Heat-related mortality data analysis project

Tetsuya Chau

2022-04-08

Problem background

Identifying and characterizing urban vulnerability to heat is a key step in designing intervention strategies to combat negative consequences of extreme heat on human health. In this study, we seek to identify the most vulnerable populations to extreme heat by analyzing the impact of socio-demographics and minimum daily temperature on the risk of mortality for the 1487 census block groups in the greater Houston area.

The dataset we consider for this analysis has the following variables collected on each of the 1487 census block groups in Houston:

Variable names and their description

*NOAC - The percent of homes without air conditioning in the block group

*MED_AGE - The median age of persons living in the block group

*HispanicPC - Percent hispanic in the block group

*BlackPCT - Percent black in the block group

*under5PCT - Percent under 5 years in the block group

*over65PCT - Percent over 65 years in the block group

*povertyPCT - Percent living below poverty line in the block group

*alonePCT - Percent living alone in the block group

*MinTemp - Average minimum summer temperature in the block group

*RR - The relative risk of mortality in the block group

To analyze this dataset, do the following:

Analysis Questions:

1. Transform the RR into log(RR) and create exploratory plots of the data by looking at the relationship between log(RR) (the response variable) and a few of the explanatory variables. Comment on any general relationships you see from the data.
2. Fit an independent MLR model with a linear effect between log(RR) and all the explanatory variables. Explore the residuals to see if there is evidence of spatial correlation by mapping them and using a Moran’s I or Geary’s C test.
3. Write out a CAR model for analyzing the mortality data in terms of parameters. Explain and interpret any parameters associated with the model.

4. Fit your spatial CAR model (using minit=maxit=1000 iterations and 250 degrees of freedom for positive spatial correlation) and validate any assumptions you made to fit the model.

5. Calculate confidence intervals for the effect of each explanatory variable included in your model. Draw conclusions about who is at greatest risk for heat-related mortality based on your estimated effects. Draw a map of the correlated residuals to try and reach conclusions about areas at risk of heat-related mortality not explained by your explanatory variables.

```
library(rgdal)
```

```
## Loading required package: sp
```

```
## Please note that rgdal will be retired by the end of 2023,
## plan transition to sf/stars/terra functions using GDAL and PROJ
## at your earliest convenience.
##
## rgdal: version: 1.5-29, (SVN revision 1165M)
## Geospatial Data Abstraction Library extensions to R successfully loaded
## Loaded GDAL runtime: GDAL 3.0.4, released 2020/01/28
## Path to GDAL shared files: /usr/share/gdal
## GDAL binary built with GEOS: TRUE
## Loaded PROJ runtime: Rel. 6.3.1, February 10th, 2020, [PJ_VERSION: 631]
## Path to PROJ shared files: /usr/share/proj
## Linking to sp version:1.4-6
## To mute warnings of possible GDAL/OSR exportToProj4() degradation,
## use options("rgdal_show_exportToProj4_warnings"="none") before loading sp or rgdal.
```

```
library(rgeos)
```

```
## rgeos version: 0.5-9, (SVN revision 684)
##  GEOS runtime version: 3.8.0-CAPI-1.13.1
##  Please note that rgeos will be retired by the end of 2023,
## plan transition to sf functions using GEOS at your earliest convenience.
##  Linking to sp version: 1.4-6
##  Polygon checking: TRUE
```

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(spdep)
```

```
## Loading required package: spData
```

```
## To access larger datasets in this package, install the spDataLarge
## package with: `install.packages('spDataLarge',
## repos='https://nowosad.github.io/drat/', type='source')`
```

```
## Loading required package: sf
```

```
## Linking to GEOS 3.8.0, GDAL 3.0.4, PROJ 6.3.1; sf_use_s2() is TRUE
```

```
library(ngspatial)
```

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

```
## Loading required package: batchmeans
```

```
## batchmeans: Consistent Batch Means Estimation of Monte Carlo Standard Errors
## Version 1.0-4 created on 2020-05-07.
## copyright (c) 2012-2020, Murali Haran, Penn State University
##                          John Hughes
## For citation information, type citation("batchmeans").
## Type help(package = batchmeans) to get started.
```

```
## ngspatial: Fitting the Centered Autologistic and Sparse Spatial Generalized
## Linear Mixed Models for Areal Data
## Version 1.2-2 created on 2020-05-08.
## copyright (c) 2013-2020, John Hughes
## For citation information, type citation("ngspatial").
## Type help(package = ngspatial) to get started.
```

```
library(pbapply)
library(car)
```

```
## Loading required package: carData
```

```
library(MASS)
sf_use_s2(FALSE)
```

```
## Spherical geometry (s2) switched off
```

```
mShp <- readOGR(dsn="HoustonHeat",layer="HoustonHeat")
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "/cloud/project/HoustonHeat", layer: "HoustonHeat"
## with 1487 features
## It has 10 fields
```

```
library(broom) #contains tidy() function which converts polygons to data.frame
mShp@data$id <- rownames(mShp@data) #Assign ID to each polygon
mShp.df <- tidy(mShp, region = "id") #Convert polygon info to data.frame()
```

```
## Warning in RGEOSUnaryPredFunc(spgeom, byid, "rgeos_isvalid"): Ring Self-
## intersection at or near point -95.288772059999999 29.646106039999999
```

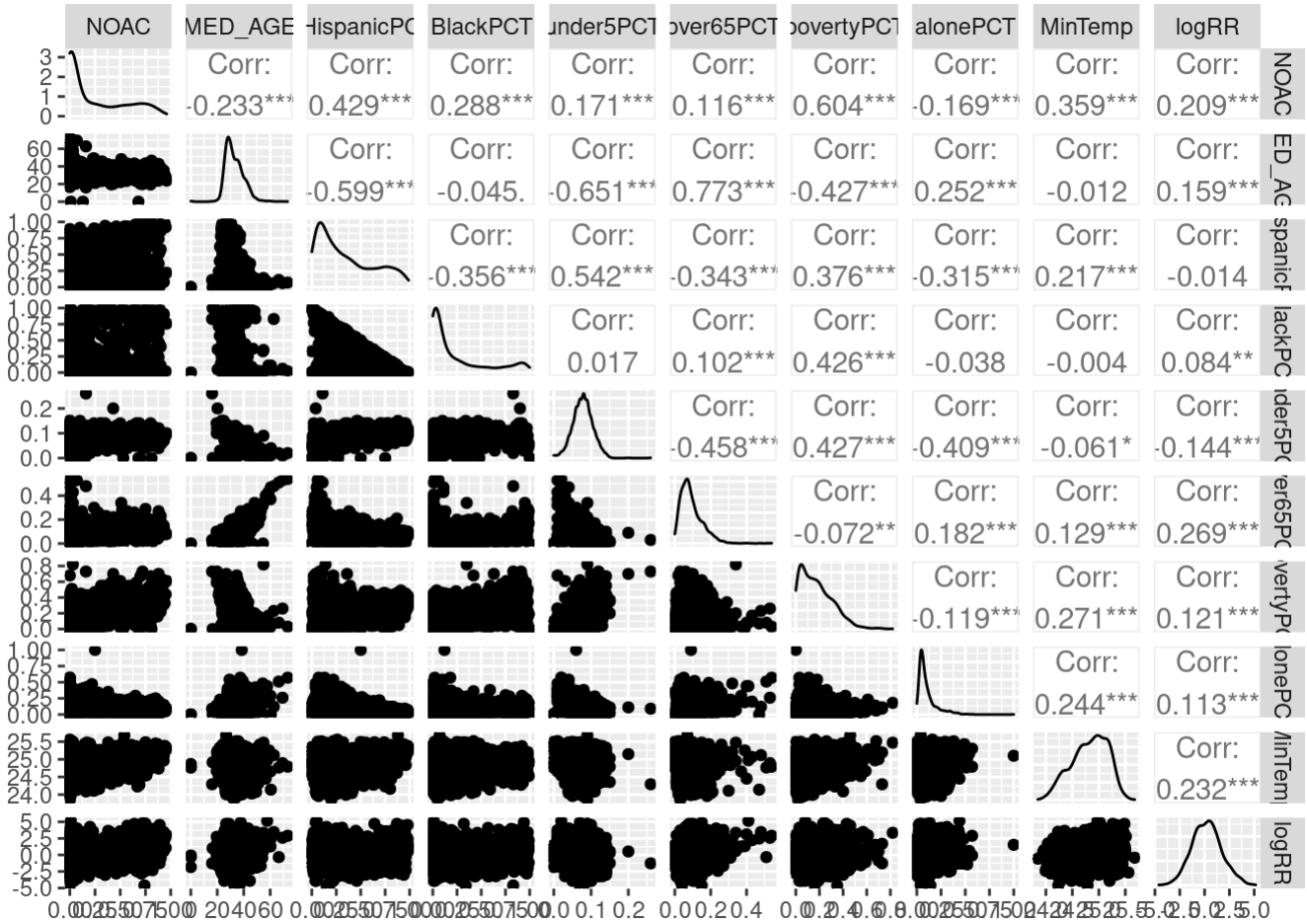
```
## SpP is invalid
```

```
## Warning in rgeos::gUnaryUnion(spgeom = SpP, id = IDs): Invalid objects found;
## consider using set_RGEOS_CheckValidity(2L)
```

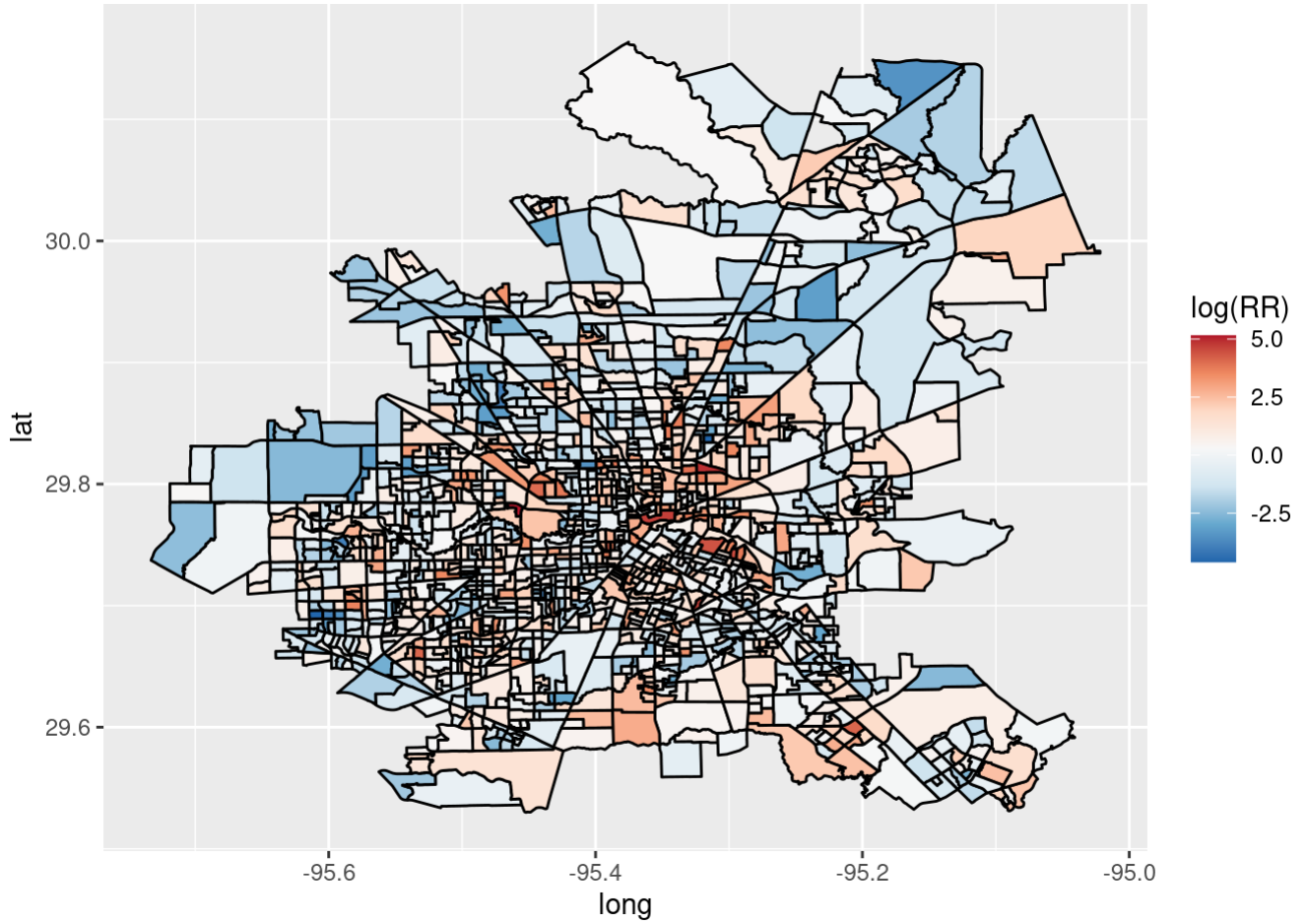
```
mShp.df <- merge(mShp.df, mShp@data, by = "id") #Merge data w/polygon data.frame
```

1. Transform the RR into log(RR) and create exploratory plots of the data by looking at the relationship between log(RR) (the response variable) and a few of the explanatory variables. Comment on any general relationships you see from the data.

```
mShp@data$logRR <- log(mShp@data$RR)
ggpairs(mShp@data[,c("NOAC", "MED_AGE", "HispanicPC", "BlackPCT", "under5PCT", "over65PCT", "pov
ertyPCT", "alonePCT", "MinTemp", "logRR")])
```




```
ggplot(data=mShp.df, aes(x=long, y=lat, group=group, fill=log(RR))) + geom_polygon(color="blac
k") + scale_fill_distiller(palette="RdBu")
```



It looks like there are correlations between the neigboring areas in Houston based on looking at the choropleth map. Aside from that, I don't see any major correlations.

2. Fit an independent MLR model with a linear effect between log(RR) and all the explanatory variables. Explore the residuals to see if there is evidence of spatial correlation by mapping them and using a Moran's I or Geary's C test.

```
model.lm1 <- lm(formula=log(RR)~NOAC+MED_AGE+HispanicPC+BlackPCT+under5PCT+over65PCT+povertyPCT+
alonePCT+MinTemp,data=mShp@data)

mShp.df$lResids <- resid(model.lm1)[match(mShp.df$id, mShp@data$id)]

moran.test(x=model.lm1$residuals , listw=nb2listw(poly2nb(mShp)))
```

```
## although coordinates are longitude/latitude, st_intersects assumes that they are planar
```

```
##
##  Moran I test under randomisation
##
## data:  model.lm1$residuals
## weights: nb2listw(poly2nb(mShp))
##
## Moran I statistic standard deviate = 20.773, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      0.3089574020     -0.0006729475      0.0002221736
```

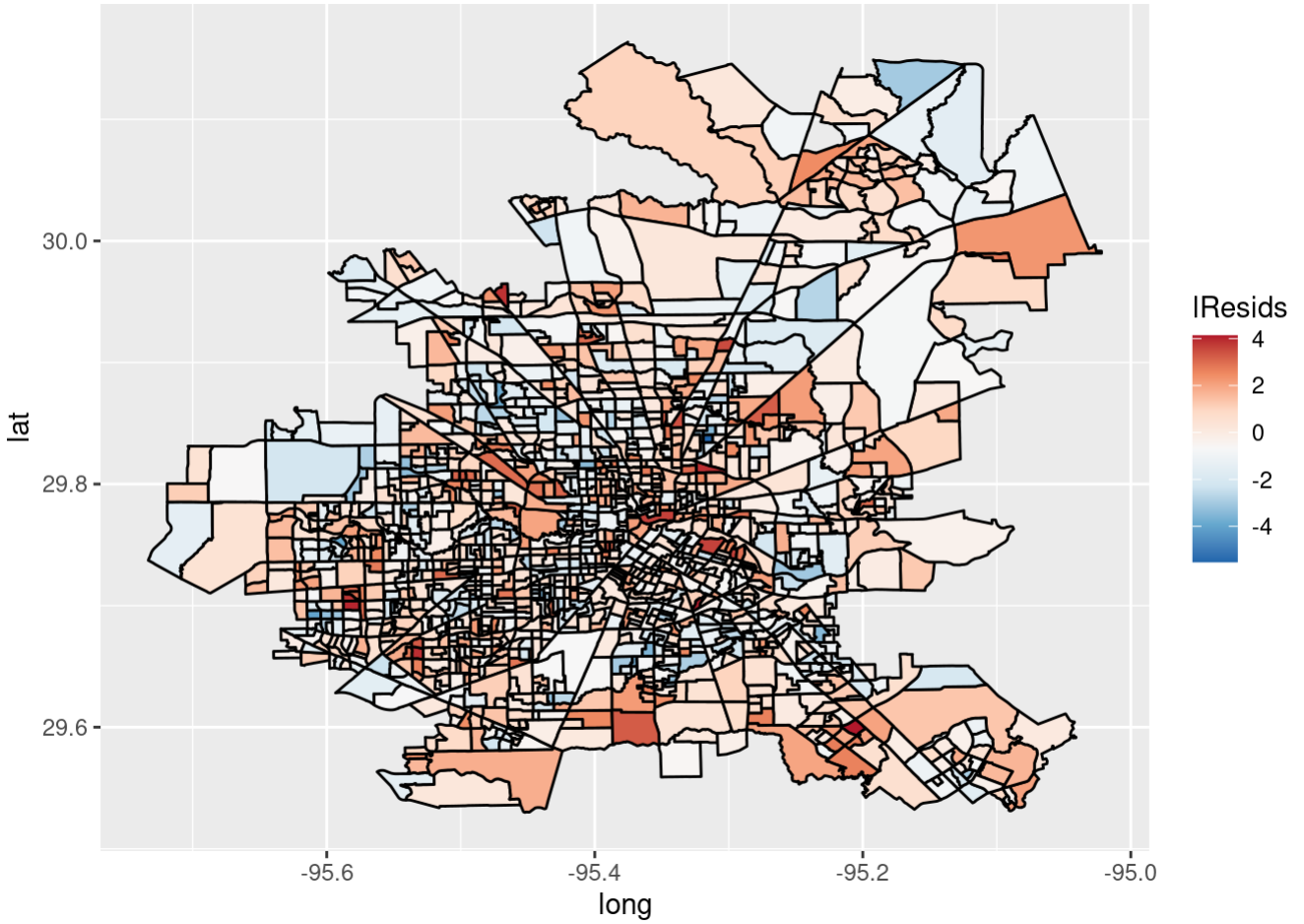
```
geary.test(x=model.lm1$residuals , listw=nb2listw(poly2nb(mShp)))
```

```
## although coordinates are longitude/latitude, st_intersects assumes that they are planar
```

```
##
##  Geary C test under randomisation
##
## data:  model.lm1$residuals
## weights: nb2listw(poly2nb(mShp))
##
## Geary C statistic standard deviate = 18.108, p-value < 2.2e-16
## alternative hypothesis: Expectation greater than statistic
## sample estimates:
## Geary C statistic       Expectation          Variance
##      0.7042613081      1.0000000000      0.0002667298
```

The p-value came out to be less than 0.05 from the Moran/Geary test so we can reject the null and fail to assume independence.

```
ggplot(data=mShp.df, aes(x=long, y=lat, group=group, fill=lResids)) + geom_polygon(color="blac
k") + scale_fill_distiller(palette="RdBu")
```



3. Write out a CAR model for analyzing the mortality data in terms of parameters. Explain and interpret any parameters associated with the model.

y=XB+epsilon; epsilon~CAR(sigma^2)

-y is the log response variable which is the log of the relative risk of mortality in the block group.

-X is the matrix of all the explanatory variables in the model.

-B is the coefficients of the explanatory variables in the model.

-epsilon is the error term.

-sigma^2 is the the constant that contributes to the variance.

N(epsilon_i,(sigma^2/#Neigh(i)))

-The effect of one state is correlated with states that share a border.

4. Fit your spatial CAR model (using minit=maxit=1000 iterations and 250 degrees of freedom for positive spatial correlation) and validate any assumptions you made to fit the model.

```
A <- nb2mat(poly2nb(mShp), style="B")
```

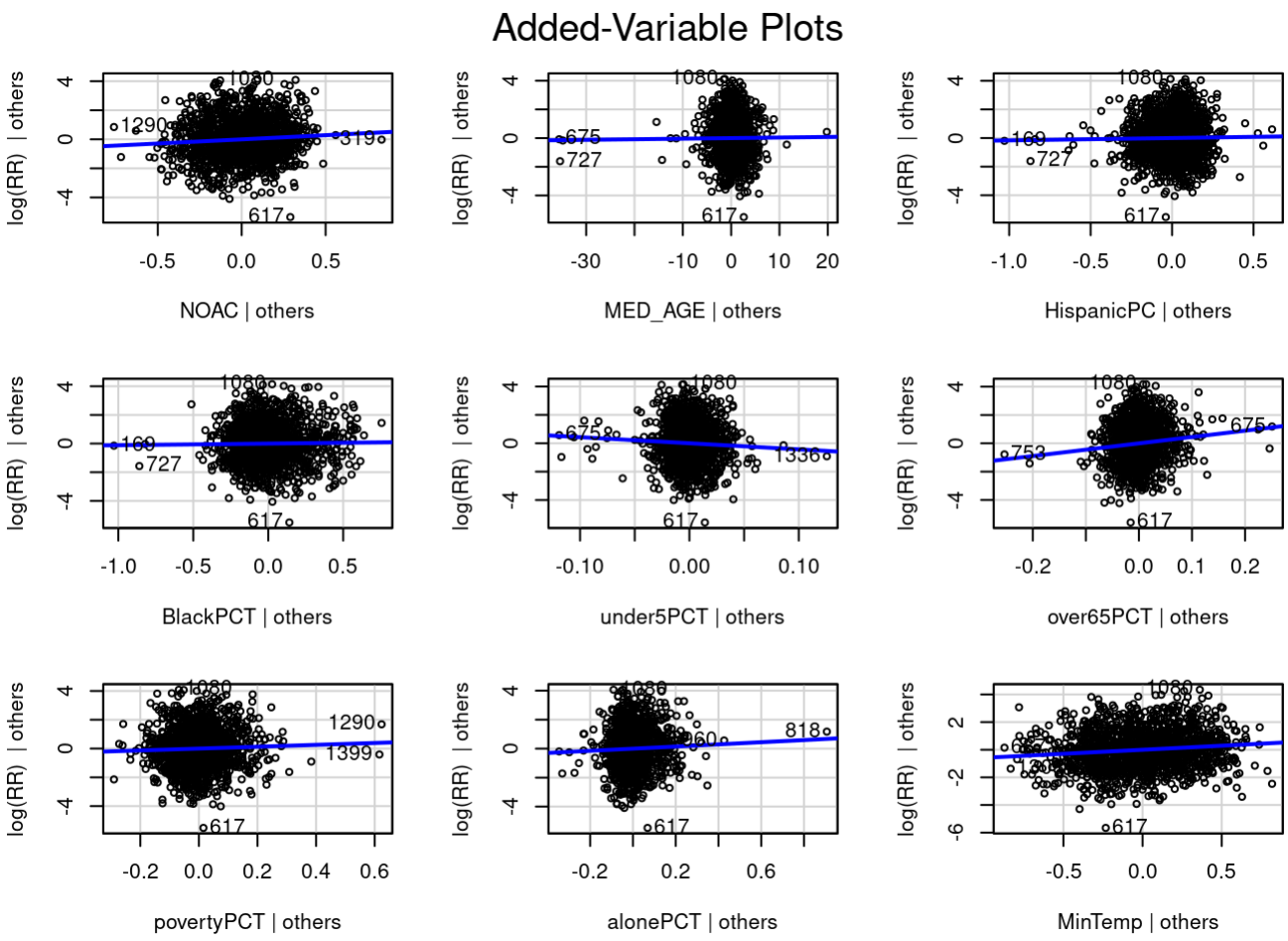
```
## although coordinates are longitude/latitude, st_intersects assumes that they are planar
```

```
colnames(A) <- rownames(A)
```

```
spatial.lm1 <- sparse.sglmm(formula=logRR~.-id-RR, data=mShp@data, A=A, attractive=250, minit=1000, maxit=1000, verbose=TRUE, method="RSR", x=TRUE)
```

```
##
## Hyperparameter 'sigma.b' must be a positive number. Setting it to the default value of 1,000.
##
## Hyperparameter 'a.h' must be a positive number. Setting it to the default value of 0.01.
##
## Hyperparameter 'b.h' must be a positive number. Setting it to the default value of 100.
##
## Warning: The Moran operator is being computed and eigendecomposed. These operations may be time consuming.
##
## Warning: MCMC may be time consuming.
```

```
avPlots(model.lm1)
```



We can assume linearity because the AV plots of the numeric variables look linear.

```
mShp.df$lmResidsZ <- resid(spatial.lm1)[match(mShp.df$id, mShp@data$id)]

moran.test(x=spatial.lm1$residuals , listw=nb2listw(poly2nb(mShp)))
```

```
## although coordinates are longitude/latitude, st_intersects assumes that they are planar
```

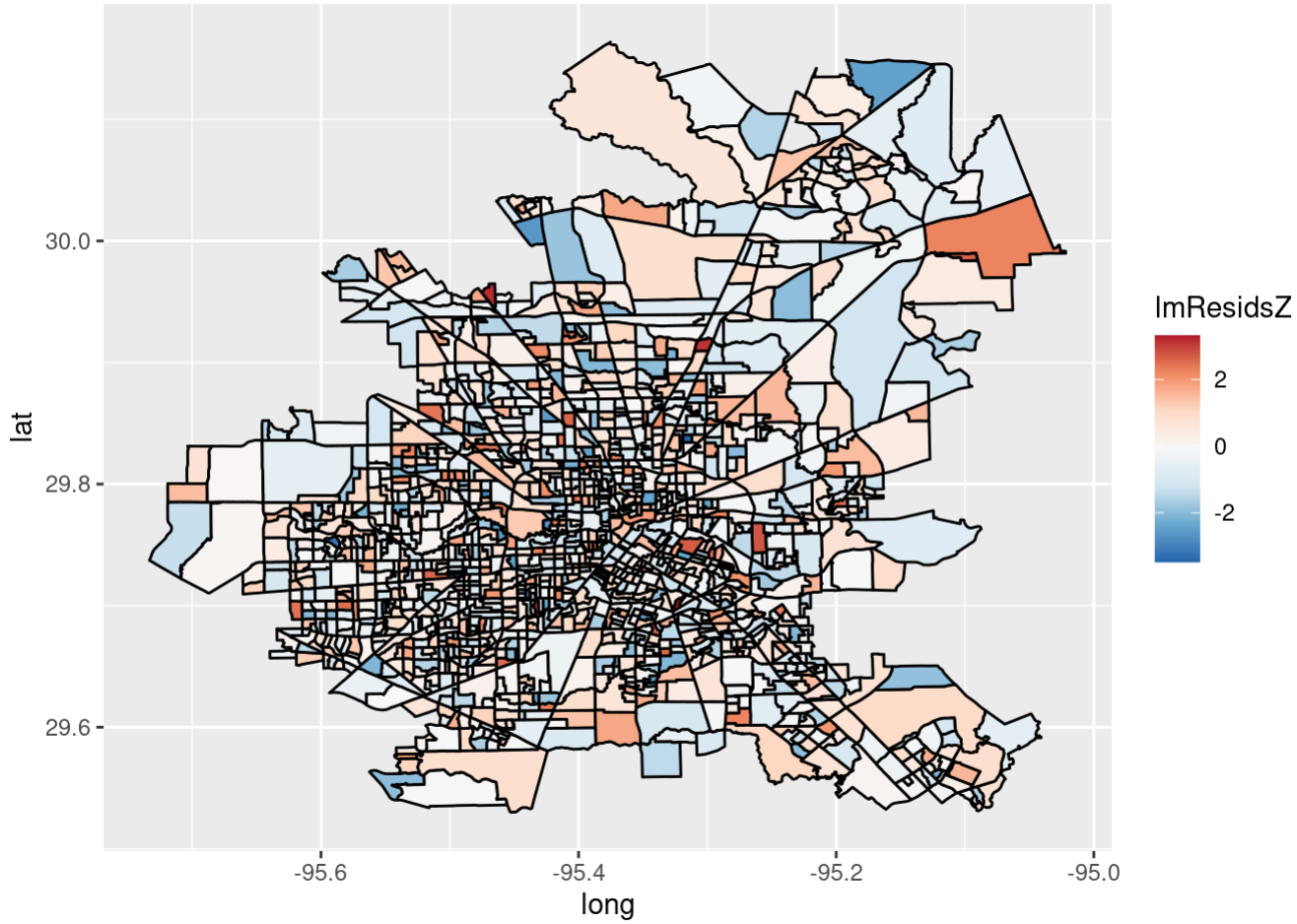
```
##
##  Moran I test under randomisation
##
## data:  spatial.lm1$residuals
## weights: nb2listw(poly2nb(mShp))
##
## Moran I statistic standard deviate = -1.8826, p-value = 0.9701
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##     -0.0287350305     -0.0006729475      0.0002221796
```

```
geary.test(x=c(spatial.lm1$residuals) , listw=nb2listw(poly2nb(mShp)))
```

```
## although coordinates are longitude/latitude, st_intersects assumes that they are planar
```

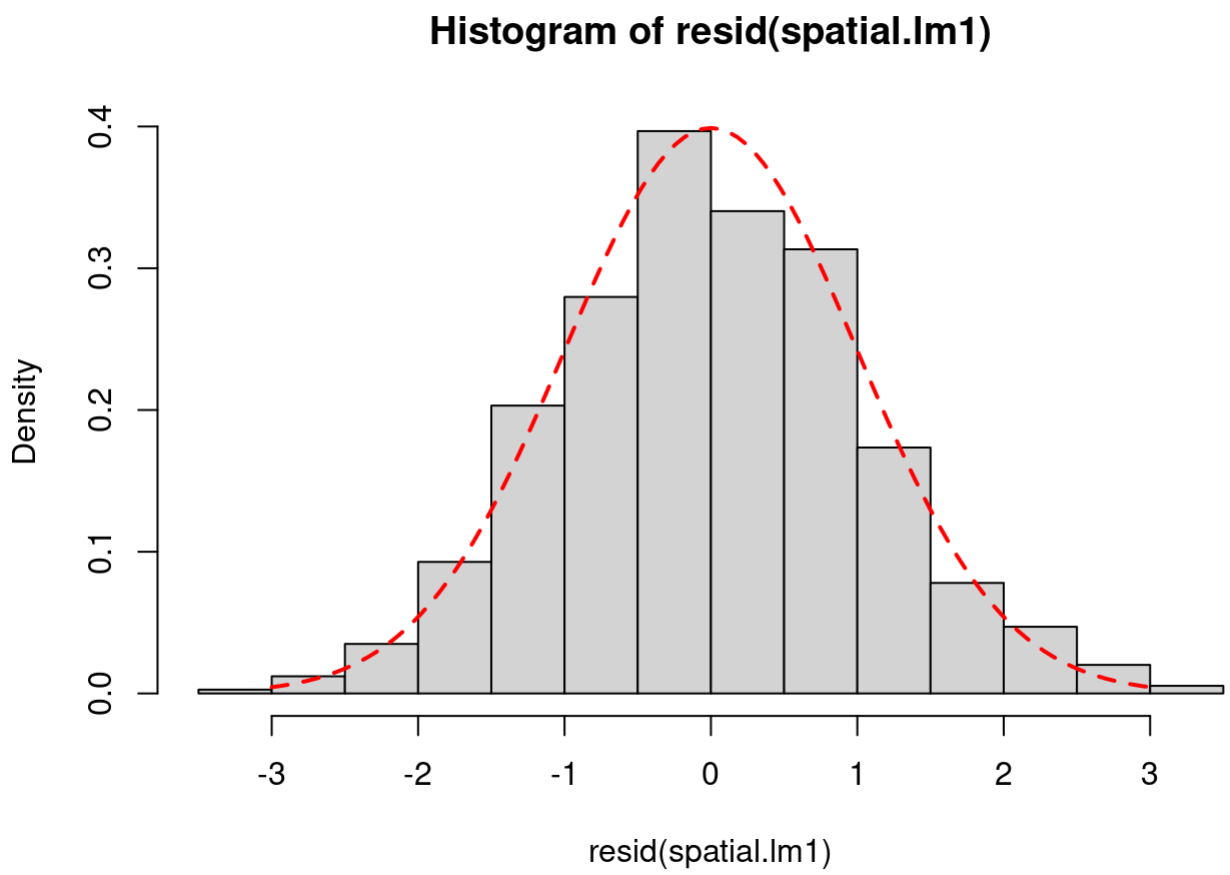
```
##
##  Geary C test under randomisation
##
## data:  c(spatial.lm1$residuals)
## weights: nb2listw(poly2nb(mShp))
##
## Geary C statistic standard deviate = -0.94562, p-value = 0.8278
## alternative hypothesis: Expectation greater than statistic
## sample estimates:
## Geary C statistic       Expectation          Variance
##      1.0154185624      1.0000000000      0.0002658624
```

```
ggplot(data=mShp.df, aes(x=long, y=lat, group=group, fill=lmResidsZ)) + geom_polygon(color="blac
k") + scale_fill_distiller(palette="RdBu")
```



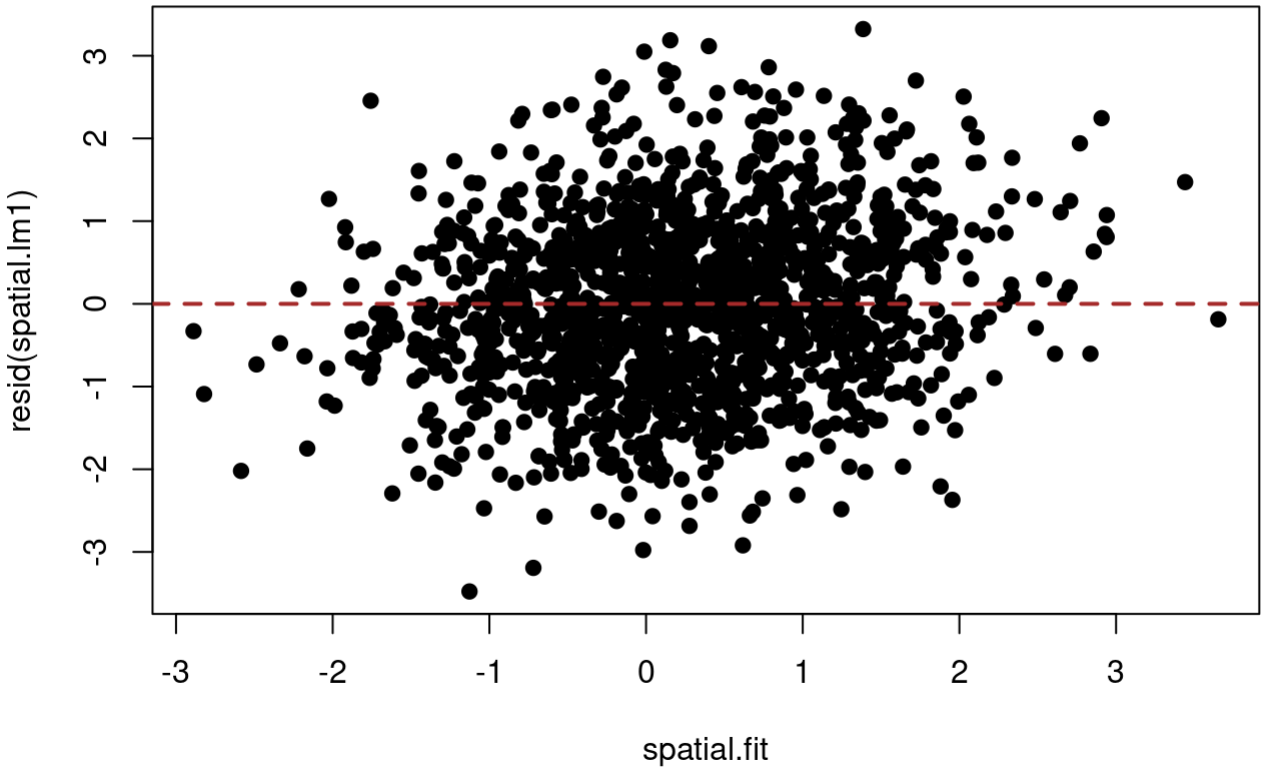
We fail to reject the null and assume independence because the p-value is greater than 0.05 in the Moran/Geary test.

```
hist(resid(spatial.lm1),freq = FALSE,breaks = 20)
curve(dnorm,from = -3,to = 3,col = "red",lwd = 2,
lty = 2,add = TRUE)
```



The histogram of the residuals of the CAR model seems to follow a Gaussian “normal” curve so we can conclude that that it meets the assumption of normality.

```
spatial.fit <- fitted(spatial.lm1)
plot(spatial.fit,resid(spatial.lm1),pch=19)
abline(0,0,col = "brown",lwd = 2,lty = 2)
```



The fitted vs residual scatter plot has most of the dots scattered evenly across the dotted line so we can conclude that it meets the equal variance assumption.

5. Calculate confidence intervals for the effect of each explanatory variable included in your model. Draw conclusions about who is at greatest risk for heat-related mortality based on your estimated effects.

```
summary(spatial.lm1)
```

```
##
## Call:
##
## sparse.sglmm(formula = logRR ~ . - id - RR, data = mShp@data,
##     A = A, method = "RSR", attractive = 250, minit = 1000, maxit = 1000,
##     x = TRUE, verbose = TRUE)
##
## Hyperparameters:
##
## sigma.b 1e+03
## a.h     1e-02
## b.h     1e+02
##
## Coefficients:
##
##               Estimate     Lower     Upper      MCSE
## (Intercept) -14.980000 -20.19000 -10.24000 0.0635000
## NOAC          0.570700   0.29070   0.84160 0.0049540
## MED_AGE       0.004184  -0.01549   0.02299 0.0003293
## HispanicPC    0.168500  -0.25770   0.57800 0.0073570
## BlackPCT      0.115000  -0.19370   0.39930 0.0045420
## under5PCT    -4.305000  -7.44300  -1.11700 0.0551500
## over65PCT     4.478000   3.01900   6.10000 0.0213100
## povertyPCT    0.640800  -0.11600   1.21800 0.0124300
## alonePCT      0.737300   0.02296   1.47800 0.0112000
## MinTemp       0.586900   0.38600   0.79320 0.0027230
##
## DIC: 4763
##
## Number of iterations: 1000
```

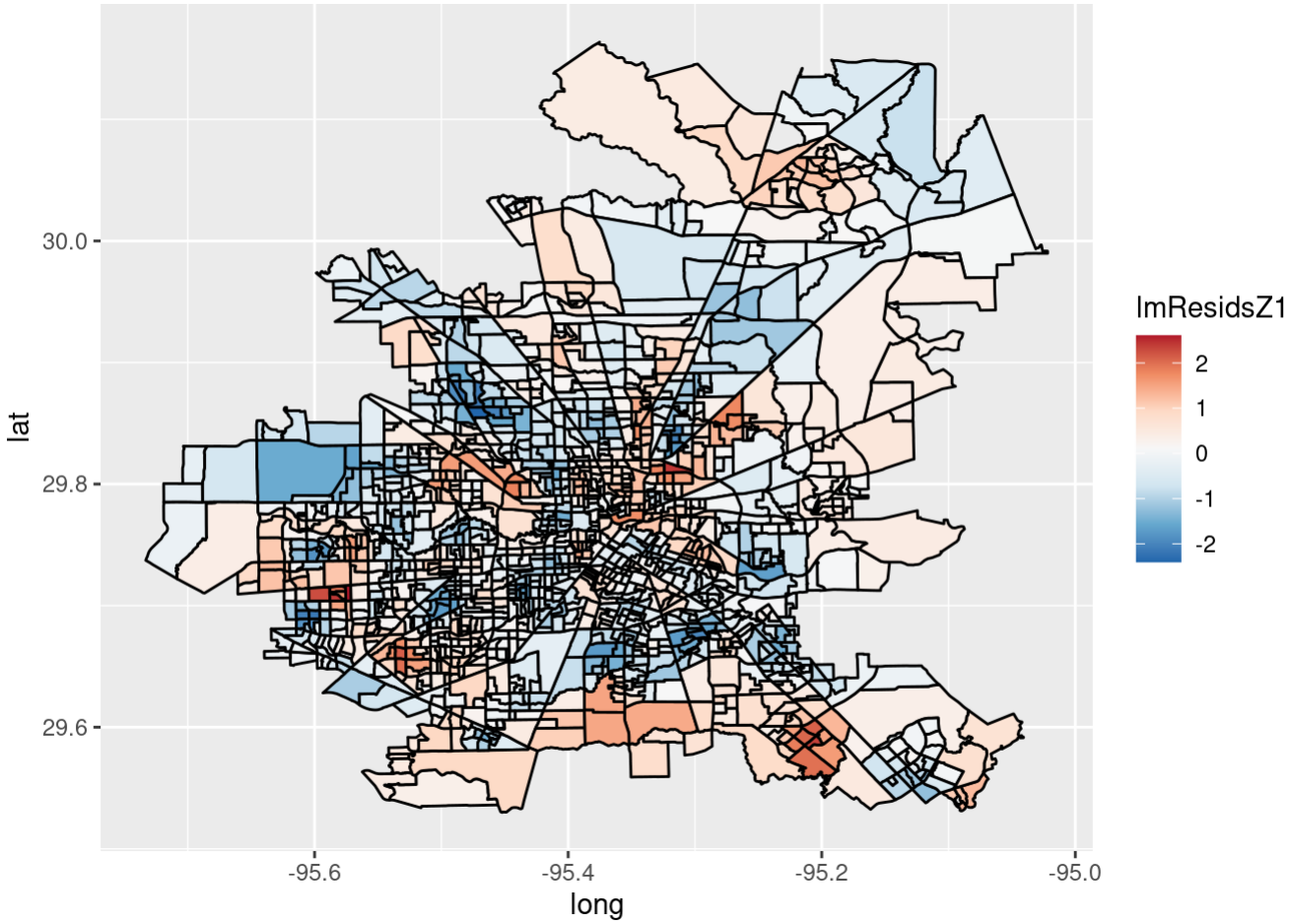
The people that are at the highest risk are people that are exposed to minimum summer temperature, people that live alone, people that are over 65 years of age, and people that lives in homes without air conditioning based on analyzing the credible intervals and picking the explanatory variables that are positive and don't include a 0 in the interval.

6. Draw a map of the correlated residuals to try and reach conclusions about areas at risk of heat-related mortality not explained by your explanatory variables.

```
spatialResid1 <- spatial.lm1$M%*%t(spatial.lm1$gamma.sample)
spatialResid1 <- rowMeans(spatialResid1)
```

```
mShp.df$lmResidsZ1 <- spatialResid1[match(mShp.df$id, mShp@data$id)]
```

```
ggplot(data=mShp.df, aes(x=long, y=lat, group=group, fill=lmResidsZ1)) + geom_polygon(color="bla
ck") + scale_fill_distiller(palette="RdBu")
```



Education level can be a variable that is not explained by the explanatory variables. Person's vulnerability to heat related mortality can also be a factor that is not explained by the explanatory variables.