Areal Data Analysis of Childhood Cancer Cases in USA in 2017

Aditi Basu Bal

Department of Statistics, Florida State University

1 Introduction

Childhood cancer is one of the leading causes of death in children in the United States. There can be a variety of factors that determine the most common occurrences of childhood cancer such as age, sex, ethnicity, etc. It is of interest to study the relationship of these covariates with the count of cases which in turn may help us understand which sub-population of children are at the highest risk. The website of Centers for Disease Control and Prevention offers a detailed report on childhood cancer statistics. Our dataset was obtained from the 'United States and Puerto Rico Cancer Statistics' report [3] available on the CDC website. Here, we study the counts of chilhood cancer cases reported in the year 2017 in the United States. The variables of interest are :

1. Count of cases
2. Age Groups: '<1 year', '1-4 years', '5-9 years', '10-14 years' and '15-19 years'
3. Sexes: 'Male' and 'Female'
4. Races: 'American Indian or Alaska Native', 'Asian or Pacific Islander', 'Black or African American', 'White', 'Other races or unknown combined'
5. US State

We consider for each state the highest count of cancer cases which is the response variable $\mathbf{Y}$ for our analysis and note which categories of the covariates this count belongs to. Age group, Sex and Race are the covariates X_1, X_2 and X_3, all of which are categorical variables. The spatial domain is the United States referenced at the 49 states and federal district(excluding Hawaii and Alaska). Thus we treat this as areal data. Several articles and books have been written on statistical analysis of spatial data. This project follows the analytical tools elucidated in [2].

2 Statistical Model

There are several ways to explore areal data to obtain an initial idea about the existence and strength of spatial association in the data. In this project we utilize the Moran's I and Geary's C statistics. We use row-normalized weights from the proximity matrix $\mathbf{W}$ to capture the neighborhood information, the i-th row and j-th column element of which is w_{ij}. Moran's I is given by:

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \bar{Y})^2}$$

Geary's C is given by:

$$C = \frac{(n-1) \sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{2(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \bar{Y})^2}$$

A very popularly used areal data model is Conditionally Autoregressive (CAR) model. It is an additive model given by:

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Y},$$
$$\mathbf{Y} \sim N(\mathbf{0}, (\mathbf{I} - \mathbf{B})^{-1}\mathbf{D})$$

In a Simultaneous Autoregressive (SAR) Model the distribution of $\mathbf{Y}$ is modified to:

$$\mathbf{Y} \sim N(\mathbf{0}, (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Sigma}_\nu(\mathbf{I} - \mathbf{B})^{-1})$$

For our purpose $\mathbf{X}$ is a matrix of size 49 by 4 (three covariates Age Group, Sex and Race and an intercept) and $\boldsymbol{\beta}$ is a 4 length vector of coefficients. $\mathbf{Y}$ is the count of cancer cases for each state and is a 49 length vector. $\mathbf{B}$=$\{b_{ij}\}$, here b_{ij}= $\frac{w_{ij}}{\sum_j w_{ij}}$ and $\mathbf{D}$ is diagonal with D_{ii}= τ_i^2.

3 Analysis

3.1 Data Processing

In the raw datafile observations or counts of cases are available for all combinations of covariates Age Group, Sex and Race for each state. Many of these are zeros if there are no cases for a particular combination of covariate values. To identify the largest number of cancer cases for each

state we sort the counts and the highest entry is taken to be the observation for a given state. This is our response variable $\mathbf{Y}$. The values of Age Group, Sex and Race for this entry are also noted as covariates. Next, since all three of our covariates are categorical variables we encode them numerically. These are the covariates X_1, X_2, and X_3. Some states are found to have reported 0 cases of childhood cancers. To distinguish the values of the covariates for these entries we add an additional code for each covariate denoting '0 count'.

1. The Age Groups '<1 year', '1-4 years', '5-9 years', '10-14 years' and '15-19 years' have a clear heirarchy and therefore can be treated as ordinal data. The encoding for Age groups is as follows 0: '<1 year', 1: '1-4 years', 2: '5-9 years', 3: '10-14 years', 4: '15-19 years' 5: '0 count'.

2. The covariate Sex is binary and naturally we use binary encoding here. 0: 'Male', 1: 'Female' and an additional 2: '0 count'.

3. Race has several categories and it is appropriate to encode it nominally. 0: 'American Indian or Alaska Native', 1: 'Asian or Pacific Islander', 2: 'Black or African American', 3: 'White', 4: 'Other races or unknown combined' and 5: '0 count'.

3.2 Exploratory Analysis

We begin with some initial exploration of the response variable, count of childhood cancer cases for each US state. A reasonable way to visualize this is by plotting a choropleth map. Figure 1 gives an idea of presence of spatial association in the data. The states in the northern half of the mid-west reported fewer chilhood cancer cases. The numbers appear to increase as one moves further away from this region to the peripheral states. The states of California and Texas are hot spots in particular followed closely by the sate of New York. Visually, it seems reasonable to incorporate spatial association in our analysis.

Furthermore, we compute the Moran's I and Geary's C statistics for the data using row-normalized weights. The Moran's I statistic is 0.00315914 with a standard deviate of 0.26833 and p-value 0.3942 which implies that we cannot reject the null hypothesis of no spatial correlation in this data. The Geary's C statistic is obtained to be 0.89445956 with standard deviate estimate of 0.91978. The Geary's C statistic being closer to 1 but not quite equal to it indicates weak positive spatial correlation in the data.

Thus we do not have any solid evidence of presence of spatial association in the data apart from our visual understanding of the choropleth map. Nevertheless, we proceed to fit an appropriate

spatial model to this data.

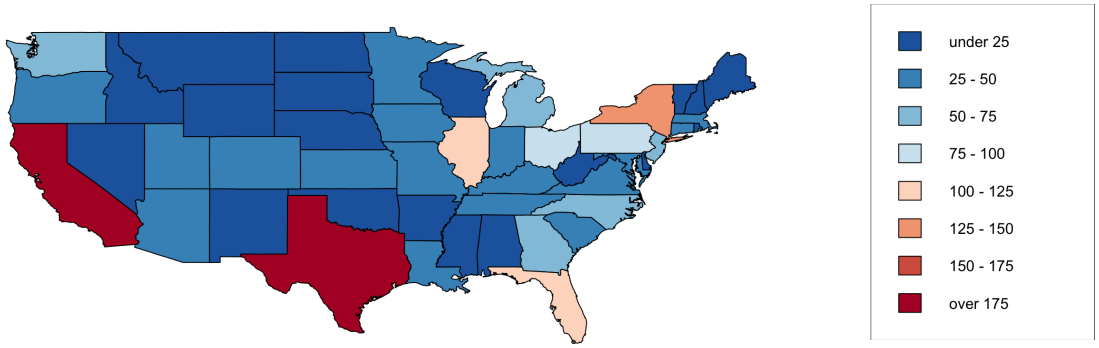


Figure 1: Choropleth map of 2017 childhood cancer cases in 49 U.S. states

3.3 Model Fitting

We first define an appropriate weight matrix to be used in model fitting. Upon initial trials we find that the row-normalized weights would be an acceptable choice. Next, we want to fit the Conditional Autoregressive (CAR) Model using these weights. But first we must check whether the normality assumption of this model is met. In Figure 2 we find that we may consider an

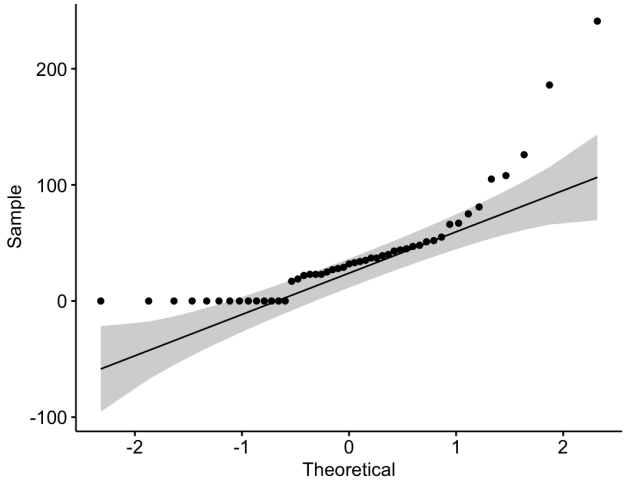


Figure 2: QQplot for childhood cancer cases in 2017 in 49 U.S. states

approximately normal model for our data with a few states on the higher end that appear non-normal. There are several states that reported 0 cases.

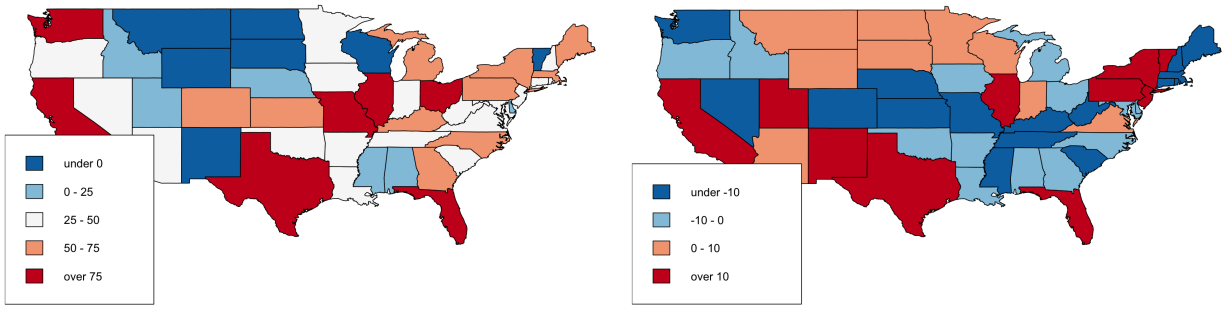
The CAR model is fitted and Table 1 shows the estimates of the regression parameters. The spatial auto-correlation parameter is estimated to -1.08513.

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	38.8775	35.4615	1.0963	0.272934
age	13.4549	6.5481	2.0548	0.039900
sex	-37.7887	12.4904	-3.0254	0.002483
race	-6.5078	11.7181	-0.5554	0.578644

Table 1: Estimates of coefficients of covariates in CAR model

The p-values for covariates Age Group and Sex are less than 0.05 and thus indicate that these are statistically significant and helpful in explaining the count of cancer cases in children while Race is not. The AIC for this model is 499.59.

Next, we look at the choropleth maps of the fitted values in (a) and the residuals in (b) in Figure 3 to visualize the spatial correlation. We observe that there is evidence of spatial correlation as the colors of neighboring states are similar with lower scores in the mid-west states and higher scores in the coastal states.



(a) Fitted Values

(b) Residuals

Figure 3: Choropleth map of Fitted Values and Residuals from CAR model

Next, for further comparison and verification we also fit the Simultaneous Autoregressive (SAR) Model and Table 2 shows the estimates of the regression parameters. The spatial auto-correlation parameter is estimated to -0.70953.

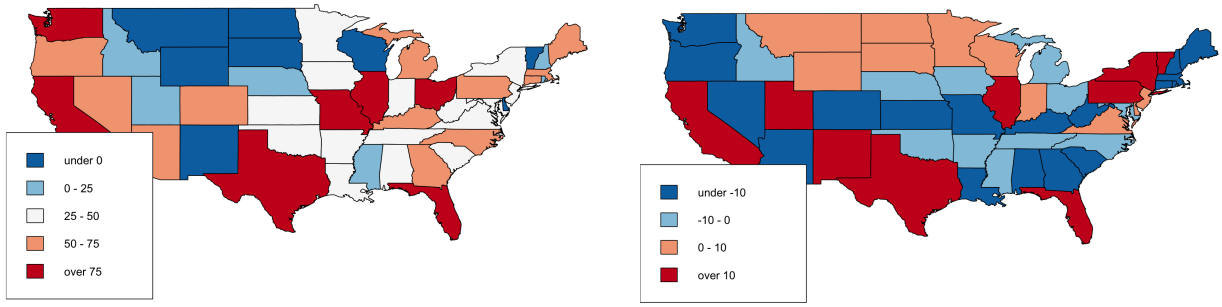
Here, the p-value for covariate Sex is less than 0.05 and thus indicates that this is statistically significant and helpful in explaining the count of cancer cases in children while Age Group and

	Estimate	Std. Error	z value	Pr($>|z|$)
(Intercept)	46.4867	33.9456	1.3694	0.170860
age	10.3389	6.1744	1.6745	0.094038
sex	-40.7932	11.9320	-3.4188	0.000629
race	-4.4193	10.9780	-0.4026	0.687273

Table 2: Estimates of coefficients of covariates in SAR model

Race are not. This is different from the results obtained from the CAR model where Age Group too was found to be significant. The AIC for this model is 498.09.

Next, we look at the choropleth maps of the fitted values in (a) and the residuals in (b) in Figure 4 to visualize the spatial correlation. We observe that the choropleth maps obtained are very similar to those obtained from the CAR model and therefore the interpretation is the same. However, the AIC is greater for the CAR model and thus we will chose the CAR model fit over that for SAR.



(a) Fitted Values

(b) Residuals

Figure 4: Choropleth map of Fitted Values and Residuals from SAR model

4 Discussion

In this project, we studied the number of cases of childhood cancer reported in the year 2017 in the 49 mainland states and federal district in the U.S. and perform spatial data analysis by adopting common areal data models like Conditional Autoregressive (CAR) and Simultaneous Autoregressive (SAR) models. Initially, some exploratory analysis using chorolopleth maps and

statistics like Moran's I and Geary's C indicated that there may be some weak spatial association in the data. This was verified through the parameter estimates of our models, particularly the spatial autocorrelation parameter and the choropleth maps of the fitted values. We included three covariates Age Group, Sex and Race in this study and found Age Group and Sex to be statistically significant. We used the values of AIC for both models to decide which one was a better fit and thus the CAR model was preferred over the SAR model.

5 Bibliography

References

[1] Jonathan Bradley, Areal Data Models Class Notes

[2] S. Banerjee, B.P. Carlin, A.E. Gelfand, Heirarchical Modeling and Analysis for Spatial Data, Second Edition, Chapter 4

[3] United States Cancer Statistics Public Information Data: Incidence United States 1999 - 2017 and Puerto Rico 2005 - June 2017, https://wonder.cdc.gov/wonder/help/cancer-v2017.html

6 R Code

```
library(spdep)
library(maps)
library(maptools)
library(classInt)
library(RColorBrewer)
library(rgdal)
library(spatialreg)
library(stats)
library(ggplot2)
library(ggpubr)
```

```
usa.state=map(database="state",fill=TRUE,plot=FALSE)
state.ID<-sapply(strsplit(usa.state$names,":"),function(x) x[1])
usa.poly=map2SpatialPolygons(usa.state,IDs=state.ID)
usa.nb=poly2nb(usa.poly)
usa.listw_binary=nb2listw(usa.nb,style="B",zero.policy = TRUE)
usa.listw_rownorm=nb2listw(usa.nb,style="W")



filename=paste0("project_data_childhood_cancer.csv")
z<-read.csv(filename, header = FALSE, sep = ",")
z<-as.matrix(z)
z<-z[2:50,]
count<-as.numeric(z[,5])
age<-as.numeric(z[,2])
sex<-as.numeric(z[,3])
race<-as.numeric(z[,4])

##choropleth map
brks=c(0,25,50,75,100,125,150,175,200)
color.pallete=rev(brewer.pal(8,"RdBu"))
class.fitted = classIntervals(var=count,n=8,style="fixed", fixedBreaks=brks,dataPr
color.code.fitted = findColours(class.fitted, color.pallete)
brks = c(-Inf,25,50,75,100,125,150,175,Inf)
plot(usa.poly,col = color.code.fitted)
title(main="Counts")
legend("bottomleft", fill=color.pallete, legend=leglabs(brks))

# Moran's I Geary's C
moran.test(count,listw=usa.listw_rownorm)
geary.test(count,listw=usa.listw_rownorm)
```

```
############ SAR
outBin = spautolm(count~age+sex+race,family="SAR", listw =usa.listw_binary, zero.p
summary(outBin)

outRowN = spautolm(count~age+sex+race,family="SAR", listw =usa.listw_rownorm)
summary(outRowN)

#plot SAR predictions for outBin
fittedBin = fitted(outBin)
brks=c(-25,0,25,50,75,100)
color.pallete=rev(brewer.pal(5,"RdBu"))
class.fitted = classIntervals(var=fittedBin,n=5,style="fixed", fixedBreaks=brks,da
color.code.fitted = findColours(class.fitted, color.pallete)
brks = c(-Inf,0,25,50,75,Inf)
plot(usa.poly,col = color.code.fitted)
title(main="Fitted_Values")
legend("bottomleft", fill=color.pallete, legend=leglabs(brks))

resBin<-residuals.spautolm(outBin)
brks=c(-56,-10,0,10,150)
color.pallete=rev(brewer.pal(4,"RdBu"))
class.fitted = classIntervals(var=resBin,n=4,style="fixed", fixedBreaks=brks,dataP
color.code.fitted = findColours(class.fitted, color.pallete)
brks = c(-Inf,-10,0,10,Inf)
plot(usa.poly,col = color.code.fitted)
title(main="Residuals")
legend("bottomleft", fill=color.pallete, legend=leglabs(brks))

#plot SAR predictions for outRowN
fittedRow = fitted(outRowN)
brks=c(-25,0,25,50,75,102)
color.pallete=rev(brewer.pal(5,"RdBu"))
```

```
class.fitted = classIntervals(var=fittedRow,n=5,style="fixed", fixedBreaks=brks,da
color.code.fitted = findColours(class.fitted, color.pallete)
brks = c(-Inf,0,25,50,75,Inf)
plot(usa.poly,col = color.code.fitted)
title(main="Fitted Values")
legend("bottomleft", fill=color.pallete, legend=leglabs(brks))

resRow<-residuals.spautolm(outRowN)
brks=c(-55,-10,0,10,140)
color.pallete=rev(brewer.pal(4,"RdBu"))
class.fitted = classIntervals(var=resRow,n=4,style="fixed", fixedBreaks=brks,dataP
color.code.fitted = findColours(class.fitted, color.pallete)
brks = c(-Inf,-10,0,10,Inf)
plot(usa.poly,col = color.code.fitted)
title(main="Residuals")
legend("bottomleft", fill=color.pallete, legend=leglabs(brks))

############ CAR
outBin = spautolm(count~age+sex+race,family="CAR", listw =usa.listw_binary, zero.p
summary(outBin)

outRowN = spautolm(count~age+sex+race,family="CAR", listw =usa.listw_rownorm)
summary(outRowN)

#plot CAR predictions for outBin
fittedBin = fitted(outBin)
brks=c(-25,0,25,50,75,100)
color.pallete=rev(brewer.pal(5,"RdBu"))
class.fitted = classIntervals(var=fittedBin,n=5,style="fixed", fixedBreaks=brks,da
color.code.fitted = findColours(class.fitted, color.pallete)
brks = c(-Inf,0,25,50,75,Inf)
plot(usa.poly,col = color.code.fitted)
```

```
title(main="Fitted Values")
legend("bottomleft", fill=color.pallete, legend=leglabs(brks))

resBin<-residuals.spautolm(outBin)
brks=c(-56,-10,0,10,150)
color.pallete=rev(brewer.pal(4,"RdBu"))
class.fitted = classIntervals(var=resBin,n=4,style="fixed", fixedBreaks=brks,dataP
color.code.fitted = findColours(class.fitted, color.pallete)
brks = c(-Inf,-10,0,10,Inf)
plot(usa.poly,col = color.code.fitted)
title(main="Residuals")
legend("bottomleft", fill=color.pallete, legend=leglabs(brks))

#plot CAR predictions for outRowN
fittedRow = fitted(outRowN)
brks=c(-26,0,25,50,75,113)
color.pallete=rev(brewer.pal(5,"RdBu"))
class.fitted = classIntervals(var=fittedRow,n=5,style="fixed", fixedBreaks=brks,da
color.code.fitted = findColours(class.fitted, color.pallete)
brks = c(-Inf,0,25,50,75,Inf)
plot(usa.poly,col = color.code.fitted)
title(main="Fitted Values")
legend("bottomleft", fill=color.pallete, legend=leglabs(brks))

resRow<-residuals.spautolm(outRowN)
brks=c(-55,-10,0,10,130)
color.pallete=rev(brewer.pal(4,"RdBu"))
class.fitted = classIntervals(var=resRow,n=4,style="fixed", fixedBreaks=brks,dataP
color.code.fitted = findColours(class.fitted, color.pallete)
brks = c(-Inf,-10,0,10,Inf)
plot(usa.poly,col = color.code.fitted)
title(main="Residuals")
```

```
legend("bottomleft", fill=color.pallete, legend=leglabs(brks))
```