Tarea 4: Problema de disease mapping con INLA

Análisis del banco de datos *Aragon.Rdata*

Irene Extremera Serrano Caterina Olaya Paparsenos Fernández

noviembre 30, 2020

El objetivo principal del presente trabajo es estudiar el *riesgo de mortalidad* por enfermedad isquemática en la comunidad de Aragón. Para ello se ha realizado un estudio del tipo **disease mapping** utilizado el modelo estadístico de Besag-York-Mollié **(BYM)** que incluye efectos aleatorios. Para ajustar este modelo se ha utilizado la aproximación anidada e integrada de Laplace **(INLA)**.

Descripción de datos

Para estudiar el riesgo relativo de mortalidad de la enfermedad en cuestión se utilizará uno de sus estimadores, la *Razón de Mortalidad Estandarizada* **(RME)** que se define como: $RME_i = \frac{Obs_i}{Esp_i}$ con i=número de municipios de Aragón, Obs_i=número de fallecidos observados en el i-ésimo municipio, Esp_i=número de fallecidos esperados en el mismo i-ésimo municipio. Cuando el valor que tola RME es superior al valor 1 ($Obs_i > Esp_i$) significa que el número de fallecidos observados es superior al esperado, y por tanto presenta un exceso de mortalidad. Sin embargo, cuando el RME de una comunidad toma valores menores a 1, significa que tiene un defecto de riesgo en comparación con el riesgo total de la comunidad de Aragón.

En los casos de disease mapping existe un problema de estimación donde las áreas pequeñas con poca población muestran valores observados altos. Esto se puede ver al representar gráficamente la razón de mortalidad estandarizada sin considerar los efectos aleatorios donde, en realidad, no se está mostrando la distribución de la enfermedad sino la distribución demográfica de la población.

Esto se puede ver en la *figura 1* donde se representa la razón de mortalidad estandarizada sin realizar el ajuste con el modelo BYM. Se puede ver como un gran número de comunidades pequeñas tienen un RME mayor a 1. Muchas de estas comunidades están aisladas por otras comunidades con RME bajo. Por lo tanto no se puede observar un patrón espacial claro para la enfermedad. Esto ocurre porque no se tienen en cuenta los posibles efectos aleatorios que influyen en la retrasmisión de esta enfermedad como puede ser las relaciones de vecindad entre los municipios.

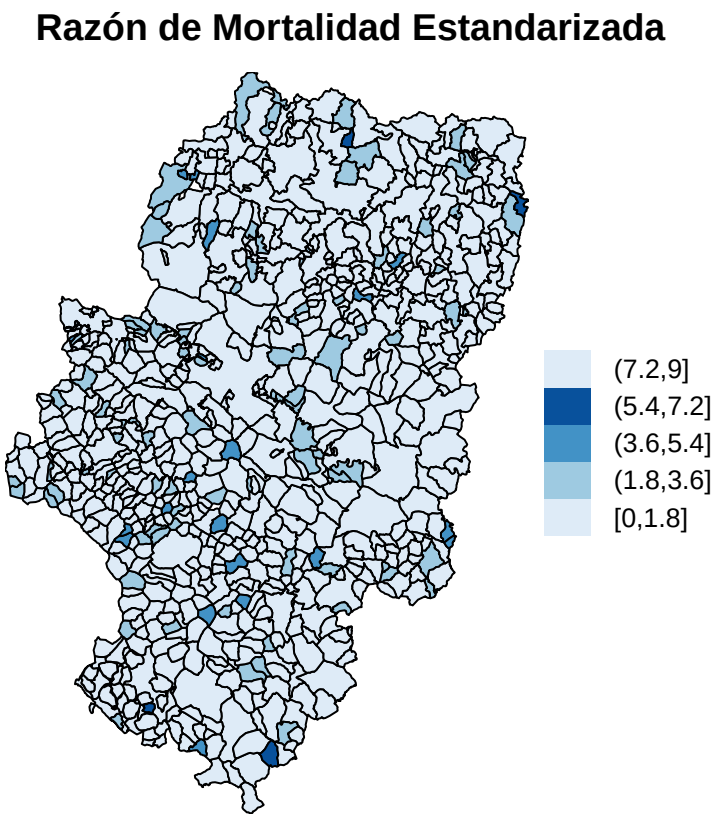

Figura 1: Razón de mortalidad estandarizada para la enfermedad isquémica sin ajustar por BYM

En la *figura 2* se muestran las relaciones de vecindades entre los diferentes municipios. Se puede observar como estas relaciones son multiples y complejas en muchos municipios por lo tanto, es preciso considerarlas en la modelización de la RME.

Modelización por BYM

Modelo Estimado

El uso del modelo de Besag, York y Mollié se usa de forma bastante habitual a la hora de estimar la distribución geográfica del riesgo de enfermedades. Este modelo considera que el número de defunciones observadas sigue una distribución de Poisson y su variabilidad depende de dos factores: un efecto aleatorio entructurado (efecto espacial dependiente de las relaciones de vecindad) y un efecto aleatorio heterogeneo independiente e idénticamente distribuido.

Para la modelización se hizo uso de *INLA* (integrated nested Laplace approximation) con el fin de facilitar el ajuste del modelo utilizando Modelos Jerarquicos Bayesianos. Por lo tanto, se ha trabajado con un modelo de varias capas: en una primera capa se encuentra la función de verosimilitud, en una segunda los campos latentes gaussianos y finalmente los hiperparámetros.

En primer lugar, como se trabaja con datos de conteo, los valores observados de las defunciones se estiman a una distribución Poisson de parametro λ. Por lo tanto la función de verosimilitud es:

$$Obs_i \sim Poisson(\lambda)$$

con

$$\lambda_i = Esp_i \rho_i,$$

donde Obs_i son los valores observados, Esp_i son los valores esperados, ρ_i es la RME e i es el número de los municipios de Aragón estudiados (i=1...729).

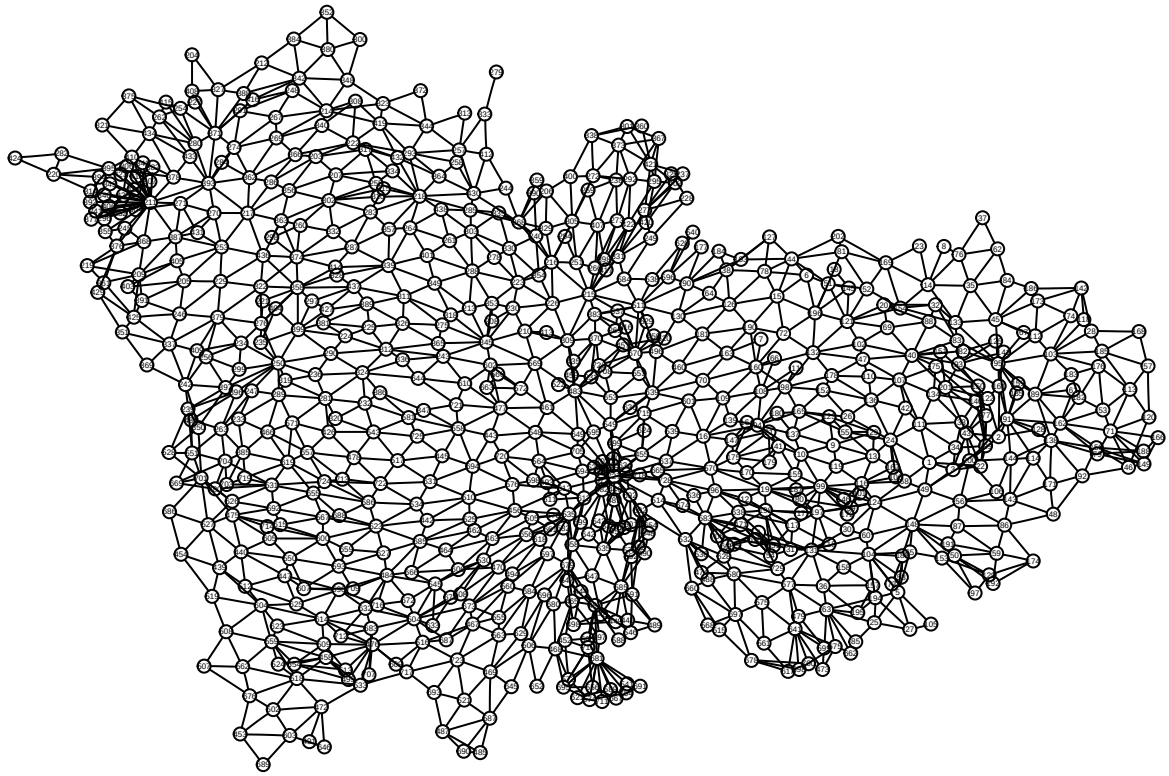

Figura 2: Relaciones de vecindad entre los diferentes municipios

Como ya se ha mencionado anteriormente, es necesario considerar los efectos aleatorios en la modelización de la RME. El estudio es de tipo disease mapping y por tanto el ρ_i se ha estimado utilizando exclusivamente los efectos aleatorios de tal manera que: $log(\rho_i) = \eta_i$.

En este caso η_i es el predictor lineal de REM donde se estima como:

$$\eta_i = \beta_0 + \upsilon_i + u_i,$$

siendo β_0 es el intercept,υ_i es el efecto aleatorio estructurado espacial y u_i el efecto aleatorio independiente idénticamente distribuido.

Al utilizar INLA se asume que los dos efectos aleatorios siguen un campo latente Gausianno que se estiman como:

$$\upsilon_i \mid \upsilon_{-i} \sim N(\frac{1}{n_i}\sum_{i\sim j}\upsilon_j, \frac{1}{n_i\tau_\upsilon})$$

$$u_i \sim N(0, \tau_{u^{-1}})$$

Finalmente, los hiperparámetros en este caso son las precisiones τ_υ y τ_u de los efectos aleatorios. Ambos se estimaron de la siguiente manera: $log(\tau_\upsilon) \sim logGamma(1, 0001)$ y $log(\tau_u) \sim logGamma(1, 0001)$.

Resultados del modelo estimado

```
##
## Call:
##    c("inla(formula = formula, family = \"poisson\", data = Aragon, E = E,
##    ", " control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE), ", "
##    control.predictor = list(compute = TRUE, cdf = c(log(1))))" )
## Time used:
##     Pre = 0.811, Running = 11.8, Post = 0.556, Total = 13.2
## Fixed effects:
##               mean    sd 0.025quant 0.5quant 0.975quant   mode kld
```

```
## (Intercept) -0.061 0.036     -0.135   -0.061      0.007 -0.059   0
##
## Random effects:
##   Name     Model
##     S Besags ICAR model
##    U IID model
##
## Model hyperparameters:
##                   mean     sd 0.025quant 0.5quant 0.975quant  mode
## Precision for S  20.35  12.84       6.78    16.85      54.29 12.30
## Precision for U 549.19 711.46       5.56   284.93    2492.20  3.51
##
## Expected number of effective parameters(stdev): 48.30(18.12)
## Number of equivalent replicates : 15.09
##
## Deviance Information Criterion (DIC) ...............: 1899.15
## Deviance Information Criterion (DIC, saturated) ....: 827.21
## Effective number of parameters .....................: 51.42
##
## Watanabe-Akaike information criterion (WAIC) ...: 1904.09
## Effective number of parameters .................: 50.80
##
## Marginal log-Likelihood:  -1246.59
## CPO and PIT are computed
##
## Posterior marginals for the linear predictor and
##  the fitted values are computed
```

Los resultados obtenidos tras la realización del modelo ajustado se pueden observar en la salida anterior. Como resultado se obtuvo que el intercepto toma un valor medio de -0.061 mientras que la precisión media para los efectos aleatorios espacial v_i e independiente u_i es de 20.35 y 549.19 respectivamente. El intervalo de credibilidad al 95 % de la media del intercept está entre -0.135 a 0.007 mientras que los mismos intervalos para las medias de las precisiones v_i e u_i son: desde 6.78 a 54.29 y desde 5.56 a 2492.20 respectivamente.

Los criterios de calidad del modelo en cuestión son: DIC toma un valor de 1910.08, WAIC toma un valor de 1917.05 y la calidad de la predicción según el CPO toma un valor de -1249.23.

Distribuciones a posteriori según el modelo estimado

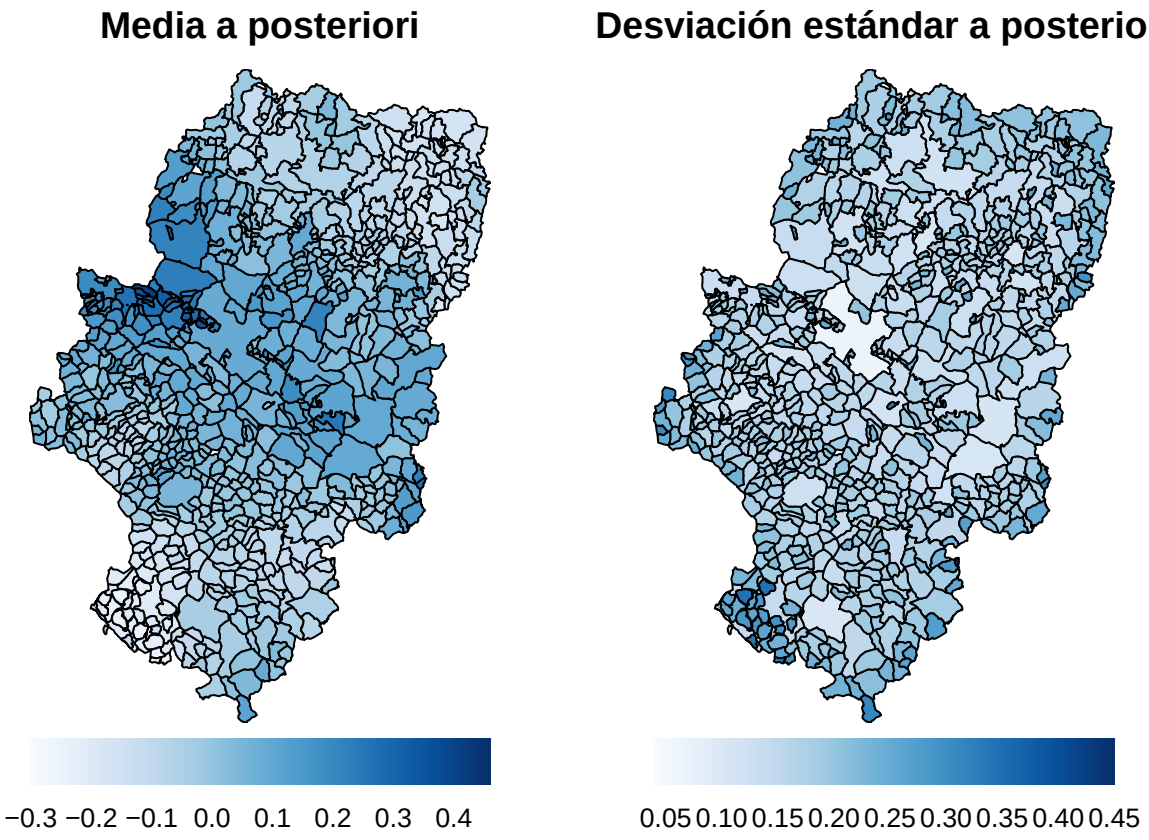

Figura 3: Distribución a posteriori del efecto aleatorio espacial del modelo ajustado.

En la *figura 3* se muestran las distribuciones a posteriori de la media y desviación típica para el efecto espacial según el modelo ajustado. En esta se puede observar claramente la existencia de un gradiente en la variabilidad espacial. Las zonas con menor variabilidad en la media de la RME se encuentran en la parte norte y sur de la comunidad. Sin embargo, esta variabilidad aumenta en zonas más céntricas. Los municipios con mayor RME parecen encontrarse en la parte central y oeste de Aragón. Por otro lado, la variabilidad de la desviación estandar es mayor en las zonas localizadas en las fronteras de Aragón.

En la *figura 4* se muestra las distribuciones a posterior de la media y la desviación estándar para la RME según el modelo que contiene los efectos aleatorios. Como era se esperar, la mayoría de los municipios que tienen una RME superior a 1 se encuentran en la parte central y oeste de la comunidad. Además, auquellas zonas con una RME>1.4 (color azul oscuro) están rodeadas por otras zonas con valores altos de la RME. Por lo tanto, se puede secir que esta figura es mucho más precisa y contiene información más fiable en comparación con la *figura 1*.

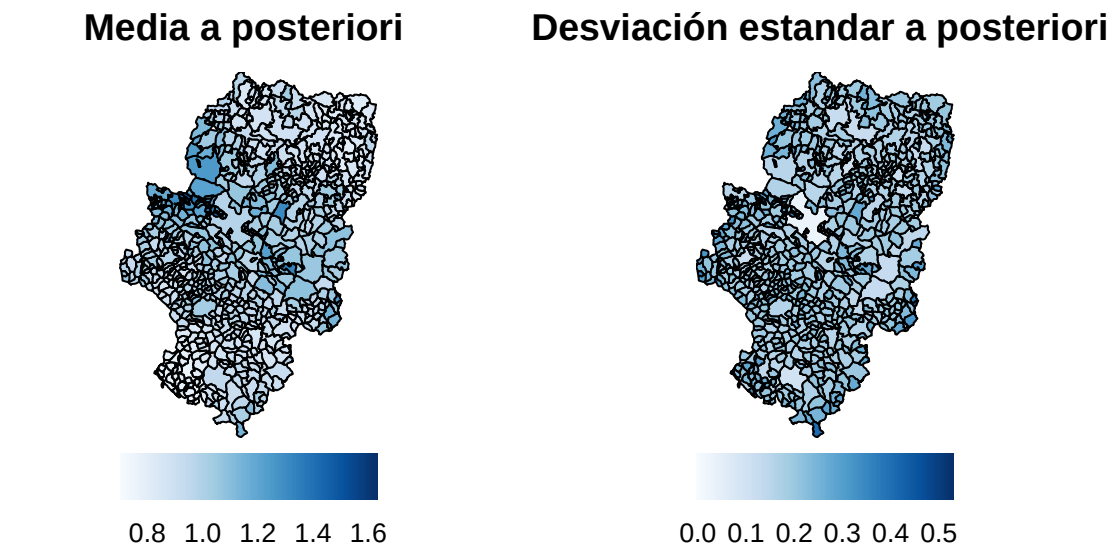

Figura 4: Distribución a posteriori de la media (mean) y desviación estándar (sd) de la mortalidad estandarizada ajustada con el modelo BYM.

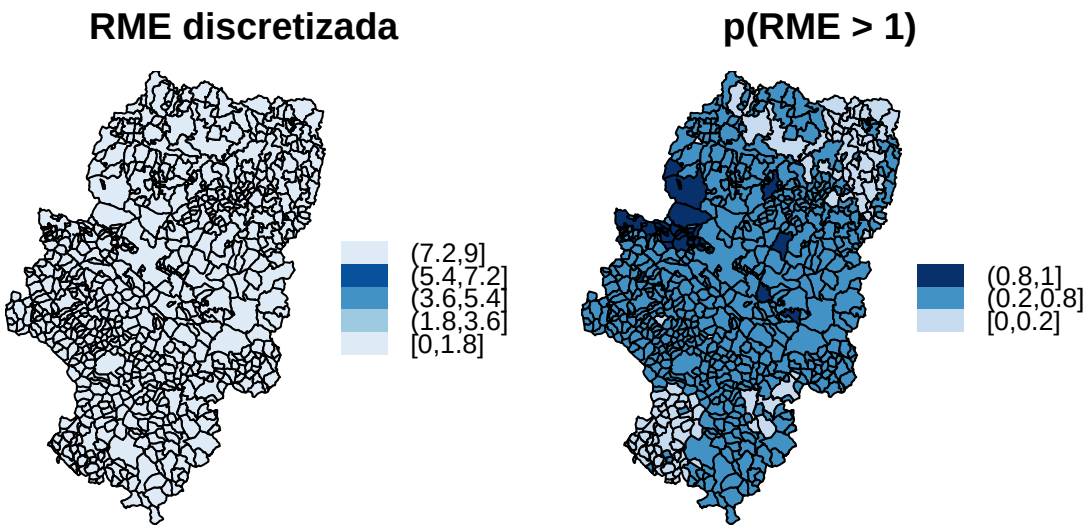

Figura 5: Probabilidad de que la RME sea mayor a 1 según el modelo ajustado.

En esta última *gráfica 5*, se muestra la RME discretizada y la probabilidad que esta tome valores superiores a 1. Como se ha comentado anteriormente, aquellas zonas con RME>1 presentan un exceso de mortalidad ya que el número de fallecidos observado es superior al esperado. Según la representación de la derecha, pocos municipios de Aragon tienen gran probabilidad (80 % o más) de presentar alto riesgo de mortalidad por enfermedad isquémica. Estos se representan en color azul oscuro y se encuentran en la parte central y oeste de la comunidad. El resto de municipios tienen entre menor probabilidad (entre 20 % y 79 %) de tener alto riesgo mientras que solo unos pocos tienen una probabilidad menor al 20 % (azul muy clarito). Estos últimos municipios se encuentran al norte y al sur de Aragón.

Anexo: código de R

```
options(tinytex.verbose = TRUE)
# Librerías

library(sp)
"rgdal_show_exportToProj4_warnings"="none"
library(rgdal)
library(spData)
library(sf)
library(spdep)
library(Matrix)
library(lattice)
library(parallel)
library(foreach)
library(INLA)
library(RColorBrewer)
library(BiocManager)
library(grid)
library(BiocGenerics)
library(graph)
library(Rgraphviz)
library(ggplot2)
library(ggpubr)
library(gridExtra)

library(bookdown)
#Antes de comenzar con el análisis, comprobamos que la información perteneciente al objeto
#espacial y la presente en el archivo Aragón coincidía y la organizamos.

#cartografia<- readOGR(dsn="", layer='aragon')
#View(cartografia) #se ve que está deshorganizado

library(raster)
cartografia <- shapefile("C:/Users/Caterina.DESKTOP-T5E1RQU/Desktop/modulo_especialización_master/Tarea


load("C:/Users/Caterina.DESKTOP-T5E1RQU/Desktop/modulo_especialización_master/Tareas/Tarea 4_INLA/Mater


#Comprobamos si está ordenado:
#head(cartografia); summary(cartografia)
aragon<- cbind(distancia,E,O)
#View(aragon)
cartografia<- cartografia[order(cartografia$CODMUNI),]

#Dibujamos los datos
cartografia$SMR_raw <- O/E #Razón de mortalidad
#min(cartografia$SMR_raw);max(cartografia$SMR_raw)

SMR_raw.cutoff<- c(0, 1.8,1.8*2,1.8*3, 1.8*4,1.8*5)
SMR_raw_disc = cut(cartografia$SMR_raw,
                   breaks         = SMR_raw.cutoff,
                   include.lowest = TRUE)
cartografia$SMR_raw_disc <- SMR_raw_disc

spplot(cartografia,
       c("SMR_raw_disc"),
       col.regions = brewer.pal(9,'Blues')[c(2,4,6,8)],
```

```
       main        = "Razón de Mortalidad Estandarizada",
       par.settings =
         list(axis.line = list(col =  'transparent')))
# Relación de vecindad entre los distintos

vecinos <- poly2nb(cartografia)
nb2INLA("aragon.shp", vecinos)
H <- inla.read.graph(filename="aragon.shp")

#Matriz de precisión
#image(inla.graph2matrix(H),xlab="",ylab="")

#matriz de relaciones de vecindad entre los municipios
plot(H)

# Dibujo los vecinos
plot_map_neig <- function(neig)
{
  plot(cartografia)
  plot(cartografia[neig, ], border="white",
       col="red", add=TRUE)

  plot(cartografia[vecinos[[neig]], ],
       border="white",
       col="blue", add=TRUE)
}

#plot_map_neig(30) #Municipio 30
# Ajuste del modelo con el efecto espacial
##S modelo estructurado
##U modelo estructurado con estructura independiente

S <- U <- seq(1,729) # Vector de índices
Aragon <- cbind(aragon, S, U)

# Fórmula
formula <- O ~ 1 + f(S,
                     model       = "besag",
                     graph       = H,
                     scale.model = TRUE,
                     hyper       =
                       list(prec = list(prior="loggamma",param = c(1,0.001)))) +
  f(U,
    model       = "iid",
    hyper       =
      list(prec = list(prior="loggamma",param = c(1,0.001))))


# Modelo
Aragon<-as.data.frame(Aragon) # Lo convertimos en data frame

modelo_aragon <- inla(formula,
                      family          = "poisson",
                      data            = Aragon,
                      E               = E,
                      control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE),
                      control.predictor = list(compute=TRUE, cdf=c(log(1))))

(summary(modelo_aragon))
#Distribución a posteriori del efecto aleatorio (media y desviación estándar)
```

```
cartografia$SPmean <- round(modelo_aragon$summary.random$S[["mean"]], 4)
cartografia$SPsd <- round(modelo_aragon$summary.random$S[["sd"]],5)

grid.arrange(spplot(cartografia, c("SPmean"),
         main = c("Media a posteriori"),
         #col.regions = rev(viridis_pal(option = "B")(101)),
         col.regions = colorRampPalette(brewer.pal(9,'Blues'))(101),
         cuts        = 100,
         colorkey=list(space="bottom", space = "bottom"),
         par.settings =
           list(axis.line = list(col =  'transparent',
                                 legend.ticks = 'black'))),
  spplot(cartografia, c("SPsd"),
         main = c("Desviación estándar a posteriori"),
         col.regions = colorRampPalette(brewer.pal(9,'Blues'))(101),
         cuts        = 100,
         colorkey=list(space="bottom", space = "bottom"),
         par.settings =
           list(axis.line = list(col =  'transparent',
                                 legend.ticks = 'black'))),
  ncol = 2)

#Distribución a posteriori del riesgo relativo en la unidad geográfica

cartografia$SMR_mean <- modelo_aragon$summary.fitted.values$mean # mean
cartografia$SMR_sd <- modelo_aragon$summary.fitted.values$sd #s
cartografia$SMR_p1 <- 1 - modelo_aragon$summary.fitted.values$`1 cdf`


grid.arrange(spplot(cartografia,
                    c("SMR_mean"),
                    col.regions = colorRampPalette(brewer.pal(9,'Blues'))(101),
                    cuts         = 100,
                    main        = "Media a posteriori ",
                    colorkey=list(space="bottom"),
                    par.settings =
                      list(axis.line = list(col =  'transparent'))),

             spplot(cartografia,
                    c("SMR_sd"),
                    col.regions = colorRampPalette(brewer.pal(9,'Blues'))(101),
                    cuts         = 100,
                    main        = "Desviación estandar a posteriori ",
                    colorkey=list(space="bottom"),
                    par.settings =
                      list(axis.line = list(col =  'transparent'))),ncol = 2)

# Rrobabilidad de que la razón de mortalidad estandarizada sea mayor que 1
cartografia$SMR_p1 <- 1 - modelo_aragon$summary.fitted.values$`1 cdf`
SMR.cutoff<- c(0, 1.8,1.8*2,1.8*3, 1.8*4,1.8*5) #Discretizamos
SMR_p1.cutoff <- c(0,0.2,0.8,1)

SMR_disc = cut(cartografia$SMR_mean,
               breaks         = SMR.cutoff,
               include.lowest = TRUE)

SMR_p1_disc = cut(cartografia$SMR_p1,
                  breaks         = SMR_p1.cutoff,
                  include.lowest = TRUE)
```

```
cartografia$SMR_disc <- SMR_disc
cartografia$SMR_p1_disc <- SMR_p1_disc

grid.arrange(spplot(cartografia,
                    c("SMR_disc"),
                    col.regions = brewer.pal(9,'Blues')[c(2,4,6,8)],
                    main        = "RME discretizada ",
                    par.settings =
                      list(axis.line = list(col =  'transparent'))),
             spplot(cartografia,
                    c("SMR_p1_disc"),
                    col.regions = brewer.pal(9,'Blues')[c(3,6,9)],
                    main        = "p(RME > 1) ",
                    par.settings =
                      list(axis.line = list(col =  'transparent'))), ncol = 2)
```