Assignment 2 Spatial Epidemiology

Ferrara Lorenzo, Lucchini Marco

9-12-2022

Description

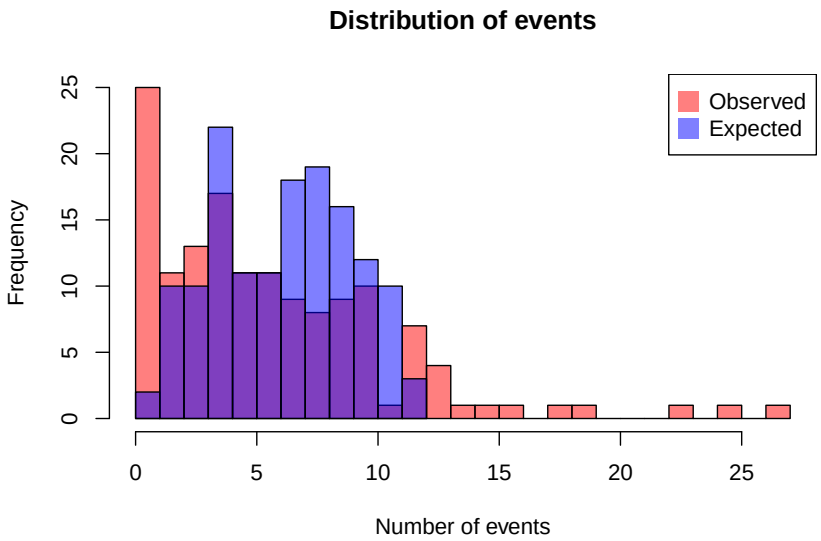
Data of the incidence of larynx cancer diagnosed during 10 years (1982-1991) in the districts of Mersey and West Lancashire in the north-east of England is in the file larynx-dates.odc. They have detected 876 cases in 144 electoral districts. The number of expected cases were calculated using internal standardization, based on the specific rates of sex and age in the zone of study (with the population of the census 1991).

1) Perform an exploratory analysis to study the possible spatial correlation of the larynx incidence (Standardized Mobility Ratio, SMR) (considering neighbors those regions that share geographic limits and the matrix of weights as the matrix standardized by rows)

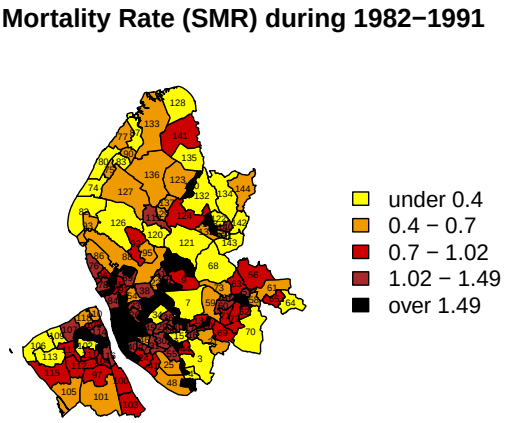
The data available to as are the following:

- O: Number of observed events
- E: Number of Expected events

In the observed regions, 0 to 27 events were registered and the median number of observations is 5. The expected value of the cases was between 0.710 and 11.700 with median 6.365. The two distribution of the data are the following:

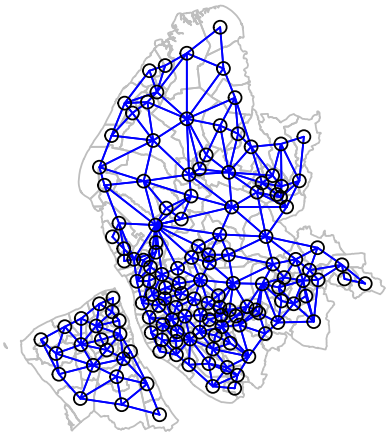


We also visualize the distribution of the data on the map of the regions:



We notice that in the center of the district there's a cluster of regions with high values of SMR, which slowly lower as we get away from the center.

Now we'll analyse the lattice structure of the data, considering neighbors those regions that share geographic limits:



After establishing the neighbors we compute the weights (standardized by row). The weights corresponding to the first four regions are:

```
## [[1]]
## [1] 0.3333333 0.3333333 0.3333333
##
## [[2]]
## [1] 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667
##
## [[3]]
## [1] 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571
##
## [[4]]
## [1] 0.2 0.2 0.2 0.2 0.2
```

Now we can perform a Moran test to prove or disprove the presence of spatial autocorrelation:

- We do it first using Moran's I statistic:

```
##
##  Moran I test under randomisation
##
## data:  NW.dataset$SMR
## weights: weight.object
##
## Moran I statistic standard deviate = 6.2244, p-value = 2.417e-10
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.313393219      -0.006993007       0.002649439
```

The I statistic is 0.313, which is quite different from the expected value (-0.0069), indeed the p-value is very low ($p = 2.4 \cdot 10^{-10}$), so we reject the null hypothesis.

- We also do it through a Monte Carlo approach:

```
##
##  Monte-Carlo simulation of Moran I
##
## data:  NW.dataset$SMR
## weights: weight.object
## number of simulations + 1: 9001
##
## statistic = 0.31339, observed rank = 9001, p-value = 0.0001111
## alternative hypothesis: greater
```

and the p-value is $p = 0.000111$, therefore also in this case we reject the null hypothesis.

Both methods lead us to the conclusion that spatial correlation is present.

2) If we consider that the number of cases of Larynx are a Poisson distribution, are data overdispersed?

We try to fit the data with a Poisson model. If the Poisson model fits the data reasonably, we would expect the residual deviance to be roughly equal to the degrees of freedom ($n - p = 143$).

```
##
## Call:
## glm(formula = O ~ 1, family = quasipoisson(link = "log"), data = data,
##     offset = E)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -12.7956   -0.0396    3.0419    5.3649   15.1299
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.0356     0.5893  -11.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 304.254)
##
##     Null deviance: 3557.1  on 143  degrees of freedom
## Residual deviance: 3557.1  on 143  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 7
```

But we can notice that the residual deviance is very large (3557.1) and that the dispersion parameter is 304.254. Therefore we can conclude that the data are overdispersed.

3) Fit the SMR taking into account the over-dispersion of the data using a heterogeneity model, a spatial (CAR intrinsic) and the convolution model. Estimate the models using Bayesian inference (Gibbs Sampling). Use three chains of initial values.

(for the sake of clarity we didn't show the estimates of RR, the Relative Risk)

- Heterogeneity model: unstructured overdispersion

```
## Inference for Bugs model at "C:/Users/lofer/OneDrive/Documenti/GitHub/Epidemiology_Assignment_2//mod
##  3 chains, each with 10000 iterations (first 1000 discarded), n.thin = 10
##  n.sims = 2700 iterations saved
##           mean   sd  2.5%   25%   50%   75% 97.5% Rhat n.eff
## alpha0    -0.1  0.1  -0.2  -0.2  -0.1  -0.1   0.0    1  2700
## tau.h      3.7  0.8   2.5   3.2   3.7   4.2   5.5    1  2700
## sigma.h    0.5  0.1   0.4   0.5   0.5   0.6   0.6    1  2700
## deviance 616.5 16.0 586.8 605.2 615.8 627.2 649.8    1  2700
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
```

```
## DIC info (using the rule, pD = Dbar-Dhat)
## pD = 80.6 and DIC = 697.1
## DIC is an estimate of expected predictive error (lower deviance is better).
```

- Spatial model: we model a structured overdispersion, using the spatial structure we have created above (the neighbors, the weights and the corresponding adjacency matrix)

```
## Inference for Bugs model at "C:/Users/lofer/OneDrive/Documenti/GitHub/Epidemiology_Assignment_2//mode
##  3 chains, each with 7000 iterations (first 1000 discarded), n.thin = 8
##  n.sims = 2250 iterations saved
##           mean   sd  2.5%   25%   50%   75% 97.5% Rhat n.eff
## alpha0    -0.2  0.0  -0.3  -0.2  -0.2  -0.1  -0.1    1  2200
## tau.s      2.1  0.6   1.2   1.7   2.0   2.4   3.4    1  1700
## sigma.s    0.7  0.1   0.5   0.6   0.7   0.8   0.9    1  1700
## deviance 619.2 13.9 593.6 609.7 618.8 628.4 647.7    1  2200
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = Dbar-Dhat)
## pD = 56.6 and DIC = 675.9
## DIC is an estimate of expected predictive error (lower deviance is better).
```

- Convolution model: we create a model which takes into account both a structured (spatial) overdispersion and an unstructured (heterogeneity) one

```
## Inference for Bugs model at "C:/Users/lofer/OneDrive/Documenti/GitHub/Epidemiology_Assignment_2//mode
##  3 chains, each with 10000 iterations (first 1000 discarded), n.thin = 8
##  n.sims = 3375 iterations saved
##           mean    sd  2.5%   25%   50%   75%  97.5% Rhat n.eff
## alpha0    -0.2   0.0  -0.3  -0.2  -0.2  -0.1   -0.1    1   770
## tau.s      2.5   0.9   1.3   1.9   2.3   2.9    4.7    1   870
## sigma.s    0.7   0.1   0.5   0.6   0.7   0.7    0.9    1   870
## tau.h    217.4 420.8   9.8  27.8  71.2 207.7 1369.8    1   570
## sigma.h    0.1   0.1   0.0   0.1   0.1   0.2    0.3    1   570
## deviance 615.3  14.2 588.3 605.9 615.1 624.6  643.5    1  3400
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = Dbar-Dhat)
## pD = 59.1 and DIC = 674.5
## DIC is an estimate of expected predictive error (lower deviance is better).
```

4) Check the convergence and study the autocorrelation of the chain. Decide which model fit better the data.

To study the convergence we look both at the traceplots of the fitted parameters (the 3 chains should stabilize after some time around the same value) and at the $\hat{R}$ coefficient of the parameters of the models (we'll assume that convergence is reached if $\hat{R} < 1.05$)

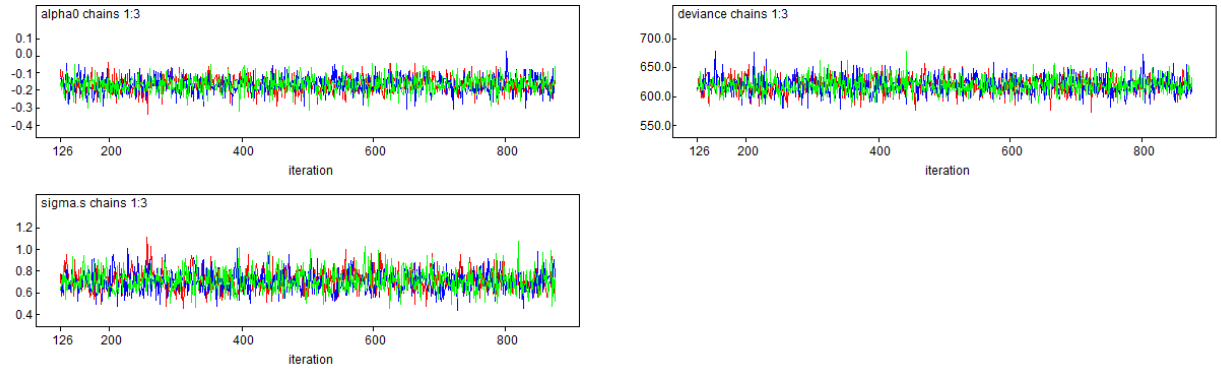
To study the autocorrelation of the chain we look at the autocorrelation plots which WinBugs gives us and verify that the values of the correlations of the chain in different iterations are low, meaning that different iterations are not or almost not correlated with each other.

- Spatial:

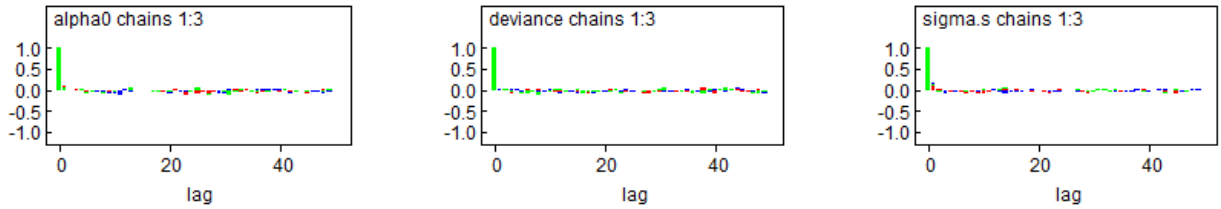
All the $\hat{R}$ coefficients are below 1.05, indeed the maximum value we get is:

```
## [1] 1.004156
```

and the traceplots are:



The autocorrelation plots are:



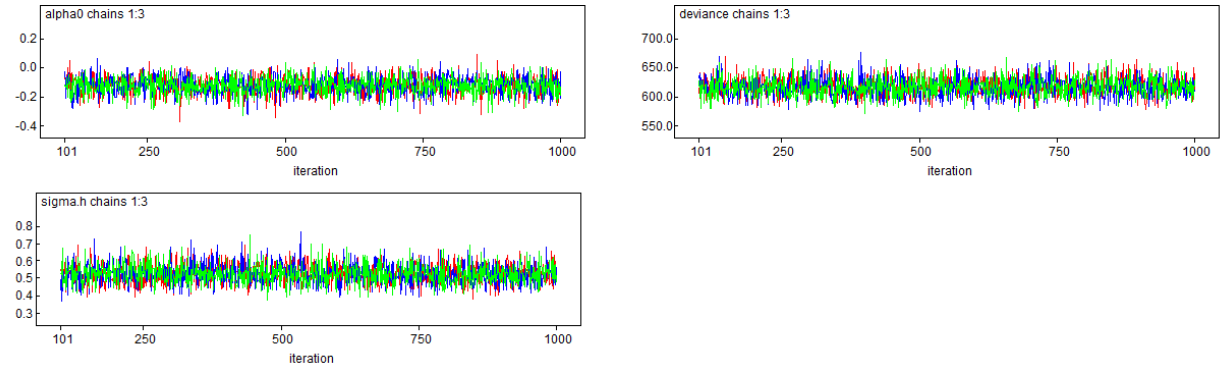
The chains of the studied parameters seem to have converged, oscillating around a value, and the correlation plots don't present any kind of correlations between contiguous iterations, since all values are low.

- Heterogeneity:

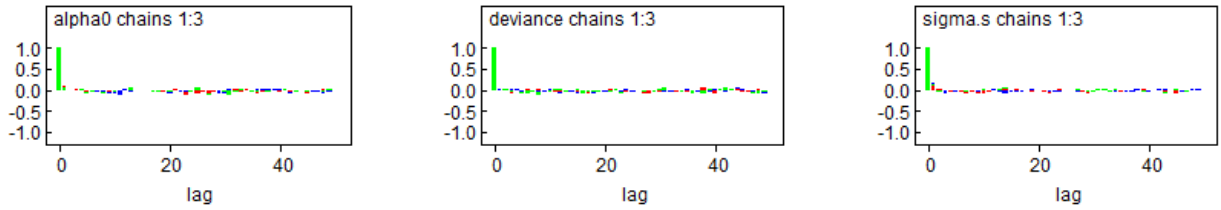
All the $\hat{R}$ coefficients are below 1.05, indeed the maximum value we get is:

```
## [1] 1.004661
```

and the traceplots are:



The autocorrelation plots are:



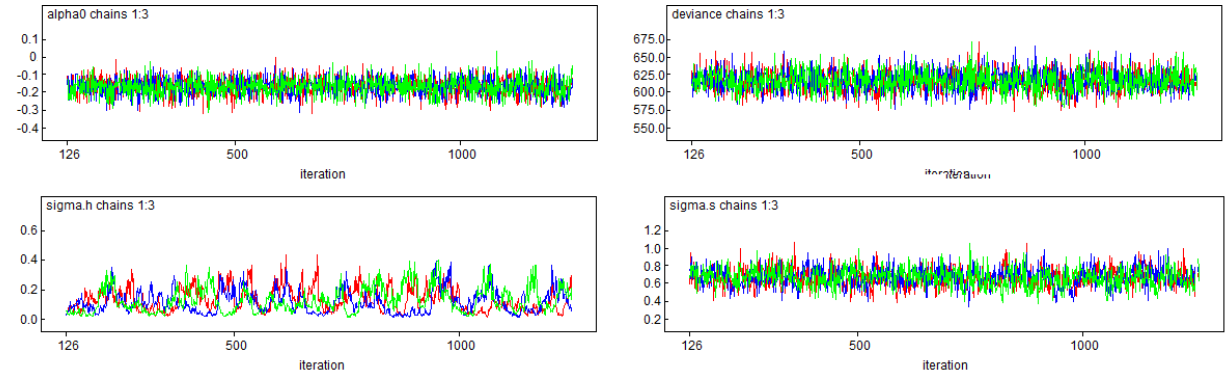
The chains of the studied parameters seem to have converged, oscillating around a value, and the correlation plots don't present any kind of correlations between contiguous iterations, since all values are low.

- Convolutional:

All the $\hat{R}$ coefficients are below 1.05, indeed the maximum value we get is:

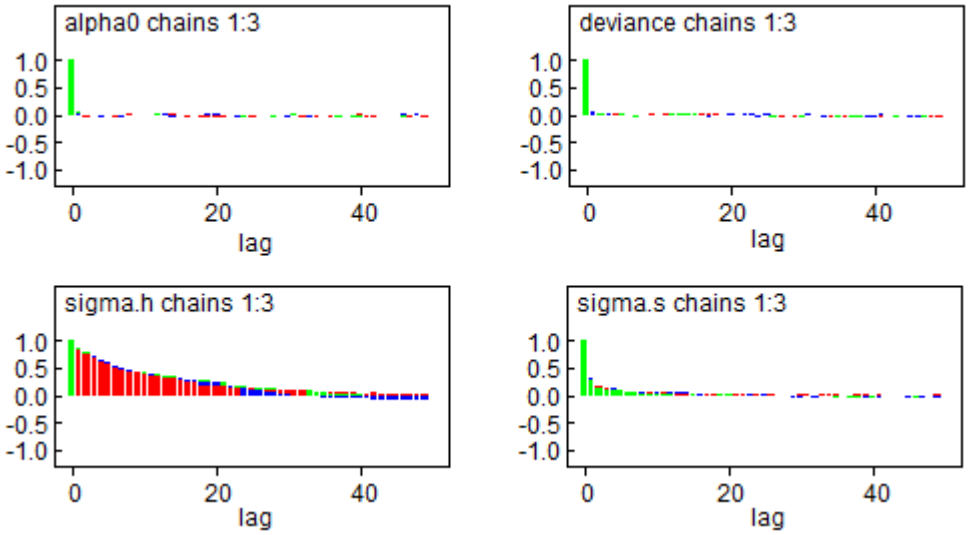
```
## [1] 1.016474
```

and the traceplots are:



Since there are more than one random effect, the convergence of the parameters is harder to reach because the Monte Carlo Markov Chain struggles to find a stable value of sigma_h to converge to.

The autocorrelation plots are:



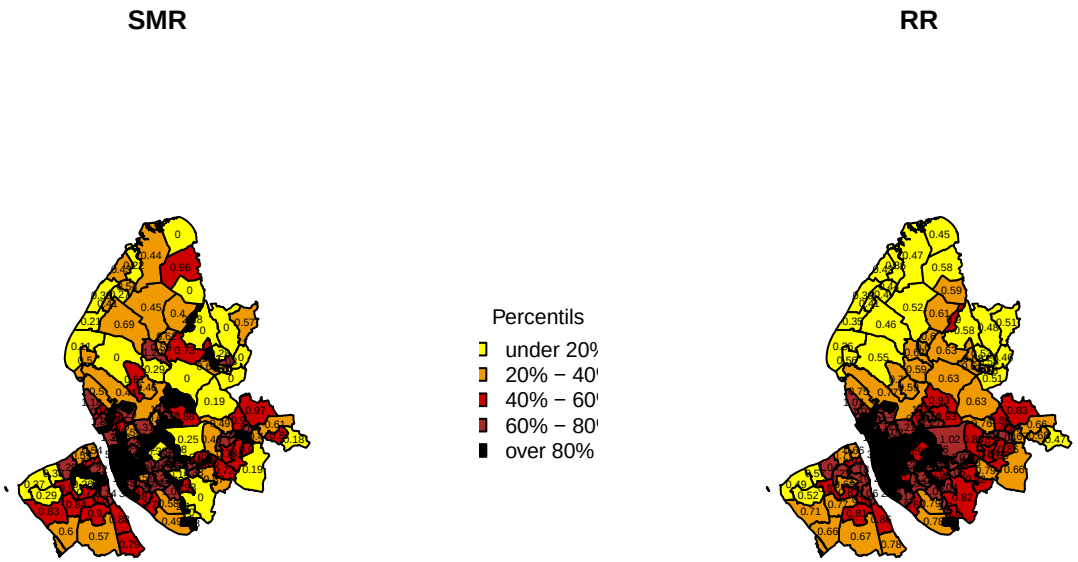
We also have issues with the autocorrelation of the chains for sigma_h: even using a high value for the thin parameter it's difficult to find a chain with no or little autocorrelation. So we'll avoid choosing the Convolutional model.

We can also evaluate the DIC (Deviance Information Criterion):

```
##            type     DIC
## 1       spatial 675.889
## 2 heterogeneity 697.117
## 3 convolutional 674.470
```

The DIC index is an estimate of expected predictive error, so we'll choose the model with the lowest DIC. The Spatial model and the Convolutional model are almost comparable, but even though the Convolutional has a slightly lower DIC we'll choose the Spatial, due to the two considerations we have made above.

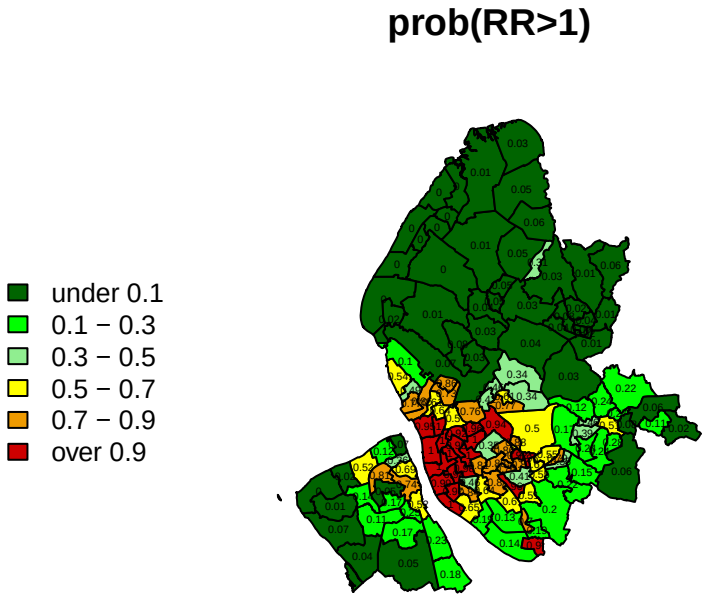
5.1) Using the model that better fit the data: Make a plot of the observed SMRs and Relative Risk estimated by the model (mean posterior of SMR), using the percentiles of the values. Compare them and discuss the differences.



The values of RR estimated by the model follow a radial pattern, with higher values in the center and gradually lower values far from the center. Its pattern is smoother than the one followed by the observed values of SMR, which presents many low-value hotspots in the center and some high-value clusters in the north.

We could say that the SMR has two main clusters (center and north), of which only one is recognized by the model (the central one). Nonetheless, from a qualitative point of view, the chosen model can describe well the phenomenon, even though just in a rough way and not so in detail.

5.2) Make a plot showing the posterior probability of relative risk of the electoral districts that exceed 1. What do you observe?



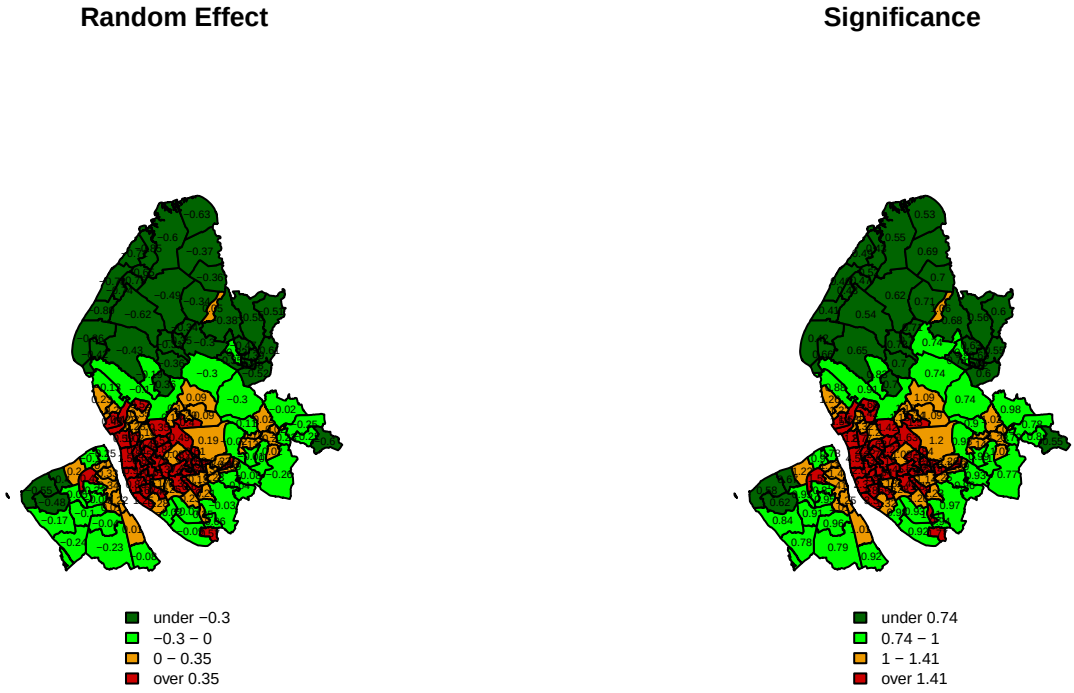
Let's analyse the meaning of the SMR: a ratio greater than 1 indicates that more mortality has occurred expected, therefore we can state with pretty high probability that the highlighted regions have an increase in the value of the observed cases ($p > 50\%$ for the yellow regions, $p > 70\%$ for the orange ones, $p > 90\%$ for the red ones).

We may suppose the presence of an underlying phenomenon which is causing this increase of the mortality due to larynx cancer. For example it may be linked to some change of the environmental/working conditions in the areas under investigation.

5.3) Make a plot showing the random effects for each districts (exponential scale) and another with the significance of this parameters. What do you observe?

The random effect is the spatial effect $s[i], \forall i \in \{1, .., 144\}$. If we want to see it in exponential scale, we'll use $resid[i] = e^{s[i]}$.

Notice that the formula for RR is $RR[i] = e^{\alpha_0 + s[i]} = e^{\alpha_0} \cdot e^{s[i]}$ and the multiplying factor is exactly $e^{s[i]}$.



These plots show how the spatial effect affects the values of the RR estimated by the model:

- In the regions where $s[i] > 0$, so $e^{s[i]} > 1$, RR is positively influenced by the spatial effect, leading to a higher number of deaths than expected. We also notice that the higher $e^{s[i]}$, the stronger the influence.
- Viceversa in the regions where $s[i] < 0$, so $e^{s[i]} < 1$, RR is negatively influenced by the spatial effect, leading to a lower number of deaths than expected.