Disease Mapping with R-INLA Under Bayesian Framework: With Application to Lung Cancer Mortality in Pennsylvania

Introduction

Disease risk mapping analyses can help people to have a better understanding of the spatial variation of the disease and allow people to identify the risk factors of important public health events. These analyses could inspire further study and policy making in the effort of disease prevention and control. Previous studies have applied various spatial data to investigate the influenza and cancer in developed countries [1] and the neglected tropical diseases [2].

Spatial data, unlike other structural data, has particular characteristics that make the modeling more complex than in conventional applications. Under the geographic background, the topological relationships and other spatial relationship are of fundamentally significance in order to define rules of spatial integrity. There may exist high spatial dependency among observations since closer objects tend to have a more positive correlation than those further apart, which violates the standard statistical assumptions of independence among observations and makes spatial data very special.

The analysis of the spatial data can be dated back to the time when Dr. John Snow investigated the outbreak of the cholera. It is a type of geographical analysis that seeks to understand the relationships between human behaviors and their patterns in the spatial context. Due to the characteristics of the spatial data, the Bayesian approaches have been very popular in spatial modelling. On the one hand, unlike the frequentist approaches, the Markov chain Monte Carlo (MCMC) methods are very general and can effectively be applied to nearly any model. In theory, MCMC can provide nearly exact inference. Besides, standard MCMC sampler are generally easy to program and are implemented in readily available software such as BUGS, WinBUGS, OpenBUGS, JAGS, etc. However, in practice, these advantages have to be balanced with model complexity and running time. Especially for large data or very complex structure (e.g. hierarchical models), the MCMC methods requires much longer computational time.

Integrated Nested Laplace Approximation (INLA)

Possible solutions to the computational efficiency include more complex model specification (blocking, overparameterization), and more complex sampling schemes (Hamiltonian Monte Carlo). Integrated Nested Laplace Approximations (INLA) is an alternative method of inference that is much more computationally efficient compared to MCMC.

INLA is a fast alternative to MCMC for the general class of latent Gaussian models (LGMs). A lot of models can be re-written to look like LGMs, such as, generalized linear mixed models, generalized linear additive models, spatio-temporal models, survival analysis, spline smoothing, etc.

INLA made the computation more efficient from 2 aspects, one is embedded in the use of LGMs and Gaussian Markov Random Fields (GMRFs). The other is the use of Laplace Approximations.

1. LGMs & GMRF

Latent Gaussian Models generally can be represented by a hierarchical structure with 3 stages. The first stage is the conditionally independent likelihood function, which can be written as $y|\theta,\psi \sim p(y|\theta,\psi) = \prod_{i=1}^{n} p(y_i|\theta,\psi)$. For the second stage, it is assumed that the parameters can be described by GMRF. GMRF is simply a Gaussian with additional conditional independence properties, meaning that any 2 elements of θ, θ_l and θ_m are independent given the remaining elements. This property makes sure the precision matrix (inverse of the variance covariance matrix) is a sparse matrix, which is a huge computational benefit, as calculations involving a dense matrix are much more costly than when a sparse matrix is used. The second stage can be formulated as: $\theta|\psi \sim Normal(0, \Sigma(\psi))$, $\theta_l \perp \theta_m|\theta_{-lm}$. The third stage is the prior distribution assigned to the hyperparameters.

2. Laplace Approximation

Another way to make the computation more efficient is the use of Laplace approximation. In order to estimate the posterior distribution $p(\theta_i|y) = \int p(\psi|y)p(\theta_i|\psi,y)d\psi$, $p(\psi|y)$ and $p(\theta_i|\psi,y)$ need to be estimated and then take numerical integration with respect to ψ. Take calculating $p(\psi|y)$ as an example, by Bayes rule, $p(\psi|y) = \frac{p(\theta,\psi|y)}{p(\theta|\psi,y)}$, and the numerator can be expanded as $\frac{p(\psi)p(\theta|\psi)p(y|\theta,\psi)}{p(\theta|\psi,y)}$. For the denominator, instead of estimating the actual PDF, Laplace approximation is used to approximate it with a normal distribution using the first 3 terms of the Taylor series at its mode. Then we can obtain the following estimate:

$$p(\psi|y) = \frac{p(\theta,\psi|y)}{p(\theta|\psi,y)} \propto \frac{p(\psi)p(\theta|\psi)p(y|\theta,\psi)}{p(\theta|\psi,y)} \approx \frac{p(\psi)p(\theta|\psi)p(y|\theta,\psi)}{\tilde{p}(\theta|\psi,y)}|_{\theta=\theta^*(\psi)} =: \tilde{p}(\psi|y)$$

Real data application

We used R-INLA to look at the association between 2012 PA lung cancer mortality and 2002 smoking percentage and Radon level.

1. Data

For smoking percentage, we used the 2002 county specific smoking percentage data obtained from the pennLC dataset in the SpatialEpi package in R. We obtained the 1989-2017 Radon level data from the PA Department of Environmental Protection (DEP). Since we only have the 2002 smoking percentage data, we only used the 2002 Radon level data. For the outcome, we obtained 1989-2016 PA lung cancer mortality data at county level from Pitt Public Health's Mortality information and Research Analytics (MOIRA) System (moira.pitt.edu). We assumed a 10-year lag of the effects of smoking and Radon level, so we used the 2012 lung cancer mortality data.

2. Model

A very general way of specifying the problem is by modeling the mean for the i-th unit by means of an additive linear predictor, defined on a suitable scale (log link for Poisson distribution, etc.). The form is as follows:

$$\eta_i = \alpha + \sum_{m=1}^{M} \beta_m x_{mi} + \sum_{l=1}^{L} f_l(z_{li}),$$

where α is the intercept, $\beta_1, \dots, \beta_M$ quantify the effect of $x_1, \dots, x_M$ on the response, $f_1(\cdot), \dots, f_L(\cdot)$ is a set of functions defined in terms of some covariates $z_1 \dots, z_L$, and assume $\theta = (\alpha, \beta, f) \sim GMRF(\psi)$.

For spatial models, we need 2 components of $f_l(\cdot)$, where $f_1(\cdot)$ follows a conditional autoregressive structure for the spatially structured effects and $f_2(\cdot)$ follows a normal distribution with mean 0 and variance $\sigma_{f_2}^2$ for the unstructured residuals.

Thus, for our lung cancer mortality application in PA, we fit the following model:

$$\begin{aligned}
&Y_i \sim Poisson(\lambda_i E_i),\\
&log(\lambda_i) = \beta_0 + \beta_1 * smoke + \beta_2 * radon + U_i + V_i,\\
&U_i \sim CAR:\ U_i | U_{j \neq i} \sim N(n_i^{-1} \textstyle\sum_{N_i} U_j,\ n_i^{-1}\sigma_U^2)\\
&V_i \sim_{iid} N(0, \sigma_V^2),
\end{aligned}$$

where Y_i and E_i are the observed and expected number of mortality cases county i, λ_i is the mortality rate for that county relative to the expected mortality. Fixed effect include intercept, β_1 for smoking percentage and β_2 for radon level. U_i are random effects from the spatial structure which we assumed a conditional independent autoregressive structure. V_i are spatially independent random effects following a normal distribution with mean 0 and variance σ_V^2. Priors used for the fixed effects is Normal(0, 0.001). For the random effects, $log\ \sigma_U^{-2} \sim logGamma(1, 0.00005)$ and $log\ \sigma_V^{-2} \sim logGamma(1,\ 0.00005)$.

3. Results

First, we looked at the descriptive statistics. From figure 1, we can see that Allegheny county and Philadelphia have the highest lung cancer mortality since these are the 2 most populated areas in PA. To exclude the effect of different population size between counties, we calculated the standardized incidence ratio (SIR) which is the ratio between observed and expected incidence at each county. The expected counts were calculated based on the assumption that there is no between county difference in lung cancer mortality, so the expected mortality is proportional the county population. The corresponding SIR of lung cancer mortality is shown in Figure 2.

Then we fitted the spatial model as described above with R-INLA. The posterior median for smoking percentage is 3.64 with 95% credible interval (1.49, 5.77). The posterior median for Radon level is 0.001 with 95% credible interval (-0.002, 0.004). Even though the 95% credible interval for Radon effect cross 0, it still leans towards the positive side. Thus, we concluded that there is likely to be a positive relationship between smoking percentage, Radon level and lung cancer mortality, which means that higher smoking percentage or higher radon level is associated with higher lung cancer mortality rate. The relative risk for lung cancer mortality for each county is shown in Figure 3, from which we can see that the 2 counties with the highest risks are Venango

county in the north west of PA with RR 1.48 and 95% credible interval (1.18, 1.84) and Fayette in the south west of PA with RR 1.44 and 95% credible interval (1.21, 1.69).

Conclusions and Limitations

In the paper, we discussed the characteristics of spatial data and Bayesian approaches in spatial modelling. Due to the complex hierarchical structure of the spatial data, the Bayesian inference may be performed using the INLA approach, which is a computational alternative to MCMC that allows to do approximate Bayesian inference in LGMs. This approach is implemented in the R package called INLA.

We then gave a practical example where we modeled the risk of lung cancer mortality in PA in 2012 with a Bayesian spatial model. The spatial correlation was accounted by a conditional independent autoregressive structure for neighboring counties. A weakly-informative prior was used in INLA for model inference. As we can see from the results, both the proportion of smokers and the level of radon in each county were positively associated with lung cancer mortality. Counties in western PA in general showed a higher risk compared to eastern counties.

However, there are a few limitations of our project. Firstly, the data quality was not good enough since the smoking data was estimated by a very small sample of population in each county (i.e. hundreds) and the degree of underreporting of lung cancer mortality was unmeasured. Secondly, there were limited covariates in our model. In our model, we only incorporated the proportion of smokers and the radon level as covariates. There may exist other potentially important demographic variables such as air pollution, levels of arsenic in the drinking water that could affect the mortality. Besides, we failed to consider the temporal effect of the exposure and naively created a 10-year time lag between exposure variables and outcome. However, this time effect could be further investigated if we could have more information on the exposure status.

References

[1] P. Moraga and M. Kulldorff. Detection of Spatial Variations in Temporal Trends with a Quadratic Function. Statistical Methods for Medical Research, 25(4):1422–1437, 2016.
[2] J. E. Hagan, P. Moraga, F. Costa, N. Capian, G. S. Ribeiro, E. A. Wunder Jr, R. D. Felzemburgh, R. B. Reis, N. Nery, F. S. Santana, D. Fraga, B. L. dos Santos, A. C. Santos, Q. A., W. Tassinari, M. S. Carvalho, M. G. Reis, P. J. Diggle, and A. I. Ko. Spatio-Temporal Determinants of Urban Leptospirosis Transmission: Four-Year Prospective Cohort Study of Slum Residents in Brazil. Public Library of Science: Neglected Tropical Diseases, 10(1):e0004275, 2016. [1] Moraga, P. (2018). Small Area Disease Risk Estimation and Visualization Using R. *The R Journal*, *10*(1), 495-506.
[3] Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, *4*, 395-421.

[4] Ferkingstad, E., Held, L., & Rue, H. (2017). Fast and accurate Bayesian model criticism and conflict diagnostics using R-INLA. *Stat*, *6*(1), 331-344.
[5] An introduction to INLA with a comparison to JAGS (retrieved on April 10th, 2019) http://www.statistica.it/gianluca/Talks/INLA.pdf
[6] A gentle INLA tutorial (retrieved on April 10th, 2019) https://www.precision-analytics.ca/blog-1/inla

Figures

Figure 1. Map of Observed PA lung cancer mortality (2012)

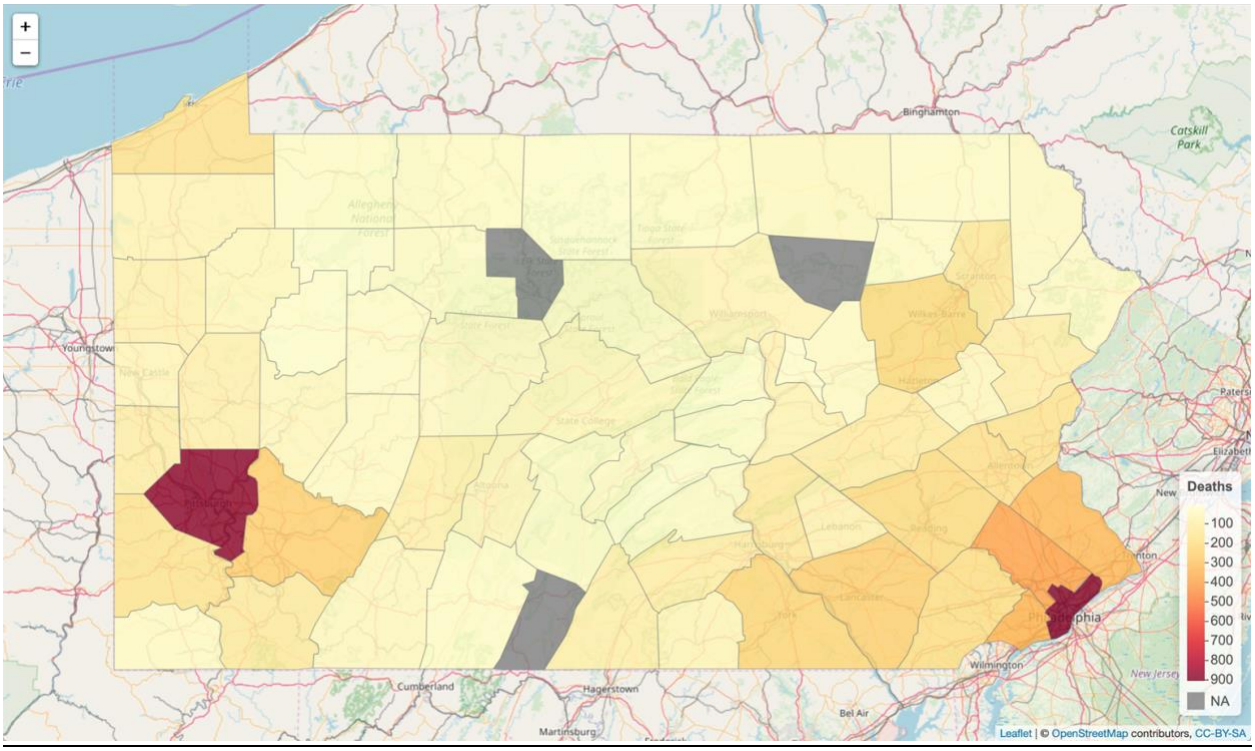


Figure 2. Map of Standardized incidence ratio of lung cancer mortality (2012)

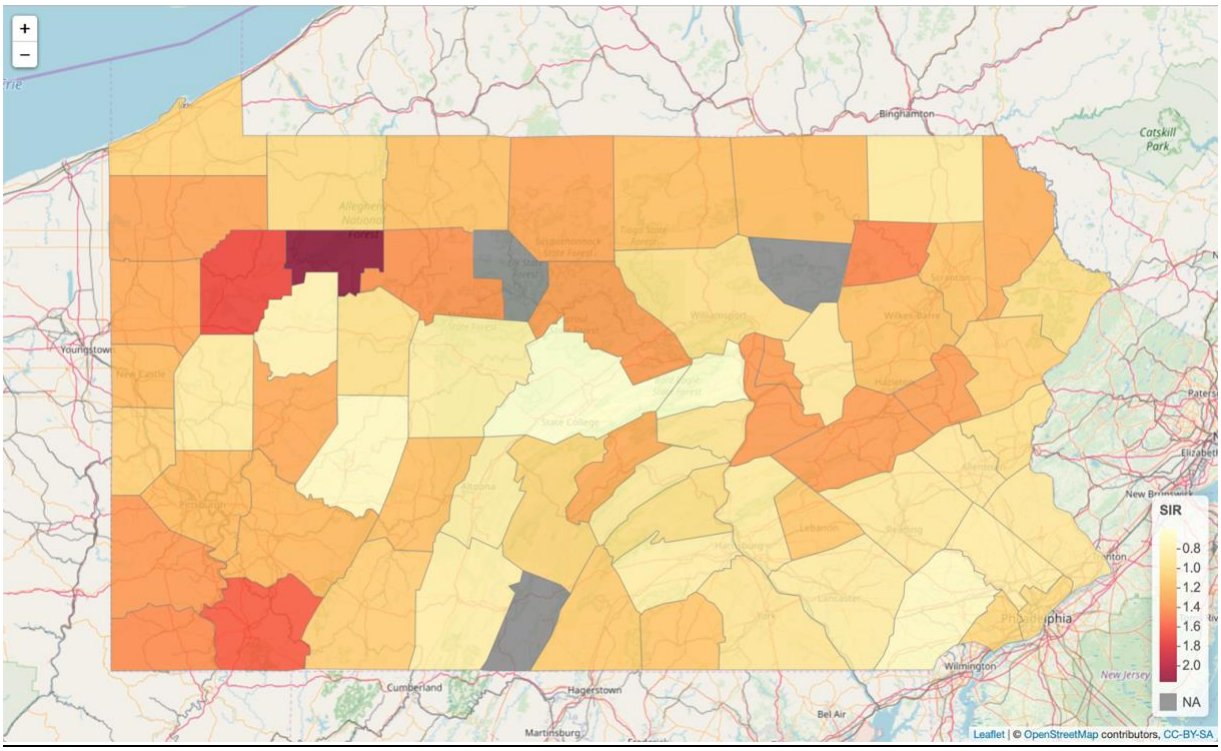
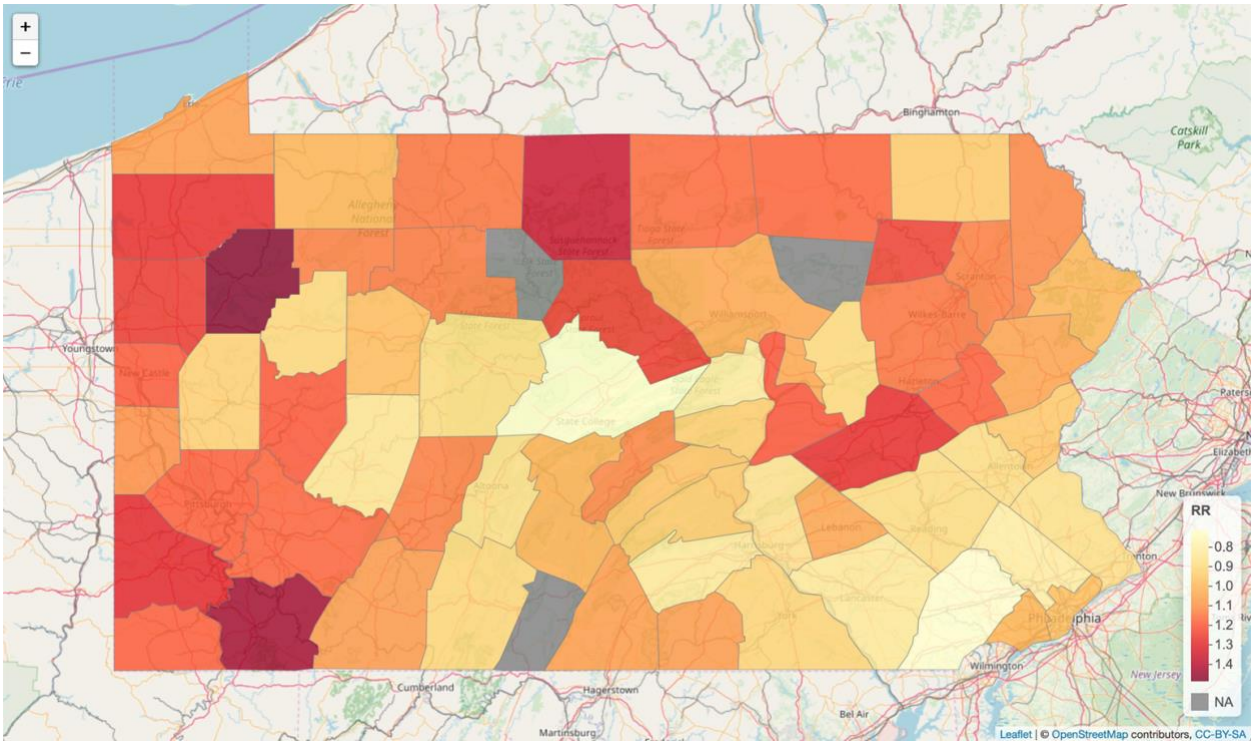


Figure 3. Map of Relative Risk for Lung Cancer Mortality



Rcode

```
install.packages("INLA", repos = "https://inla.r-inla-download.org/R/stable",
            dep = TRUE)
library(SpatialEpi)
library(INLA)
library(sp)
library(leaflet)
require(spdep)
library(dplyr)
library(ggplot2)

load('DataWithRadon.RData')


#extract data
d <- d1[d1$Year==2012,]
d$Age_Adjusted_Rate <- as.numeric(d$Age_Adjusted_Rate)
d$Population <- as.numeric(d$Population)
d[is.na(d$Deaths),]$Population <- NA
d$Deaths <- as.numeric(d$Deaths)

#expected count
E <- d$Population*sum(d$Deaths, na.rm=TRUE)/sum(d$Population, na.rm=TRUE)
d$E <- E
d$SIR <- d$Deaths/d$E


#add to map
d <- as.data.frame(d)
rownames(d)<- d$id
map <- SpatialPolygonsDataFrame(pennLC$spatial.polygon, d, match.ID = TRUE)
head(map@data)


#mapping variable
######## raw count
l <- leaflet(map) %>% addTiles()
pal <- colorNumeric(palette = "YlOrRd", domain = map$Deaths)
l %>% addPolygons(color = "grey", weight = 1, fillColor = ~pal(Deaths),
              fillOpacity = 0.8) %>%
  addLegend(pal = pal, values = ~Deaths, opacity = 0.8, title = "Deaths",
          position = "bottomright")

######## expected count
l <- leaflet(map) %>% addTiles()
pal <- colorNumeric(palette = "YlOrRd", domain = map$E)
l %>% addPolygons(color = "grey", weight = 1, fillColor = ~pal(E),
              fillOpacity = 0.8) %>%
  addLegend(pal = pal, values = ~E, opacity = 0.8, title = "Expected",
          position = "bottomright")

######### SIR
l <- leaflet(map) %>% addTiles()
```

```
pal <- colorNumeric(palette = "YlOrRd", domain = map$SIR)
l %>% addPolygons(color = "grey", weight = 1, fillColor = ~pal(SIR),
              fillOpacity = 0.8) %>%
  addLegend(pal = pal, values = ~SIR, opacity = 0.8, title = "SIR",
            position = "bottomright")


#neighborhood matrix
nb <- poly2nb(map)
head(nb)
nb2INLA("map.adj", nb)
g <- inla.read.graph(filename = "map.adj")

#inference using INLA
map$re_u <- 1:nrow(map@data)
map$re_v <- 1:nrow(map@data)

########### SIR
formula <- Deaths ~ smoking + Radon + f(re_u, model = "besag", graph = g) + f(re_v, model = "iid")
res <- inla(formula, family = "poisson", data = map@data[!is.na(E),], E=E,
            control.predictor = list(compute = TRUE))
summary(res)

marginal <- inla.smarginal(res$marginals.fixed$smoking)
marginal <- data.frame(marginal)
png('posterior distribution of smoking.png', res=300, width=1200, height=1000)
ggplot(marginal, aes(x = x, y = y)) + geom_line() +
  labs(x = expression(beta[1]), y = "Density") +
  geom_vline(xintercept = 0, col = "blue") + theme_bw() +
  ggtitle('Posterior distribution of Beta1')
dev.off()

marginal <- inla.smarginal(res$marginals.fixed$Radon)
marginal <- data.frame(marginal)
png('posterior distribution of radon.png', res=300, width=1200, height=1000)
ggplot(marginal, aes(x = x, y = y)) + geom_line() +
  labs(x = expression(beta[2]), y = "Density") +
  geom_vline(xintercept = 0, col = "blue") + theme_bw() +
  ggtitle('Posterior distribution of Beta2')
dev.off()



# mapping disease Risk
map$RR <- c(res$summary.fitted.values[, "mean"][1:11],NA,
            res$summary.fitted.values[, "mean"][12:27],NA,
            res$summary.fitted.values[, "mean"][28:54],NA,
            res$summary.fitted.values[, "mean"][55:64])
map$LL <- c(res$summary.fitted.values[, "0.025quant"][1:11],NA,
            res$summary.fitted.values[, "0.025quant"][12:27],NA,
            res$summary.fitted.values[, "0.025quant"][28:54],NA,
            res$summary.fitted.values[, "0.025quant"][55:64])
map$UL <- c(res$summary.fitted.values[, "0.975quant"][1:11],NA,
```

```
        res$summary.fitted.values[, "0.975quant"][12:27],NA,
        res$summary.fitted.values[, "0.975quant"][28:54],NA,
        res$summary.fitted.values[, "0.975quant"][55:64])

pal <- colorNumeric(palette = "YlOrRd", domain = map$RR)
labels <- sprintf("<strong> %s </strong> <br/> Observed: %s <br/> Expected: %s <br/> Smokers proportion: %s
 <br/> Radon level: %s <br/> SIR: %s <br/> RR: %s (%s, %s)", map$County, map$Deaths,  round(map$E, 2),
 map$smoking, round(map$Radon,2), (map$SIR, 2), round(map$RR, 2), round(map$LL, 2), round(map$UL,
 2)) %>%
 lapply(htmltools::HTML)
leaflet(map) %>% addTiles() %>%
 addPolygons(color = "grey", weight = 1, fillColor = ~pal(RR),  fillOpacity = 0.8, highlightOptions =
 highlightOptions(weight = 4), label = labels, labelOptions = labelOptions(style = list("font-weight" = "normal",
 padding = "3px 8px"), textsize = "15px", direction = "auto")) %>%
 addLegend(pal = pal, values = ~RR, opacity = 0.8, title = "RR",
        position = "bottomright")
```