Disease risk modelling and visualization using R

Paula Moraga

Lancaster University

CHICAS centre for health informatics, computing, and statistics

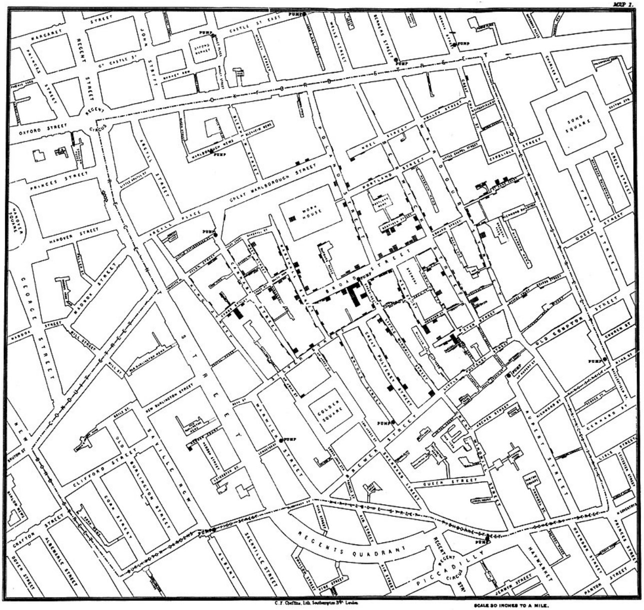
RaukR Summer School

Visby, 18 June 2018

Outline

Introduction to disease mapping

John Snow's map of cholera deaths in Soho, London, 1854

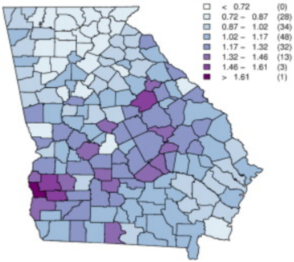


Disease mapping

Disease maps help understand the spatial patterns of disease and its determinants. This information can guide decision makers and programme managers to better allocate limited resources and to design strategies for disease prevention and control

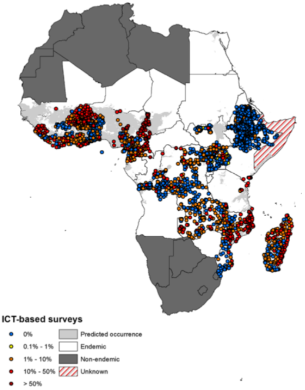
Types of spatial data

1. Areal data



Moraga and Lawson 2012

2. Geostatistical data



Moraga et al. 2015

3. Point patterns

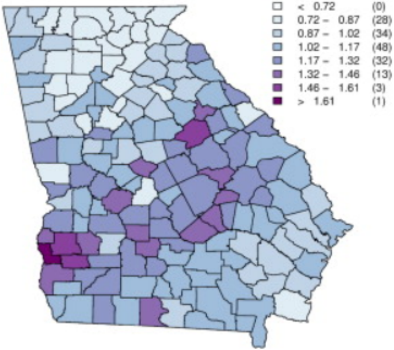
Moraga and Montes 2011

Modelling

- Disease risk predictions are based on the observed disease cases, the number of individuals at risk, and risk factors information such as demographic and environmental factors

- Models describe the variability in the response variable as a function of the risk factors covariates and random effects to account for unexplained variability

Areal data



Moraga and Lawson 2012

Areal data

Disease risk is often estimated by the Standardized Mortality Ratio:

$$SMR = \frac{Y}{E}$$

- Y number of observed cases
- E number of expected cases if the study population had the same disease rate as the standard population
- $SMR > 1$: more cases observed than expected

- Expected cases calculated using indirect standardization

$$E = \sum_{j=1}^{m} r_j^{(s)} n_j$$

- $r_j^{(s)}$ =(number of events)/(number of individuals at risk). Rate in strata j (e.g. age group, sex) in the standard population
- n_j population in stratum j of the observed population

Areal data

- SMRs may be misleading and insufficiently reliable in areas with small populations
- In contrast, model-based approaches enable to incorporate covariates and borrow information from neighboring areas to improve local estimates, resulting in the smoothing of extreme rates based on small sample sizes

Areal data

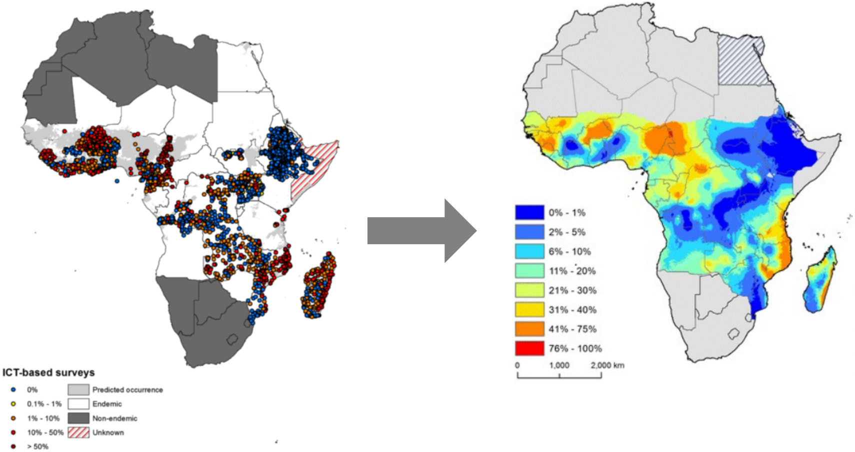
Model to estimate disease risks θ_i in areas $i = 1, \ldots, n$

$$Y_i | \theta_i \sim Po(E_i \times \theta_i),$$

$$\log(\theta_i) = \boldsymbol{z}_i' \boldsymbol{\beta} + u_i + v_i$$

- u_i is an structured spatial effect to account for the spatial dependence between relative risks (areas that are close show more similar risk than areas that are not close)
- v_i is an unstructured spatial effect to account for independent area-specific noise

Geostatistical data



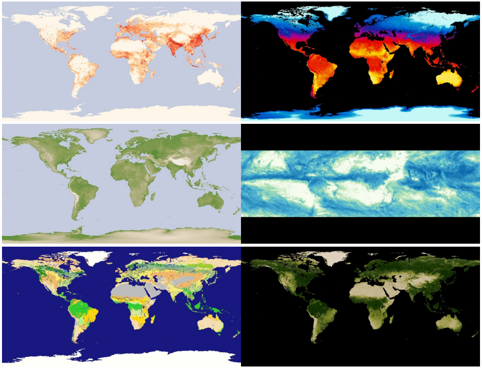
Moraga et al. 2015

Geostatistical data

$$Y_i | P(\boldsymbol{x}_i) \sim \text{Binomial}(N_i, P(\boldsymbol{x}_i)),$$
$$\text{logit}(P(\boldsymbol{x}_i)) = \boldsymbol{z}_i' \boldsymbol{\beta} + S(\boldsymbol{x}_i) + v_i$$

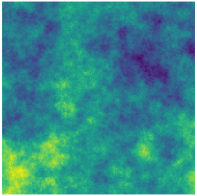
Risk factors covariates

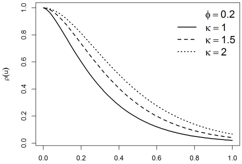
(e.g. temperature, precipitation, vegetation, etc)



NASA Earth Observations

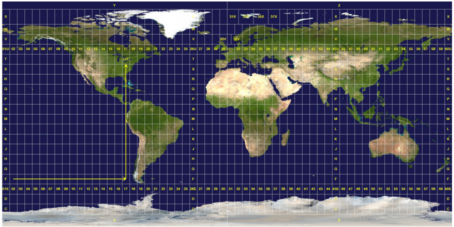
Gaussian Random Field





Coordinate Reference Systems (CRS)

1. **unprojected or geographic**: Latitude/Longitude for referencing location on the ellipsoid Earth
2. **projected**: Easting/Northing for referencing location on 2-dimensional representation of Earth. Common projection: Universal Transverse Mercator (UTM)



Tutorials

Install R packages

```
install.packages(c("dplyr", "ggplot2", "leaflet",
                   "geoR", "rgdal", "raster",
                   "sp", "spdep", "SpatialEpi",
                   "SpatialEpiApp"))


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

Tutorial: areal data

Areal data. Lung cancer in Pennsylvania

https://paula-moraga.github.io/tutorial-areal-data/

5 Mapping disease risk

We show the estimated disease risk in an interactive map using `leaflet`. In the map, we add labels that appear when mouse hovers over the counties showing information about observed and expected counts, SMRs, smokers proportions, RRs, and lower and upper limits of 95% credible intervals.

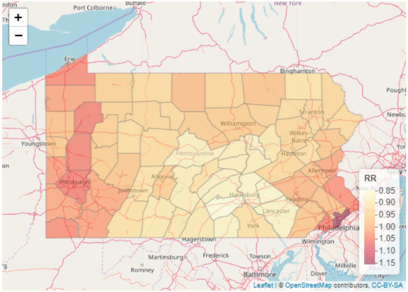
We observe counties with greater disease risk are located in the west and south east of Pennsylvania, and counties with lower risk are located in the center. The 95% credible intervals indicate the uncertainty in the risk estimates.

Hide

```
pal <- colorNumeric(palette = "YlOrRd", domain = map$RR)

labels <- sprintf("<strong> %s </strong> <br/> Observed: %s <br/> Expected: %s <br/>
                  Smokers proportion: %s <br/>SMR: %s <br/>RR: %s (%s, %s)",
                  map$county, map$Y,  round(map$E, 2),  map$smoking, round(map$SMR, 2),
                  round(map$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 = 0.5,
    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 = 0.5, title = "RR", position = "bottomright")
```



Tutorial: geostatistical data

Geostatistical data. Malaria in The Gambia

https://paula-moraga.github.io/tutorial-geostatistical-data/

2.4 Environmental covariates

To model malaria prevalence we will use a covariate that indicates the elevation in The Gambia. This covariate can be obtained with the `getData()` function of the `raster` package which can be used to obtain geographic data from anywhere in the world. In order to get the elevation values in The Gambia, we need to call `getData()` with the three following arguments:

- `name` of the data set to `'alt'`,
- `country` set to the 3 letters of the International Organization for Standardization (ISO) code of The Gambia (`GMB`), and
- `mask` set to `TRUE` so the neighbouring countries are set to NA.

Hide

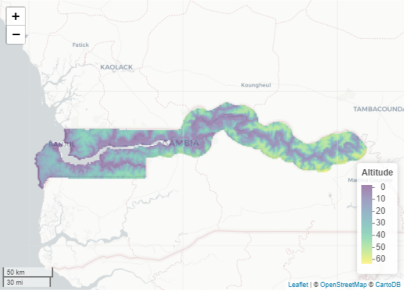
```
library(raster)
r <- getData(name = 'alt', country = 'GMB', mask = TRUE)
```

We make a map with the elevation raster using the `addRasterImage()` function of the `leaflet` package. First we create a palette function `pal` using the values of the raster (`values(r)`) and specifying that the NA values are transparent.

Hide

```
pal <- colorNumeric("viridis", values(r), na.color = "transparent")

leaflet() %>% addProviderTiles(providers$CartoDB.Positron) %>%
  addRasterImage(r, colors = pal, opacity = 0.5) %>%
  addLegend("bottomright", pal = pal, values = values(r), title = "Altitude") %>%
  addScaleBar(position = c("bottomleft"))
```



Presentations options: interactive dashboards and Shiny apps

Interactive dashboards with `flexdashboard`

- https://rmarkdown.rstudio.com/flexdashboard/
- Uses R Markdown to publish a group of related data visualizations as a dashboard
- Components that can be included include plots, tables, value boxes and `htmlwidgets`

Layout

```
---
title: "Row Orientation"
output:
  flexdashboard::flex_dashboard:
    orientation: rows
---

Row
-------------------------------------

### Chart 1

```{r}
```

Chart 2

```{r}
```

Row

Chart 3

```{r}
```

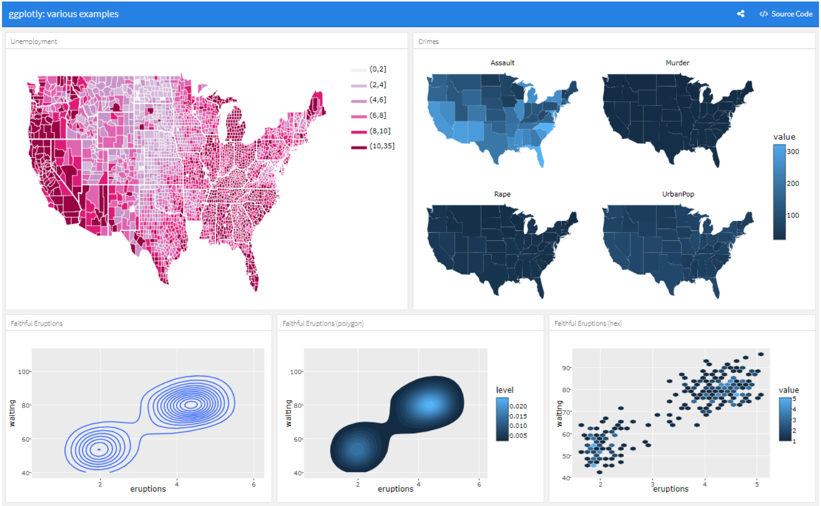
Chart 4

```{r}
```

```

# Example

https://rmarkdown.rstudio.com/flexdashboard/examples.html



# Interactive Shiny web applications

- https://shiny.rstudio.com/
- Shiny is a web application framework for R that enables to build interactive web applications

# SpatialEpiApp

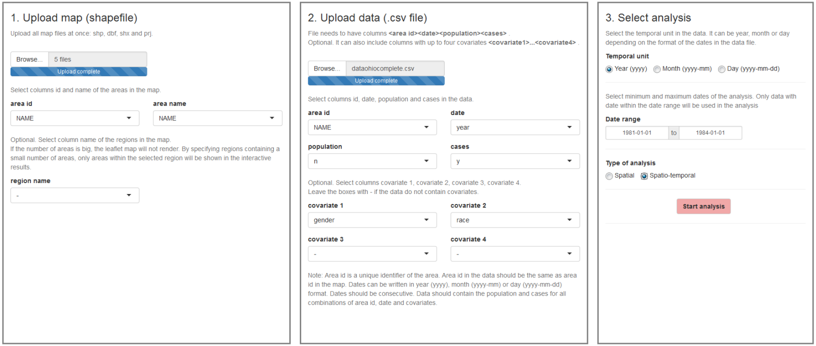
# R package SpatialEpiApp

- Shiny web application that allows to visualize spatial and spatio-temporal disease data, estimate disease risk and detect clusters
- Risk estimates by fitting Bayesian models with INLA
- Detection of clusters by using the scan statistics in SaTScan

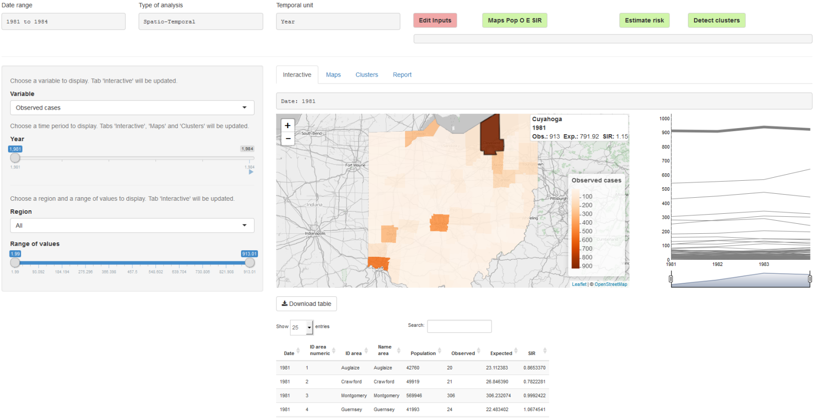
Launch SpatialEpiApp:

```
install.packages("SpatialEpiApp")
library(SpatialEpiApp)
run_app()
```

Data entry


1. Upload map (shapefile)
Upload all map files at once: shp, dbf, shx and prj.
Browse... 5 files
Upload complete
Select columns id and name of the areas in the map.
area id
NAME
area name
NAME
Optional. Select column name of the regions in the map.
If the number of areas is big, the leaflet map will not render. By specifying regions containing a small number of areas, only areas within the selected region will be shown in the interactive results.
region name
-
2. Upload data (.csv file)
File needs to have columns <area id><date><population><cases> .
Optional. It can also include columns with up to four covariates <covariate1>...<covariate4> .
Browse... dataohiocomplete.csv
Upload complete
Select columns id, date, population and cases in the data.
area id
NAME
date
year
population
n
cases
y
Optional. Select columns covariate 1, covariate 2, covariate 3, covariate 4.
Leave the boxes with - if the data do not contain covariates.
covariate 1
gender
covariate 2
race
covariate 3
-
covariate 4
-
Note: Area id is a unique identifier of the area. Area id in the data should be the same as area id in the map. Dates can be written in year (yyyy), month (yyyy-mm) or day (yyyy-mm-dd) format. Dates should be consecutive. Data should contain the population and cases for all combinations of area id, date and covariates.
3. Select analysis
Select the temporal unit in the data. It can be year, month or day depending on the format of the dates in the data file.
Temporal unit
Year (yyyy)
Month (yyyy-mm)
Day (yyyy-mm-dd)
Select minimum and maximum dates of the analysis. Only data with date within the date range will be used in the analysis
Date range
1981-01-01
to
1984-01-01
Type of analysis
Spatial
Spatio-temporal
Start analysis

Interactive



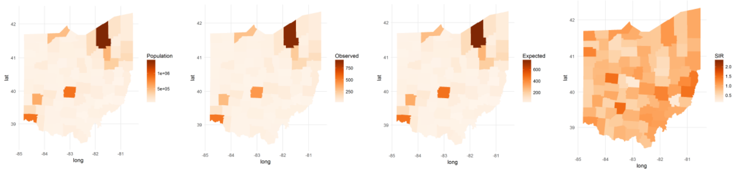
Maps

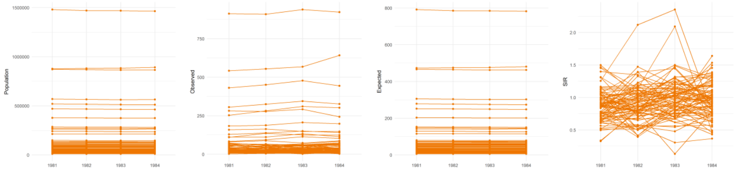
Interactive | Maps | Clusters | Report

```
Date: 1981
```

```
   Population        Observed          Expected            SIR
 Min.   :  11253   Min.   :  2.00   Min.   :  6.108   Min.   :0.3274
 1st Qu.:  32643   1st Qu.: 14.00   1st Qu.: 17.723   1st Qu.:0.7179
 Median :  54588   Median : 22.50   Median : 29.381   Median :0.8641
 Mean   : 122617   Mean   : 62.58   Mean   : 66.037   Mean   :0.8709
 3rd Qu.: 104815   3rd Qu.: 50.00   3rd Qu.: 56.468   3rd Qu.:1.0115
 Max.   :1481287   Max.   :913.00   Max.   :791.923   Max.   :1.4989
```

Population | Observed | Expected | SIR

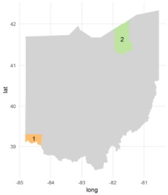


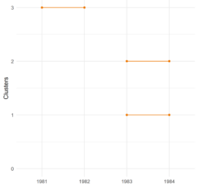


Clusters

Interactive | Maps | Clusters | Report

```
Date: 1984
```





Show 25 entries

Search:

Cluster	Central area	No. areas	Start date	End date	Risk in / Risk out	LLR	p-value	Areas
1	Hamilton	1	1983	1984	1.32	41.75818	1.23e-14	Hamilton
2	Cuyahoga	1	1983	1984	1.21	28.87297	1.04e-09	Cuyahoga
3	Belmont	5	1981	1982	1.30	10.54458	1.06e-02	Guernsey, Monroe, Harrison, Belmont, Jefferson
Cluster	Central area	No. areas	Start date	End date	Risk in / Risk ou	LLR	p-value	Areas

Showing 1 to 3 of 3 entries

Previous 1 Next

Report

Interactive | Maps | Clusters | Report

Download report

Choose the variables to include in the report. Variables that have not been calculated will not be included.

Maps

☑ Population ☑ Observed ☑ Expected ☑ SIR ☑ Risk ☑ 2.5 percentile ☑ 97.5 percentile ☑ Clusters

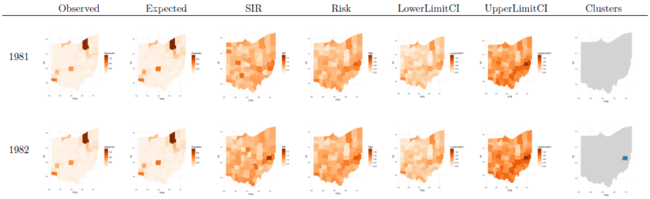
Tables summary

☑ Population ☑ Observed ☑ Expected ☑ SIR ☑ Risk ☑ 2.5 percentile ☑ 97.5 percentile

Table clusters

☑ Clusters

- Date range: 1981 to 1984
- Type of analysis: Spatio-Temporal
- Temporal unit: Year



References

- Paula Moraga. SpatialEpiApp: A Shiny Web Application for the analysis of Spatial and Spatio-Temporal Disease Data, (2017), Spatial and Spatio-temporal Epidemiology, 23:47-57
- Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2017). shiny: Web Application Framework for R. https://CRAN.R-project.org/package=shiny
- Barbara Borges and JJ Allaire (2017). flexdashboard: R Markdown Format for Flexible Dashboards. https://CRAN.R-project.org/package=flexdashboard

Thanks!

https://Paula-Moraga.github.io

Twitter @_PaulaMoraga_