Case Study 3: Geographic distribution of crime in Mexico

David Buil-Gil and Reka Solymosi

2024-01-25

CONTEXT

We take police-recorded crime data made available on the webside of the Mexican Government <https://www.gob.mx/sesnsp/acciones-y-programas/datos-abiertos-de-incidencia-delictiva?state=published> to visualise kidnapping across the different states (entidades) of Mexico.

We begin by loading the required packages in R:

library(here) # to identify the path to the data  
library(readr) # to read in CSV data  
library(dplyr) # for data wrangling  
library(ggplot2) # for data visualisation  
library(sf) # for spatial data manipulation and visualisation  
library(viridis) # for colour schemes

The data are saved in a CSV (comme separated values) which make it easy load into an object called data\_Mexico. We use the read\_csv() function from the readr package.

#Read csv file with crime data  
data\_Mexico <- read\_csv(here("data/IDM\_nov2023.csv"))

We can have a look at the data and see there are a few variables of interest:

* TIPO (crime type)
* ANO (year)
* ENTIDAD (state)

To visualise kidnappings we filter on the TIPO (crime type) variable for SECUESTRO (kidnapping). We filter for incidents in the year 2017 on the ANO variable.

data\_Mexico <- data\_Mexico %>%  
 filter(TIPO == "SECUESTRO") %>% # select only kidnappings  
 filter(ANO == 2017) # filter for year 2017 only

To plot the number of kidnappings per state on a map, we need to create a table which contains two columns: the state name (we can keep this as ENTIDAD) and the number of kidnappings for each one. We can see that the number of kidnappings are broken down by month: there is a new column for each month of the year. To get a total score for the year for each state, we must sum the numbers across these columns (these are columns 8 through 19, inclusive). We can use the rowSums() function, and specify the columns we need in the accross() function. Remember to set the na.rm parameter to TRUE to make sure that NAs (missing data) are just counted as 0 (no kidnappings in that month).

data\_Mexico <- data\_Mexico %>%  
 mutate(sum\_secuestro = rowSums(across(8:19), # create new variable of sums  
 na.rm = TRUE)) # treat NA as 0 here

Now we can count the number of total kidnappings in each entidad in 2017 by using the group\_by() and summarise() functions to create a frequency table where each row represents one state.

#Calculate number of crimes in each state  
data\_Mexico\_states <- data\_Mexico %>%  
 group\_by(ENTIDAD) %>% #group by state  
 summarise(secuestro = sum(sum\_secuestro)) # sum all kidnappings in the state

We have now saved this frequency table in a new object called data\_Mexico\_states which includes the count of kidnappings for each state in 2017.

To explore the data before mapping we can look at the top 3 states with the most kidnappings:

top\_n(data\_Mexico\_states, 3, secuestro)

## # A tibble: 3 × 2  
## ENTIDAD secuestro  
## <chr> <dbl>  
## 1 MEXICO 173  
## 2 TAMAULIPAS 140  
## 3 VERACRUZ 172

Observe that the State of Mexico concentrates the largest number of kidnappings, 173, followed by Veracruz, with 172.

We can look at the mean, median, standart deviation as well:

mean(data\_Mexico\_states$secuestro)

## [1] 35.90625

median(data\_Mexico\_states$secuestro)

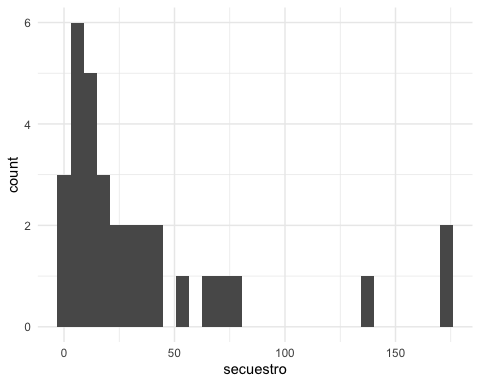
## [1] 17.5

sd(data\_Mexico\_states$secuestro)

## [1] 46.07733

On average, 36 kidnappings were recorded in each of the 32 states of Mexico, however we see this is not evenly distributed. A histogram can help look at the skew:

ggplot(data\_Mexico\_states, aes(x = secuestro)) +   
 geom\_histogram() +  
 theme\_minimal()



We can see the top three states () all the way on the right there, far away from the rest of the country. Are these states particularly risky for kidnappings? To answer this we need to consider the rate of kidnappings per population, as it is possible that these are just the more populous states. Where there are more people, there are more of everything, even kidnappings!

To account for this we can calculate rates of kidnappings per 100,000 residents. To do so we can download data about the population size for each state from the website of the National Institute of Statistics and Geography (INEGI): <https://www.inegi.org.mx/app/tabulados/default.html?nc=mdemo02>

#Read csv file with population data  
population <- read\_csv(here("data/Population2010.csv"))

## Rows: 32 Columns: 2  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): STATE  
## dbl (1): Population2010  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

To link with our crime data, we can link the two tables using their common column for state, which is still ENTIDAD in our crime data, and is STATE in INEGI’s population statistics. We use the left\_join() function:

#Merge with crime data and calculate crime rates  
data\_Mexico\_states <- data\_Mexico\_states %>%  
 left\_join(population, by = c("ENTIDAD" = "STATE")) %>%  
 mutate(secuestro\_rate = secuestro / Population2010 \* 100000)

Let’s look again:

top\_n(data\_Mexico\_states, 3, secuestro\_rate)

## # A tibble: 3 × 4  
## ENTIDAD secuestro Population2010 secuestro\_rate  
## <chr> <dbl> <dbl> <dbl>  
## 1 TABASCO 77 2238603 3.44  
## 2 TAMAULIPAS 140 3268554 4.28  
## 3 ZACATECAS 67 1490668 4.49

We can see new states appear according to calculated crime rates. However, Tamaulipas is here again, indicating it is high count and high rate for kidnapping. This might be a state of interest to consider for problem solving interventions aimed to reduce kidnapping.

But how are these rates distributed in space? Are these ares near each other. It is important to realise that these places do not exist independently of one another, and spatial relationships can be best considered initially through a nice clear exploratory visualisation, such as a thematic map.

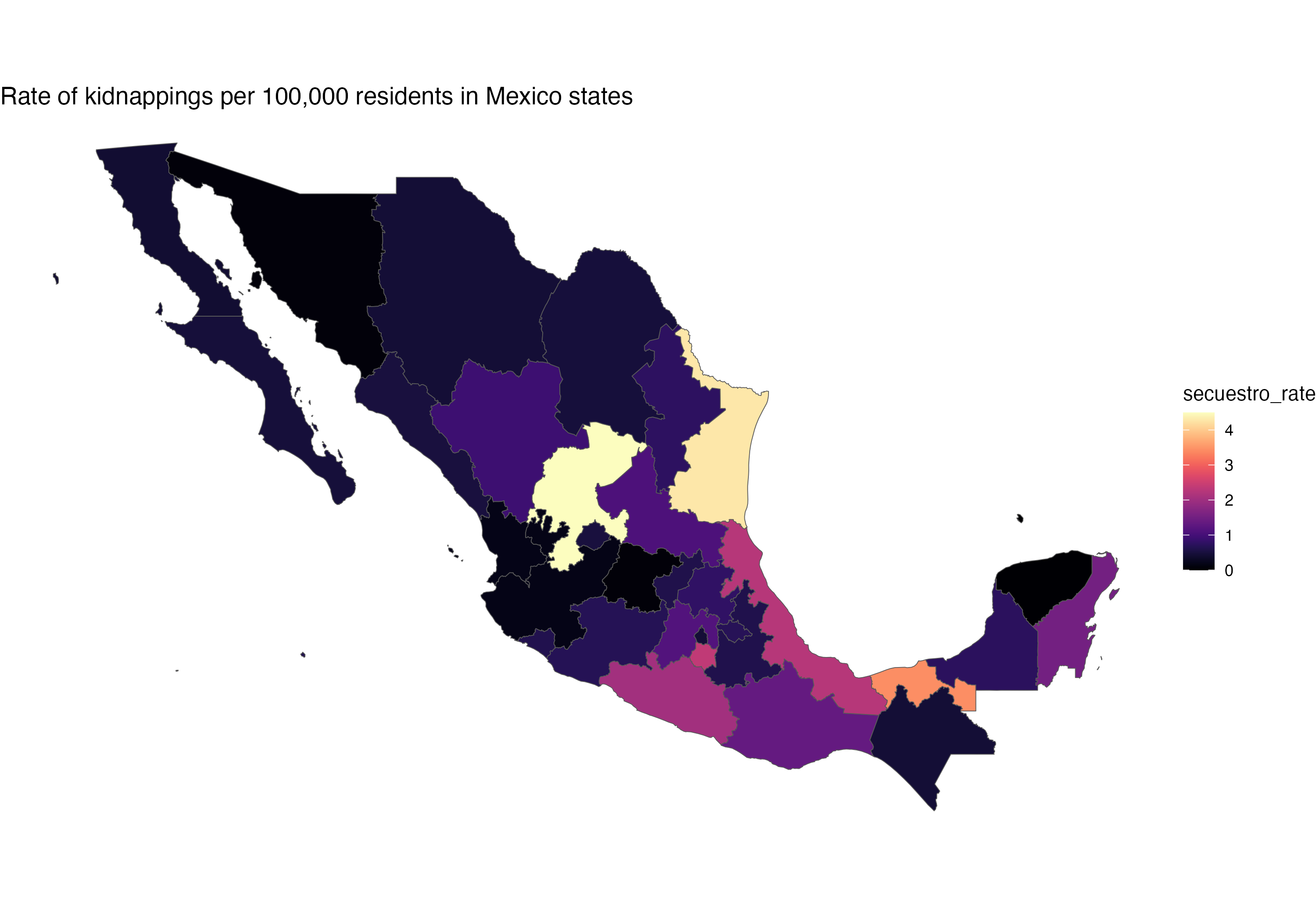
Source of shapefile: <https://github.com/strotgen/mexico-leaflet/>

#Read geojson of Mexico states  
#states\_geojson <- st\_read("https://github.com/strotgen/mexico-leaflet/blob/master/states.geojson")  
states\_geojson <- st\_read(here("data/states.geojson"))

## Reading layer `states' from data source   
## `/Users/user/Dropbox (The University of Manchester)/BA\_sg/experiments/analysis/eda\_python/crim-data-south/data/states.geojson'   
## using driver `GeoJSON'  
## Simple feature collection with 32 features and 3 fields  
## Geometry type: MULTIPOLYGON  
## Dimension: XY  
## Bounding box: xmin: -118.4 ymin: 14.5321 xmax: -86.72404 ymax: 32.71865  
## Geodetic CRS: WGS 84

#Merge crime rates with geojson file  
states\_geojson <- states\_geojson %>%  
 mutate(state\_name = toupper(state\_name), #capital letters for consistency  
 state\_name = recode(state\_name, #rename some states for consistency  
 'DISTRITO FEDERAL' = 'CIUDAD DE MEXICO',  
 'MÉXICO' = 'MEXICO',  
 'MICHOACÁN DE OCAMPO' = 'MICHOACAN',  
 'QUERÉTARO' = 'QUERETARO',  
 'SAN LUIS POTOSÍ' = 'SAN LUIS POTOSI',  
 'VERACRUZ DE IGNACIO DE LA LLAVE' = 'VERACRUZ',  
 'NUEVO LEÓN' = 'NUEVO LEON',  
 'COAHUILA DE ZARAGOZA' = 'COAHUILA',  
 'YUCATÁN' = 'YUCATAN')) %>%   
 left\_join(data\_Mexico\_states, by = c("state\_name" = "ENTIDAD"))

ggplot(data = states\_geojson) +  
 ggtitle("Kidnappings in Mexico by states") +  
 geom\_sf(aes(fill = secuestro\_rate)) +  
 scale\_fill\_viridis(option = "magma", name= "Number of kidnappings per 100,000 population")+  
 theme\_void() +  
 theme(legend.position="top")



XXXX interpret

**References**

<https://link.springer.com/article/10.1007/s12117-012-9185-x> <https://doi.org/10.1080/17440572.2011.632499> <https://doi.org/10.1016/j.ijlcj.2021.100479>