Modelling spatial data in `R` with `CARBayes`

Part 3: Modelling spatio-temporal data

Duncan Lee and Eilidh Jack

1. Introduction

This practical session will show you how to undertake spatio-temporal modelling in `R`. Specifically, this session will cover:

- Fitting and checking convergence of spatio-temporal correlation models using the `CARBayesST` software package.
- Summarising the result from a fitted model.

2. Working example

The data analysed here consist of the numbers of admissions to hospital due to coronary heart disease (CHD) in each Intermediate Zone (IZ) in the Greater Glasgow and Clyde health board between 2002 and 2012. The data are stored in the file `Scotland space time data.csv` and contain the following columns

- `IZ -` The code for each intermediate zone, which is the unique identifier for this dataset.
- `year -` The year the data relate to.
- `Y -` The number of admissions to hospital due to coronary heart disease.
- `E -` The expected number of admissions to hospital due to coronary heart disease.

For these data we aim to address the following two questions of interest.

1. What is the overall temporal trend in CHD risk across the Greater Glasgow and Clyde health board?
2. What are the changing spatial dynamics in CHD risk across the Greater Glasgow and Clyde health board?

3. Reading in and formatting the data

The first step is to read in the data and shapefiles (same shapefiles as in the previous practical sessions), which can be done using the following code. For this to work ensure that the `R` code file you write your analysis in is saved in the same directory as the data. Then set the working directory to the current directory. This latter step is done via the `Rstudio Session` menu, and then selecting `Set Working Directory` and then `To Source File Location`.

```
#### Data
dat <- read.csv(file="Scotland space time data.csv")
head(dat)
```

```
##          IZ year  Y        E
## 1 S02000260 2002 30 25.29314
## 2 S02000261 2002 13 10.71034
## 3 S02000262 2002 27 25.92405
## 4 S02000263 2002 18 21.42882
## 5 S02000264 2002 37 36.59268
## 6 S02000265 2002 11 15.73416
```

```
#### Shapefiles
library(shapefiles)
```

```
## Loading required package: foreign
```

```
##
## Attaching package: 'shapefiles'
```

```
## The following objects are masked from 'package:foreign':
##
##     read.dbf, write.dbf
```

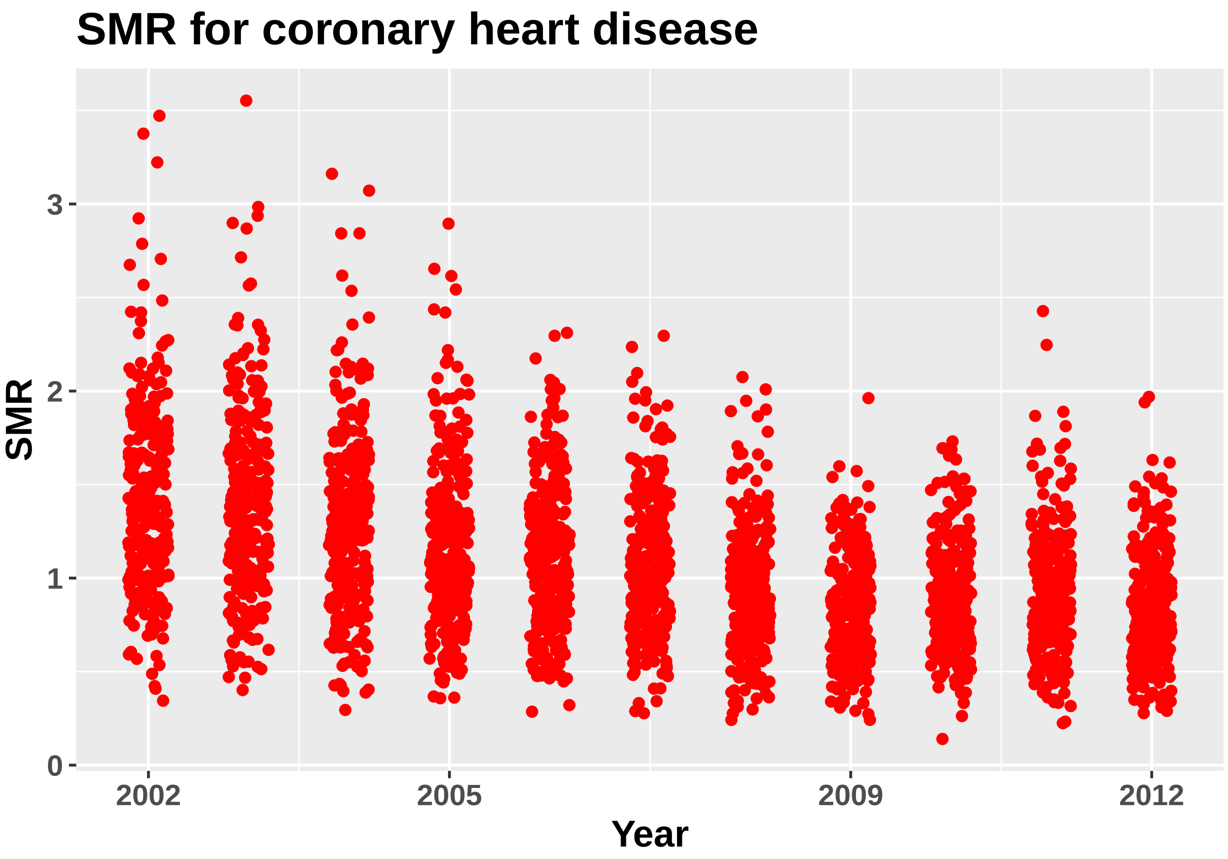
```
shp <- read.shp(shp.name = "ScotlandIZ.shp")
dbf <- read.dbf(dbf.name = "ScotlandIZ.dbf")
```

Now the data are read in, creating and then plotting the SMR (observed / expected) in each IZ and year would give a good initial look at the data. The SMR can be added to the dataset using the following code:

```
dat$smr <- dat$Y / dat$E
```

We now plot the SMR using the code below.

```
library(ggplot2)
ggplot(dat, aes(x=jitter(year), y=smr)) +
  geom_point(colour="red") +
  xlab("Year") +
  scale_x_continuous(breaks=c(2002, 2005, 2009, 2012 )) +
  ylab("SMR") +
  ggtitle("SMR for coronary heart disease") +
  theme(text=element_text(face="bold", size=14))
```



Here, the `jitter()` function jitters the data in the x (year) direction so that all the points can be seen. Additionally, the `scale_x_continuous()` function specifies the locations of the axis labels in the x (year) direction.

The plot shows a clear decreasing temporal trend over the entire time period. However, to get an estimate and a 95% credible interval for this overall temporal effect we need to fit the `ST.CARanova()` model to the data. To do this we first need to create the neighbourhood matrix, which is computed from a `spatialPolygonsDataFrame` object. To create this we first need to combine a subset of the data (just one years worth of data, below uses 2002 in `dat.2002`) with the shapefiles, which can be done using the following code.

```
library(sp)
library(CARBayes)
```

```
## Loading required package: MASS
```

```
## Loading required package: Rcpp
```

```r
dat.2002 <- dat[dat$year==2002, ]
rownames(dat.2002) <- dat.2002$IZ
sp.dat <- combine.data.shapefile(data=dat.2002, shp=shp, dbf=dbf)
```

Note, we cannot combine the entire `dat` object with the shapefiles because the `combine.data.shapefile()` function requires each area in the shapefiles to be matched by at most one row in the data set (here we have 9 time points (rows) per area). Then the neighbourhood matrix `W` and its list object variant (`W.list`) can be constructed from the following code.

```r
library(spdep)
```

```
## Loading required package: Matrix
```

```r
W.nb <- poly2nb(sp.dat, row.names = rownames(sp.dat@data))
W <- nb2mat(W.nb, style = "B")
W.list <- nb2listw(W.nb, style = "B")
```

Then before we undertake any modelling we assess the presence of spatial correlation in the SMR, to determine if a spatial correlation model is appropriate. To illustrate this we compute Moran's I statistic for 2002 using the code:

```r
moran.mc(dat$smr[dat$year==2002], listw = W.list, nsim = 10000)
```

```
##
##  Monte-Carlo simulation of Moran I
##
## data:  dat$smr[dat$year == 2002]
## weights: W.list
## number of simulations + 1: 10001
##
## statistic = 0.31993, observed rank = 10001, p-value = 9.999e-05
## alternative hypothesis: greater
```

4. Estimating the overall temporal trend in CHD risk

To estimate the overall temporal trend in CHD risk we need to fit the `ST.CARanova()` model to the data. Given the number of admissions to hospital is a discrete and rare count, the following Poisson data model is appropriate for area k and time period t:

$$Y_{kt} \sim \text{Poisson}(E_{kt}\theta_{kt})$$

Here (Y_{kt}, E_{kt}) respectively denote the number of admissions to hospital and the expected number of admissions computed using indirect standardisation. Thus θ_{kt} represents the estimated relative risk of admission to hospital relative to the expected counts. This risk θ_{kt} is modelled by 3 terms:

- An overall temporal trend common to all areas.
- An overall spatial trend common to all time periods.
- An interaction effect that quantifies discrepancies from the above.

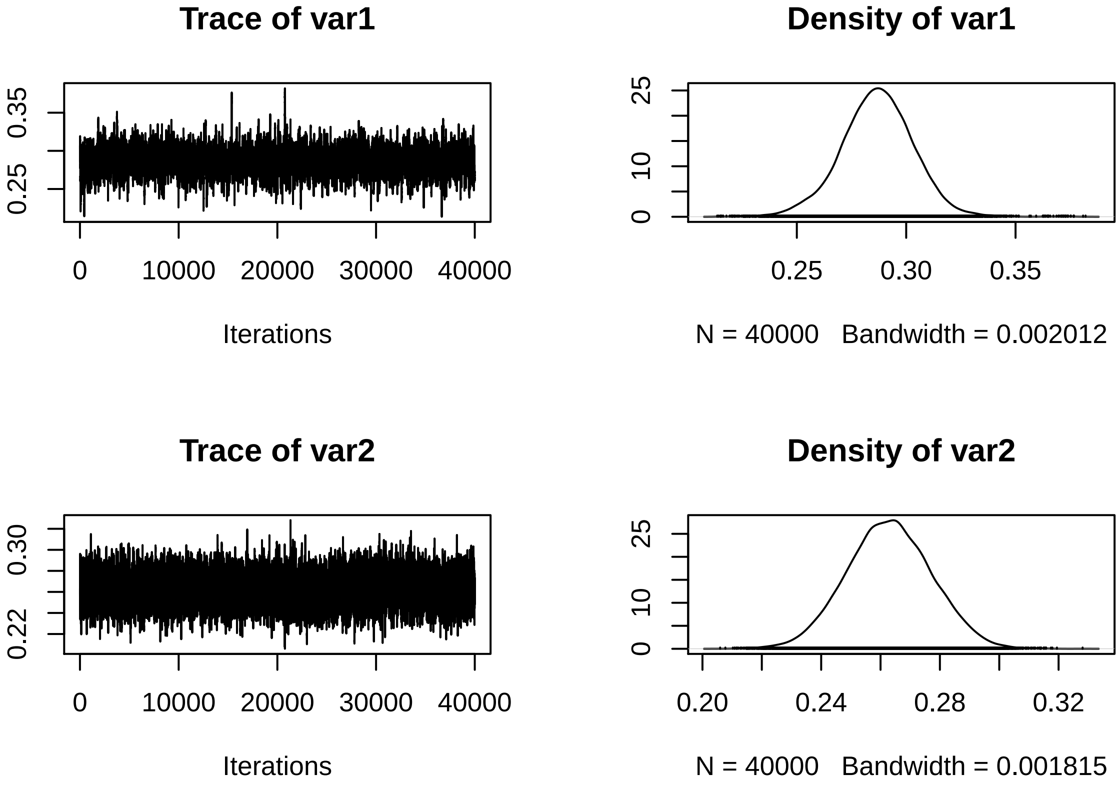
This model can be fitted using the following code:

```r
library(CARBayesST)
model1 <- ST.CARanova(formula=Y~offset(log(E)), family="poisson", data=dat,
            W=W, burnin=10000, n.sample=50000, verbose=FALSE)
print(model1)
```

```
##
## #################
## #### Model fitted
## #################
## Likelihood model - Poisson (log link function)
## Latent structure model - spatial and temporal main effects and an interaction
## Regression equation - Y ~ offset(log(E))
##
## ############
## #### Results
## ############
## Posterior quantities for selected parameters and DIC
##
##             Median   2.5%  97.5% n.sample % accept n.effective Geweke.diag
## (Intercept) 0.0225 0.0143 0.0304    40000     35.2      9713.4         0.9
## tau2.S      0.1387 0.1055 0.1802    40000    100.0      5273.2         0.0
## tau2.T      0.0089 0.0039 0.0258    40000    100.0     11499.9        -0.6
## tau2.I      0.0261 0.0228 0.0297    40000    100.0      1298.7         1.5
## rho.S       0.6375 0.4127 0.8673    40000     44.7      3342.9         0.2
## rho.T       0.9129 0.5926 0.9926    40000     44.5      4911.8         0.2
##
## DIC =  19329.59       p.d =  1323.988       LMPL =  -8485.893
```

Convergence of the MCMC samples can be checked by plotting traceplots of sample parameters. For example, samples from the temporal trend terms (denoted `delta`) for the first two time periods can be plotted by the code below. The parameters appear to have converged in both cases.

```r
plot(model1$sample$delta[ ,1:2])
```



To estimate the underlying temporal trend in the risk $\{\theta_{kt}\}$, we recall that the model has the following structure

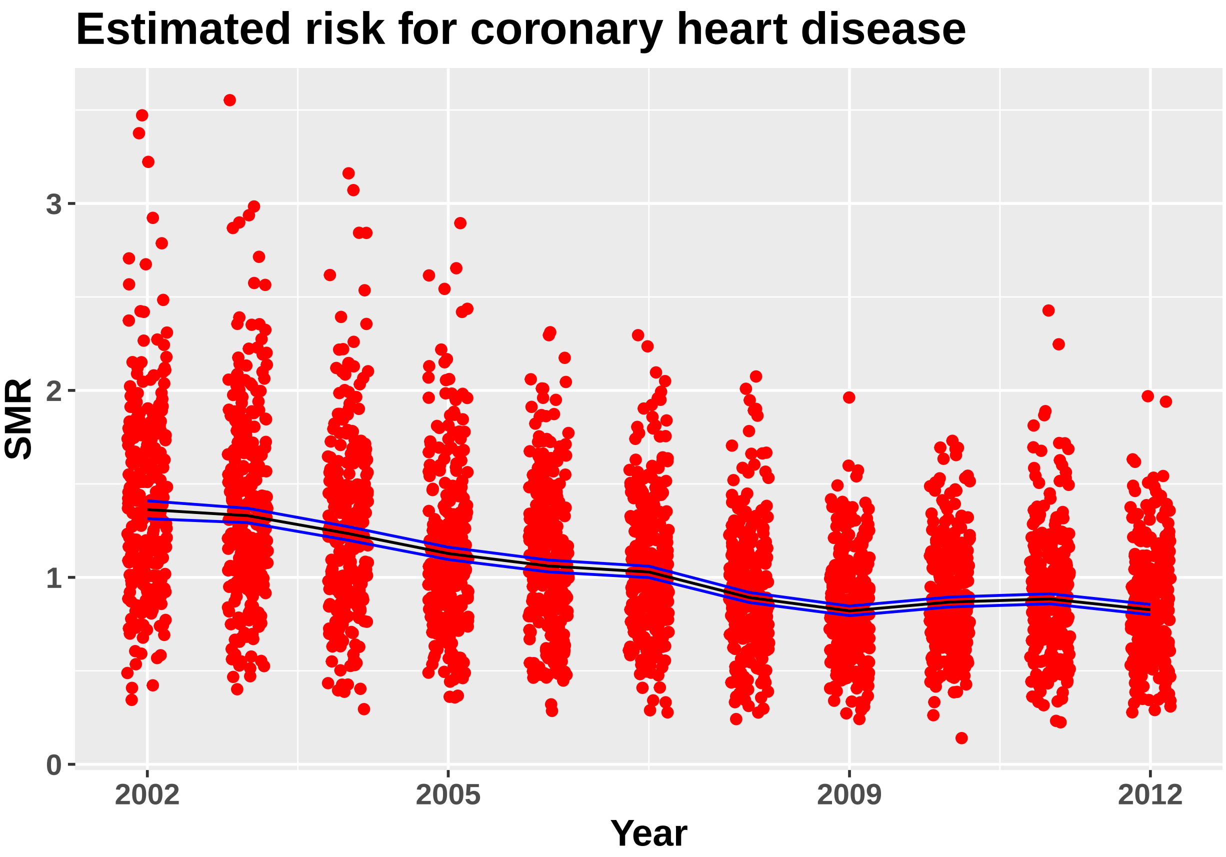
$$\ln(\theta_{kt}) = \beta_0 + \delta_t + \phi_k + \psi_{kt}$$

where β_0 is the intercept term and δ_t is the average (over all spatial areas) temporal trend at time t. Thus estimates and 95% credible intervals for the average temporal trend can be computed using the following code.

```r
theta.trend <- as.data.frame(array(NA, c(11,4)))
colnames(theta.trend) <- c("year", "median", "LCI", "UCI")
theta.trend$year <- 2002:2012
  for(i in 1:11)
  {
  temp <- quantile(model1$samples$beta[ ,1] + model1$samples$delta[ ,i],
                   c(0.5, 0.025, 0.975))
  theta.trend[i,2:4] <- exp(temp)
  }
```

Here the log transformation $\ln(\theta_{kt})$ above is undone by the inverse transformation $\exp(\beta_0+\delta_t)$. Here the `quantile()` function computes quantiles of a distribution, and the 95% credible interval is computed from the 2.5th percentile and the 97.5th percentile. These elements can then be combined to produce a graph of the estimated temporal trend and a 95% credible interval. The code below produces this graph, where the red dots are the SMR values as before.

```
ggplot(dat, aes(x=jitter(year), y=smr)) +
  geom_point(colour="red") +
  xlab("Year") +
  scale_x_continuous(breaks=c(2002, 2005, 2009, 2012 )) +
  ylab("SMR") +
  ggtitle("Estimated risk for coronary heart disease") +
  theme(text=element_text(face="bold", size=14)) +
  geom_line(mapping=aes(x=year, y=median), data=theta.trend) +
  geom_line(mapping=aes(x=year, y=LCI), colour="blue", data=theta.trend) +
  geom_line(mapping=aes(x=year, y=UCI), colour="blue", data=theta.trend)
```



Here the last 3 lines add the posterior median (black line) and 95% credible interval (blue lines) for the estimated region-wide temporal trend. This graph shows a clear and significant decline in SMR between 2002 and 2009, as a horizontal line will not fit within the black 95% credible intervals. The last part of the period from 2009 to 2012 shows a fairly constant risk trend.

5. Estimating the evolving spatial surface in CHD risk

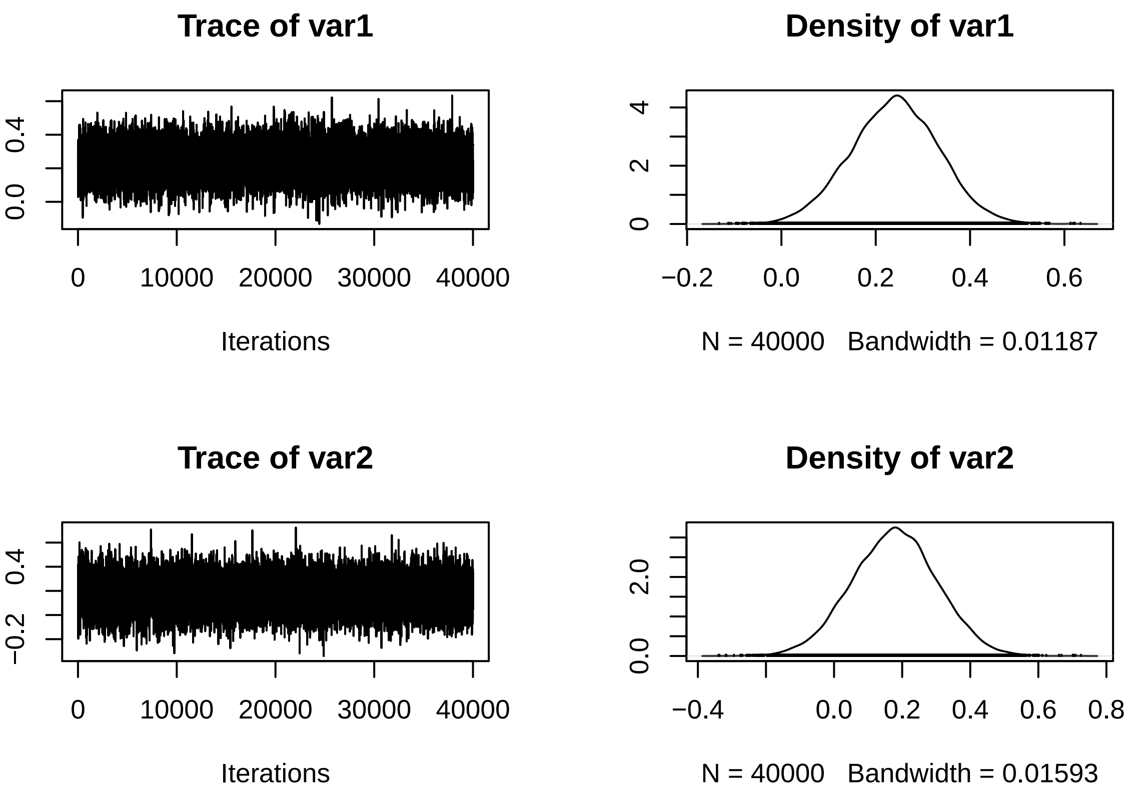
The ST.CARanova() model is designed to estimate overall spatial and temporal effects, but is not appropriate for estimating an evolving spatial surface, as it assumes the same spatial surface for all time periods. Therefore we fit the ST.CARar() model to estimate the extent to which the spatial surface evolves over time. The model assumes the spatial surface at a given time period is equal to a proportion of the surface at the previous time period and spatially correlated error. This model can be fitted using the following code:

```
model2 <- ST.CARar(formula=Y~offset(log(E)), family="poisson", data=dat,
    W=W, burnin=10000, n.sample=50000, verbose=FALSE)
print(model2)
```

```
##
## #################
## #### Model fitted
## #################
## Likelihood model - Poisson (log link function)
## Latent structure model - Autoregressive CAR model
## Regression equation - Y ~ offset(log(E))
##
## ############
## #### Results
## ############
## Posterior quantities for selected parameters and DIC
##
##              Median   2.5%  97.5% n.sample % accept n.effective Geweke.diag
## (Intercept) 0.0238 0.0162 0.0315    40000     35.3      5183.8         1.9
## tau2        0.0782 0.0676 0.0907    40000    100.0       768.7        -0.2
## rho.S       0.9223 0.8744 0.9571    40000     44.0      2195.9        -0.2
## rho.T       0.8662 0.8311 0.8995    40000    100.0      1372.5        -0.6
##
## DIC =  19395.18       p.d =  1103.598       LMPL =  -8699.956
```

Again, convergence of the MCMC samples can be checked by plotting traceplots of sample parameters. For example, samples from the first 2 spatio-temporal effects (denoted phi) for areas 1 and 2 in time period 1 can be plotted by the code below. The resulting plots are shown below and the parameters appear to have converged in both cases.

```
plot(model2$sample$phi[ ,1:2])
```



To show the evolution in the spatial surface over time, we map the estimated proportion susceptible in 2002, 2005, 2009 and 2012. These four years of risks are computed using the code below.

```
risk.median.all <- model2$fitted.values / dat$E
risk.2002 <- risk.median.all[dat$year==2002]
risk.2005 <- risk.median.all[dat$year==2005]
risk.2009 <- risk.median.all[dat$year==2009]
risk.2012 <- risk.median.all[dat$year==2012]
```

Then add them to the spatial data set using the code:

```
sp.dat@data$risk.2002 <- risk.2002
sp.dat@data$risk.2005 <- risk.2005
sp.dat@data$risk.2009 <- risk.2009
sp.dat@data$risk.2012 <- risk.2012
```

These four sets of risks can then be plotted, but before that the spatial data object `sp.dat` needs to transformed into a `data.frame` as before using the code below.

```
#### Load the libraries required
library(rgeos)
```

```
## rgeos version: 0.3-23, (SVN revision 546)
##  GEOS runtime version: 3.6.1-CAPI-1.10.1 r0
##  Linking to sp version: 1.2-4
##  Polygon checking: TRUE
```

```r
library(maptools)
```

```
## Checking rgeos availability: TRUE
```

```r
#### Turn into a data.frame
sp.dat@data$id <- rownames(sp.dat@data)
temp1 <- fortify(sp.dat, region = "id")
sp.dat2 <- merge(temp1, sp.dat@data, by = "id")


#### Change the scale to kilometres
sp.dat2$long <- sp.dat2$long / 1000
sp.dat2$lat <- sp.dat2$lat / 1000
```

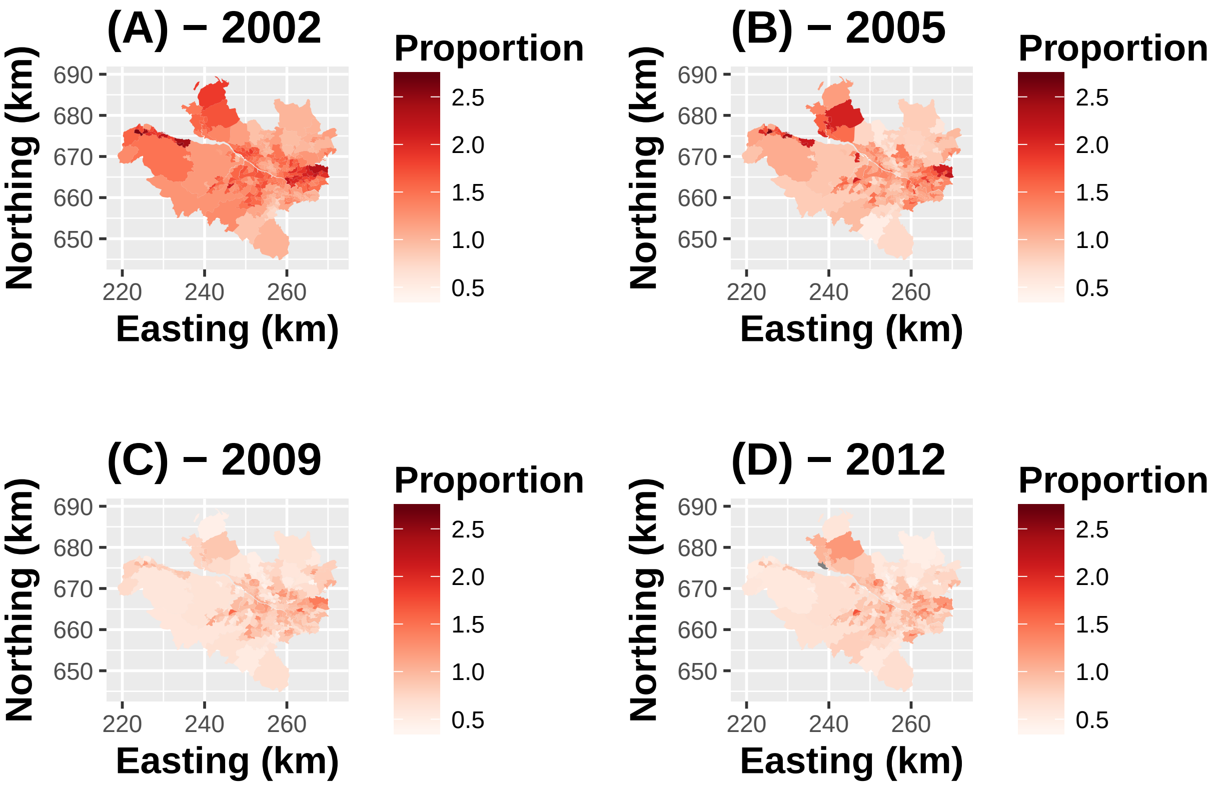
Finally, the four spatial surfaces can be plotted using the code below. Note, that here to put each plot on the same page, we first create each plot separately and save them to variables called `map2002`, `map2005`, `map2009`, `map2012`.

```r
#### Create each map separately
library(RColorBrewer)

## Map 2002
map2002 <-  ggplot(data = sp.dat2, aes(x=long, y=lat, group=group,
            fill = risk.2002)) +
  geom_polygon() +
  coord_equal() +
  xlab("Easting (km)") +
  ylab("Northing (km)") +
  labs(title = "(A) - 2002", fill = "Proportion") +
  theme(title = element_text(face="bold", size=14)) +
  scale_fill_gradientn(colors=brewer.pal(n=9, name="Reds"),
            limits=c(0.4,2.7))

## Map 2005
map2005 <-  ggplot(data = sp.dat2, aes(x=long, y=lat, group=group,
            fill = risk.2005)) +
  geom_polygon() +
  coord_equal() +
  xlab("Easting (km)") +
  ylab("Northing (km)") +
  labs(title = "(B) - 2005", fill = "Proportion") +
  theme(title = element_text(face="bold", size=14)) +
  scale_fill_gradientn(colors=brewer.pal(n=9, name="Reds"),
            limits=c(0.4,2.7))
```

```
## Map 2009
map2009 <-  ggplot(data = sp.dat2, aes(x=long, y=lat, group=group,
            fill = risk.2009)) +
  geom_polygon() +
  coord_equal() +
  xlab("Easting (km)") +
  ylab("Northing (km)") +
  labs(title = "(C) - 2009", fill = "Proportion") +
  theme(title = element_text(face="bold", size=14)) +
  scale_fill_gradientn(colors=brewer.pal(n=9, name="Reds"),
            limits=c(0.4,2.7))

## Map 2012
map2012 <-  ggplot(data = sp.dat2, aes(x=long, y=lat, group=group,
            fill = risk.2012)) +
  geom_polygon() +
  coord_equal() +
  xlab("Easting (km)") +
  ylab("Northing (km)") +
  labs(title = "(D) - 2012", fill = "Proportion") +
  theme(title = element_text(face="bold", size=14)) +
  scale_fill_gradientn(colors=brewer.pal(n=9, name="Reds"),
            limits=c(0.4,2.7))
```

Then finally we combine the maps into a single plot using the code below via the `gridExtra` package.

```
library(gridExtra)
grid.arrange(map2002, map2005, map2009, map2012, ncol=2, nrow=2)
```



The resulting plots show the evolution in the SMR over space and time. The figure shows a clear negative region-wide trend from 2002 until 2009, and more static behaviour thereafter. However, spatially, the pattern is largely unchanged, with the highest risk areas for each year being largely similar and in the east of the city.

References