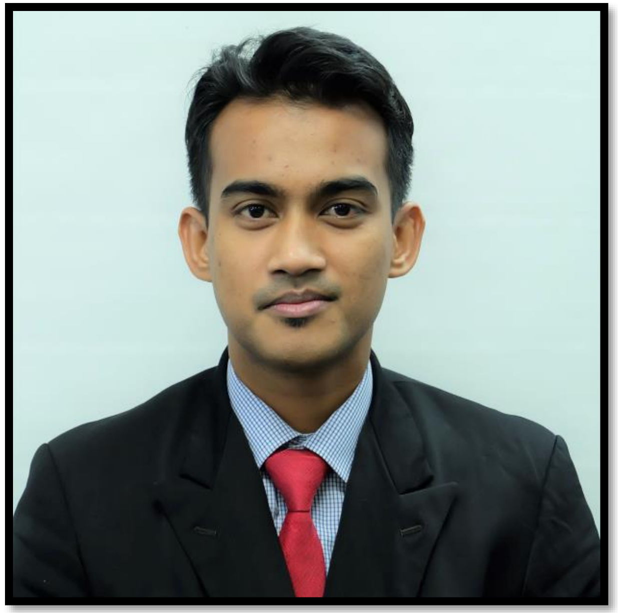
R CONFERENCE

ORGANIZED BY MALAYSIAN R USER GROUP (MYRUG)

R-INLA for Spatial Data Analysis



Dr Mohamad Afiq Amsyar

MBBS (MAHE)
MPH (USM)

Department of Community Medicine

Universiti Sains Malaysia


ASIA
NORTH POLE
90°
PRIME MERIDIAN
60°
EUROPE
30°
AFRICA
0°
39° PARALLEL
NORTH AMERICA
PACIFIC OCEAN
WASHINGTON, D.C.
39° North Latitude
77° West Longitude
ATLANTIC OCEAN
150°
120°
EQUATOR
90°
60°
30°
77° MERIDIAN
SOUTH AMERICA

All About Relations

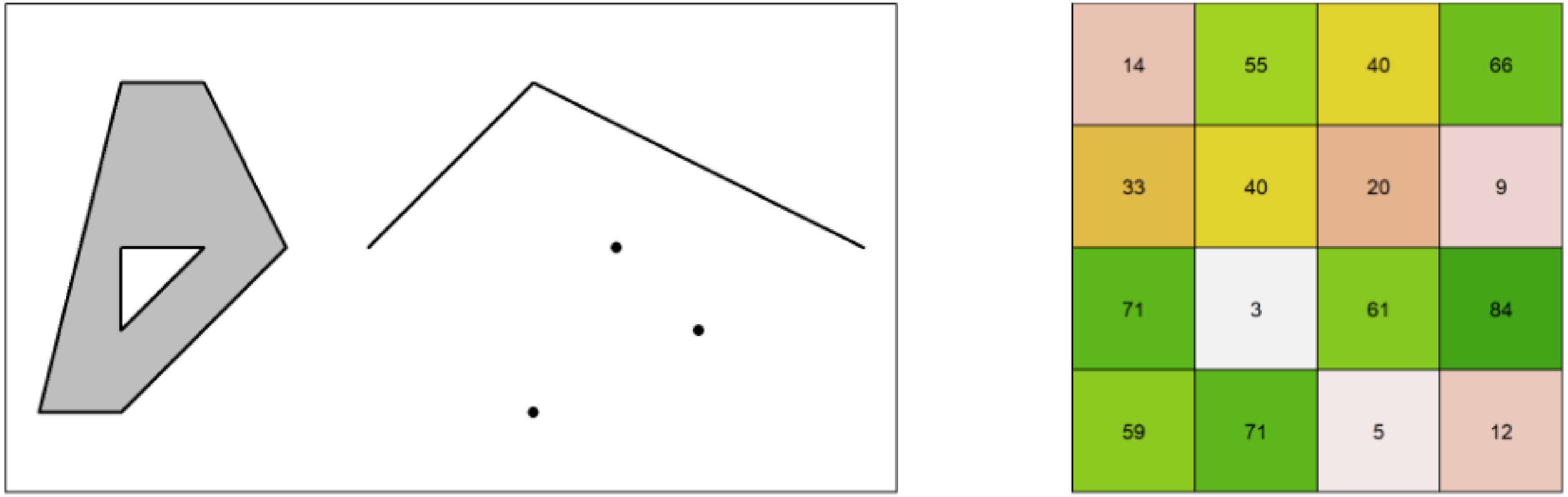
- Points
- Lines
- Polygon

Today’s Goal

- Draw Maps using R
- Compute Relations using R-INLA
- Model and Mapping the relations

Spatial data in R

- Spatial data can be represented using vector and raster data
- Vector data displays points, lines and polygons, and associated information
  Examples: locations of monitoring stations, road networks, municipalities
- Raster data are regular grids with cells of equal size that are used to store values of spatially continuous phenomena
  Examples: elevation, temperature, air pollution values



R packages: **sf** (vector data) and **terra** (raster and vector data)

CRS in R

- We need to know if the data is projected
  - which projection system is used
- if not, you need to project it
  - use the correct system
- most common projection is WGS84
  - good for mapping global data
- When data with different CRS are combined
  - need to transform them to a common CRS so they align with one another

Integrated Nested Laplace Approximation (INLA)

- **Purpose**: INLA is a fast computational method for **approximate Bayesian inference** in complex models, such as:
  - **Generalized Linear Mixed Models (GLMMs)**
  - **Spatial and Spatio-Temporal Models**
- **How it Works**:
  - INLA combines **analytical approximations** and **numerical integration** to approximate posterior distributions.
  - **Speed Advantage**: Much faster than traditional **Markov Chain Monte Carlo (MCMC)** methods, especially for large datasets.

R-INLA Package

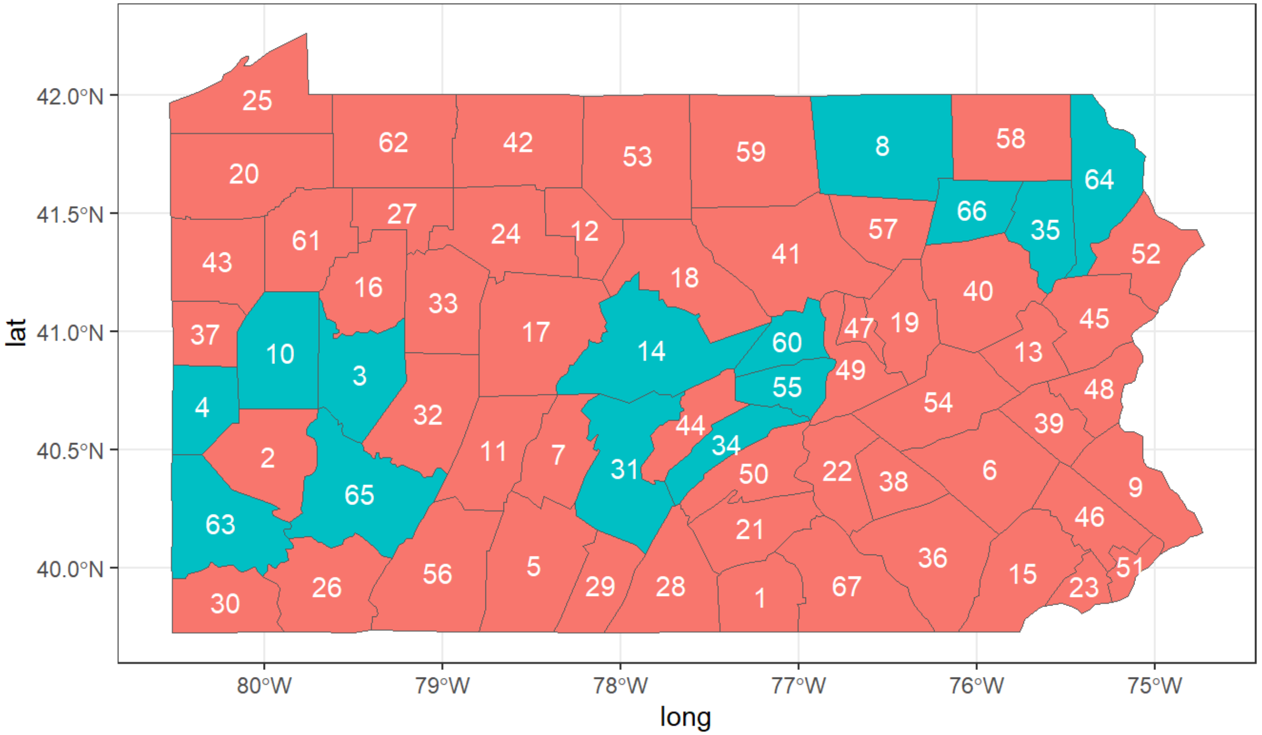
```
install.packages("INLA",
repos = "https://inla.r-inla-download.org/R/stable",
dep = TRUE)

library(INLA)
```

Areal Data

- Areal data is common in disease mapping applications where often, for confidentiality reasons
- Disease data can be used to construct atlases that show the geographic distribution of aggregated outcomes to understand spatial patterns, identify high-risk areas, and reveal inequalities

Spatial neighborhood matrices

```r
library(SpatialEpi)
map <- pennLC$spatial.polygon
plot(map)
```



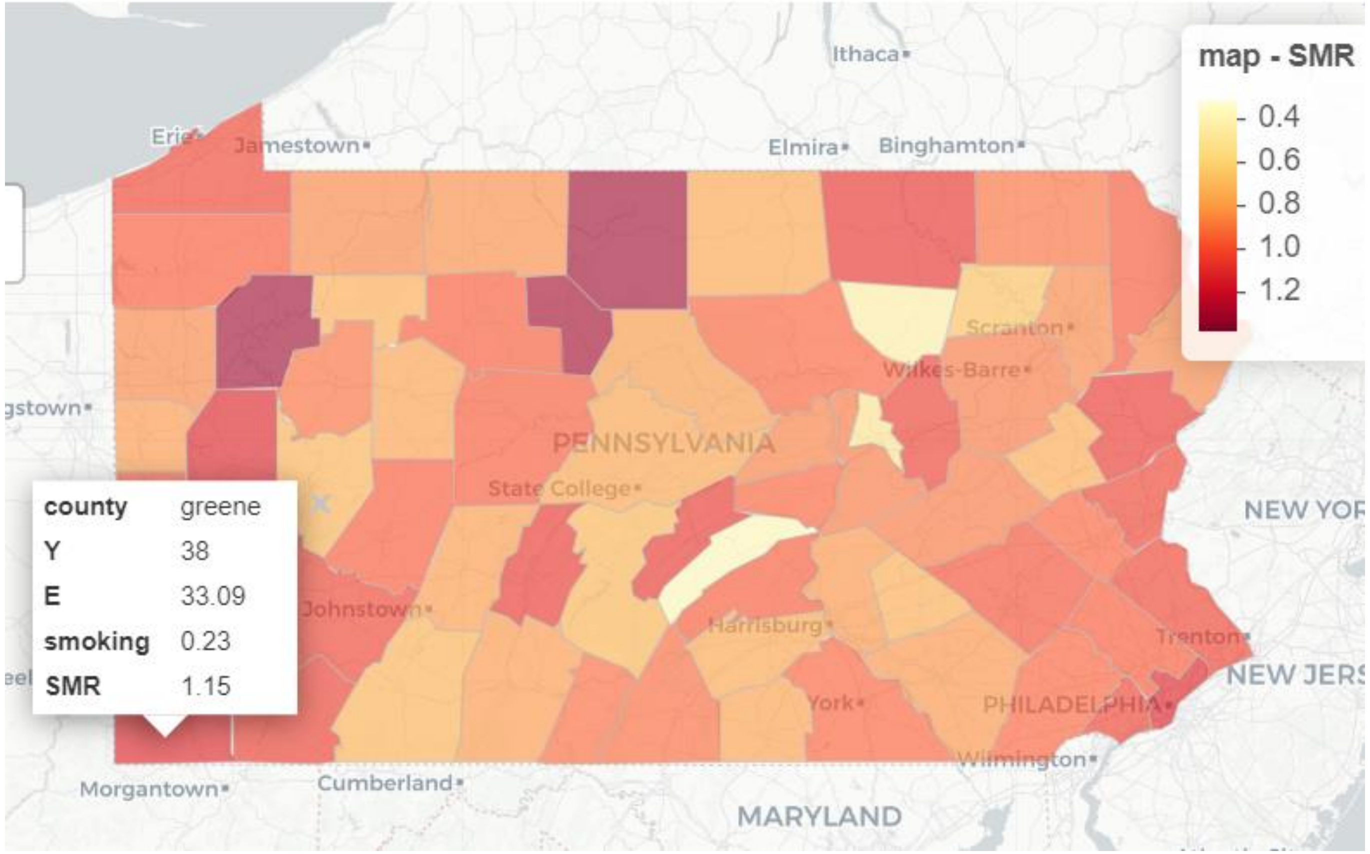
Map of Pennsylvania counties

Standardized Mortality Ratio (SMR)

```r
head(d)
```

```
# A tibble: 6 × 5
  county        Y      E smoking   SMR
  <fct>     <int>  <dbl>   <dbl> <dbl>
1 adams        55   69.6   0.234 0.790
2 allegheny  1275 1182.    0.245 1.08
3 armstrong    49   67.6   0.25  0.725
4 beaver      172  173.    0.276 0.997
5 bedford      37   44.2   0.228 0.837
6 berks       308  301.    0.249 1.02
```

Map the SMR



map - SMR
0.4
0.6
0.8
1.0
1.2
Ithaca
Erie
Jamestown
Elmira
Binghamton
Scranton
Wilkes-Barre
PENNSYLVANIA
State College
Johnstown
Harrisburg
York
PHILADELPHIA
Trenton
Wilmington
NEW YOR
NEW JERS
Morgantown
Cumberland
MARYLAND
county greene
Y 38
E 33.09
smoking 0.23
SMR 1.15

$$Y_i|\theta_i \sim Poisson(E_i \times \theta_i),\ i = 1, \ldots, n,$$

$$\log(\theta_i) = \beta_0 + \beta_1 \times smoking_i + u_i + v_i.$$

Here, β_0 is the intercept and β_1 is the coefficient of the covariate smokers proportion. u_i is a structured spatial effect modeled with an intrinsic conditionally autoregressive model (CAR), $u_i|\mathbf{u_{-i}} \sim N(\bar{u}_{\delta_i} \frac{1}{\tau_u n_{\delta_i}})$. Finally, v_i is an unstructured effect, $v_i \sim N(0, 1/\tau_v)$.

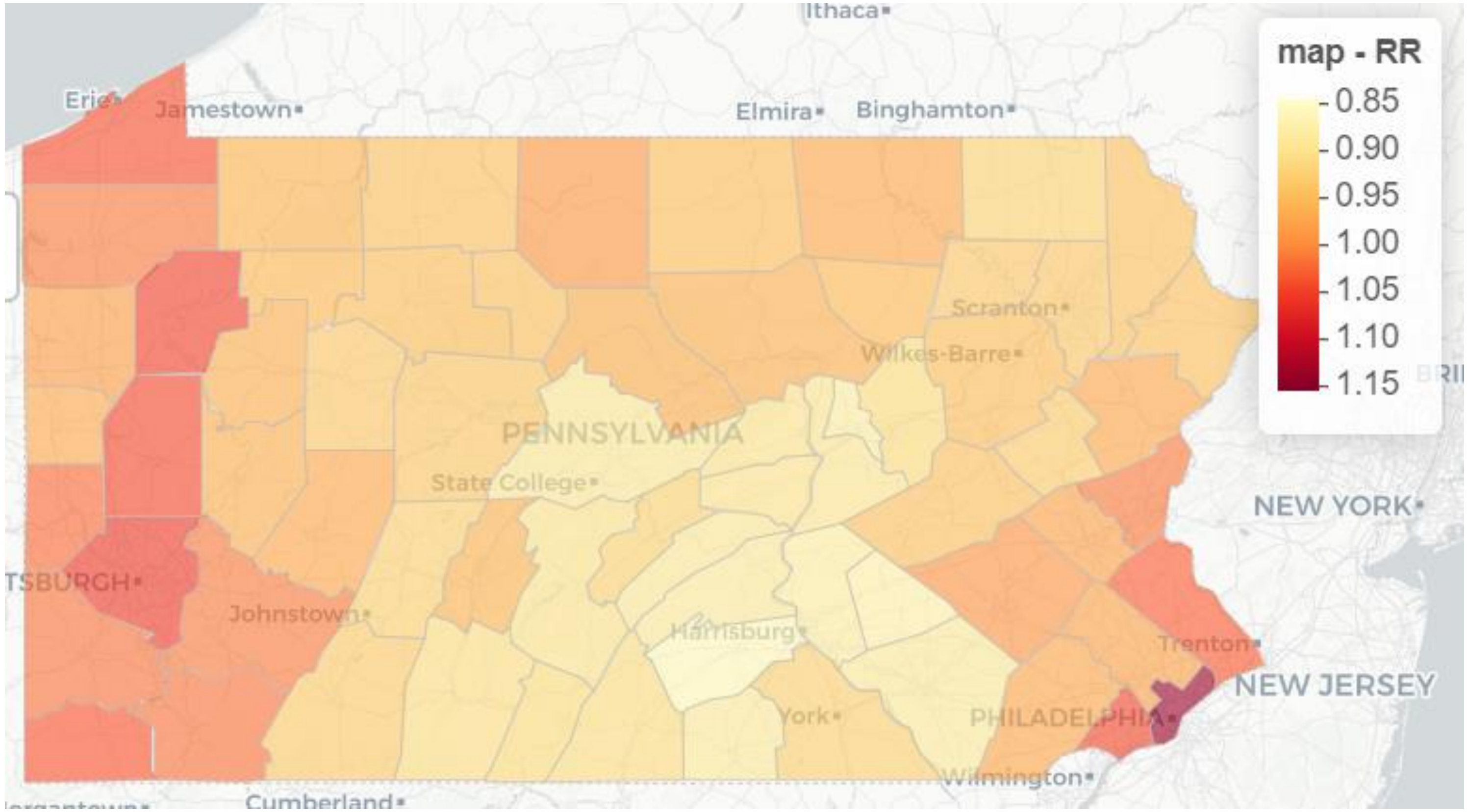
```r
formula <- Y ~ smoking +
  f(re_u, model = "besag", graph = g, scale.model = TRUE) +
  f(re_v, model = "iid")
```

fit the model using the inla() function

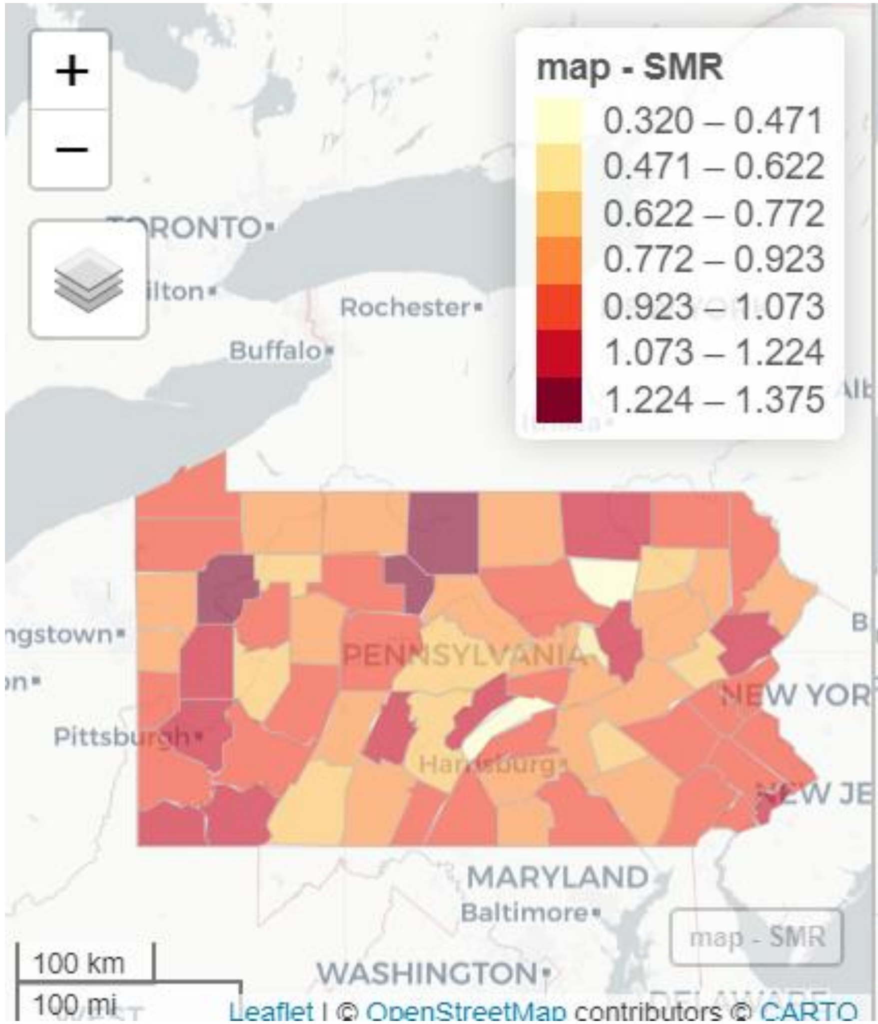
```r
res <- inla(formula, family = "poisson", data = map, E = E,
control.predictor = list(compute = TRUE),
control.compute = list(return.marginals.predictor = TRUE))
```

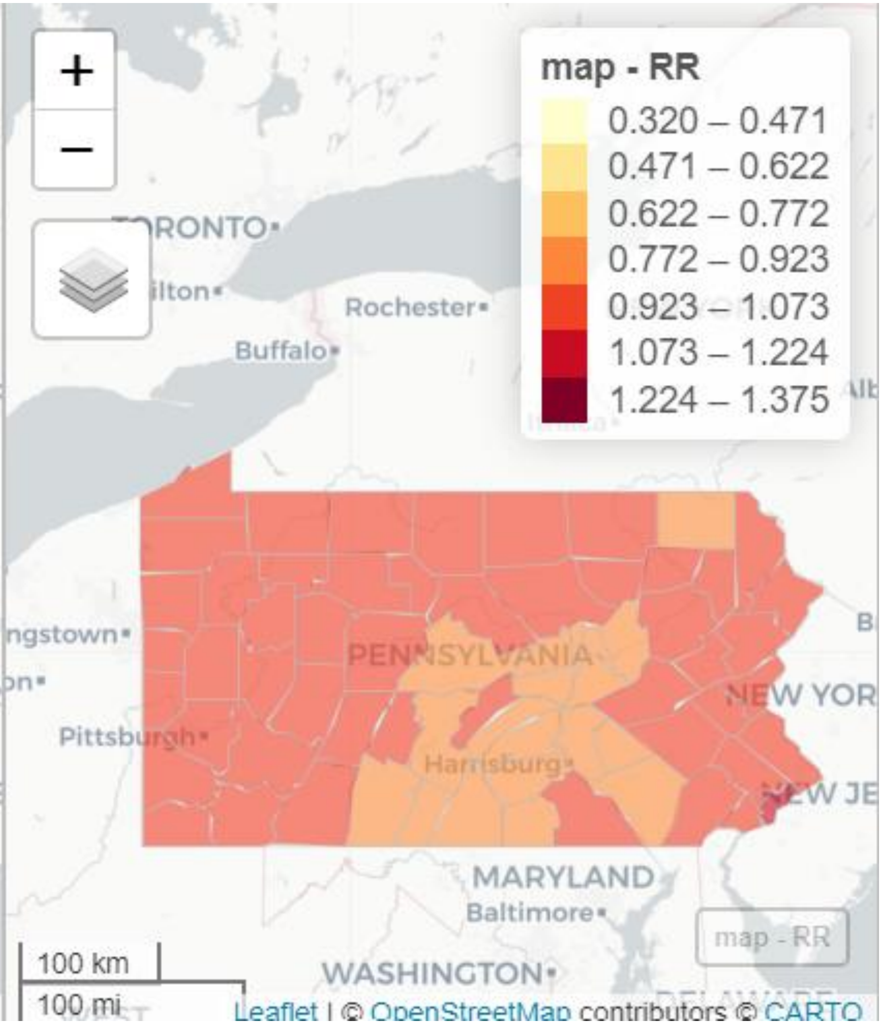
```
                mean     sd 0.025quant 0.5quant
(Intercept) -0.3235 0.1498   -0.61925  -0.3233
smoking      1.1546 0.6226   -0.07569   1.1560
            0.975quant    mode       kld
(Intercept)   -0.02877 -0.3234 3.534e-08
smoking        2.37845  1.1563 3.545e-08
```

Mapping disease risk



Comparing SMR and RR maps





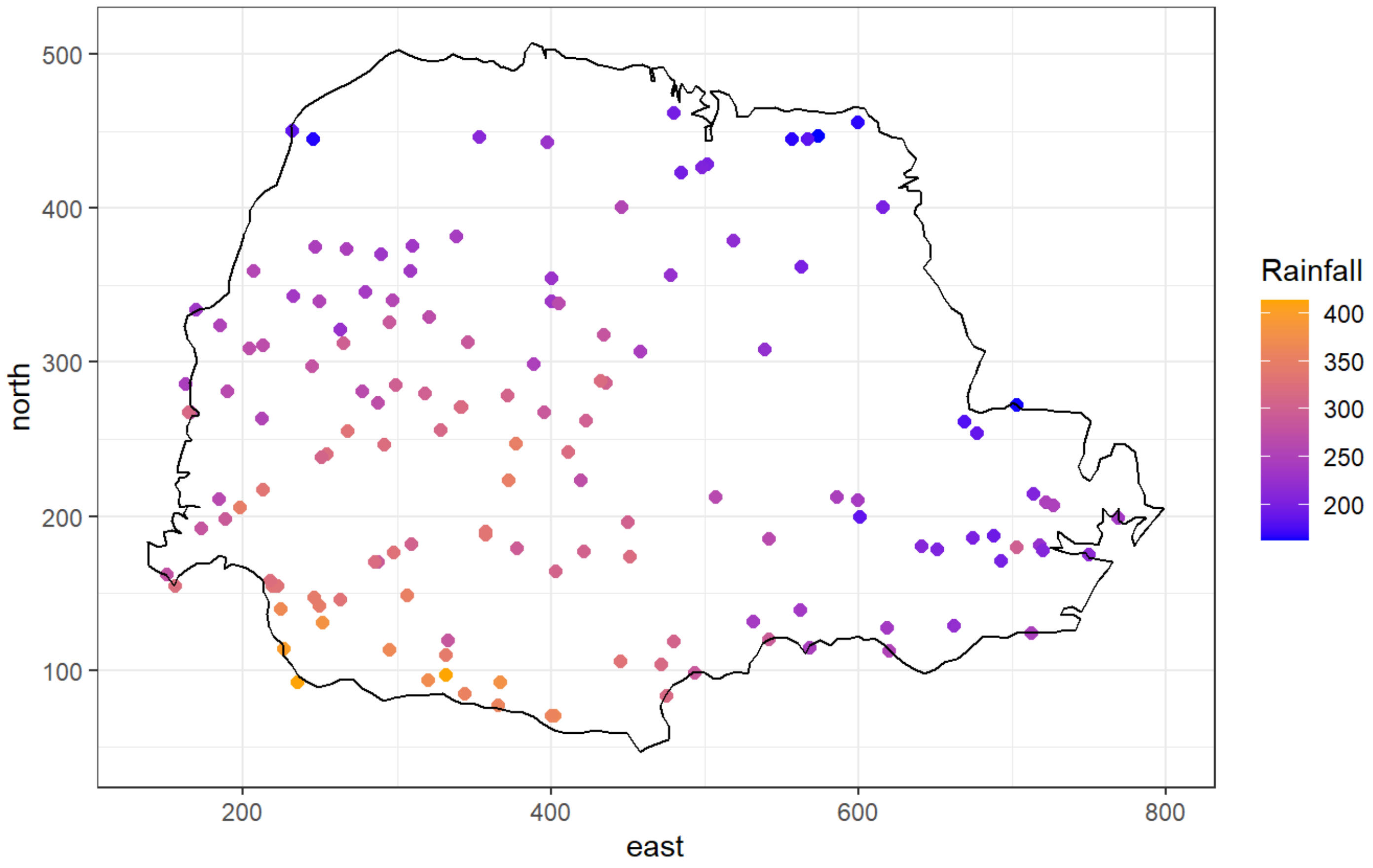
Geostatistical data

- Geostatistical data provide information of a spatially continuous phenomenon that has been measured at particular sites.
- air pollution levels taken at a set of monitoring stations
- disease prevalence survey data at a collection of sites
- density of mosquitoes responsible for disease transmission measured using traps placed at different locations

INLA and the Stochastic Partial Differential Equation (SPDE)

- A framework that helps model spatial dependencies in data, particularly useful for point-level (geostatistical) data.
- Combining INLA with SPDE:

- Enables analysis of continuous spatial data by converting it into a discrete spatial model.
- Effective for modelling large spatial datasets with a continuous domain
- GRF over large data sets



data from the **geoR** package

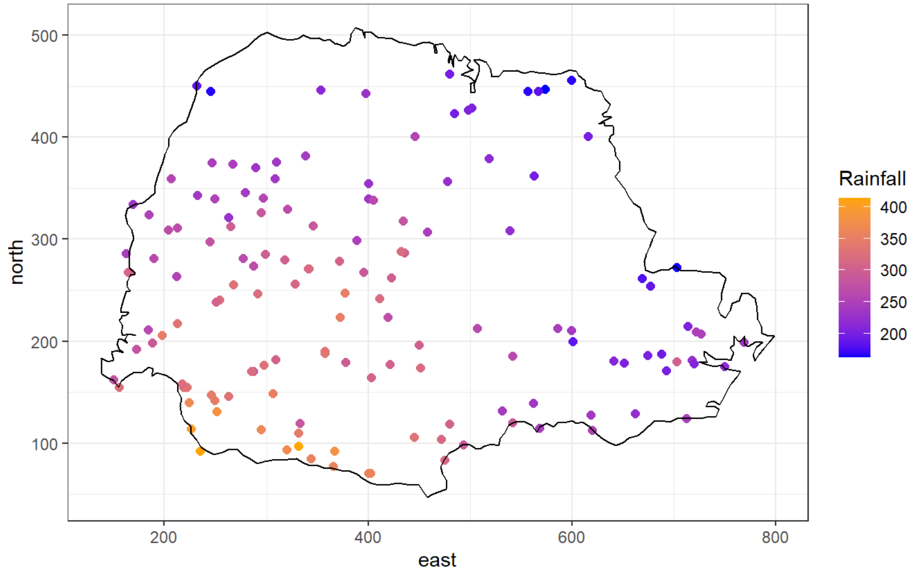
Model

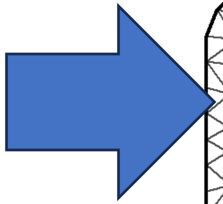
$$Y_i \sim N(\mu_i, \sigma^2),\ i = 1, 2, \ldots, n,$$

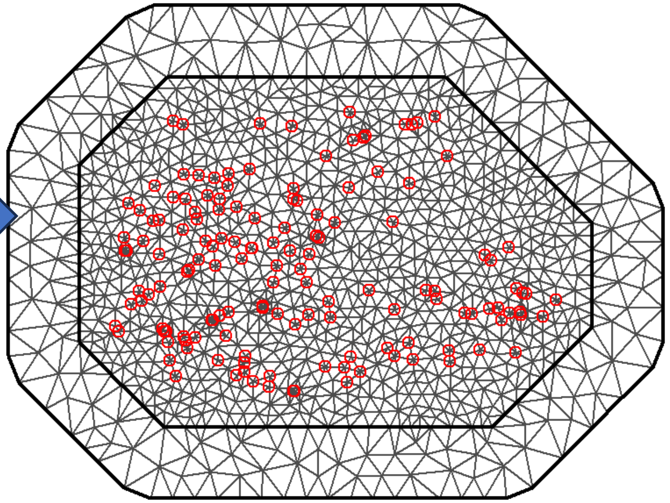
$$\mu_i = \beta_0 + Z(\boldsymbol{s_i}).$$

We can assume that rainfall at location si, Yi, follows a Gaussian distribution with mean μi and variance σ2, and the mean μi is expressed as the sum of an intercept β0 and a spatially structured random effect that follows a zero-mean Gaussian process.

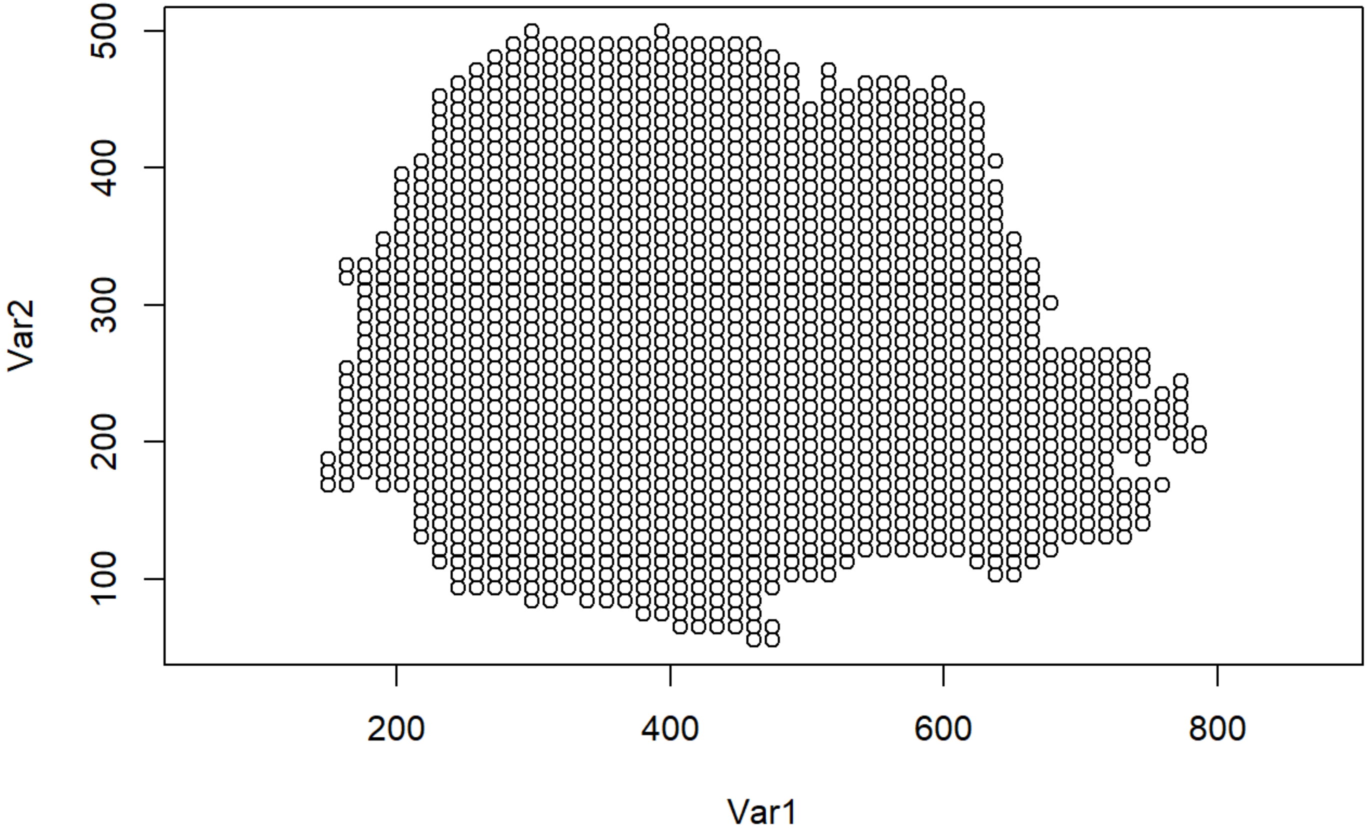
Mesh to build the SPDE model.



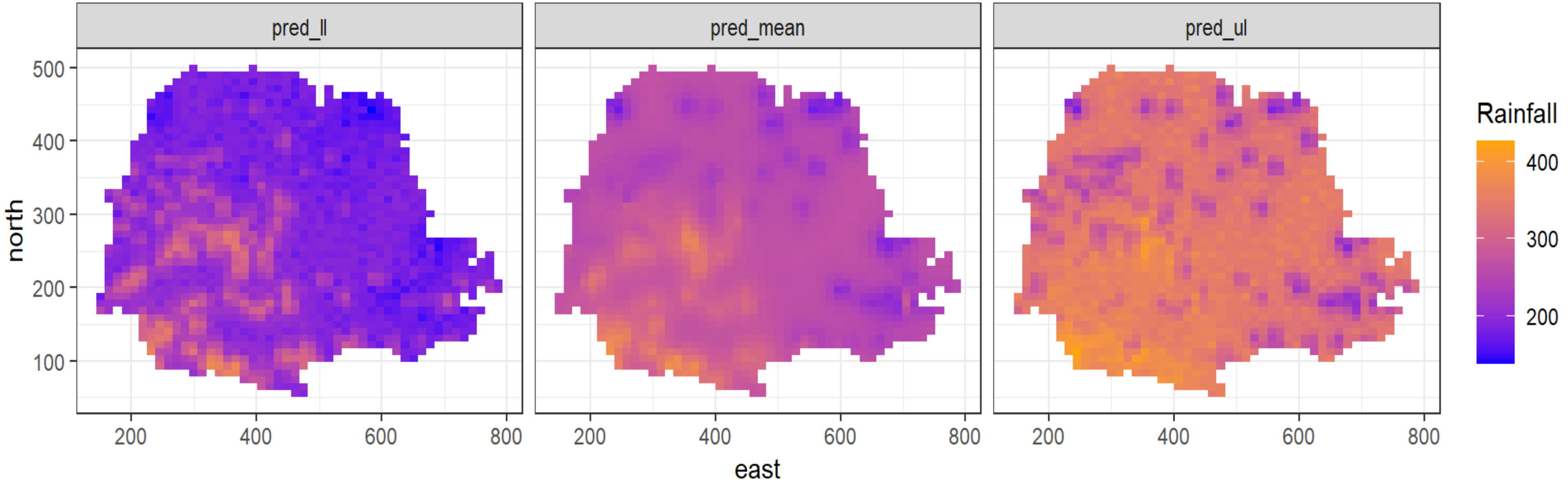




Prediction locations



Output



Rainfall predictions and lower and upper limits of 95% CI

Local Data

SUMMARY

- Efficiency: INLA is much faster than traditional MCMC methods, making it ideal for large and complex spatial datasets.
- Accuracy: By combining INLA with the SPDE approach, we can accurately model continuous spatial processes, providing detailed insights into spatial patterns.
- Flexibility: R-INLA handles a range of spatial and spatio-temporal models, making it suitable for various applications in public health, ecology, environmental studies, and beyond.
- Stabilized Estimates: With Bayesian hierarchical modeling, R-INLA produces reliable estimates even in areas with small sample sizes by borrowing information from neighboring regions, reducing the impact of outliers and extreme values.

References

- Tennekes, M. (2018). tmap: Thematic Maps in R. *Journal of Statistical Software*, *84*(6), 1–39. https://doi.org/10.18637/jss.v084.i06
- Paula Moraga (2021) Handbook of Spatial Epidemiology, Journal of the American Statistical Association, 116:533, 451-453, DOI: 10.1080/01621459.2021.1880230
- https://cran.r-project.org/web/packages/tmap/index.html

Thank You

- Personal Email : amsyarafiq90@gmail.com
- Student Email: afiqamsyar90@student.usm.my