Practical 5: R-INLA: Disease mapping and ecological regression

In this practical you will work on disease mapping (DM) and ecological regression models using INLA. To do so you will use a dataset related to stroke mortality in Sheffield (this example is also included in the "Spatial and Spatio-Temporal Bayesian Models with R-INLA" book, used in Chapter 5). The data includes: (i) `Stroke.csv` with information about stroke deaths, population at risks and covariates (NOx concentration and social deprivation measured through Townsend index); (ii) Shapefile with the enumeration districts for the city of Sheffield.

2 Data preparation

1. First you need to import the data and the shapefile:

```
> #Data
> Stroke <- read.csv("Stroke.csv")
> #Shapefile
> library(maptools)
> sheffield.gen <- readShapePoly("Sheffield.shp")
>
> #Remember to set the correct working directory
```

Typing

```
> head(stroke)
```

```
     SP_ID stroke_exp     pop y Townsend.class   NOx.class Townsend NOx
1 05CGFA01   2.620953  931.65 3  [-6.54,-2.07] (41.4,53.6]        1   2
2 05CGFA02   1.989744  710.36 1  (-2.07,0.086]  (82.1,243]        2   5
3 05CGFA03   2.286056  654.12 3   (0.086,3.11]  (82.1,243]        3   5
4 05CGFA04   2.151746 1109.37 0  (-2.07,0.086]  (82.1,243]        2   5
5 05CGFA05   4.380834 1159.31 1  (-2.07,0.086] (53.6,64.8]        2   3
6 05CGFA06   3.094954 1020.21 5  [-6.54,-2.07] (53.6,64.8]        1   3
     Offset
1 -2.549570
2 -2.551463
3 -2.455050
4 -2.711442
5 -2.420999
6 -2.516716
```

and

```
> names(sheffield.gen)
```

```
 [1] "SP_ID"      "ED8"        "wbID"       "stroke_o"   "stroke_exp"
 [6] "strexpdep"  "Townsend"   "rnq"        "proprnq1"   "proprnq2"
[11] "proprnq3"   "proprnq4"   "proprnq5"   "propinassQ" "pop"
```

we can see that the dataset has 10 columns and includes an ID (`SP_ID`), information on expected deaths for stroke and population at risk (`stroke_exp`, `pop`), information on the counts of deaths for stroke (`y`), two covariates - in categories - (`Townsend`, `NOx`) and an offset (`Offset`) (which we will focus on later). Note that also the shapefile has a column called `SP_ID` which we will use later to merge with the data before mapping.

3 Data preparation

1. As the data are already in a `data.frame` format we can directly use it for the disease mapping and ecological regression models. However we first need to create the adjacency matrix (see slides 19-21 of Lecture 7). As we have the shapefile we can use the third method described in slide 21 to build the adjacency matrix:

```
> library(spdep)
> nb2INLA(paste(my.dir,"Sheffield.graph",sep=""),poly2nb(sheffield.gen))
> sheffield.adj <- paste(my.dir,"/sheffield.graph",sep="")
```

2. IMPORTANT: We need to make sure that the order of the areas in the data frame follows the one in the shapefile otherwise the adjacency matrix which we are going to build will not be valid! To do so we can order the data based on the shapefile:

```
> stroke<- stroke[match(sheffield.gen$SP_ID,stroke$SP_ID),]
> #Check for the first 10 areas
> stroke$SP_ID[1:10]

 [1] 05CGGD02 05CGFT44 05CGGF35 05CGGD29 05CGFH02 05CGGD31 05CGFM08
 [8] 05CGGD22 05CGGD17 05CGGD16
1030 Levels: 05CGFA01 05CGFA02 05CGFA03 05CGFA04 05CGFA05 ... 05CGGF36

> sheffield.gen$SP_ID[1:10]

 [1] 05CGGD02 05CGFT44 05CGGF35 05CGGD29 05CGFH02 05CGGD31 05CGFM08
 [8] 05CGGD22 05CGGD17 05CGGD16
1030 Levels: 05CGFA01 05CGFA02 05CGFA03 05CGFA04 05CGFA05 ... 05CGGF36
```

4 Disease mapping

1. Specify a BYM model to evaluate the spatial variation of stroke mortality across Sheffield's enumeration districts. The model we want to use is the following (note that in slide 13

we presented a Poisson likelihood while here we want to use a Binomial likelihood):

$$\begin{aligned} y_i &\sim \text{Binomial}(\pi_i, n_i) \\ \text{logit}(\pi_i) &= b_0 + v_i + u_i \\ v_i &\sim \text{Normal}(0, \sigma_v^2) \\ \boldsymbol{u} &\sim \text{ICAR}(\boldsymbol{W}, \sigma_u^2) \end{aligned}$$

To do so in INLA we need to follow these steps:

(a) Add an area identifier (from 1 to 1030 which is the total number of areas) which will be used for the BYM specification

```
> stroke$ID<- seq(1,1030)
```

(b) Write the formula

```
> formula.DM <- y ~  f(ID,model="bym", graph=sheffield.adj,
                       hyper=list(prec.spatial=list(param=c(1,1))))
```

(c) Run the model (remember that here we need to specify the Binomial distribution instead of the Poisson). Include the DIC estimate as we will use it to compare this model with the hierarchical model which does not include a spatially structured component.

```
> library(INLA)
> model.DM <- inla(formula.DM,family="binomial",
            data=stroke, offset=Offset, Ntrials=pop,
            control.compute=list(dic=TRUE))
```

Here the `Offset` must be on the logit scale of the probability of death (as we are using the Binomial distribution for the $\boldsymbol{y}$). Luckily we have a variable in the `stroke` dataset called `Offset` which is calculated as

$$\text{offset}_i = \text{logit}(E_i/\text{Pop}_i)$$

so we can directly include it in the `formula`. Note that for the DIC we need to include `control.compute=list(dic=TRUE)`

2. Obtain the summary statistics for the model (random effects and hyperparameters). Remember (as you have seen yesterday) that the fixed effects here include only the intercept so they are not interesting.

```
> #Random effect
> summary.random <- model.DM$summary.random$ID
> dim(summary.random)

> head(summary.random)

[1] 2060    8
```

```
   ID   mean    sd 0.025quant 0.5quant 0.975quant   mode kld
1   1 -0.347 0.497     -1.372   -0.330      0.582 -0.296   0
2   2 -0.304 0.570     -1.456   -0.293      0.785 -0.271   0
3   3 -0.260 0.498     -1.288   -0.243      0.671 -0.209   0
4   4  0.516 0.307     -0.122    0.529      1.084  0.554   0
5   5 -0.051 0.597     -1.249   -0.042      1.098 -0.024   0
6   6 -0.060 0.474     -1.038   -0.044      0.827 -0.011   0
```

Note that the dimension of `summary.random` is $2 \times N$ (N is the number of areas). This is because the BYM specification includes two parameters, u_i and v_i. INLA parametrises $u_i + v_i$ (first N rows) and u_i (rows $N+1$ to $2N$).

```
> #Hyperparameters
> summary.hyper<- model.DM$summary.hyper
> round(summary.hyper,3)
```

```
                                       mean    sd 0.025quant 0.5quant
Precision for ID (iid component)     2.869 0.271      2.377    2.855
Precision for ID (spatial component) 5.715 1.739      3.043    5.470
                                     0.975quant  mode
Precision for ID (iid component)          3.440 2.826
Precision for ID (spatial component)      9.808 5.011
```

3. Evaluate the proportion of the variance explained by the spatially structured component ($\boldsymbol{u}$). To do so we will follow slide 31 and use the `inla.hyperpar.sample` command:

```
> marg.hyper <- inla.hyperpar.sample(100000,model.DM)
> colnames(marg.hyper)
```

```
[1] "Log precision for ID (idd component) in user-scale"
[2] "Log precision for ID (spatial component) in user-scale"
```

which returns a matrix of dimension 100000×2 with a sample from the joint posterior distribution of the hyperparameters.

Then we simply build the proportion as

```
> perc.var.u1 <- mean(marg.hyper[,1] / (marg.hyper[,1]+marg.hyper[,2]))
> perc.var.u1
```

```
[1] 0.3469765
```

so it seems that the spatial component is explaining about 34% of the total variance.

Note that

$$\text{frac}_{\text{spatial}} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2}$$

but here we have the precisions τ_u and τ_v; following these simple steps:

$$\begin{aligned} \text{frac}_{\text{spatial}} &= \frac{1/\tau_u}{1/\tau_u + 1/\tau_v} \\ &= \frac{1/\tau_u}{\frac{\tau_u+\tau_v}{\tau_u\tau_v}} \\ &= \frac{\tau_v}{\tau_v + \tau_u} \end{aligned}$$

4. Map the u_i and $u_i + v_i$ effects as an additional way of comparing the role of spatially structured and unstructured components.

   (a) First we transform u_i and $u_i + v_i$ into natural scale and into categories (we use user-defined categories to make the two maps comparable)

```
> #Natural scale
> exp.uv <- lapply(model.DM$marginals.random$ID[1:1030], function(x)
                  inla.emarginal(exp,x))
> exp.u <- lapply(model.DM$marginals.random$ID[1031:2060], function(x)
                  inla.emarginal(exp,x))
> #Categories
> cutoff=c(0.5,0.8,0.95,1.05,1.2,18.5)
> exp.uv.cat <- cut(unlist(exp.uv),breaks=cutoff,include.lowest=TRUE)
> exp.u.cat <- cut(unlist(exp.u),breaks=cutoff,include.lowest=TRUE)
```

   (b) Then we create a data frame with u_i, $u_i + v_i$ and the area identifier and we merge it with the shapefile

```
> data.exp.BYM <- data.frame(SP_ID=stroke$SP_ID,
                             exp.u=exp.u.cat, exp.uv=exp.uv.cat)
> row.names(data.exp.BYM) <- seq(1,1030)
> head(data.exp.BYM)
     SP_ID      exp.u     exp.uv
1 05CGGD02 (0.8,0.95]  [0.5,0.8]
2 05CGFT44 (0.8,0.95] (0.8,0.95]
3 05CGGF35 (0.8,0.95] (0.8,0.95]
4 05CGGD29 (0.8,0.95] (1.2,18.5]
5 05CGFH02 (1.05,1.2] (1.05,1.2]
6 05CGGD31 (0.8,0.95] (1.05,1.2]
> #Merge exp(v), exp(v+u) and the sheffield shapefile
> sheffield <- sheffield.gen
> data.sheffield <- attr(sheffield, "data")
> attr(sheffield, "data") <- merge(data.sheffield,
                                   data.exp.BYM, by="SP_ID")
```

   (c) Finally we map u_i and $u_i + v_i$

```
> library(RColorBrewer)
> spplot(obj=sheffield, zcol=c("exp.u","exp.uv"),
        col.regions= brewer.pal(5, "BrBG"), main="")
```

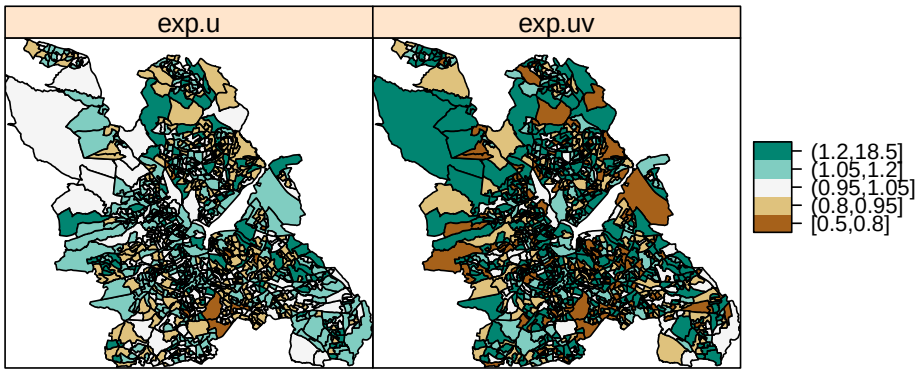

Figure 1: u_i and $u_i + v_i$ spatial distribution (posterior mean with user-defined categories).

5. Compare the results of the DM model with the simpler hierarchical model (without the spatially structured component).

   (a) To do so we need to specify and run the hierarchical model

   ```
   > formula.hier <- y ~  f(ID,model="iid")
   > model.hier <- inla(formula.hier,family="binomial",
                data=stroke, offset=offset, Ntrials=pop,
                control.compute=list(dic=TRUE))
   ```

   (b) A first comparison can be done in terms of the τ_v hyperparameter (or even better in terms of $1/\tau_v$). We can plot the posterior marginal for $1/\tau_v$ under the two model specifications as follows:

   ```
   > plot(density(inla.tmarginal(function(x) 1/x,
                      model.hier$marginals.hyper[[1]])), main="")
   > lines(density(inla.tmarginal(function(x) 1/x,
                      model.DM$marginals.hyper[[1]])),col="red")
   ```

   and we do not see large differences between the two specifications.

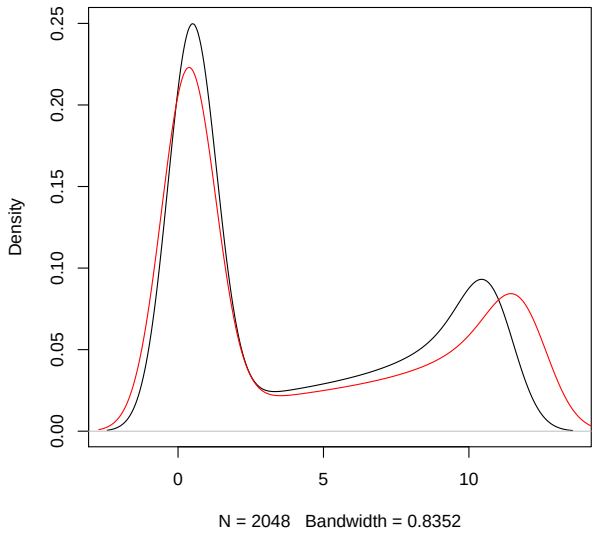

Figure 2: Posterior distribution of $1/\tau_v$ for the hierarchical (black) and DM model (red).

(c) An additional comparison can be done in terms of model fitting using the DIC which we can accessed easily from the `inla` command (note that `control.compute=list(dic=TRUE)`).

```
> compareDIC=cbind(model.hier$dic[1:4],model.DM$dic[1:4])
> colnames(compareDIC) = c("Hier", "DM")
> compareDIC
```

```
                Hier     DM
dic             4006.685 3955.784
p.eff           518.8074 483.1586
mean.deviance   3487.878 3472.625
deviance.mean   2969.07  2989.467
```

Again we do not see large differences between the two models in terms of model fitting.

5 Ecological Regression

In this section we move from disease mapping to ecological regression and we evaluate the effect of social deprivation (`Townsend`) and air pollution (`NOx`) on the risk of death for stroke:

$$\begin{aligned} y_i &\sim \text{Binomial}(\pi_i, n_i) \\ \text{logit}(\pi_i) &= b_0 + v_i + u_i + \beta_1 \text{NOx}_i + \beta_2 \text{Townsend}_i \\ v_i &\sim \text{Normal}(0, \sigma_v^2) \\ \boldsymbol{u} &\sim \text{ICAR}(\boldsymbol{W}, \sigma_u^2) \\ \beta_1 &\sim N(0, 0.5) \\ \beta_2 &\sim N(0, 0.5) \end{aligned}$$

1. Change the `formula` specification to include the two additional covariates and run INLA using the prior for β_1 and β_2 as specified above:

```
> formula.reg <- y ~  f(ID,model="bym", graph=sheffield.adj,
                        hyper=list(prec.spatial=list(param=c(1,1)))) +
                        Townsend + NOx
> model.reg <- inla(formula.reg,family="binomial",
                data=stroke, offset=offset, Ntrials=pop,
                control.compute=list(dic=TRUE),
                control.fixed=list(prec=list(mean=0, prec=0.5)))
```

2. Obtain the estimates of the effect of air pollution and social deprivation on the risk of stroke mortality:

```
> round(model.reg$summary.fixed,3)
```

```
              mean    sd 0.025quant 0.5quant 0.975quant   mode kld
(Intercept) -6.434 0.119     -6.669   -6.433     -6.201 -6.432   0
Townsend     0.050 0.027     -0.002    0.050      0.103  0.050   0
NOx          0.068 0.032      0.005    0.068      0.130  0.068   0
```

3. Calculate the posterior probability of a "positive" (i.e. above zero, not good for health!) effect of the two covariates on stroke mortality

```
> names(model.reg$marginals.fixed)
```

```
[1] "(Intercept)" "Townsend"    "NOx"
```

```
> log.Townsend <- model.reg$marginals.fixed[[2]]
> log.NOx <- model.reg$marginals.fixed[[3]]
> prob.Townsend <- 1-inla.pmarginal(0,log.Townsend)
> prob.NOx <- 1-inla.pmarginal(0,log.NOx)
> prob.Townsend
```

```
[1] 0.9698826

> prob.NOx

[1] 0.9833312
```

which returns values above 0.95 for both effects - so we can conclude that there is strong evidence of increased stroke mortality when air pollution or social deprivation increase.

4. Now compare the spatially structured residuals $\boldsymbol{u}$ from the ecological regression model and from the disease mapping model:

```
> #Natural scale
> exp.u.reg <- lapply(model.reg$marginals.random$ID[1031:2060],
                      function(x) inla.emarginal(exp,x))
> #Categories
> cutoff=c(0.5,0.8,0.95,1.05,1.2,18.5)
> exp.u.reg.cat <- cut(unlist(exp.u.reg),breaks=cutoff,
                       include.lowest=TRUE)
> data.exp.BYM <- data.frame(SP_ID=stroke$SP_ID,
                             exp.u=exp.u.cat, exp.u.reg=exp.u.reg.cat)
> row.names(data.exp.BYM) <- seq(1,1030)
> head(data.exp.BYM)

     SP_ID      exp.u   exp.u.reg
1 05CGGD02 (0.8,0.95]  (0.8,0.95]
2 05CGFT44 (0.8,0.95]  (0.8,0.95]
3 05CGGF35 (0.8,0.95] (0.95,1.05]
4 05CGGD29 (0.8,0.95] (0.95,1.05]
5 05CGFH02 (1.05,1.2]  (0.8,0.95]
6 05CGGD31 (0.8,0.95]  (0.8,0.95]

> #Merge exp(u) for the two models with the sheffield shapefile
> sheffield <- sheffield.gen
> data.sheffield <- attr(sheffield, "data")
> attr(sheffield, "data") <- merge(data.sheffield,data.exp.BYM,
                                   by="SP_ID")
```

and use `spplot` to map both quantities side by side

```
> spplot(obj=sheffield, zcol=c("exp.u","exp.u.reg"),
        col.regions= brewer.pal(5, "BrBG"), main="")
```

It seems that the $\boldsymbol{u}$ under the ecological regression model are slightly less variable than under the disease mapping model (expected as some of the spatial variability will be explained by the covariates in the ecological regression).

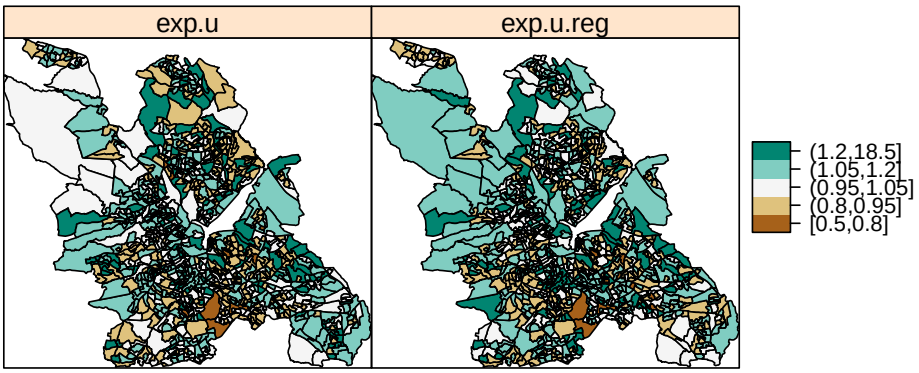

Figure 3: u_i spatial distribution under disease mapping and ecological regression (user-defined categories).

5. Has the percentage of variance explained by the spatially structured component ($\boldsymbol{u}$) gone down?

```
> marg.hyper2 <- inla.hyperpar.sample(100000,model.reg)
> perc.var.u2 <- mean(marg.hyper2[,1] /(marg.hyper2[,1]+marg.hyper2[,2]))
> perc.var.u2

[1] 0.3321348
```

It goes down just a little: remember that it was around 35% for the disease mapping while now it is about 33%, confirming what we see from the maps of $\boldsymbol{u}$.