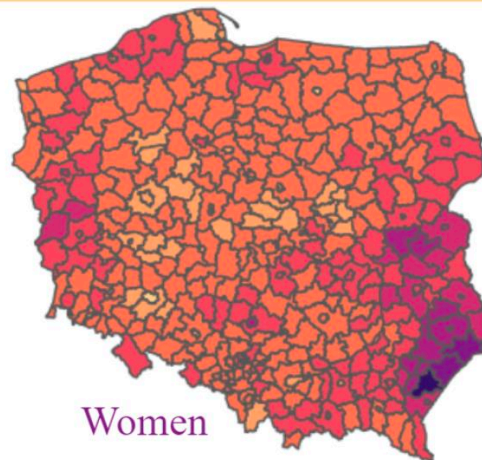
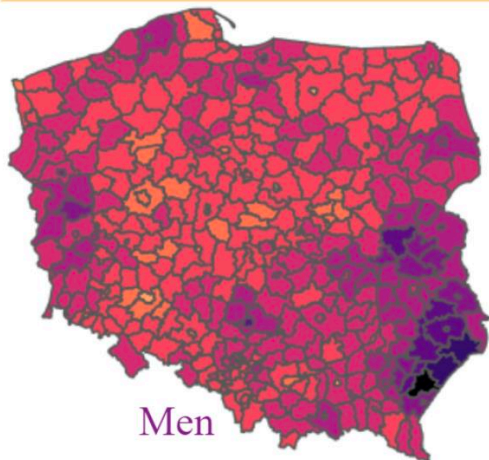
covariable		LOGIT	SURVIVAL
intercept	mean	-5.925	-5.925
	CI	(-6.042, -5.811)	(-6.041, -5.811)
Woman	mean	-0.217	-0.214
	CI	(-0.291, -0.142)	(-0.288, -0.14)
Group Age2 (58-68 y.o.)	mean	0.933	0.931
	CI	(0.81, 1.057)	(0.809, 1.055)
Group Age3 (>68 y.o.)	mean	1.729	1.722
	CI	(1.612, 1.847)	(1.605, 1.839)
City county	mean	0.07	0.07
	CI	(-0.104, 0.242)	(-0.1, 0.24)
Diabetes TRUE	mean	0.238	0.235
	CI	(0.149, 0.326)	(0.147, 0.322)
Antithrombotics TRUE	mean	0.236	0.234
	CI	(0.141, 0.329)	(0.14, 0.326)
Cardiovascular TRUE	mean	0.325	0.323
	CI	(0.224, 0.426)	(0.223, 0.424)
Deprivation index	mean	0.128	0.129
	CI	(0.012, 0.243)	(0.015, 0.242)

# Results

## WITHOUT CARDIOVASCULAR TREATMENT



## WITH CARDIOVASCULAR TREATMENT



# Limitations of INLA

- Certain models do not fall within the latent GMRF framework:
  - Mixture models.
  - Double-hierarchical models.
  - Zero-inflated models.
- Also, models not implemented in the R-INLA package may be difficult to fit, i.e., models with a particular structure of the latent effects.
- INLA focus is on making marginal inference (but sampling from the approximate posterior can be done).

# Conditional latent GMRF

- However, sometimes conditioning on some of the hyperparameters the model can become a *conditional* latent GMRF.
- Let's take  $\theta = (\theta_1^*, \theta_2^*)$  so that the model can be fit with INLA conditional on  $\theta_2^*$ .
- Hence, INLA will estimate:
  - *Conditional* posterior marginals:  $\pi(\theta_{1,\cdot}^* | y, \theta_2^*)$  and  $\pi(x_{\cdot} | y, \theta_2^*)$ .
  - *Conditional* marginal likelihood:  $\pi(y | \theta_2^*)$ .

# Estimation of the marginals

- Note that:

$\pi(\theta_2^*|y) \propto \pi(y|\theta_2^*)\pi(\theta_2^*)$ ;  $\pi(\theta_2^*)$  is the prior of  $\theta_2^*$ , which is known.

- The marginals of  $\theta_1^*$  can be estimated by integrating over  $\theta_2^*$ :

$$\pi(\theta_{1,\cdot}^*|y) = \int \pi(\theta_{1,\cdot}^*|y, \theta_2^*)\pi(\theta_2^*|y) d\theta_2^* \simeq \sum_{\theta_2^* \in \Omega} \pi(\theta_{1,\cdot}^*|y, \theta_2^*)w_{\theta_2^*}$$

where  $\Omega$  defines a set of integration points with associated weights:

$$w_{\theta_2^*} = \frac{\pi(y|\theta_2^*)\pi(\theta_2^*)}{\sum_{\theta_2^* \in \Omega} \pi(y|\theta_2^*)\pi(\theta_2^*)}$$

# “INLA con cosas”



Source: Wikipedia.

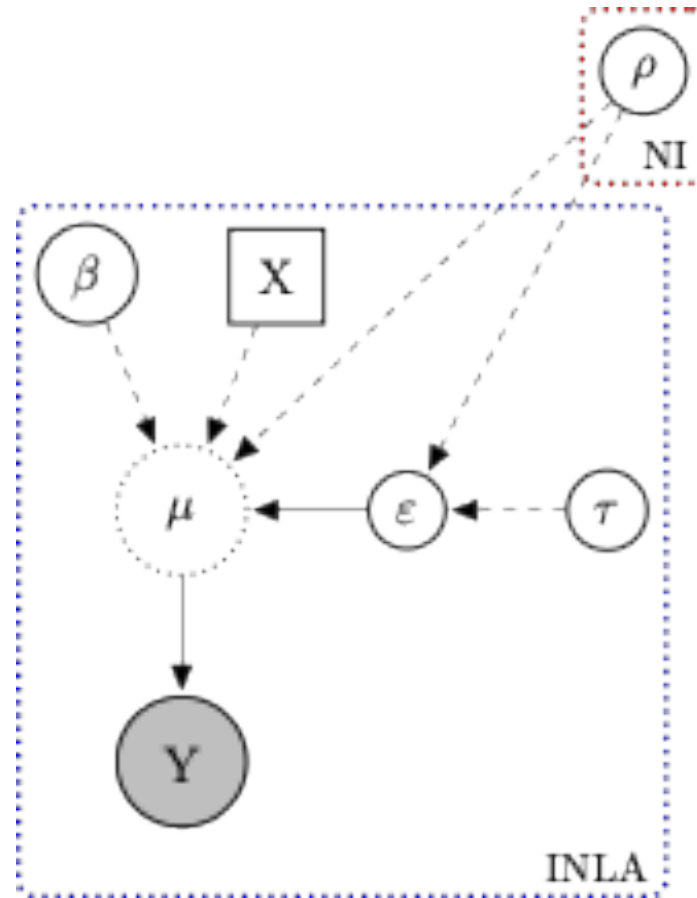


Source: Las Provincias.

# INLA and numerical integration

- The question is how to choose the integration points in  $\Omega$ .
- Bivand et al. (2014) use numerical integration in a model with one nuisance hyperparameter:
  - Grid of values defined on the line.
  - Numerical integration of a single variable.
- Application to spatial econometrics:
  - $y = (I - \rho W)^{-1} X\beta + \varepsilon$
  - $\varepsilon$  is distributed as a MVN with zero mean and precision matrix:  
 $\tau(I - \rho W^\top)(I - \rho W)$ .

# INLA and numerical integration



# INLA within MCMC

- The question is how to choose the integration points in  $\Omega$ .
- Bivand et al. (2014) use numerical integration in a model with one nuisance hyperparameter:
  - Grid of values defined on the line.
  - Numerical integration of a single variable.
- Gómez-Rubio and Rue (2019) use MCMC (M-H algorithm) to estimate  $\pi(\theta_2^* | y)$  and obtain integration points.
- Berild et al. (2022) propose the use of Importance Sampling (IS):
  - Simple to implement.
  - Easy to parallelize.

# Importance Sampling with INLA

- Importance sampling is a Monte Carlo method to estimate integrals:

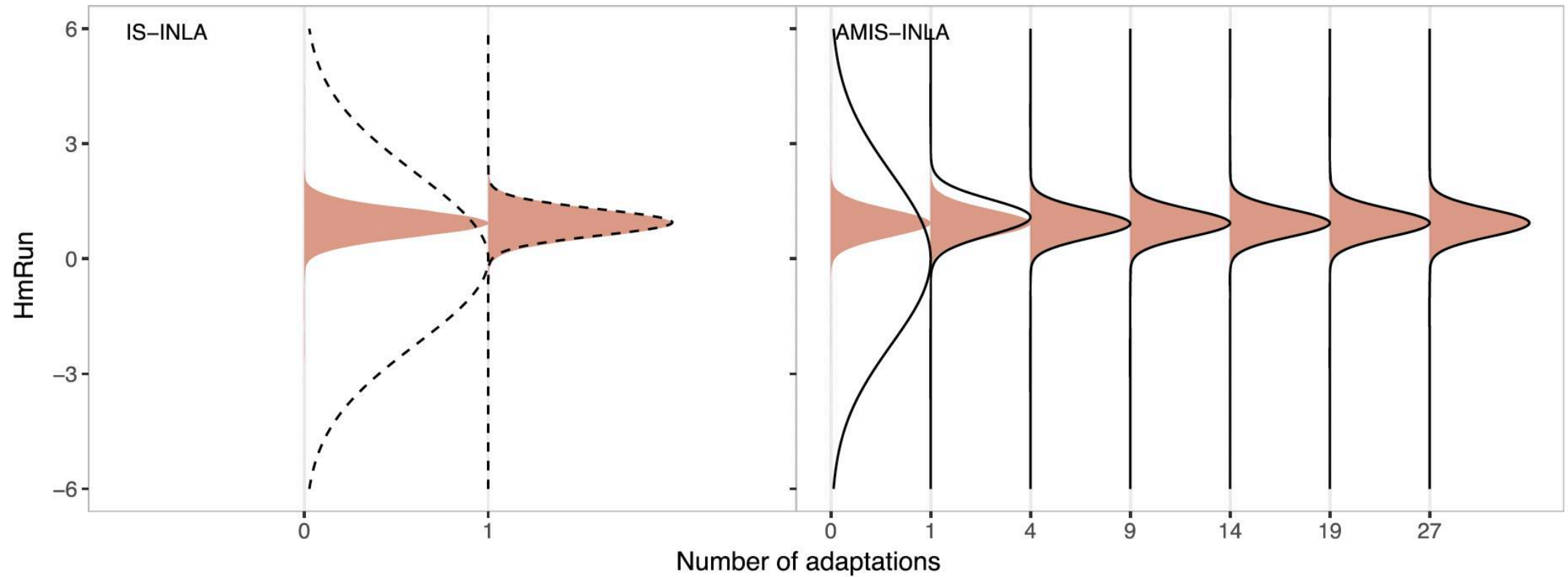
$$\mu_{\pi} = E_{\pi}[h(X)] = \int_D h(x)\pi(x)dx$$

- This integral can be estimated by sampling from a density  $g(x)$ :

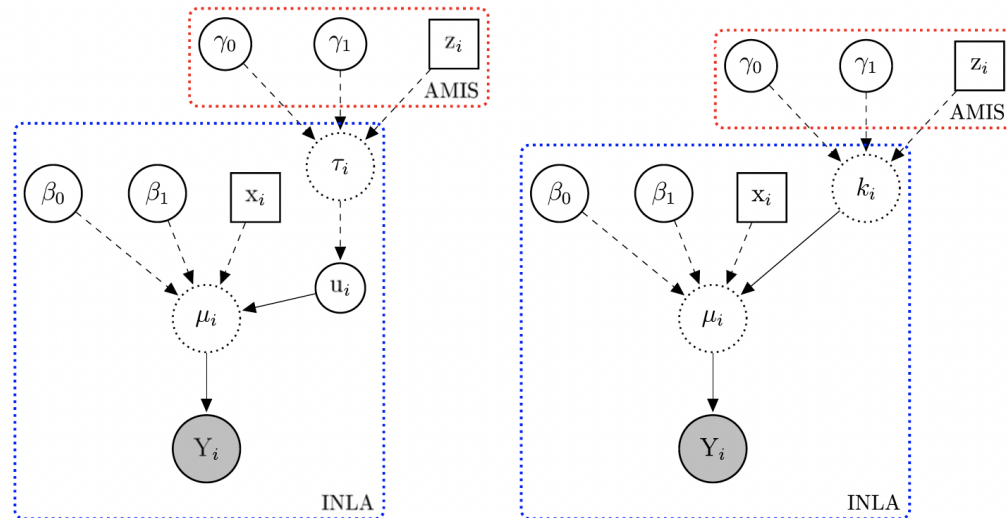
$$\mu_{\pi} = \int_D h(x) \frac{\pi(x)}{g(x)} g(x) dx \simeq \sum_{i=1}^N h(x^{(i)}) \frac{\pi(x^{(i)})}{g(x^{(i)})} = \sum_{i=1}^N h(x^{(i)}) w_i$$

- Weights  $w_i = \frac{\pi(x^{(i)})}{g(x^{(i)})}$
- $x^{(1)}, \dots, x^{(N)}$  is a random sample from  $g(x)$ .

# Importance Sampling with INLA

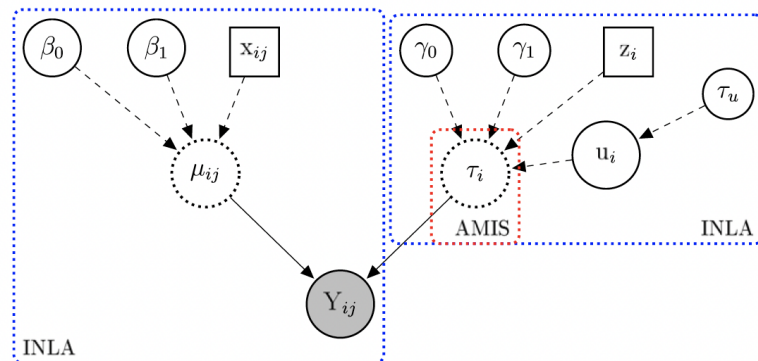


# Double hierarchical GLMMs



(a)

(b)



(c)

## a) Linear model:

- Mean:  $\mu_i = \beta_0 + \beta_1 x_i$ .
- Precision:  $\log(\tau_i) = \gamma_0 + \gamma_1 z_i$ .

## b) Negative binomial model:

- Parameter:  $\log(\mu_i) = \beta_0 + \beta_1 x_i$ .
- Size:  $\log(k_i) = \gamma_0 + \gamma_1 z_i$ .
- Probability:  $p_i = \frac{k_i}{k_i + \mu_i}$ .

## c) Multilevel Gaussian model:

- Mean:  $\mu_i = \beta_0 + \beta_1 x_i$ .
- Precision:  $\log(\tau_i) = \gamma_0 + \gamma_1 z_i + u_i$ .
  - $u_i$  i.i.d. random effects:
    - Precision  $\tau_u$ .

# Variable and model selection with INLA

- INLA provides (accurate) approximations to the marginal likelihood of a model.

- This can be used for computing Bayes Factors:

$$BF(M_i, M_j) = \frac{P(Y|M_i)}{P(Y|M_j)}$$

- Posterior probabilities can also be computed:

$$P(M_i|Y) = \frac{P(Y|M_i)P(M_k)}{\sum_{k \in \mathcal{M}} P(Y|M_k)P(M_k)}, \text{ with } \mathcal{M} \text{ the model space.}$$

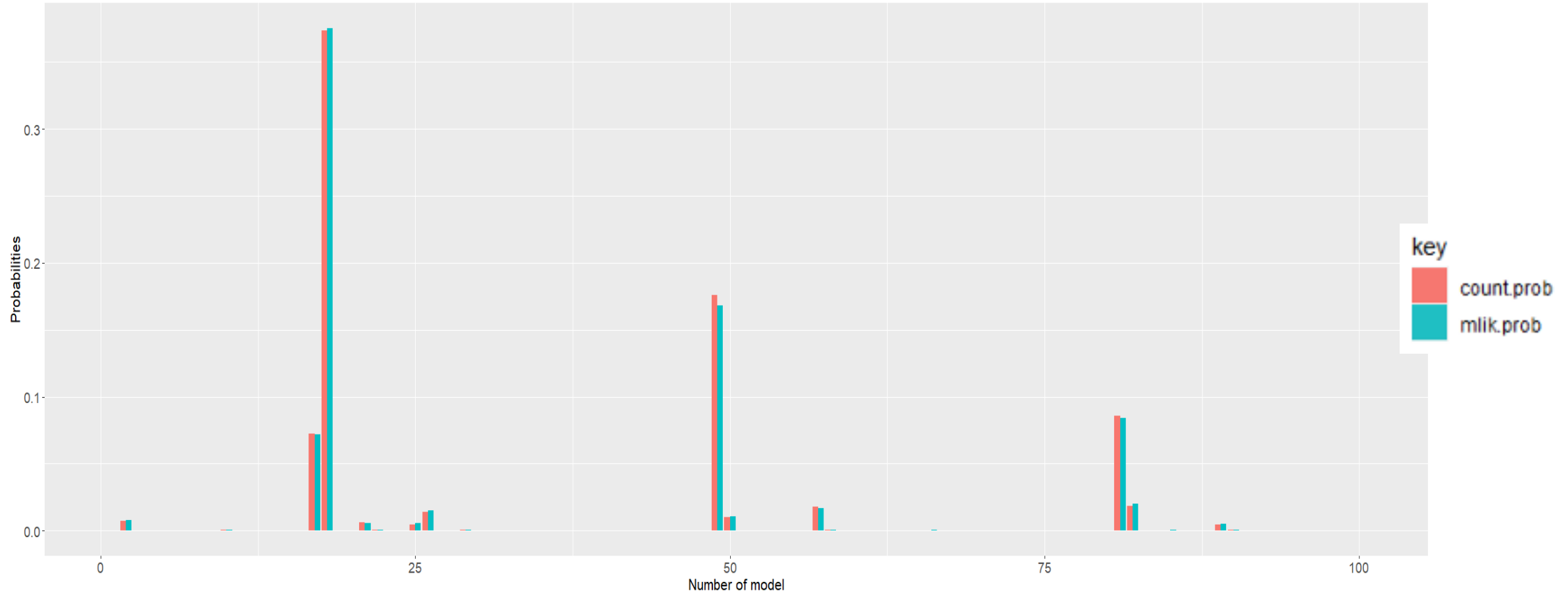
- If the model space is too large, a Metropolis-Hastings approach can be used to explore it.

# Example: Motor Trend Car Road Tests

- The mtcars (R package) provides data from car fuel consumptions in 1974.
- Consumption is measured in miles per US gallon (mpg).
- There are 10 covariates.
- Altogether  $2^{10} = 1024$  different possible models.
- Posterior probabilities:
  - Computed using MCMC.
  - Computed using the marginal likelihood from INLA.
- Other quantities of interest:
  - Posterior probability of including a covariate.
  - Averaged distribution of coefficients (via Bayesian model averaging, BMA).

# Example: Motor Trend Car Road Tests

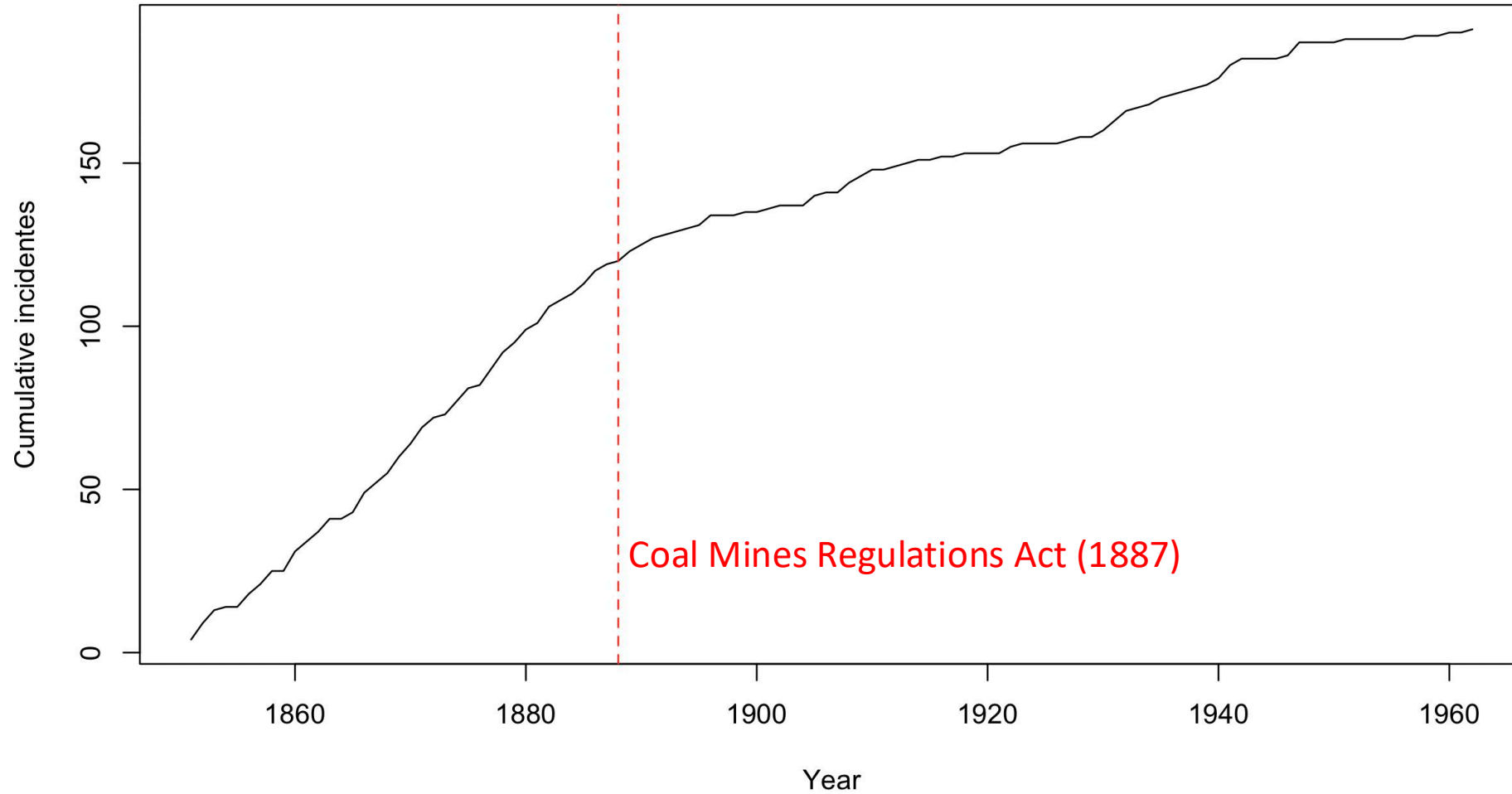
Comparison of marginal likelihood and number of times in the Markov Chain probabilities, models 1 to 100



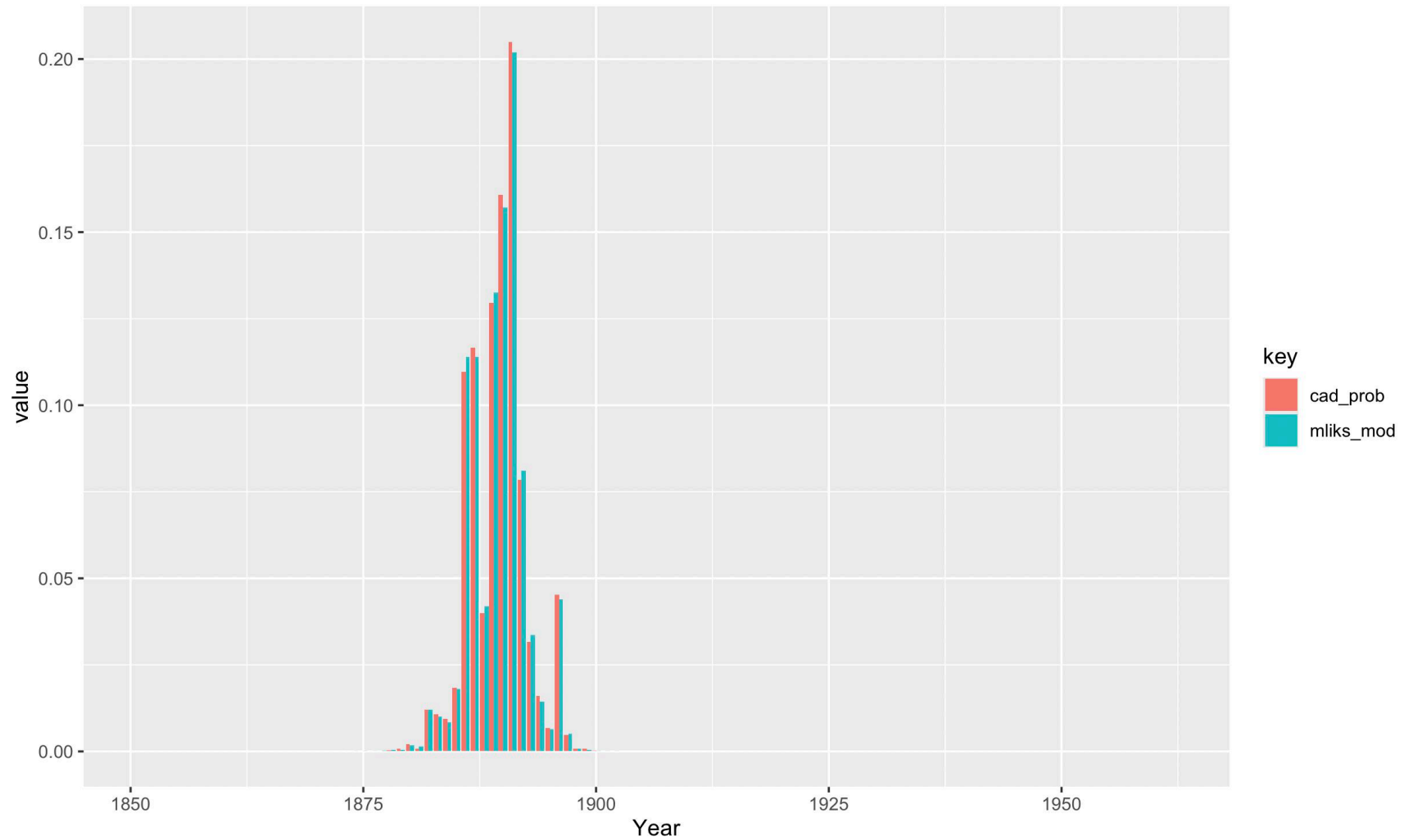
# Models for changepoint detection

- Models for changepoint detection can be approached as:
  - Variable selection (by adding an adequate “dummy covariate”).
  - Models with several likelihoods (when models’ structures are very different).
- A Metropolis-Hastings algorithm can be used to make inference about the changepoint (a discrete variable).
- *Conditional* on the changepoint the model becomes a model with several likelihoods.

# Yearly Coal Mining Disasters in the U.K.



# Yearly Coal Mining Disasters in the U.K.

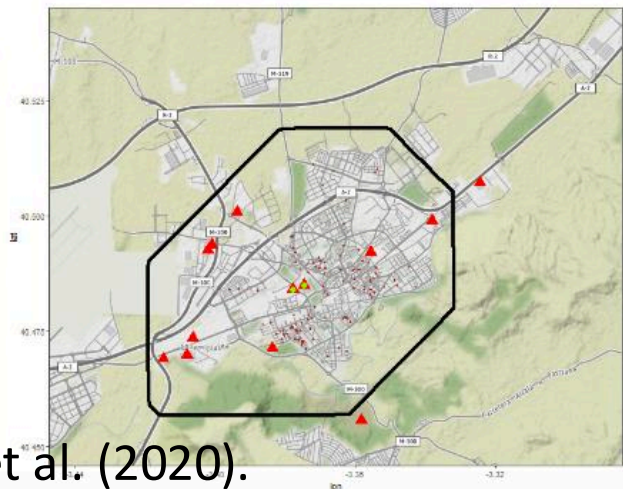
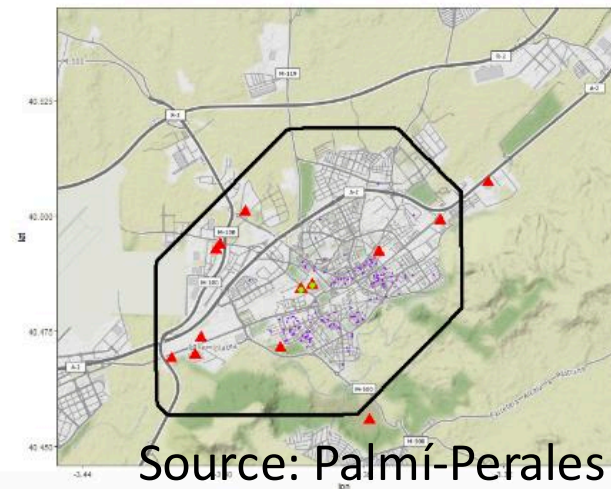
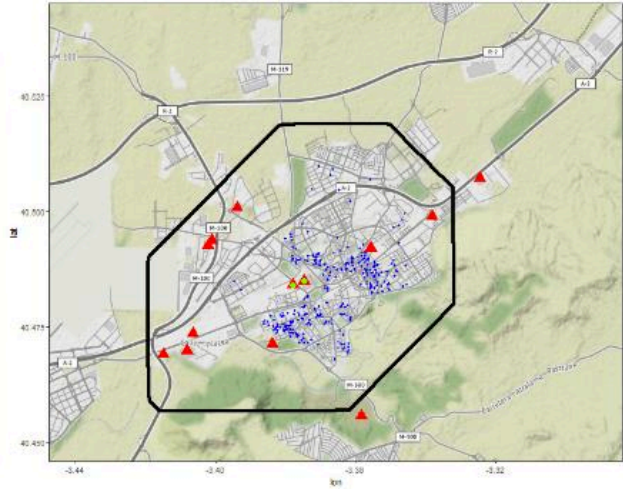
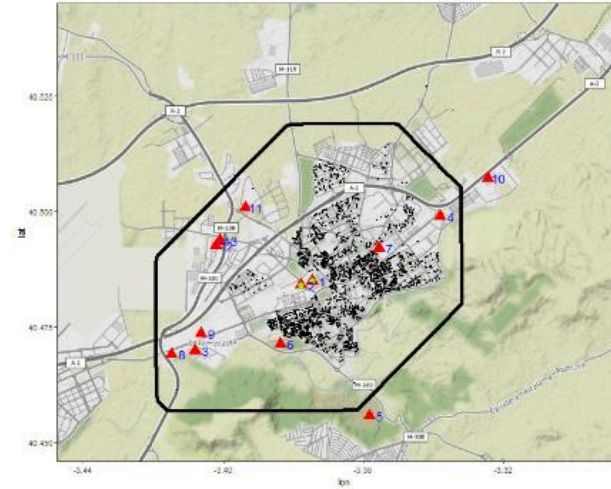


# Exposure to pollution sources

- Study conducted in Alcalá de Henares (Madrid) about three types of cancer.
- People aged > 39 years old and diagnosed with lung, stomach and kidney cancer between January 2012 and June 2014.
- ICD-10 codes available.
- Number of cases:
  - Lung cancer: 313.
  - Stomach cancer: 136
  - Kidney cancer: 115.
- 3000 controls.
- Data provided by the main hospital in Alcalá de Henares (*Minimum Basic Data Set, MBDS*).

# Exposure to pollution sources

- Controls (in black) and cases of lung (blue), stomach (blue) and kidney (Brown).
- Study area (black polygon).
- Location of 13 polluting industries that emit polluting gases (red triangles).
- Two industries also work with heavy metals (red triangle with green dot).



Source: Palmí-Perales et al. (2020).

# Exposure to pollution sources

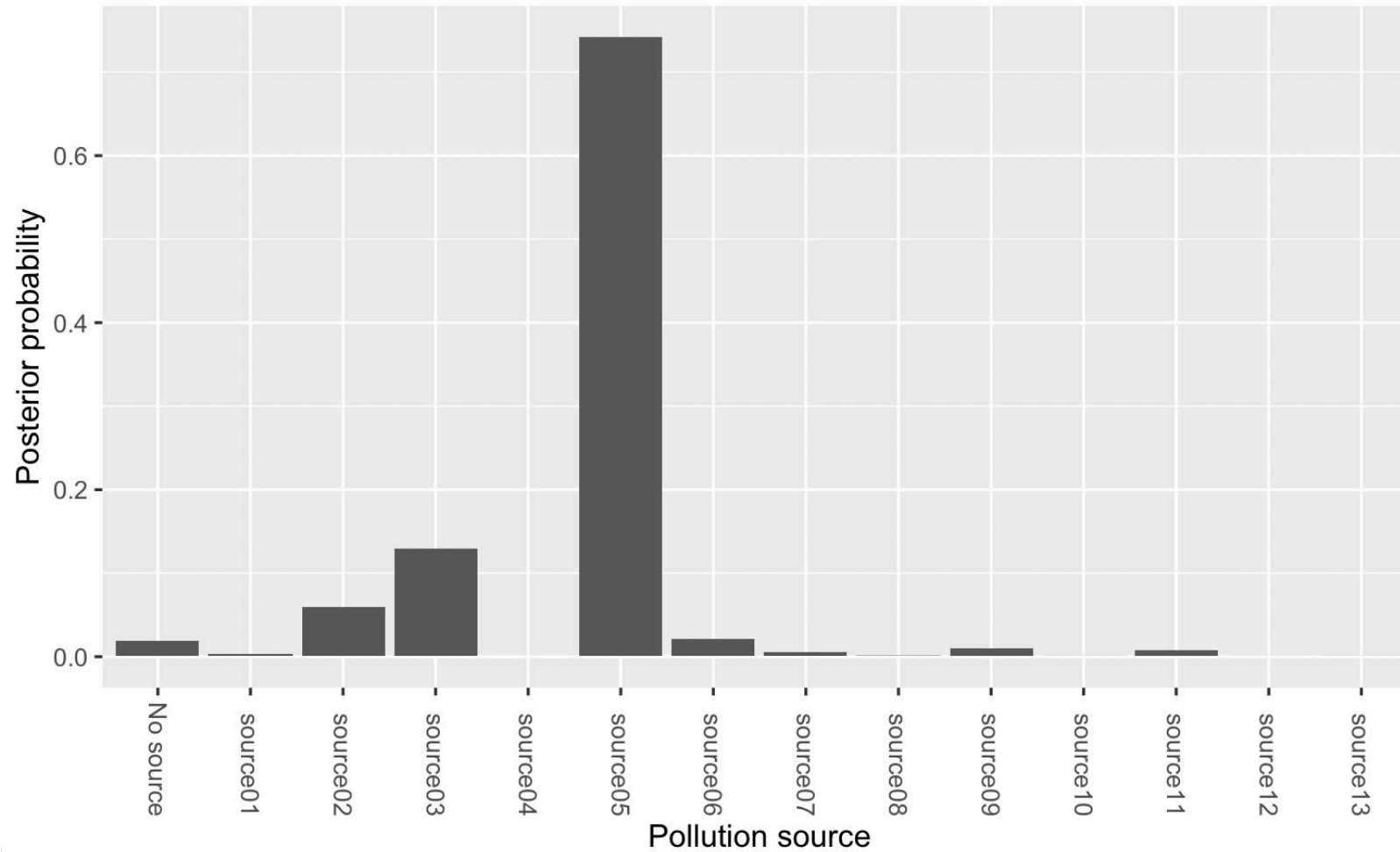
- Models are based on log-Gaussian Cox Processes:
  - Controls.
  - Cases.

$$\log(\lambda_0(x)) = \alpha_0 + S_0(x); x \in \mathcal{D},$$

$$\log(\lambda_i(x)) = \alpha_i + S_0(x) + S_i(x), i = 1, \dots, K; x \in \mathcal{D},$$

- Socio-economic covariates can be included
- Effect of pollution source (using a non-linear effect on distance).

# Exposure to pollution sources (lung cancer)





**THANK YOU**