Geospatial Code Replication Workflow

Jyotishka Datta

6/17/2019

Contents

Goals

To create a supervised learning framework for the child maltreatment data in Little Rock between 2015 and 2018. The framework I am trying to follow is similar to the one here: https://pennmusa.github.io/MUSA_801.io/project_5/.

The first task is to perform an exploratory analysis similar to the one here https://pennmusa.github.io/MUSA_801.io/project_5/

Package Dependencies and Preamble

```r
# install.packages("devtools")
# devtools::install_github("thomasp85/patchwork")

library("sf")            # Spatial data objects and methods
library("mapview")       # Interactive Map Viewing
library("ggmap")         # ggplot2 addon for base maps
library("cowplot")
library("spatstat")      # KDE and other spatial functions
library("raster")        # cell-based spatial operations
library("tidyverse")     # data manipulation framework
library("Hmisc")         # using cut2() functions for ggplot legends
library("fitdistrplus")  # Distribution fitting functions
library("lubridate")     # Power tools for handling dates
library("tidycensus")
library("lwgeom")
library("Hmisc")
library("hrbrthemes")
library("gridExtra")
library("patchwork")
library("spdep")         # KNN functions
library("foreach")
library("doParallel")
library("corrplot")
library("ranger")        # randomforest implimentation
library("glmnet")        # for Ridge and Lasso Regression
library("knitr")          # for kable table
library("kableExtra")
library("FNN")            # KNN for CPS vs. NN plots
library("groupdata2")
library("htmltools")
library("viridis")
library("viridisLite")
```

```
mapviewOptions(basemaps = c("Stamen.TonerLite", "OpenStreetMap.DE"))

base_dir = "C:/Users/jd033/Box/Child Maltreatment"
fishnet_grid_dim = 1000
k_direction = 8 # 4 = rook, 8 = queen
k_nearest_neighbors = 5
# Either k (e.g. 5 or 10) or "LOOCV"
n_folds = "LOOCV"
# threshold quntile for statArea grouping
stat_area_quantile = 0.60
# Number of simulations for CPS vs. NN
simulations = 1000
# Number of neighbors for CPS vs. NN
k = 5
# random seed
set.seed(11235)
```

```
source('C:/Users/jd033/Documents/GitHub/PAP-child/FUNCTIONS_VAPAP_LR.R', echo = FALSE, keep.source = TRU
# source('C:/Users/jd033/Documents/GitHub/PAP-child/FEA_CREATE_VARIABLES_LR.R', echo = TRUE, keep.sourc
```

Report Appendix 1: Data wrangling

Step 1. Detect all xls/xlsx and csv files in a directory and read them into a list

This Code example of data import function. The inputs are `.xls`, `.xlsx`, and `.csv` files in a folder. The output is a list data type where each element of the list if one of the input data sets in the sf spatial data format.

Reading Variables

```
# requires all data in *.csv or *.xls files containing coordinate field names "X" and "Y"
# `crs` in the call to `st_as_sf()` needs to be set to the ESPG code of your data projection
# `base_dir` file path and many feature names are specified for the current project.

##1.1 Global Variables
# mapviewOptions(basemaps = c("Stamen.TonerLite", "OpenStreetMap.DE"))
base_dir = "C:/Users/jd033/Box/Child Maltreatment"


##2.1 Load Data
files <-list.files(file.path(base_dir,"/Little Rock Data/CSV"), pattern = "*\\.xlsx$|*\\.csv$")
var_list <- vector(mode = "list")
var_names <- NULL
for(i in seq_along(files)){
  filename <- str_sub(files[i], start = 1, end = -5)
```

Table 1: List of Variables

x
Banks.
BarberAndBeautyShops.
CM_geocoded
HighSchoolsPublic.
HotelMotel.
LiquorStores.
MajorDeptRetailDiscount.
Rental_MobileHomes.
Rental_SingleToQuad.
TattooPiercing.

```r
  sf_i <- tryCatch({
    if(tools::file_ext(files[i]) == "xlsx"){
      dat <- readxl::read_xlsx(file.path(base_dir,"/Little Rock Data/CSV",files[i]))
    } else if(tools::file_ext(files[i]) == "csv"){
      dat <- read.csv(file.path(base_dir,"/Little Rock Data/CSV",files[i]))
    }
    dat %>%
      filter(!is.na(X) | !is.na(Y)) %>%
      st_as_sf(., coords = c("X", "Y"), crs = 2765)
  }, error = function(e){
    cat(filename, "error = ",e$message,"\n")
    return(e)
  }
  )
  if(!inherits(sf_i, "error")){
    var_list[[length(var_list)+1]] <- sf_i
    var_names[length(var_list)] <- filename
  }
}
names(var_list) <- var_names
knitr::kable(var_names, caption = "List of Variables")
```

Question: what was the var_list for the Richmond child maltreatment data analysis, i.e. list of variables as .csv or .xls files?

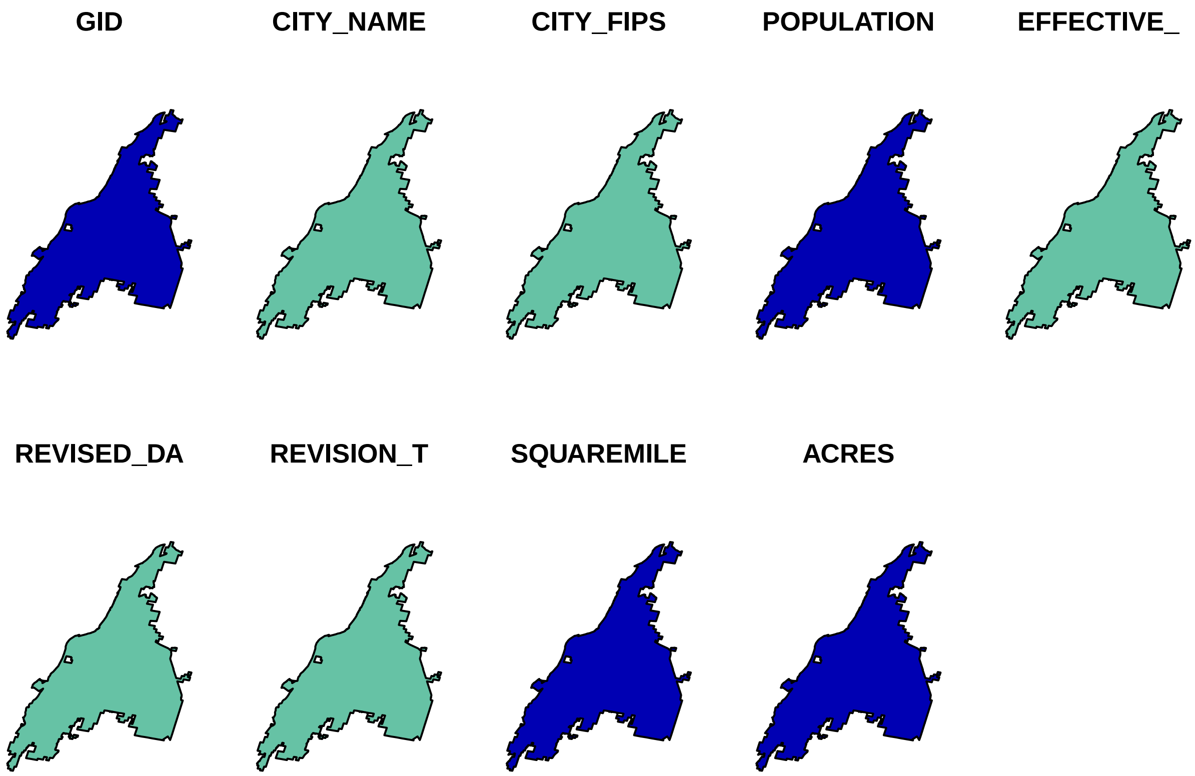
Read LR shapefile

```r
# Load new packages (might be redundant)
pacman::p_load(lubridate, sf, raster, rgdal, broom, rgeos, GISTools)

# Load old packages
pacman::p_load(dplyr, ggplot2, ggthemes, magrittr, viridis)

setwd("C:/Users/jd033/Box/Child Maltreatment/Little Rock Data")

# nbr =readOGR(
#   dsn = paste0("Shapefile_LR"),
#   layer = "LR_Municipal_Boundary_SF")
```

```
nbr = st_read("C:/Users/jd033/Box/Child Maltreatment/Little Rock Data/Shapefile_LR/LR_Municipal_Boundar
              st_transform(2756)
```

```
## Reading layer `LR_Municipal_Boundary_SF' from data source `C:\Users\jd033\Box\Child Maltreatment\Lit
## Simple feature collection with 1 feature and 9 fields
## geometry type:  POLYGON
## dimension:      XYZ
## bbox:           xmin: 1155745 ymin: 106599.4 xmax: 1267049 ymax: 178167
## epsg (SRID):    NA
## proj4string:    +proj=lcc +lat_1=34.93333333333333 +lat_2=36.23333333333333 +lat_0=34.33333333333334
```

```
plot(nbr)
```



```
# Class
class(nbr)
```

```
## [1] "sf"         "data.frame"
```

```
# Dimensions
dim(nbr)
```

```
## [1]  1 10
```

```
# Info in shapefile
names(nbr)
```

```
##  [1] "GID"        "CITY_NAME"  "CITY_FIPS"  "POPULATION" "EFFECTIVE_"
##  [6] "REVISED_DA" "REVISION_T" "SQUAREMILE" "ACRES"      "geometry"
```

Question: Should we project immediately using st_transform? (like you've done here https://pennmusa.github.io/MUSA_801.io/project_5/#61_set_up)

```
nbr_diss <- nbr %>%
  mutate(dissolve = 1) %>%
  # get rid of slivers
  st_buffer(., dist = 0.1) %>%
  group_by(dissolve) %>%
  summarise()

nbr_rast_SP <- raster(as(nbr_diss, "Spatial"), nrows = 2000, ncol = 2000)
```

```
### get CPS_Accepted values (add 1 column for dissolving)
cps_dissolve <- var_list[["CM_geocoded"]] %>%
  mutate(value = 1) %>%
  dplyr::select(value)
```

Next steps

- Step 2: Import spatial neighborhood data with `read_sf()` and `get_decennial()`
- Step 3: Create spatial fishnet grid with `st_make_grid()` and calculate spatial weights with `poly2nb()` and `nb2listw()`
- Step 4: Intersect fishnet and census blocks to create populations estimates and weights per fishnet cell.

To do: R code chunks

Questions about R chunks that do not work correctly, so these are mostly related to using some specific functions inside these packages.

Question: the get_map is not working, could be because it's expecting lat-lon but getting (X,Y) coordianates? The error message says Error: scale must be a positive integer 0-18. Stackexchange isn't much of a help.

Question: How did the `CPS Accepted` file look like? Are these all child maltreatment cases or a subset of them? The one that I have here is the same data-set that Sherri Jo shared with us.

```
# nbr <- read_sf("https://data.richmondgov.com/resource/7juf-nwis.geojson") %>%
#   st_transform(crs = 102747)

cm_bbox = unname(st_bbox(ll(st_buffer(var_list[["CM_geocoded"]],dist = 0.1))))

cm_bbox

cps_base_map   <- get_map(location = cm_bbox,
                          source = "google",
                          maptype = "toner")
```

Question: `st_make_grid` seems to be working, but not generating a fishnet over the LR boundaries as I want. It's creating a regular rectangle. As a result, the next lines are not working !

```
net <- st_make_grid(nbr, cellsize = fishnet_grid_dim)

# count CPS incidents per net cell - really just to get net raster into sf polygon format
net_agg <- aggregate(cps_dissolve, net, sum) %>%
  tibble::rowid_to_column(.,"net_id")
```

```
# list of net cells IDs that intersect with Richmond
net_intersect <- st_intersects(nbr, net_agg)

# extract Richmonds net cells based on intersect ID
net_littlerock <- net_agg[unique(unlist(net_intersect)),]
net_hood <- st_join(net_littlerock, nbr, largest = TRUE)
listw <- nb2listw(poly2nb(as(net_littlerock, "Spatial"), queen = TRUE))
```

Question: Do we need the following, or can we use the population information from the shapefile that we read at the very beginning?

```
vars10 <- c("P0010001") # total population (correct, I checked the web)
## get total 2010 census pop for blocks & calculate area
#littlerock_block <- get_decennial(geography = "block", variables = vars10, year = 2010,

#summary_var = "P0010001", state = 51, county = 760, geometry = TRUE) %>%
#st_transform(crs = 2756)

# calc area
# littlerock_block <- st_read("littlerock_block.shp")    #was having issues with above code when I knit

littlerock_block <- littlerock_block %>%
  mutate(acre = as.numeric(st_area(littlerock_block)*2.29568e-5),
         # acre = units::set_units(acre, acre),
         pop_acre_rate = value / acre)
```