Bayesian Spatial Analysis on Global Suicide Rates

Members: Minh 33077769 Jana 87884193

```
options(repos = c(CRAN = "https://cloud.r-project.org"))
knitr::opts_chunk$set(warning = FALSE, message = FALSE, error = TRUE)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
theme_set(theme_bw())
require(extraDistr) #need for rdunif
```

```
## Loading required package: extraDistr
##
## Attaching package: 'extraDistr'
##
## The following object is masked from 'package:purrr':
##
##     rdunif
```

```
library(dplyr)
install.packages("webshot")
```

```
##
## The downloaded binary packages are in
##  /var/folders/h2/0z07kqqn1n99gzcftq2ldtv40000gp/T//RtmpsqZf3I/downloaded_packages
```

```
webshot::install_phantomjs()
```

```
## It seems that the version of `phantomjs` installed is greater than or equal to the requested version
```

```
suppressPackageStartupMessages(require(rstan))

required_packages <- c("sf", "spdep", "terra", "dplyr", "readr", "rnaturalearth", "rnaturalearthdata")
installed_packages <- rownames(installed.packages())

for (pkg in required_packages) {
  if (!(pkg %in% installed_packages)) {
    install.packages(pkg)
  }
}
```

```r
library(sf)
```

```
## Linking to GEOS 3.10.2, GDAL 3.4.2, PROJ 8.2.1; sf_use_s2() is TRUE
```

```r
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')`
```

```r
library(terra)
```

```
## terra 1.8.42
##
## Attaching package: 'terra'
##
## The following object is masked from 'package:rstan':
##
##     extract
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
library(dplyr)
library(readr)
library(rnaturalearth)
library(rnaturalearthdata)
```

```
##
## Attaching package: 'rnaturalearthdata'
##
## The following object is masked from 'package:rnaturalearth':
##
##     countries110
```

```r
library(ggplot2)
library(bayesplot)
```

```
## This is bayesplot version 1.11.1
## - Online documentation and vignettes at mc-stan.org/bayesplot
## - bayesplot theme set to bayesplot::theme_default()
##    * Does _not_ affect other ggplot2 plots
##    * See ?bayesplot_theme_set for details on theme setting
```

```r
library(mapview)
library(webshot)
library(htmlwidgets)

install.packages("fuzzyjoin")
```

```
##
## The downloaded binary packages are in
##  /var/folders/h2/0z07kqqn1n99gzcftq2ldtv40000gp/T//RtmpsqZf3I/downloaded_packages
```

```r
library(fuzzyjoin)
```

Introduction

Mental health has become an increasingly important topic and suicide remains a significant public health concern worldwide, with rates varying across regions due to complex social, economic, and cultural factors. Therefore, understanding the geographic distribution of suicide rates may be important for the development of targeted mental health policies and preventative measures. Although place of habitation clearly affect mental and physical lifestyles, there have few studies conducted on the geographical relationship between suicide rates and mental well-being.

In this study, we apply a Bayesian hierarchical model with a conditionally autoregressive (CAR) prior to investigate spatial patterns in suicide rates across countries. We model the suicide rate as a continuous outcome using a Gaussian likelihood, with a global intercept, a temporal effect comparing 2019 and 2021, and spatial random effects that capture regional deviationas.

Our analysis focuses specifically on the years2019 and 2021, to investigate any observable changes in suicide patterns potentially influenced by global events like the COVID-19 pandemic. Inference is performed using MCMC sampling in Stan. Our main research question is: Are there identifiable spatial patterns that persist after taking global and temporal effects into account, and did suicide rates change significantly between 2019 and 2021 at a global level?

This approach allows us to identify high-risk regions, quantify uncertainty, and better understand how suicide rates are spatially structure, with similarities among neighboring countries. Valuable insights from this analysis may provide the opportunity to create more data-informed mental health interventions.

The GitHub repository can be found at the following link: https://github.com/minhVu03/Bayesian-Data-Analysis-Project

Literature Review

In a recent press release, the CDC stated that suicide rates in the US decreased from 2019 to 2020, but then increased from 2020 to 2021 (Centers for Disease Control and Prevention, 2022). Additionally, another paper stated that factors such as domestic violence, financial strain, and mental health conditions were prominent issues during the pandemic, and "the social restriction practices and policies imposed by different countries secondary to the COVID-19 pandemic might have negatively influenced the fore-said risk factors that has been indirectly led increased rates of suicidal attempts and deaths" (Pathirathna et al., 2022). This information leads us to believe that it would be worthwhile to study the difference in suicide rates in 2019 and 2021.

A similar study on the relationship between location and suicide has been conducted, but the data was limited to regions in London and with the rise of social media and the global pandemic, the information may now be outdated (Congdon, P., 1997). In comparison, our dataset contains over 150 countries and the analysis focuses on 2019 and 2021, which may show the impact of the COVID-19 pandemic on suicide rates.

Dataset and Data Cleaning

Dataset Name: Crude Suicide Rate (Per 100,000 Population)

Source: https://www.who.int/data/gho/data/themes/mental-health/suicide-rates

Description: The raw dataset has notable features like country, age group, sex, and suicide rate (per 100,000 people) that can be extracted.

Location: Country name

Period: Year (2019, 2021)

Dim1: Sex ("Female", "Both sexes", "Male)

FactValueNumeric: Number of suicide deaths in a year, divided by the population and multiplied by 100 000 (as indicated in the original data source)

FactValueNumericLow: Low estimate

FactValueNumericHigh: High estimate

Note: The FactValueNumeric data are estimates of the number of suicides. The data was obtained from the WHO Global Health Estimates (GHE), but some countries may not have an accurate way of recording the exact number of deaths, potentially leading to inaccurate estimations. Hence there is a high and low in the death rates. The source states, "for countries without high-quality death registration data, cause of death estimates are calculated using other data, including household surveys with verbal autopsy, sample or sentinel registration systems, special studies" (World Health Organization, n.d.).

```
data_raw = read.csv("suicide_rate_raw.csv", header = TRUE)

#filter out "both sexes" to avoid duplication
data = as.data.frame(data_raw |> select(Location, Period, Dim1, FactValueNumeric, FactValueNumericLow, F
unique(data$Period)
```

```
## [1] 2021 2019
```

```
max(data$FactValueNumeric)
```

```
## [1] 48
```

```
min(data$FactValueNumeric)
```

```
## [1] 0
```

```
nrow(data)
```

```
## [1] 740
```

```
head(data)
```

```
##                                                   Location Period   Dim1
## 1                         Saint Vincent and the Grenadines   2021 Female
## 2                                                     Oman   2021 Female
## 3 occupied Palestinian territory, including east Jerusalem   2021 Female
## 4                                                   Jordan   2021 Female
## 5                                                   Kuwait   2021 Female
## 6                                     Syrian Arab Republic   2021 Female
##   FactValueNumeric FactValueNumericLow FactValueNumericHigh
## 1             0.00                0.00                 0.00
## 2             0.21                0.12                 0.34
## 3             0.23                0.14                 0.36
## 4             0.29                0.18                 0.43
## 5             0.34                0.25                 0.39
## 6             0.26                0.16                 0.42
```

The dataset after filtering consists of 740 observations.

Data Analysis

As we have obtained the cleaned data for suicide rates in 2019 and 2021, we can now declare a prior model from information obtained historically.

Model

Priors:

$\mu \sim \mathcal{N}(9.2, 3)$

$\beta \sim \mathcal{N}(0.1, 0.05)$

$\sigma_\phi \sim \text{Exp}(1)$

$\sigma \sim \text{Exp}(1)$

$\phi_{\text{node1}[i]} - \phi_{\text{node2}[i]} \sim \mathcal{N}\left(0, \sigma_\phi^2\right) \quad \text{for } i = 1, \ldots, N_{\text{edges}}$

$\sum_{r=1}^{R} \phi_r^2 \sim \mathcal{N}\left(0, R \cdot \sigma_\phi^2\right)$

Likelihood:

$y_n \sim \mathcal{N}(\mu + \beta \cdot t_n + \phi_{r_n}, \sigma) \quad \text{for } n = 1, \ldots, N$

In 2020, the global average suicide rate was 9.2 people per 100,000 people (World Health Organization, n.d.). Therefore, we've chosen this as the mean for our prior on the estimate of the global suicide rate μ. Additionally, a standard deviation of 3 allows for reasonable uncertainty around the average without being overly tight.

The β parameter represents the effect of time. In the US, the suicide rate in 2019 was 13.9 people per 100,000, then decreased to 13.5 people per 100,000 in 2020. It increased to 14 people per 100,000 in 2021, so we've used the overall change in suicide rate, 0.1, from 2019 to 2021 as the mean parameter.

The prior on both the standard deviation of spatial effects σ_ϕ and the observation noise σ is $Exp(1)$, which allows for smaller, more reasonable standard deviations. We've chosen to use weakly informative parameters here.

Get Adjacency Pairs

Firstly we need to know which countries are neighbors of each other https://cran.r-project.org/web/packages/rnaturalearth/vignettes/rnaturalearth.html

During the data analysis process we realized that the country names in our dataset did not match with the country names of the rnaturalearthdata dataset that we are using to model the spatial data. This led to the model mistaking the countries as having no neighbors and producing nodes with values 0. To solve this, we renamed the country names in our dataset to match that of rnaturalworld's.

```r
data_cleaned <- data %>%
  mutate(Location = case_when(
    Location == "Viet Nam" ~ "Vietnam",
    Location == "Türkiye" ~ "Turkey",
    Location == "Iran (Islamic Republic of)" ~ "Iran",
    Location == "Russian Federation" ~ "Russia",
    Location == "Republic of Korea" ~ "South Korea",
    Location == "Syrian Arab Republic" ~ "Syria",
    Location == "Brunei Darussalam" ~ "Brunei",
    Location == "Netherlands (Kingdom of the)" ~ "Netherlands",
    Location == "Republic of Moldova" ~ "Moldova",
    Location == "Lao People's Democratic Republic" ~ "Laos",
    Location == "United Kingdom of Great Britain and Northern Ireland" ~ "United Kingdom",
    Location == "Venezuela (Bolivarian Republic of)" ~ "Venezuela",
    Location == "Bolivia (Plurinational State of)" ~ "Bolivia",
    Location == "Democratic People's Republic of Korea" ~ "North Korea",
    Location == "Micronesia (Federated States of)" ~ "Federated States of Micronesia",
```

```
    Location == "Cote d'Ivoire" ~ "Ivory Coast",
    Location == "Eswatini" ~ "eSwatini",
    Location == "Timor-Leste" ~ "East Timor",
    Location == "occupied Palestinian territory, including east Jerusalem" ~ "Palestine",
    Location == "Sao Tome and Principe" ~ "São Tomé and Principe",
    Location == "Bahamas" ~ "The Bahamas",
    Location == "Congo" ~ "Republic of the Congo",
    Location == "Serbia" ~ "Republic of Serbia",
    TRUE ~ Location  # keep all other names unchanged
  ))
```

Now we can join the two datasets so our original dataset will have adjacency parameters from world_sf

```
# From rnaturalearth dataset
world_sf <- ne_countries(scale = "medium", returnclass = "sf") %>%
  st_make_valid() %>%
  filter(admin %in% data_cleaned$Location) %>%
  arrange(admin)  # ensure a consistent order

world_sf$region_id <- 1:nrow(world_sf)

data_matched <- data_cleaned %>%
  filter(Location %in% world_sf$admin) %>%
  left_join(world_sf %>% st_drop_geometry() %>% select(admin, region_id),
            by = c("Location" = "admin"))

stopifnot(all(!is.na(data_matched$region_id)))

any(world_sf$region_id == 0)
```

```
## [1] FALSE
```

Convert Neighbor List to Adjacency Pairs

```
world_sp <- as(world_sf, "Spatial")

neighbors <- poly2nb(world_sp, row.names = world_sf$region_id)

num_neighbors <- sapply(neighbors, length)
R <- length(neighbors)

#regions with at least one neighbor -> we want to leave out countries with no neighbors
valid_indices <- which(num_neighbors > 0)

node1 <- c()
node2 <- c()

for (i in valid_indices) { #only make nodes for countries with neighbors
  for (j in neighbors[[i]]) {
    if (j != 0 && world_sf$region_id[j] != 0) { #purposefully excluded zeros so node2 doesnt
                                               # have 0 "indexing" from region_id in world_sf
      node1 <- c(node1, world_sf$region_id[i])
      node2 <- c(node2, world_sf$region_id[j])
    }
```

```
  }
}

stopifnot(!any(node2 == 0))
length(node1)
```

```
## [1] 610
```

```
length(node2)
```

```
## [1] 610
```

```
any(node2==0)
```

```
## [1] FALSE
```

STAN Data List

```
nrow(data_matched)
```

```
## [1] 740
```

```
stan_data <- list(
  N = nrow(data_matched),
  y = data_matched$FactValueNumeric,
  time = as.integer(data_matched$Period == 2021),
  R = R,
  region = data_matched$region_id,
  N_edges = length(node1),
  node1 = node1,
  node2 = node2,
  num_neighbors = num_neighbors
)
```

Extract posterior data from STAN code file Code reference: https://ubc-stat-ml.github.io/web447/w08_mcmc1/topic06_hands_on.html Why use iter = 2000 and chains = 4:

```
model <- stan_model(file = "model.stan")
fit <- sampling(model, data = stan_data, iter = 4000, warmup = 2000, chains = 4, seed = 123)
print(fit)
```

!!! TODO- evaluation of posterior, e.g. "An appropriate combination of diagnostics, synthetic datasets and other validation strategies."

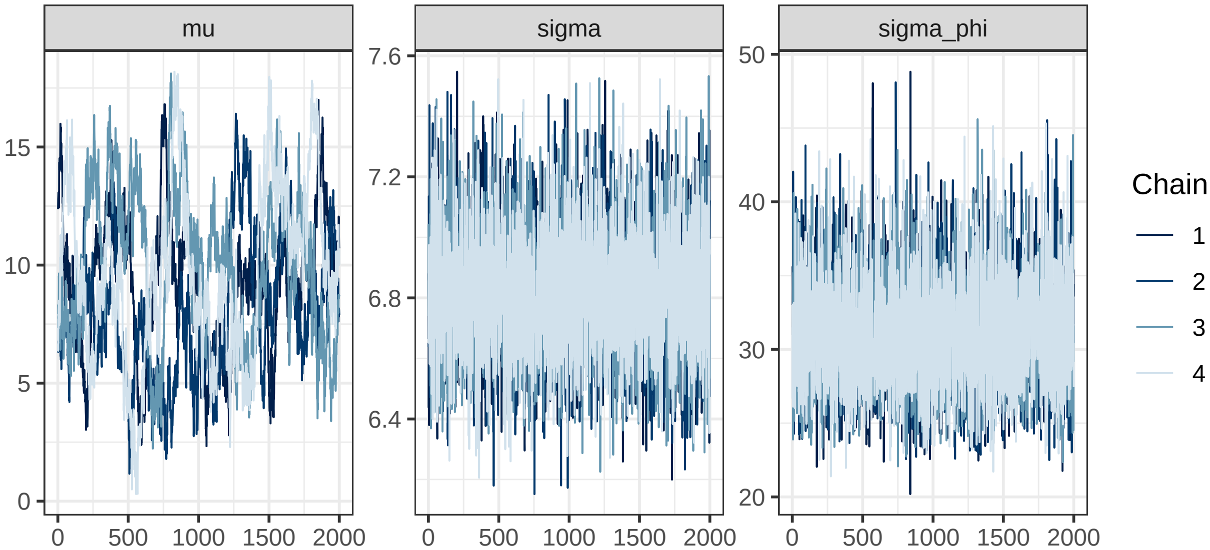
Model Diagnostics

```
head(summary(fit)$summary)
```

```
##                mean      se_mean        sd          2.5%          25%
## mu       9.21328283 0.4121980218 3.2329925    3.504476632   6.81115523
## beta     0.09573789 0.0006022686 0.0494647   -0.001462817   0.06271374
## phi[1]  -5.35047226 0.4215234074 4.5361666 -14.335074442  -8.48537360
## phi[2]  -5.64014370 0.4177089407 4.6207632 -14.608047263  -8.78799909
## phi[3]  -6.98719067 0.4048469159 4.5206663 -16.086907444 -10.04065942
## phi[4]  -1.40416306 0.4185862317 4.5760501 -10.648498596  -4.43054666
##                 50%        75%      97.5%      n_eff     Rhat
## mu       9.14504666 11.4043095 15.7602152   61.51735 1.074959
```

```
## beta    0.09521792  0.1286139  0.1922375 6745.44155 1.000046
## phi[1] -5.30809944 -2.1912208  3.3495637  115.80694 1.036607
## phi[2] -5.68149566 -2.4109242  3.4240582  122.37139 1.037162
## phi[3] -6.89638643 -3.8206648  1.5795925  124.68759 1.035359
## phi[4] -1.35629479  1.6606395  7.4720218  119.51204 1.035097
```

```
mcmc_trace(as.array(fit), pars = c("mu", "sigma", "sigma_phi"))
```




```
summary(fit, pars = "mu")$summary
```

```
##        mean  se_mean       sd     2.5%      25%      50%      75%    97.5%
## mu 9.213283 0.412198 3.232993 3.504477 6.811155 9.145047 11.40431 15.76022
##       n_eff     Rhat
## mu 61.51735 1.074959
```

Posterior Visualization

```
# Extract posterior samples for phi
phi_samples <- rstan::extract(fit)$phi

# Calculate posterior mean of phi for each region
phi_mean <- apply(phi_samples, 2, mean)

# Add phi_mean to the world_sf dataset
world_sf$phi_mean <- phi_mean

# Visualize the posterior mean of phi on the map
#m <- mapview(world_sf, zcol = "phi_mean")
#m

# uncomment the following lines to convert to png
#mapshot(m, file = "phi_map.png")
#knitr::include_graphics("phi_map.png")
```

Discussion

Results

TODO!!

Limitations

The spatial prior assumes that nearby countries have similar suicide rates, so if there are sharp differences between neighboring regions, the model may over-smooth and underrepresent the true variation.

Additionally, the focus of this analysis is whether or not there is a spatial relationship between location and suicide rate, but it does not consider underlying factors, such as culture, mental health resources, and economic state. For example, the topic of mental health is considered to be taboo in many countries, resulting in limited access to mental health resources.

The next step would be to look further into these underlying factors and determine whether or not there is a relationship between the factor and suicide rate (e.g. is there a relationship between suicide rate and lower income households in the United States and Canada?). This can be combined with information about distribution of suicide rates across sex and age to give more insight into which subset of groups should be targeted for suicide prevention methods in certain countries.

Member Contributions

!!!TODO - small paragraph discussing what each member did

Appendix

TODO!! - move all code here at the end

References

Brunsdon, C. (2019). *Using rstan and spdep for spatial modelling.* RPubs by RStudio.https://rstudio-pubs-static.s3.amazonaws.com/243277_01730c1f0a984132bce5d5d25bec62aa.html

Centers for Disease Control and Prevention. (2022, September 30). *Suicide increases in 2021 after two years of decline.*https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2022/20220930.html

Congdon, P. (1997). Bayesian models for spatial incidence: A case study of suicide using the BUGS program. *Health & Place,* 3(4), 229–247.https://doi.org/10.1016/s1353-8292(97)00017-8

Donegan, C. (2023, October 2). *Custom spatial models with RStan and geostan.* The Comprehensive R Archive Network.https://cran.r-project.org/web/packages/geostan/vignettes/custom-spatial-models.html

Garnett, M. F., & Curtin, S. C. (2023). Suicide mortality in the United States, 2001–2021. *NCHS Data Brief No. 464.*https://doi.org/10.15620/cdc:125705

Moraga, P. (n.d.). *Bayesian Spatial Models.* Spatial Statistics for Data Science: Theory and Practice with R.https://www.paulamoraga.com/book-spatial/bayesian-spatial-models.html

Morris, M. (n.d.). *Spatial Models in Stan: Intrinsic Auto-Regressive Models for Areal Data.* Stan: Software for Bayesian Data Analysis.https://mc-stan.org/users/documentation/case-studies/icar_stan.html

Pathirathna, M. L., Nandasena, H. M., Atapattu, A. M., & Weerasekara, I. (2022). Impact of the COVID-19 pandemic on suicidal attempts and death rates: A systematic review. *BMC Psychiatry,* 22.https://doi.org/10.1186/s12888-022-04158-w

U.S. Department of Health and Human Services. (2025, March). *Suicide.* National Institute of Mental Health.https://www.nimh.nih.gov/health/statistics/suicide

World Health Organization. (n.d.). *Suicide Rates.*https://www.who.int/data/gho/data/themes/mental-health/suicide-rates