

STK4150/9150 - Environmental and Spatial Statistics

Spatial lattice processes, Part IV

AUTHOR
Thordis Thorarinsdottir

PUBLISHED
February 29, 2024

Example 2: Lung cancer data from Pennsylvania

This example is taken from Duncan *et al.* (2019).

Read in the data and packages:

```
library(SpatialEpi)
library(leaflet)
library(maptools)
library(spdep)
library(CARBayes)

data(pennLC)
names(pennLC)
```

```
[1] "geo"          "data"         "smoking"      "spatial.polygon"
```

```
head(pennLC$data, n=3)
```

	county	cases	population	race	gender	age
1	adams	0	1492	o	f	Under.40
2	adams	0	365	o	f	40.59
3	adams	1	68	o	f	60.69

```
head(pennLC$smoking, n=3)
```

	county	smoking
1	adams	0.234
2	allegHENy	0.245
3	armstrong	0.250

Data preprocessing: We want the observed and expected number of cases per county across all strata.

```
# Order the data
pennLC$data <- pennLC$data[order(
  pennLC$data$county,
  pennLC$data$race,
  pennLC$data$gender,
  pennLC$data$age), ]

# Observed cases
y <- aggregate(
```

```
x = pennLC$data$cases,
by = list(county = pennLC$data$county),
FUN = sum)

# Expected cases
E <- expected(
  population = pennLC$data$population,
  cases = pennLC$data$cases,
  n.strata = 16) # 2 races, 2 genders, 4 age groups

head(y, n=5)
```

```
  county    x
1   adams   55
2 alleghe 1275
3 armstro  49
4   beaver 172
5  bedford  37
```

```
round(E[1:5])
```

```
[1] 70 1182 68 173 44
```

```
# Merge observed and expected cases
d <- cbind(y, E)
names(d)[2] <- "y" # Rename counts as "y"

# Compute raw standardized incidence ratios (SIR)
d$SIR <- d$y/d$E

# Add the proportion of smokers variable
d <- merge(d, pennLC$smoking, by = "county")

head(d)
```

```
  county    y      E      SIR smoking
1   adams   55 69.62730 0.7899200  0.234
2 alleghe 1275 1182.42804 1.0782897  0.245
3 armstro  49  67.61012 0.7247435  0.250
4   beaver 172 172.55806 0.9967660  0.276
5  bedford  37  44.19013 0.8372910  0.228
6   berks  308 300.70598 1.0242563  0.249
```

Create an interactive map of the data:

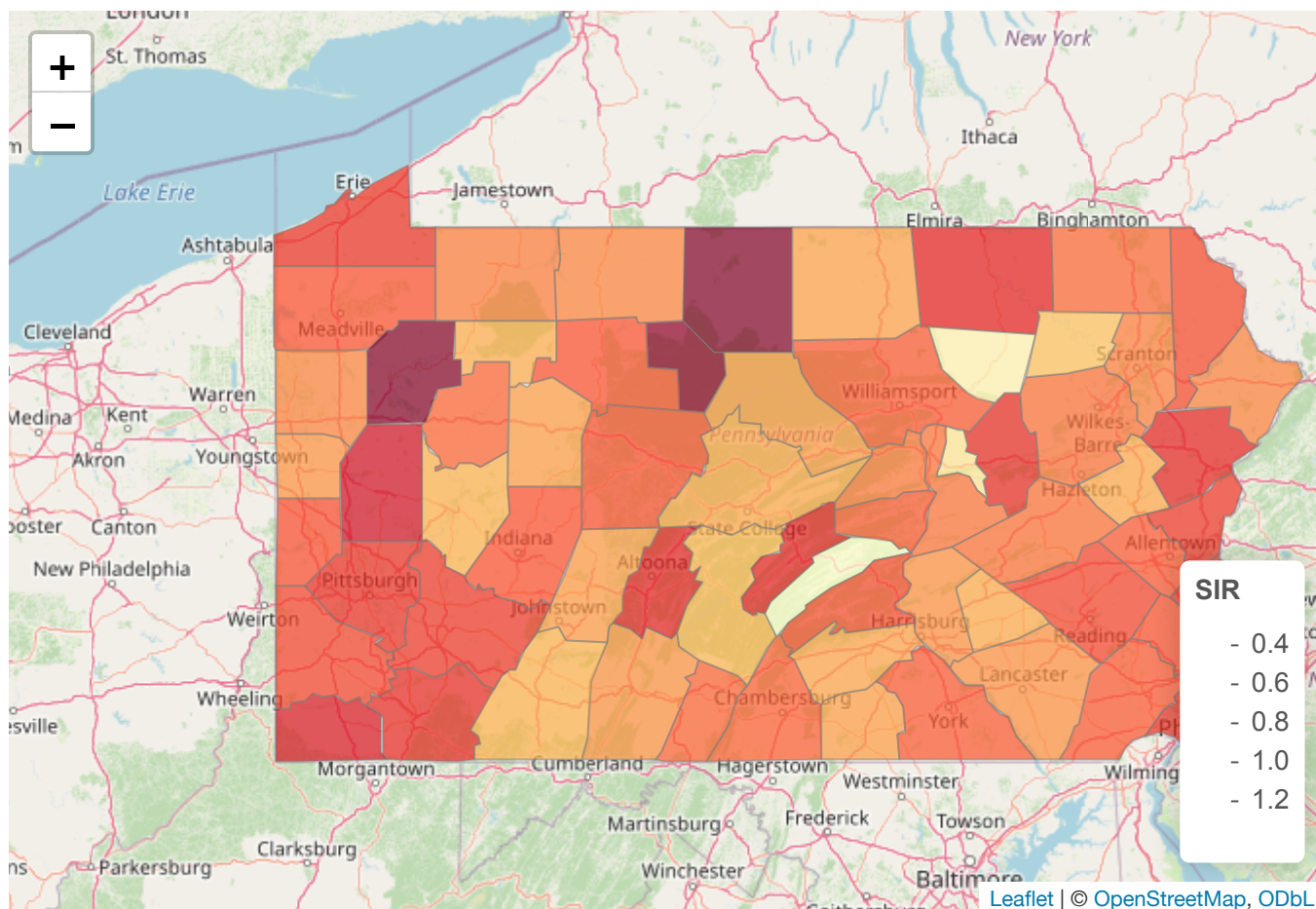
```
# Create map
rownames(d) <- d$county # Use county names as unique IDs
map <- SpatialPolygonsDataFrame(pennLC$spatial.polygon, d, match.ID = TRUE)

# Set colour palette
pal <- colorNumeric(palette = "YlOrRd", domain = map$SIR)

# Create interactive map
```

```
lmap <- leaflet(map) %>%
  addTiles() %>%
  addPolygons(., color = "grey", weight = 1,
    fillColor = ~pal(SIR), fillOpacity = 0.7) %>%
  addLegend(., pal = pal, values = ~SIR, opacity = 0.7,
    title = "SIR", position = "bottomright")

# View map
lmap
```



We can add information to the map by highlighting the counties when the mouse hovers over them, and showing information about observed and expected counts. This is done by adding arguments to and using HTML syntax.

```
# Labels
labels <- sprintf(
  "<strong>%s</strong><br>Observed: %s <br>Expected: %s
  <br>Smokers proportion: %s <br>SIR: %s",
  map$county, map$y, round(map$E, 2), map$smoking, round(map$SIR, 2))
labels <- lapply(labels, htmltools::HTML)

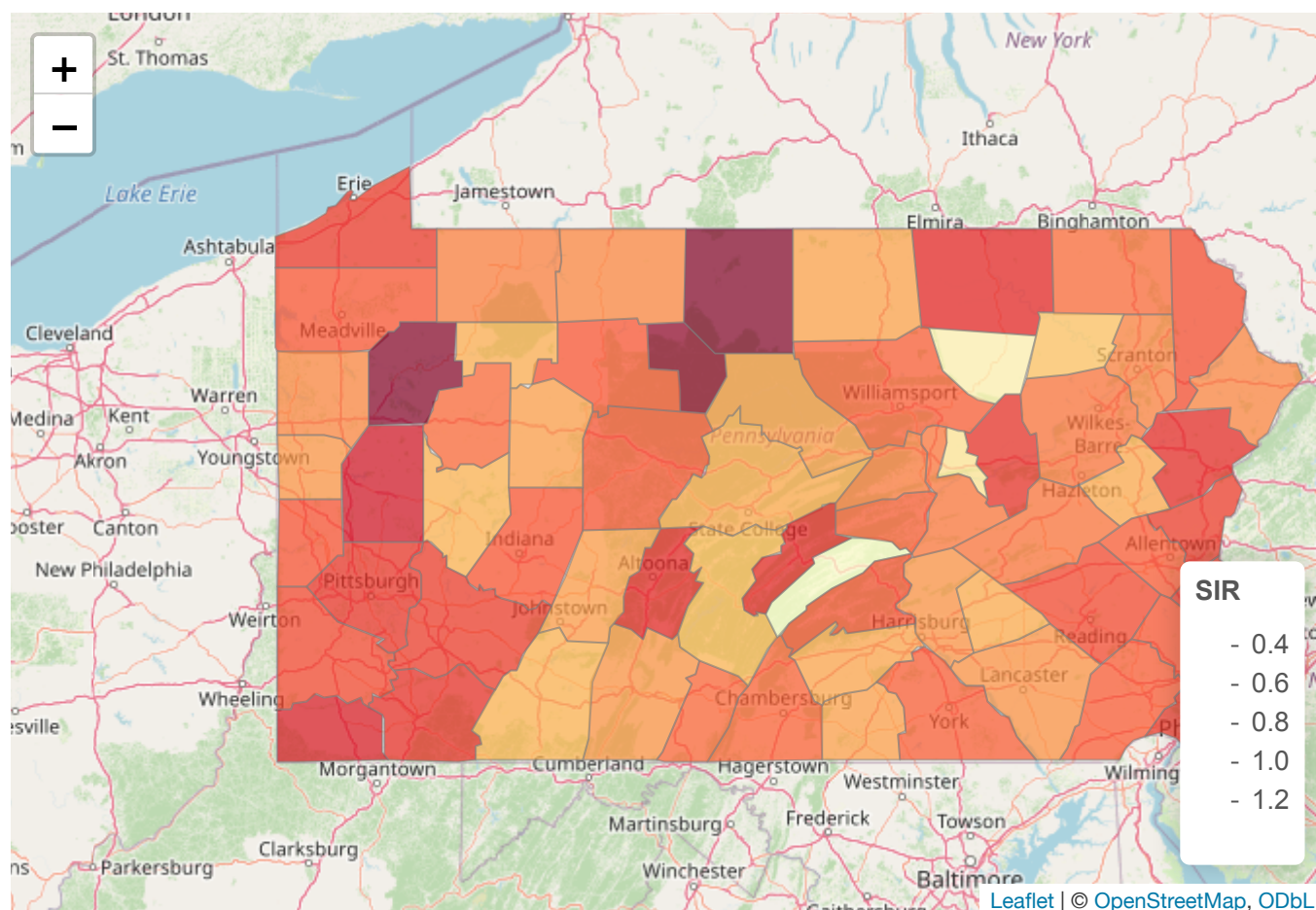
# Create interactive map
lmap <- leaflet(map) %>%
  addTiles() %>%
  addPolygons(.,
    color = "grey",
    weight = 1,
    fillColor = ~pal(SIR),
    fillOpacity = 0.7,
```

```

highlightOptions = highlightOptions(weight = 4),
label = labels,
labelOptions = labelOptions(
  style = list("font-weight" = "normal", padding = "3px 8px"),
  textsize = "15px",
  direction = "auto"
)) %>%
addLegend(.,pal = pal, values = ~SIR, opacity = 0.7,
  title = "SIR", position = "bottomright")

# View interactive map
lmap

```



Now we can examine the map and see which areas have

- $SIR = 1$ which indicates observed cases are the same as expected,
- $SIR > 1$ which indicates observed cases are higher than expected,
- $SIR < 1$ which indicates observed cases are lower than expected.

This map gives a sense of the disease risk across Pennsylvania. However, observed (raw) SIRs may be misleading and insufficiently reliable in counties with small populations. Here, we can use model-based approaches to incorporate covariates and borrow information from neighboring counties to improve local estimates, resulting in the smoothing of extreme values based on small sample sizes.

```

# Create binary, first-order, adjacency spatial weights matrix
W <- nb2mat(poly2nb(map), style = "B")

```

```
# MCMC parameters
M.burnin <- 10000      # Number of burn-in iterations (discarded)
M <- 100000           # Number of iterations retained

# Fit BYM model using CARBayes
set.seed(111)          # For reproducibility
MCMC <- S.CARbym(
  formula = y ~ 1 + offset(log(E)),
  data = d,
  family = "poisson",
  W = W,
  burnin = M.burnin,
  n.sample = M.burnin + M,    # Total iterations
  verbose = FALSE
)

# Broad summary of results
print(MCMC$summary.results)
```

	Mean	2.5%	97.5%	n.sample	% accept	n.effective	Geweke.diag
(Intercept)	-0.0526	-0.0838	-0.0215	1e+05	32.9	7682.2	0.8
tau2	0.0123	0.0033	0.0306	1e+05	100.0	1574.0	-0.5
sigma2	0.0051	0.0018	0.0115	1e+05	100.0	1950.1	-0.8

To obtain the modelled SIR, we need to divide the fitted values by the expected values.

```
# Compute posterior SIR
y.fit <- MCMC$samples$fitted
SIR <- t(t(y.fit) / E)
```

Because we have a posterior sample size of $M=100000$ iterations, SIR is a $M \times N$ matrix where N is the number of counties. To visualise these estimates on the map, we will plot the median, and then add the 2.5th and 97.5 percentiles to the information that pops up when you hover the mouse over an area. We use RR to denote relative risk, which is the smoothed SIR (specifically, the median of the smoothed SIR), to distinguish it from the observed SIR. The posterior probability of the relative risk being greater than 1 (i.e. the proportion of MCMC iterations where the smoothed SIR is greater than 1) is also included, denoted by PP.

```
# Add summary statistics of the posterior SIR to map
map$SIR.50 <- apply(SIR, 2, median)
map$SIR.025 <- apply(SIR, 2, quantile, 0.025)
map$SIR.975 <- apply(SIR, 2, quantile, 0.975)
map$PP <- apply(SIR, 2, function(x) length(which(x > 1))) / M

# Optional: Update colour palette (harder to compare risk surfaces)
# pal <- colorNumeric(palette = "YlOrRd", domain = map$SIR.50)

# Update labels
labels <- sprintf(
  "<strong> %s </strong> <br/> Observed: %s <br/> Expected: %s <br/>
  Smoker proportion: %s <br/> SIR: %s <br/> RR: %s (%s, %s) <br/> PP: %s",
  map$county, map$y, round(map$E, 2), map$smoking, round(map$SIR, 2),
  round(map$SIR.50, 2), round(map$SIR.025, 2), round(map$SIR.975, 2),

```

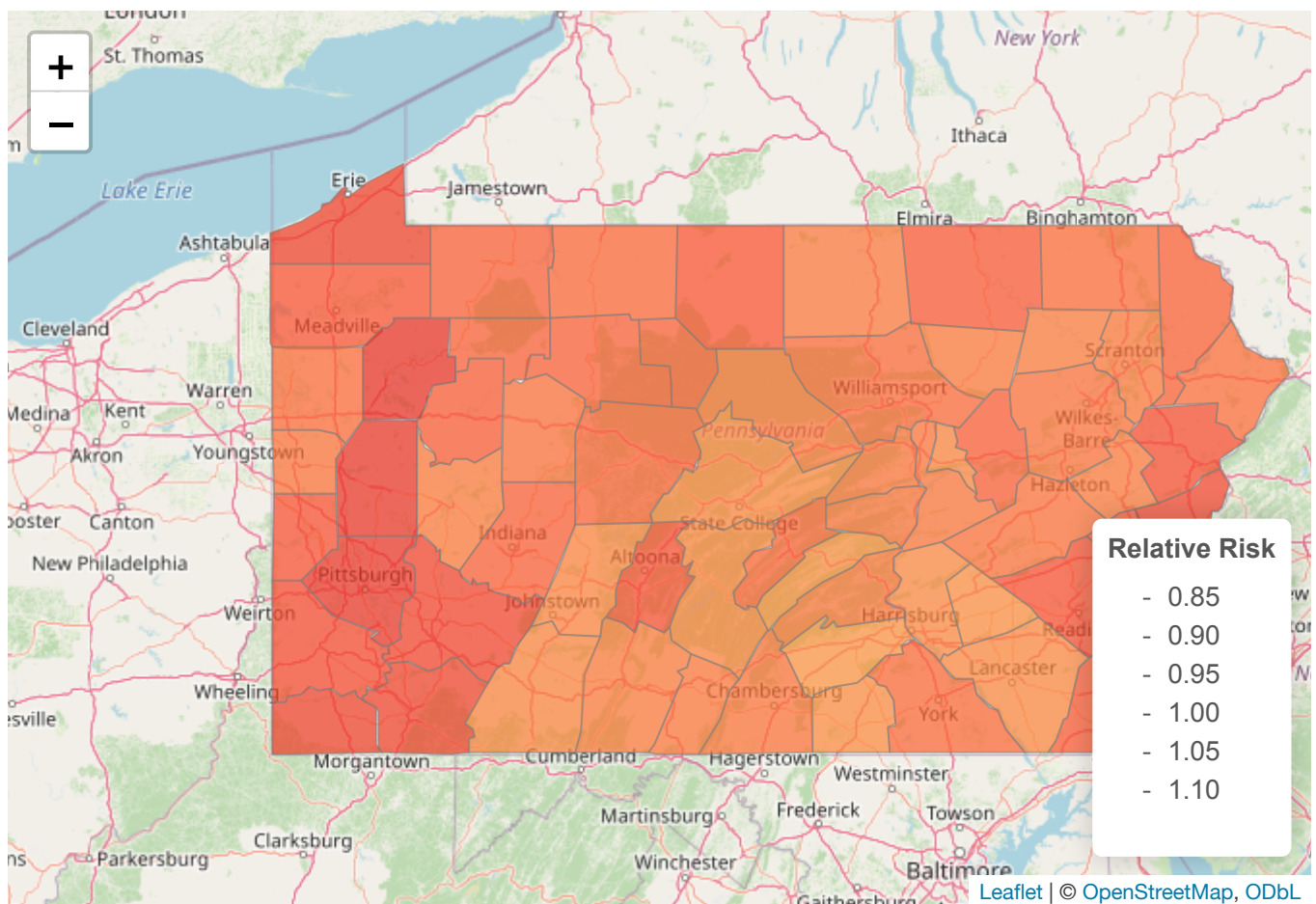
```

round(map$PP, 3))
labels <- lapply(labels, htmltools::HTML)

# Create interactive map
lmap <- leaflet(map) %>%
addTiles() %>%
addPolygons(.,
  color = "grey",
  weight = 1,
  fillColor = ~pal(SIR.50),
  fillOpacity = 0.7,
  highlightOptions = highlightOptions(weight = 4),
  label = labels,
  labelOptions = labelOptions(
    style = list("font-weight" = "normal", padding = "3px 8px"),
    textsize = "15px",
    direction = "auto"
  )
) %>%
addLegend(.,
  pal = pal,
  values = ~SIR.50,
  opacity = 0.7,
  title = "Relative Risk",
  position = "bottomright"
)

# View interactive map
lmap

```



References

Duncan, E. W., S. M. Cramb, P. D. Baade, K. L. Mengersen, T. Saunders, and J. F. Aitken. 2019. *Developing a Cancer Atlas using Bayesian Methods: A Practical Guide for Application and Interpretation*. Brisbane: Queensland University of Technology (QUT) and Cancer Council Queensland. Available at <https://atlas.cancer.org.au/developing-a-cancer-atlas/> (Accessed on February 28, 2024).

Moraga, P. (2018). Small area disease risk estimation and visualization using R. *The R Journal*, 10(1), 495-506.