Kristina White
Professor Castruccio
ACMS 60855
May 18, 2021

Final: Scotland Lip Cancer Data

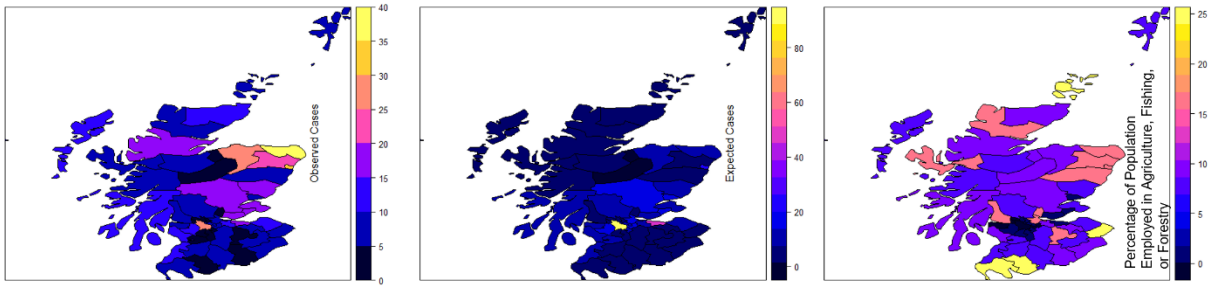
The Scotland Lip Cancer dataset contains data on the incidence of male lip cancer in Scotland from 1975-1980 for each county. The data is for each of the 56 districts/counties following The Local Government (Scotland) Act passed by the British Parliament in 1973 that created and divided nine regional authorities into 53 districts as well as three single-tier Island Authorities (Western, Orkney, and Shetland), which are considered districts in this dataset (Cressie, 1993).

Lip cancer for this dataset is considered cancer of the vermillion border of the lip, or the light roll that forms the border between the lip or mouth and the surrounding skin, as opposed to caner of the skin of the lip (Cressie, 1993). According to descriptive epidemiological studies, lip cancer usually arises in the lower lip, is more common in rural areas versus urban ones, and occurs more frequently in males than in females (Cressie, 1993). Exposure to sunlight is thought to contribute to higher incidence of lip cancer among rural populations, and there are higher rates of lip cancer among people who work outdoors exposed to the sun (Cressie 1993). Smoking is also thought to contribute to lip cancer (Cressie 1993).

The Scotland Lip Caner dataset contains three variables for each county *i*. The first is the number observed cases of male lip cancer in each county, denoted O_i. The second is the expected number of cases in each county, denoted E_i. The expected number of cases for each county are based on the MLEs of the age effects (ξ_j) in a simple multiplicative model $\theta_i E_i = \theta_i(\sum_j Y_{ij}\xi_j)$, where Y_{ij} is the number of person-years of observation giving rise to the observed number of cases in the jth age group in the i-th county and θ_i is the relative risk for that county, and calculated using a recursive algorithm (Clayton and Kaldor, 1987). Essentially, the expected number of cases in a county is based on the age and sex distributions of that particular county. The third variable in the dataset, denoted X_i, is the percentage of the population (from 1975-1980) employed in agriculture, fishing, or forestry. This is an important covariate because workers in these occupations spend a lot of time outside exposed to sunlight, which is known to increase risk of lip cancer. If certain counties employ a lot of people in these industries, it is possible these counties will see higher incidence of lip cancer. It is likely the data are likely reliable as they are included in the International Agency for Research on Cancer (an intergovernmental agency that is part of the World Health Organization) scientific publication *Atlas of Cancer in Scotland, 1975-1980: Incidence and Epidemiological Perspective*.

Looking at the data, observed cases range from 0 to 39 cases per county, expected cases range from 1.1 to 88.7, and the percentage of the population employed in agriculture, fishing, or

forestry ranges from 0 to 24% in a county. Figure 1 shows spatial plots for each of the three variables. There appears to be higher numbers of observed cases in the central west to northwest part of Scotland. This area includes largely agricultural Aberdeenshire and the Gampian mountain range, which would make sense since rural areas tend to have higher incidences of lip cancer. The percentage of the population employed in agriculture, fishing, or forestry is lowest in the area immediately around Glasgow, which makes sense as this is Scotland's largest city. There does not appear to be much spatial dependence, however, for expected cases, as they are on the lower end of the spectrum throughout most of the country, though this number is high close to Glasgow and in one coastal county.



(a) Observed Cases

(b) Expected Cases

(c) Percentage of Population Employed in Agriculture, Fishing, or Forestry

Figure 1 Exploratory Plots of Variables

While the covariates do not appear normally distributed based on their spread, the normality assumption for the covariates is not entirely unreasonable. Using a normal distribution as a prior distribution for the fixed effects is natural choice, and we use a mean of 0 and a large variance since we have no actual prior information. It does makes sense, however, to use a Poisson distribution for O_i since the observed number of cases of lip cancer is count data, and the Poisson distribution is often used to count the number of successes (in this case, cases of male lip cancer) during a specified time or space interval. We will now assume $O_i \; ind \sim P(E_i \mu_i)$, where μ_i is a number representing the region risk, such as if the expected number of observed cases is higher or smaller than E_i. There could be potential problems with this model. Depending on how the relative risk is calculated, the geographic distribution of cancer incidence may be misrepresented, with factors such as different population size among the counties being unaccounted for or altering the effect (Clayton and Kaldor, 1987). Additionally, the data may not be completely independent. For example, someone may live in one county, but work in another.

We first conduct a non-spatial analysis on the data using the Poisson loglinear model:

$$\log(\mu_i) = \beta_0 + X_i\beta_1$$

The model is summarized in Table 1. X_i has a significant (p-value < 2e-16) positive coefficient. Thus, the percentage of a county's population employed in agriculture, fishing, or forestry increases the odds of that county having a higher number of observed cases of male lip cancer.

Table 1: Summary of Non-Spatial Model Coefficients	
Coefficient	Estimate
β_0	-0.542268*
β_1	0.073732*

*Significant at the $\alpha = 0.05$ level

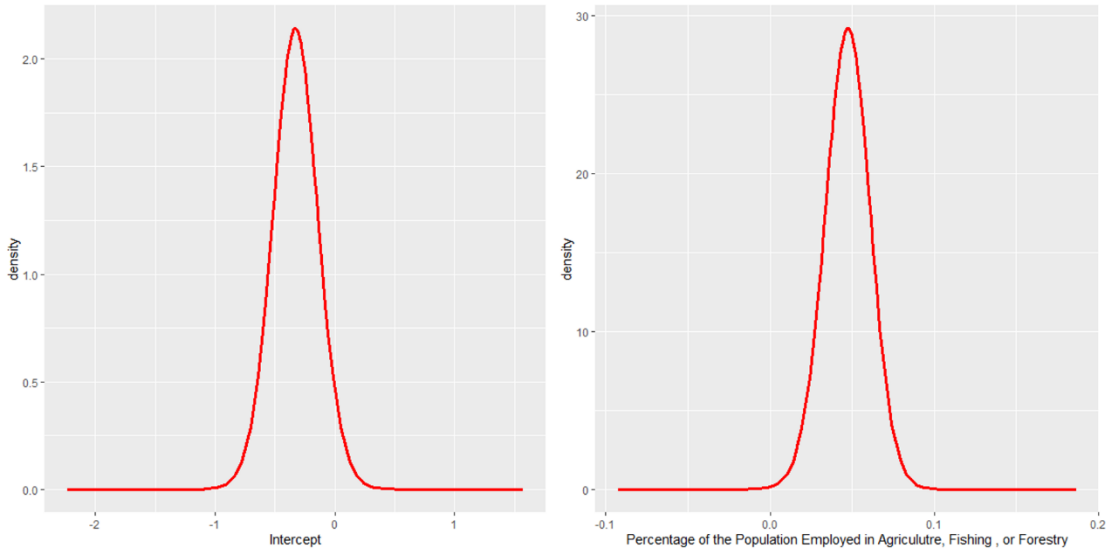
Looking at Moran's I and Geary's C, which are computed from the model residuals, both have very significant p-values (0.001998 and 0.000999, respectively). This is strong indication there is spatial dependence in the data and further spatial analysis is needed. We perform this analysis with the following Bayesian model:

$$O_i \; ind \sim P(E_i\mu_i)$$

$$\log(\mu_i) = \beta_0 + X_i\beta_1 + \varphi_i$$

where $\varphi = (\varphi_1, \ldots, \varphi_n)$ is a CAR process with parameters ρ and τ^2. For these parameters, the defaults for INLA in R were used. The model is summarized in Table 2, and plots of the marginal posteriors for the fixed effects are shown in Figure 2. From the plots, the marginal posteriors for the fixed affects appear to follow a normal distribution.

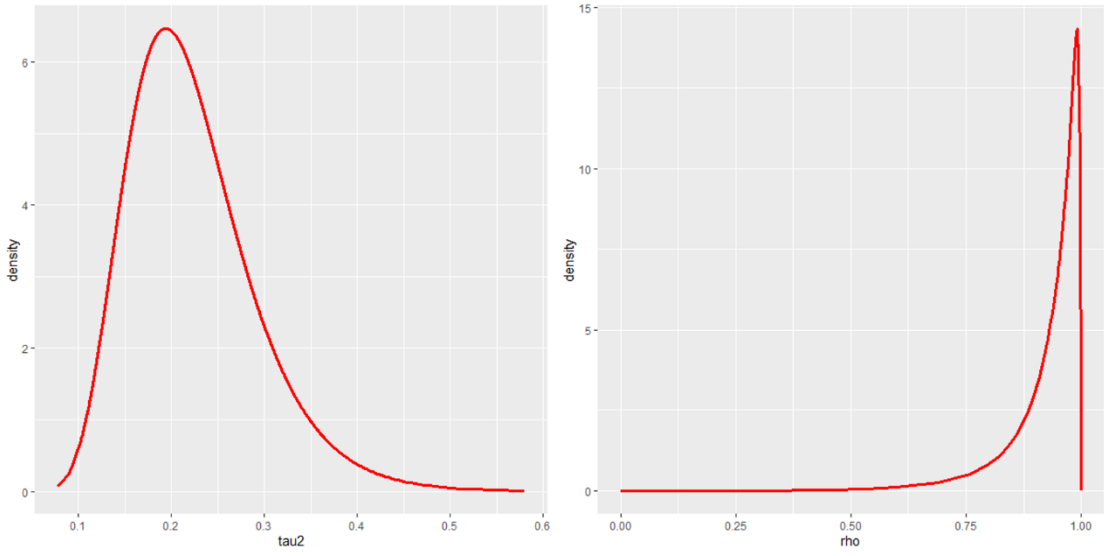
Table 2: Summary of Spatial Model Fixed Effects			
Fixed Effect	Mean	0.025quant	0.975quant
β_0	-0.330	-0.703	0.046
β_1	0.047	0.019	0.074



(a) β_0 (a) β_1

Figure 2 Marginal Posteriors for Fixed Effects

We can also see in Figure 3 what ρ and τ^2 look like. They are summarized in Table 3.



(a) Plot for τ^2 (b) Plot for ρ

Figure 3 Posterior Plots of Parameters

Table 2: Summary of Spatial Model Parameters			
Parameter	Mean	0.025quant	0.975quant
τ^2	0.22261	0.11739	0.384833
ρ	0.9251844	0.6891356	0.9963058

Figure 4 shows the fitted values for the spatial model. They represent the average incidence, or number of cases, of male lip cancer by county. We can see higher incidence of lip cancer rural Aberdeenshire and the Gampian mountain range. We also see high incidence of cancer near Glasgow, which is interesting given lip cancer is usually less frequent in urban areas and the area does not have a high percentage of people employed in industries like agriculture, fishing, and forestry. There is also higher incidence in one coastal county near Edinburgh, even though this county also has a low percentage of employed in agriculture, fishing, and forestry. Several counties with higher percentages of the population employed in these industries have higher predicted levels of incidence than close-by counties with lower percentages of the population working in those industries.

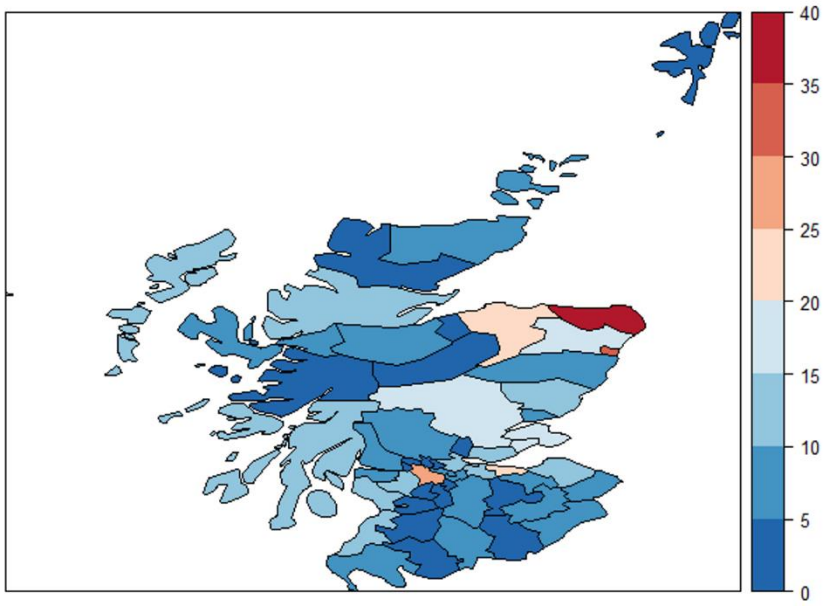


Figure 4 Average Incidence of Male Lip Cancer by County

Finally, this analysis looked at Scotland Lip Caner Data, which contains the observed number of cases of male lip cancer for 56 counties in Scotland from 1975-1980, the expected number of cases for each county, and the percentage of each county's population employed in agriculture, fishing, or forestry. The data are contained in the International Agency for Research on Cancer (IARC) scientific publication *Atlas of Cancer in Scotland, 1975-1980: Incidence and Epidemiological Perspective*. As the IARC is part of the World Health Organization, the data is likely reliable. Lip cancer is considered cancer of the vermillion border (light area that forms the border between the lip or mouth and the surrounding skin) rather than cancer of the skin of the lip. It is more common in males than females and in rural areas than urban areas, and it most often occurs in the lower lip. Sunlight exposure is thought to contribute to higher incidence of lip cancer among rural populations. An analysis of the Scotland Lip Cancer data found the percentage of the population employed in agriculture, fishing, or forestry affected lip cancer incidence. This would be expected because people in these industries often spend their days outdoors exposed to sunlight. Counties with higher percentages of people working in these industries had more predicted cases of male lip cancer than counties with lower percentages employed in these industries. There was also high incidence of male lip cancer close to the major cities of Glasgow and Edinburgh. While lip cancer tends to be less prevalent in urban areas, that is not true of cancer overall, and given the larger populations of these cities, it is not unreasonable to see a higher number of observed cases there.

Appendix A

R Code

```
library(sp)
library(spdep)
library(splines)
library(grid)
library(INLA)
library(tidyverse)
library(RColorBrewer)

load("lipscotland.Rdata")

summary(lipscotland$observed)

lipscotland$observed.plot=lipscotland$observed
brks.observed=c(0,5,10,15,20,25,30,35,40)
spplot(lipscotland, "observed.plot", at = brks.observed)
grid.text("Observed Cases", x=unit(0.8, "npc"), y=unit(0.50, "npc"), rot=90)

summary(lipscotland$expected)

lipscotland$expected.plot=lipscotland$expected
spplot(lipscotland, "expected.plot")
grid.text("Expected Cases", x=unit(0.8, "npc"), y=unit(0.50, "npc"), rot=90)

summary(lipscotland$pcaff)
```

```
lipscotland$pcaff.plot=lipscotland$pcaff

spplot(lipscotland, "pcaff.plot")

#grid.text("Percentage of the Population Employed in Agriculutre, Fishing , or Forestry",
x=unit(0.8, "npc"), y=unit(0.50, "npc"), rot=90)


### Analysis

# nonspatial model


form <- observed ~ offset(log(expected))+pcaff

model <- glm(formula=form, data=lipscotland, family = poisson)

summary(model)


# Moran's I and Geary's C


resid.model <- residuals(model)

moran.mc(x=resid.model, listw=W.list, nsim=1000)

geary.mc(x=resid.model, listw=W.list, nsim=1000)


############ spatial model (with INLA)


# first we need to standardize W so that each entry is not 1, but 1/N_i, N_i being the total number
of entries for each row

Wsd=matrix(0,nrow=dim(W)[1],ncol=dim(W)[1])

for (i in 1:dim(W)[1]){

  Wsd[i,]=W[i,]/sum(W[i,])

}

# then we multiply by the maximum eigenvalue, so that generic1 allows the usual CAR
representation

Wsd=Wsd*max(eigen(Wsd)$value)
```

```
data.inla=lipscotland@data
data.inla$ID=1:dim(W)[1]

mod.car <- inla(update(form, . ~. +
                    f(ID, model = "generic1", Cmatrix = Wsd)),
                family="poisson",data = data.inla,
                control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE),
                control.predictor = list(compute = TRUE))
summary(mod.car)


mod.car$summary.fixed
#let's get the marginals posteriors for the fixed effects (the betas)
#marginal posterior intercept
df=data.frame(X=mod.car$marginals.fixed[[1]][,1],Y=mod.car$marginals.fixed[[1]][,2])
gp=ggplot()+geom_line(data=df,aes(X,Y),color=c("red"),size=1.2)+xlab("Intercept")+ylab("den
sity")
gp


df=data.frame(X=mod.car$marginals.fixed[[2]][,1],Y=mod.car$marginals.fixed[[2]][,2])
gp=ggplot()+geom_line(data=df,aes(X,Y),color=c("red"),size=1.2)+xlab("Percentage of the
Population Employed in Agriculutre, Fishing , or Forestry")+ylab("density")
gp


mod.car$summary.hyperpar
```

```
# spatial variance
post.tau2 = inla.tmarginal(function (x) exp(-x), mod.car$internal.marginals.hyperpar[[1]])
inla.zmarginal(post.tau2)

# posterior plot for tau2
df=data.frame(X=post.tau2[,1],Y=post.tau2[,2])
gp=ggplot()+geom_line(data=df,aes(X,Y),color=c("red"),size=1.2)+xlab("tau2")+ylab("density"
)
gp

# rho of the CAR for the spatial effect
df=data.frame(X=mod.car$marginals.hyperpar[[2]][,1],Y=mod.car$marginals.hyperpar[[2]][,2])
gp=ggplot()+geom_line(data=df,aes(X,Y),color=c("red"),size=1.2)+xlab("rho")+ylab("density")
gp



# summary of the posterior means for the random effects

summary(mod.car$summary.random$ID[,"mean"])
#they are really small compared to the mean, indicating again small spatial dependence

# plotting the posterior means for the fitted values now

lipscotland$fitted.car = mod.car$summary.fitted.values[,"mean"]
brks.observed=c(0,5,10,15,20,25,30,35,40)
spplot(lipscotland, "fitted.car",  at = brks.observed, col.regions = rev(brewer.pal(8,"RdBu")))
```

References

Clayton D. and Kaldor J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43(3), 671–681.

Cressie, N.A.C (1993) Statistics for Spatial Data. New York: Wiley.