svi

Kexin Wang, Yiming Miao

12/21/2022

```
library(svi)
library(dplyr)
library(purrr)
library(ggplot2)
```

Background

The package `svi` includes two datasets: `vulnerability` and `diabetes`, and 6 relevant functions to do basic data cleaning and visualizations.

Dataset: `vulnerability` Social vulnerability measures a community's ability to prevent suffering and financial loss in case of disasters, by exhibiting certain social conditions, e.g. poverty rate, household composition, and educational attainment. By studying the distribution of vulnerability across the country, the government may better identify the disadvantaged group in a specific community or region, and hence make a more well-designed plan regarding emergency events or long-term social benefits.

The social vulnerability index 2018 dataset (`vulnerability`) in this package was achieved from the Centers for Disease Control and Prevention (CDC). The dataset records the relative vulnerability estimates of the US at a county-level by ranking the 15 census variables in 4 major themes: socioeconomic status (SES), household composition and disability (HCD), minority status and language (MSL), housing type and transportation (HTT).

The SES theme comprises 4 variables: percentage below poverty, percentage unemployed, per capita income, and percentage with no high school diploma. The HCD theme constitutes 4 variables: the percentage of people aged *geq* 65 years, percentage of people aged ≤ 17 years, percentage of disabled people, and percentage of single-parent households. The MSL theme comprises 2 variables: percentage minority and the percentage who speak English less than well. Finally, the HTT theme includes 5 social variables: percentage multi-unit structures, percentage mobile homes, percentage crowding, percentage no vehicle, and percentage group quarters. The vulnerability index (percentile ranking) of 15 variables, 4 theme summary, and 1 overall summary are scored from 0 to 1, with higher values denoting higher vulnerability.

Dataset: `diabetes` The diabetes dataset (`diabetes`) in this package was obtained from CDC WONDER Online Database. It contains the population and number of deaths due to diabetes of each county in 2018. Crude rates are expressed as the number of deaths per 100,000 population. In the original dataset, crude rates were represented as "unreliable" for counties with diabetes deaths less than 20.

Functions: In this package, there are 3 functions regarding data wrangling: `mnnn_to_na`, `cr_interpolate`, `prepare`; and 3 functions regarding visualizations: `svi_map`, `mortality_map`, `mortality_vs_svi_scatter`. There functionalities are listed as below:

- `mnnn_to_na`: substitutes a target value in a column into `NA`
- `cr_interpolate`: substitutes the `unreliable` items in `diabetes$crude` into `NA` or the calculated results
- `prepare`: join the vulnerability dataset with mortality dataset, and select svi in themes only
- `svi_map`: plot a given vulnerability index per county on a US map
- `mortality_map`: plot the mortality of a disease per county or per state on a US map
- `mortality_vs_svi_scatter`: visualize the relationship between mortality rate and a vulnerability index in a scatter plot

With this package, users can achieve the CDC SVI 2018 dataset and the diabetes mortality data, and use the functions to conduct primary data cleaning and visualizations. Users may also download other mortality data of other diseases from CDC WONDER Online Database. The functions are managed to handle datasets in the same format.

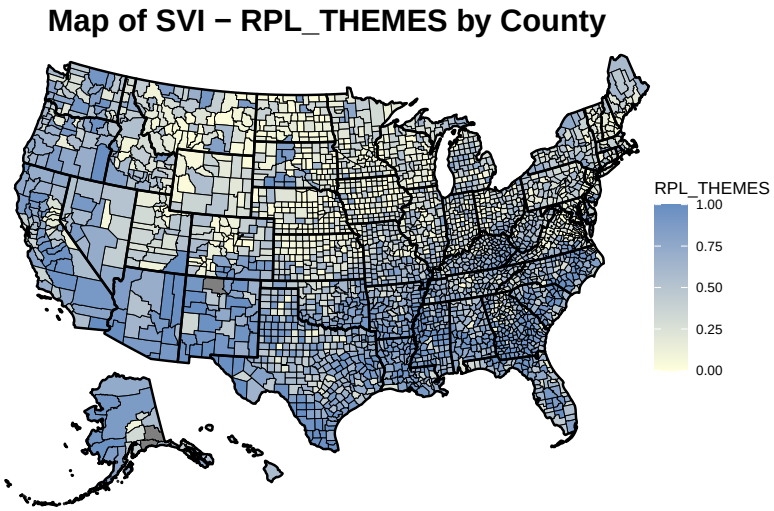
Research Question and Motivation

Diabetes, a chronic and metabolic disease, was the seventh leading cause of deaths in the United States in 2018. Both environmental and genetic factors can lead to diabetes. Some specific causes include family history, race, age, obesity and diet. Meanwhile, social determinants such as education, income and housing have also been shown to be responsible for the increased incidence. By effectively identifying high-risk areas and populations for diabetes, the government can provide more appropriate social support services and reduce health disparities.

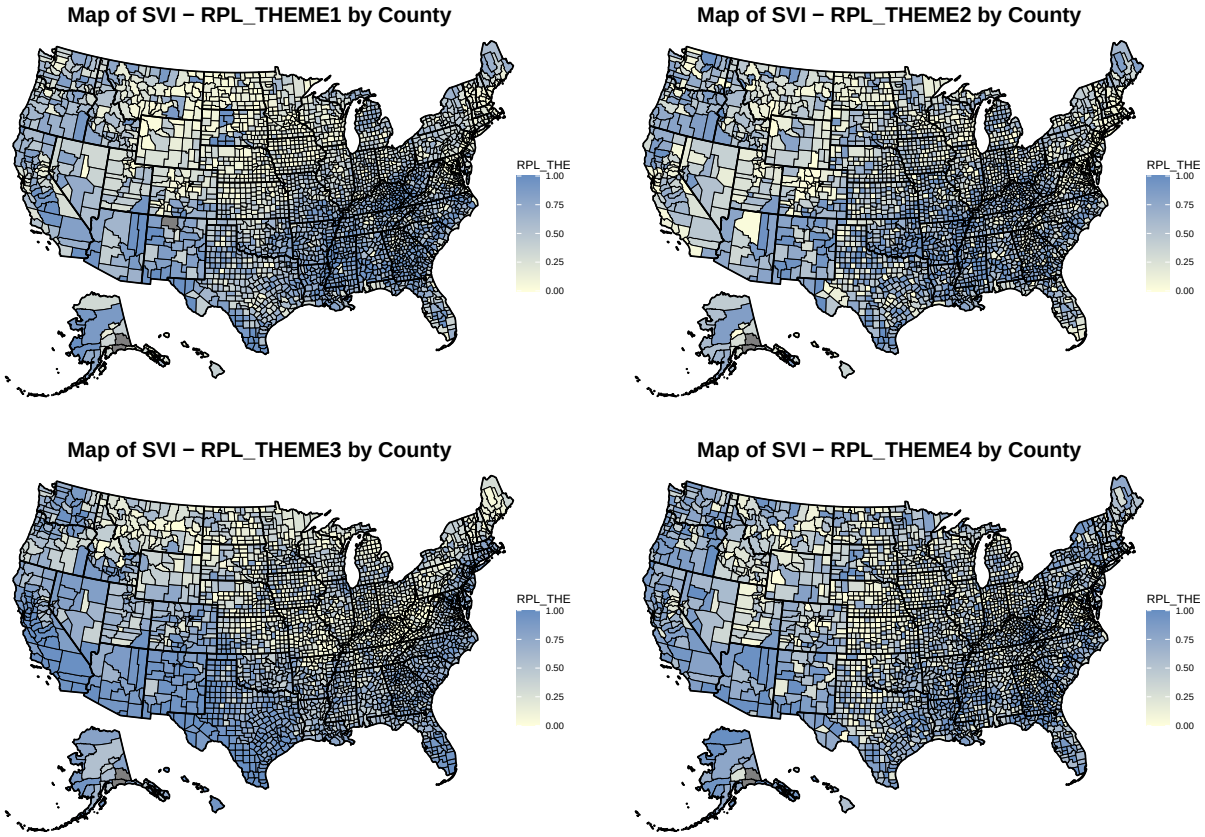
To address this pressing health concern, a growing number of studies have scrutinized the impact of multi-level social factors on diabetes. In this R package project, with the built-in data and functions, we aim to investigate the impact of four themes of social vulnerability on US diabetes deaths, and identify hot spots where social vulnerability index was positively associated with death rates.

Data cleaning and exploration

```
vulnerability |> mnnn_to_na(names(which(map_lgl(vulnerability, is.double))), -999) |>
                 rename(fips = FIPS) |>
                 svi_map("RPL_THEMES")
```



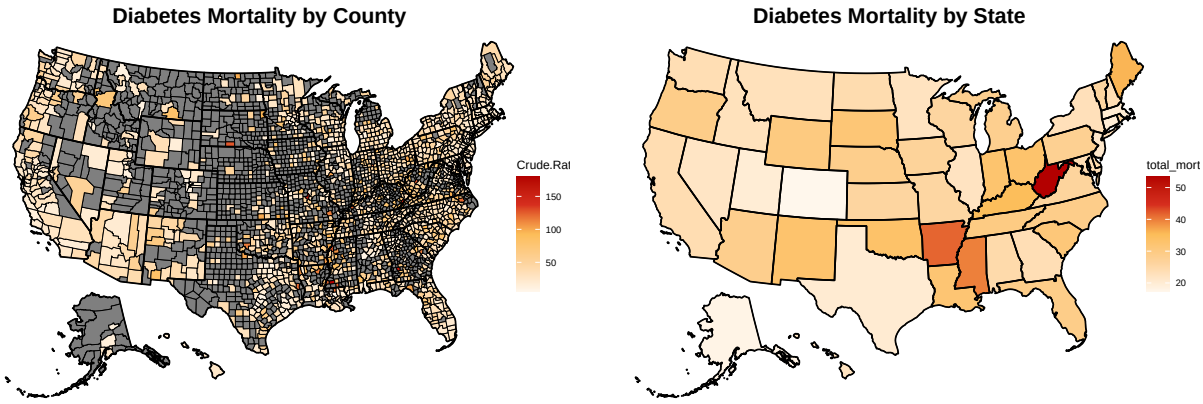
```
temp <- vulnerability |>
        mnnn_to_na(names(which(map_lgl(vulnerability, is.double))), -999) |>
        rename(fips = FIPS)
svi_map(temp, "RPL_THEME1")
svi_map(temp, "RPL_THEME2")
svi_map(temp, "RPL_THEME3")
svi_map(temp, "RPL_THEME4")
```



The five plots above visualizes the social vulnerability indices in the US at county-level, in the order of overall index, SES theme(1), HCD theme(2), MSL theme(3) and HTT theme(4). We can find that the overall vulnerability index is relatively higher in the west and south than in the north part. For theme1 (SES) and theme3 (MSL), SVI is higher in the southeast part; whereas for theme 3 (MSL) and theme 4 (HTT), SVI showed a higher trend in the west part.

The plots illustrate that there should be a higher proportion of minorities in the south and west part of the US, like California, Arizona, New Mexico, and Texas; and this may be because of the closeness to Mexico. Also, in some southeast states, e.g. Arkansas and Mississippi, there might be a higher proportion of people under poverty or without high school education. This suggests that upon social emergencies, the federal government may focus more on the minority group in the west states, and the low-income group in the southeast states for better resource-allocation.

```
diabetes |> cr_interpolate(reliable = FALSE) |>
                              rename(fips = County.Code) |>
                              mortality_map("Crude.Rate")
diabetes |> mortality_by_state() |>
                             mortality_map("total_mortality", "State", "Diabetes")
```



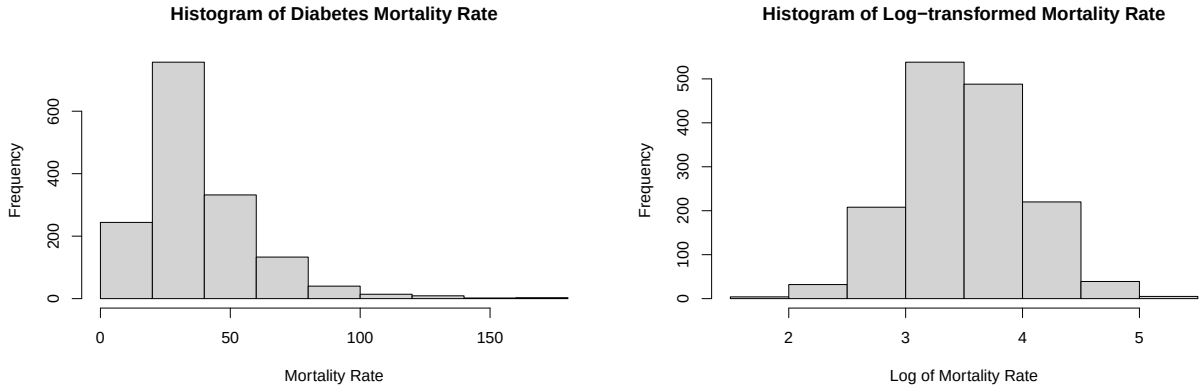
These two plots visualize diabetes mortality at county and state levels respectively. Counties with missing data are filled in gray. To address the issue of sparsity, we averaged the total death rates across states based on population size. The eastern region has more states with higher mortality rates, with West Virginia averaging more than 50 deaths per 100,000 population due to diabetes. This is followed by Arkansas and Mississippi (around 40). These results suggest that the government needs to pay more attention to the policies supporting improved lifestyles and dietary choices in areas with high diabetes mortality. However, the central region has serious missing data problems that the remaining counties may not be adequately representative of the entire state. Comparing the map of social vulnerability and diabetes mortality, we can find that the mid-east part of the US showed relatively higher vulnerability index and higher mortality rate of diabetes, but a similar pattern was unobvious to detect in other regions.

Analysis

In this part, we are going to evaluate the relationship between SVI and diabetes mortality rate quantitatively with a linear model. Since the distribution of mortality is right-skewed, we log-transformed the diabetes mortality as the outcome variable to make skewed data to approximately conform to normality.

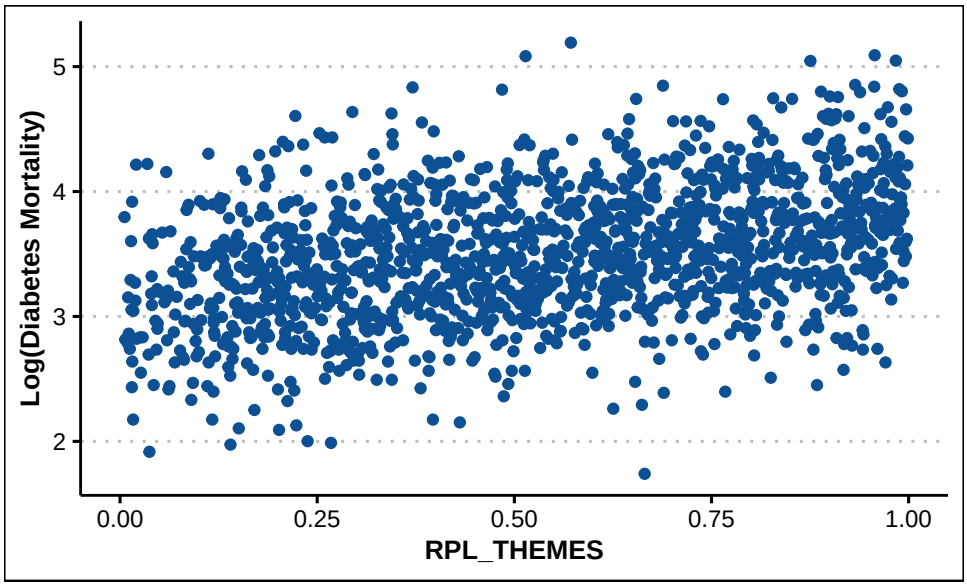
```
df <- prepare(vulnerability, diabetes, reliable = FALSE)
hist(df$MORTALITY, breaks = 9,
       main = "Histogram of Diabetes Mortality Rate",
       xlab = "Mortality Rate")

hist(log(df$MORTALITY), breaks = 9,
       main = "Histogram of Log-transformed Mortality Rate",
       xlab = "Log of Mortality Rate")
```

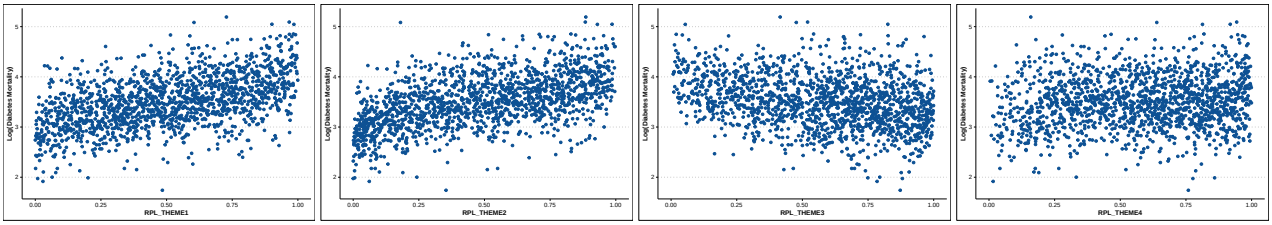


We firstly create scatter plots of diabetes mortality versus the overall SVI and four thematic indices.

```
df |> mortality_vs_svi_scatter("RPL_THEMES", "Diabetes")
```




```
df |> mortality_vs_svi_scatter("RPL_THEME1", "Diabetes")
df |> mortality_vs_svi_scatter("RPL_THEME2", "Diabetes")
df |> mortality_vs_svi_scatter("RPL_THEME3", "Diabetes")
df |> mortality_vs_svi_scatter("RPL_THEME4", "Diabetes")
```



From the scatter plots, overall SVI, socioeconomic status and household composition are positively correlated to diabetes mortality while minority is negatively associated. There is no obvious relationship between housing type and mortality.

We then further fit a multiple linear regression to model the relationship between log mortality and the vulnerability indices in 4 themes:

```
model1 <- lm(log(MORTALITY) ~ RPL_THEME1 + RPL_THEME2 + RPL_THEME3 + RPL_THEME4, data = df)
summary(model1)
#>
#> Call:
#> lm(formula = log(MORTALITY) ~ RPL_THEME1 + RPL_THEME2 + RPL_THEME3 +
#>     RPL_THEME4, data = df)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -1.49434 -0.23672  0.00216  0.24230  1.59552
#>
```

```
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  3.31626    0.03109 106.660   <2e-16 ***
#> RPL_THEME1   0.67307    0.06022  11.177   <2e-16 ***
#> RPL_THEME2   0.47509    0.05051   9.405   <2e-16 ***
#> RPL_THEME3  -0.63363    0.03976 -15.936   <2e-16 ***
#> RPL_THEME4  -0.02946    0.04808  -0.613     0.54
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.3891 on 1528 degrees of freedom
#>   (1609 observations deleted due to missingness)
#> Multiple R-squared:  0.4276, Adjusted R-squared:  0.4261
#> F-statistic: 285.3 on 4 and 1528 DF,  p-value: < 2.2e-16
```

The results of regressing log-transformed mortality on the overall index are consistent with the findings of the scatter plots.

The coefficients show that an increase of one unit in socioeconomic status (household composition) ranking is associated with 67.3 (47.5) percent increase in mortality; whereas a decrease of one unit in minority ranking is related to 63.4 percent increase in mortality. A p-value of 0.54 indicates that the relationship between housing type and mortality is not statistically significant.

The negative coefficient of `RPL_THEME3` is slightly counterintuitive here, since it is usually thought that being minority is a risk factor of limited access to health service, and hence higher mortality rate. However, the negative coefficient in the model above illustrates that being a minority is protective in diabetes death, with other social circumstances conditioned. For this issue, there might be two explanations. First, minority ethnicities are not having higher mortality rates, but the lower socioeconomic status and other factors that are usually correlated with it lead to the higher mortality. Given the same socioeconomic status and educational attainment, it may be the truth that racial ethnicity itself contributes to lower diabetes mortality rate. Second, "minority" is a broad concept, which includes various ethnicities, e.g. hispanic, African American, and asian. The given data only records the total percentage of minorities, but without further details on the composition of specific races. However, the prevalence of diabetes among ethnicities do vary a lot, so this issue may lead to the model results being slightly biased.

Furthermore, the R-squared (0.43) of the model is not high enough, which indicates there may be other factors besides SVI that can explain the variability in diabetes mortality.

Interpretations and Conclusion

Limitations

There are several limitations to this study. Our unit of analysis is a geographical unit involving state and county rather than an individual. The findings can only represent the general population within the geographical area and cannot be linked to an individual case. In addition, some researchers show that assessing social vulnerability using CDC's SVI is associated with lower accuracy and weaker validity, and is sensitive to weighting schemes. Despite these inherent limitations, this measurement provides a relatively comprehensive hierarchical index for emergency management compared with other indicators. Finally, missing diabetes death counts in some central counties may affect our assessment of the statewise situation.