Report for Japanese covid-19 death cases

Tsuyoshi Akiyama

2022-06-04

Abstract

This report's purpose is to analyze the Japanese time-series covid-19 data, including confirmed cases, deaths, and vaccinated. The main goal is to find which factors affect a lot to the people who died of covid-19 in Japan.

I will first collect covid-19's and vaccinated data from the 2 github repos. Next, I will clean them up to combine both to the one `global` table, then show 2 visualizations. I utilize random forest model to find a set of predictors that best explains the covid-19 new death cases. I will see 3 independent variables, new confirmed cases, partial vaccinated, fully vaccinated toward 1 dependent new death variable. Lastly, I will conclude the outcome and consider 2 possible biases behind this analysis.

The data sources are publicly operated by Johns Hopkins University in the github repo. The sources are inside https://github.com/CSSEGISandData/COVID-19, licenced under the Creative Commons Attribution 4.0 International. The data times pan is between 01/2020 to current day.

Prerequisite

- Install and load necessary libraries
- Define source URLs
- Set `country` variable. If you wanna see the different country case, other than Japan, change this variable here.

```
# Install dependencies
if (!require(tidyverse)) {
  install.packages("tidyverse", repos = "http://cran.us.r-project.org")
}
if (!require(lubridate)) {
  install.packages("lubridate", repos = "http://cran.us.r-project.org")
}
if (!require(randomForest)) {
  install.packages("randomForest", repos = "http://cran.us.r-project.org")
}
# Load dependencies
library(tidyverse)
library(lubridate)
library(randomForest)
# URLs
url_in = "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covic
file_names = c("time_series_covid19_confirmed_global.csv",
               "time_series_covid19_deaths_global.csv"
               )
urls <- str_c(url_in, file_names)
uid_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/UID_ISO_
vaccine_data_url <- "https://raw.githubusercontent.com/govex/COVID-19/master/data_tables/vaccine_data/gl
country <- "Japan"
```

Load and Tidy

Load the data

I don't get recovered data here because it wasn't recorded since 09/2021.

```
# Get the covid-19 confirmed case and death data
cases <- read_csv(urls[1])
deaths <- read_csv(urls[2])
# UIDs
uids <- read_csv(uid_url)
# Vaccines
vaccines <- read_csv(vaccine_data_url)
```

See the data's dimension, spec, and summary.

```
dim(cases)
# See original data specification
# spec(cases)
```

Tidy the data

- Reformat characterized date to lubridate:datetime.
- Select necessary columns for later analysis.
- Combine all collected data into 1 `global` table.

```
cases <- cases %>%
  pivot_longer(cols = -c(`Province/State`,
                         `Country/Region`,
                         Lat,
                         Long),
               names_to = "date",
               values_to = "cases"
               ) %>%
  filter(cases > 0) %>%
  select(-c(Lat, Long)) %>%
  mutate(date = mdy(date))

deaths <- deaths %>%
  pivot_longer(cols = -c(`Province/State`,
                         `Country/Region`,
                         Lat,
                         Long),
               names_to = "date",
               values_to = "deaths"
               ) %>%
  select(-c(Lat, Long)) %>%
  mutate(date = mdy(date))

uids <- uids %>%
  select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2, UID, FIPS))

vaccines <- vaccines %>%
  select(-c(Doses_admin, Report_Date_String, UID)) %>%
  rename(date = Date) %>%
  mutate(date = ymd(date))
```

I operate left-joined all data by its Country_Region, Province_State, and date columns. Now we have 1 time-series `global` table, having confirmed cases, deaths, population, and vaccinated cases.

```
global <- cases %>%
  left_join(deaths, by = c("Country/Region", "Province/State", "date")) %>%
  rename(Country_Region = `Country/Region`,
         Province_State = `Province/State`) %>%
  left_join(uids, by = c("Province_State", "Country_Region")) %>%
  left_join(vaccines, by = c("Province_State", "Country_Region", "date"))

dim(global)
```

```
## [1] 226939      8
```

```
summary(global)
```

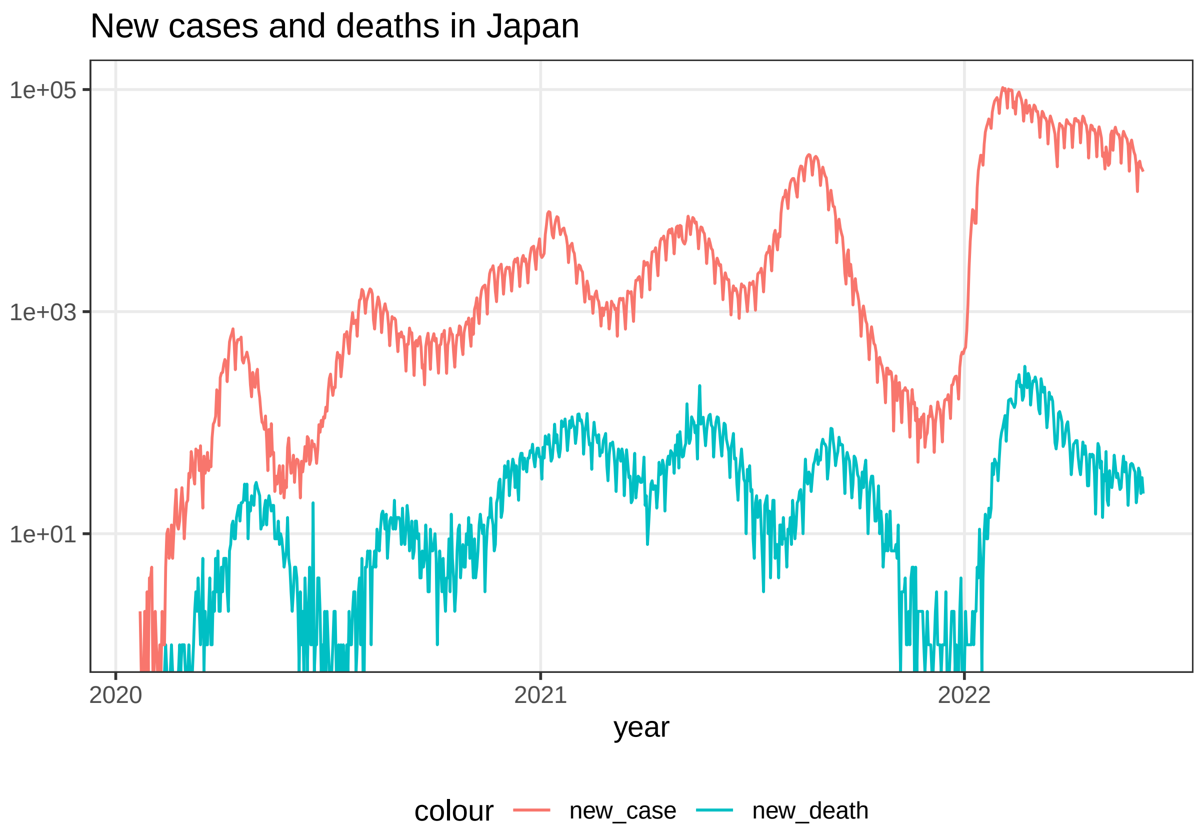
```
##  Province_State     Country_Region          date                cases
##  Length:226939      Length:226939      Min.   :2020-01-22   Min.   :       1
##  Class :character   Class :character   1st Qu.:2020-09-29   1st Qu.:     712
##  Mode  :character   Mode  :character   Median :2021-04-25   Median :   10997
##                                        Mean   :2021-04-22   Mean   :  636237
##                                        3rd Qu.:2021-11-15   3rd Qu.:  158411
##                                        Max.   :2022-06-04   Max.   :84748884
##
##      deaths           Population        People_partially_vaccinated
##  Min.   :      0   Min.   :8.090e+02   Min.   :0.000e+00
##  1st Qu.:      6   1st Qu.:8.696e+05   1st Qu.:2.647e+05
##  Median :    131   Median :7.133e+06   Median :1.360e+06
##  Mean   :  11325   Mean   :2.930e+07   Mean   :1.313e+07
##  3rd Qu.:   2492   3rd Qu.:2.914e+07   3rd Qu.:6.386e+06
##  Max.   :1008567   Max.   :1.380e+09   Max.   :1.101e+09
##                    NA's   :4505        NA's   :148548
##  People_fully_vaccinated
##  Min.   :0.000e+00
##  1st Qu.:9.412e+04
##  Median :8.872e+05
##  Mean   :9.840e+06
##  3rd Qu.:4.821e+06
##  Max.   :1.070e+09
##  NA's   :148548
```

Visualize the data

New cases and deaths in Japan

- 1st image is for new covid-19 confirmed cases and deaths in Japan.
- The red line shows new cases, while blue one for new deaths.
- There are correlation between the two variables.

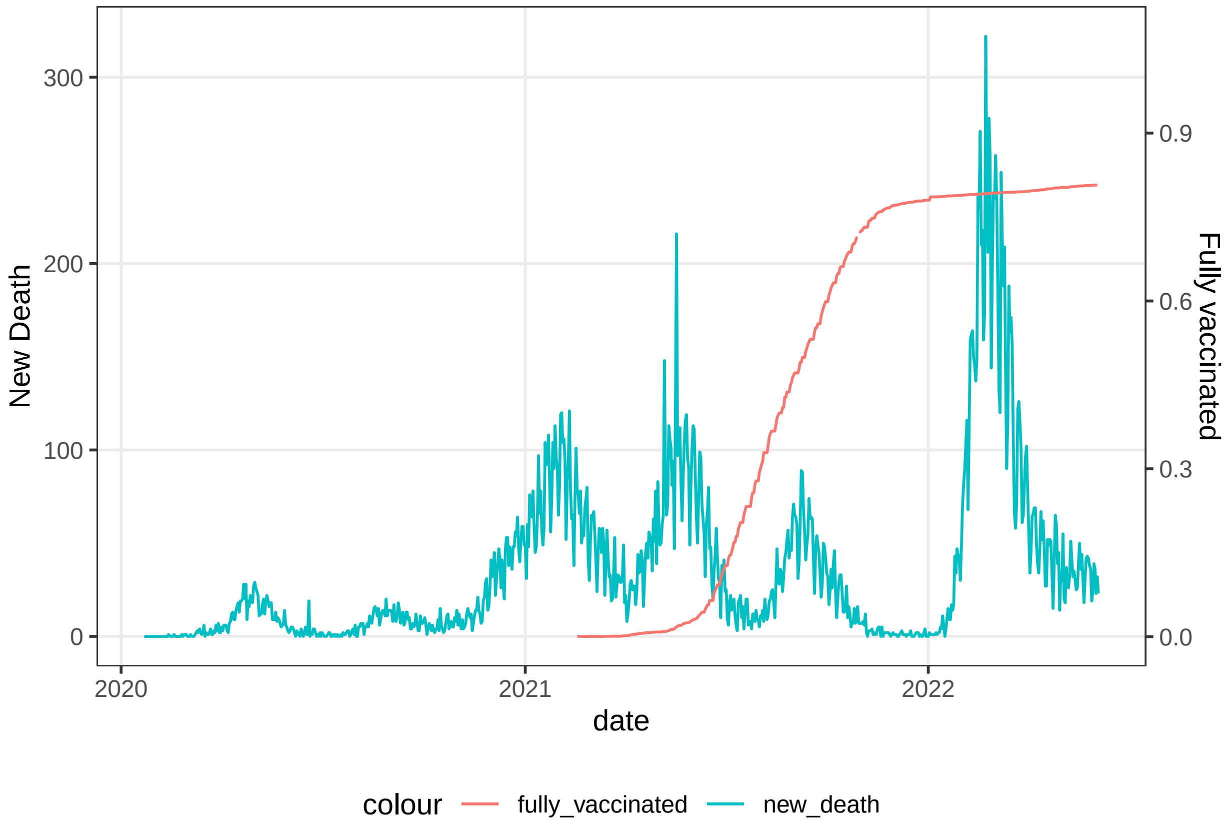
```
global %>%
  filter(Country_Region %in% c(country)) %>%
  mutate(new_death = c(deaths[1], diff(deaths)), new_case = c(cases[1], diff(cases))) %>%
  ggplot(aes(x = date)) +
  geom_line(aes(y = new_death, color = "new_death", label = new_death)) +
  geom_line(aes(y = new_case, color = "new_case", label = new_case)) +
  theme_bw() +
  scale_y_log10() +
  theme(legend.position = "bottom", panel.grid.minor = element_blank()) +
  labs(title = "New cases and deaths in Japan", y = "", x = "year")
```



New deaths and fully vaccinated in Japan

- 2nd image is for new covid-19 deaths and fully vaccinated ratio in Japan.
- The red line shows fully vaccinated ratio, while blue one for new deaths.
- The takeaway is no relationships between them, just by seeing those 2 variables.

```
vaccinatd_scale <- 300
global %>%
  filter(Country_Region %in% c(country)) %>%
  mutate(new_death = c(deaths[1], diff(deaths)), fully_vaccinated =  (People_fully_vaccinated / Populati
  ggplot(aes(x = date)) +
  geom_line(aes(y = new_death, color = "new_death", label = new_death)) +
  geom_line(aes(y = vaccinatd_scale * fully_vaccinated, color = "fully_vaccinated", label = fully_vaccin
  theme_bw() +
  theme(legend.position = "bottom", panel.grid.minor = element_blank()) +
  scale_y_continuous(
    name = "New Death",
    sec.axis = sec_axis(~./vaccinatd_scale, name="Fully vaccinated")
  )
```



In above visualization steps, The new cases seems the most influential factor to new death in Japan. Is it true? I don't yet decide its the biggest factor since vaccine could decrease the covid-19 death rate. So I try to make the model in the next section.

Model the data

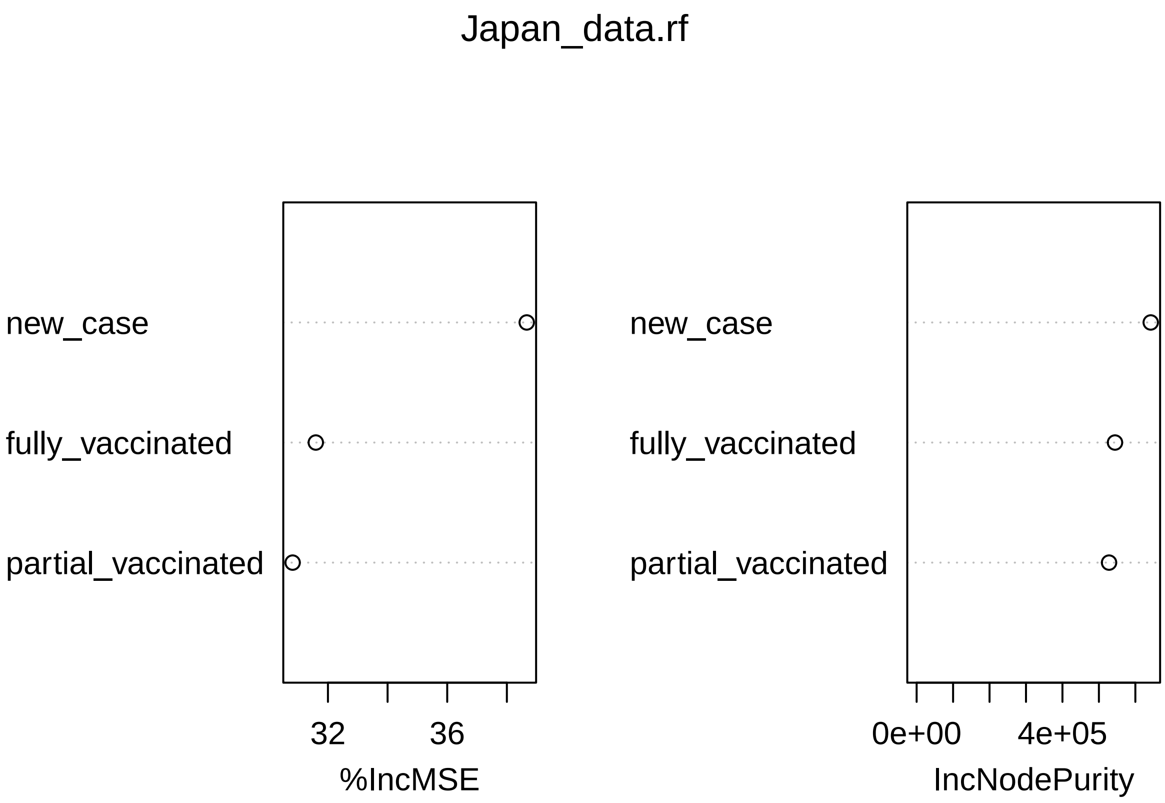
I will apply Random Forest Model to see the most influential factors to new death cases in Japan.

Feature Selection

- Find which one affects the most to new death cases among `new_case`, `partial_vaccinated`, and `fully_vaccinated`.
- Random Forest Model fits well in this case, having no hypothesis.

```
Japan_data <- global %>%
  filter(Country_Region %in% c(country)) %>%
  mutate(new_death = c(deaths[1], diff(deaths)), partial_vaccinated =  (People_partially_vaccinated / Pe

Japan_data.rf <- randomForest(new_death ~ new_case + partial_vaccinated + fully_vaccinated, data = Japan
imp = varImpPlot(Japan_data.rf)
```

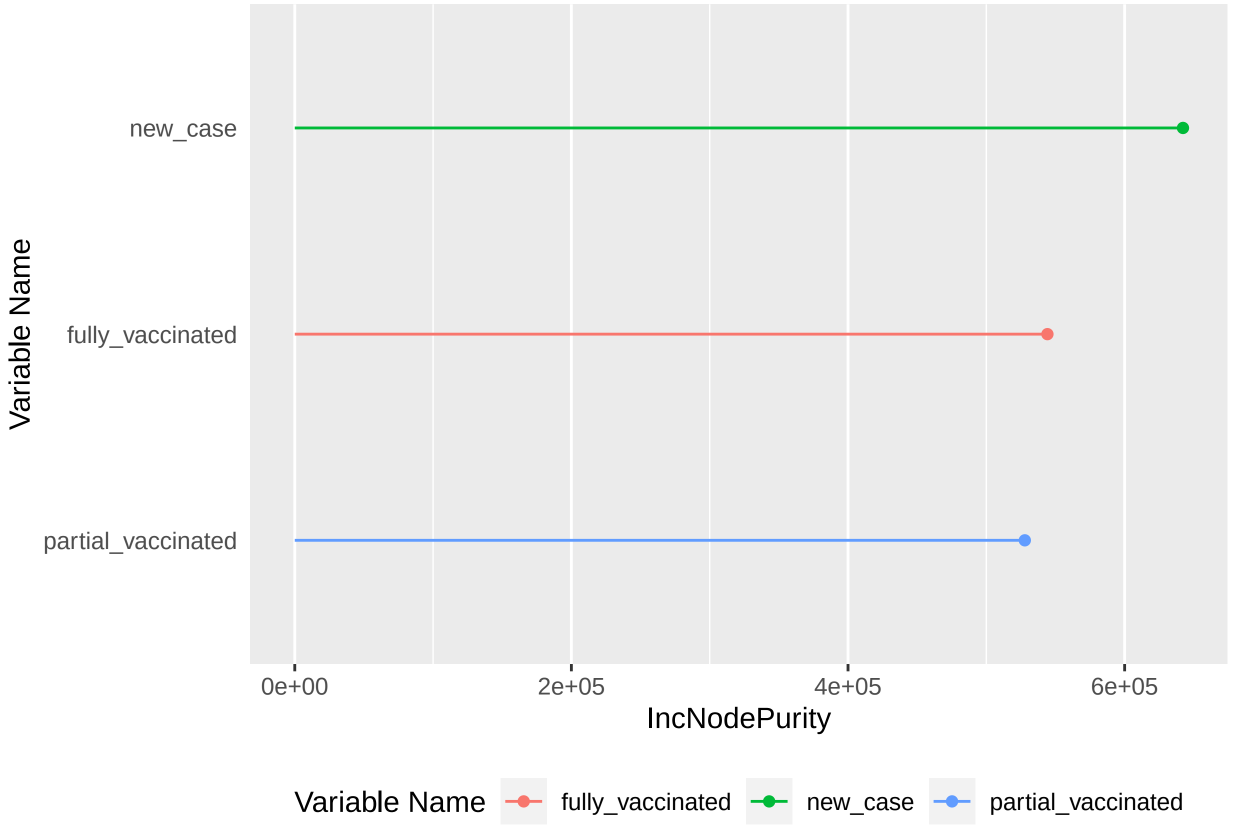


```
imp <- as.data.frame(imp)
imp$varnames <- rownames(imp)
rownames(imp) <- NULL
```

Visualize the result

- Find which one affects the most to new death cases among `new_case`, `partial_vaccinated`, and `fully_vaccinated`.
- Most influential variable is `new_case`, whereas `fully_vaccinated` are the second factor.

```
ggplot(imp, aes(x=reorder(varnames, IncNodePurity), y=IncNodePurity, color=as.factor(varnames))) +
  geom_point() +
  geom_segment(aes(x=varnames,xend=varnames,y=0,yend=IncNodePurity)) +
  scale_color_discrete(name="Variable Name") +
  ylab("IncNodePurity") +
  xlab("Variable Name") +
  coord_flip() +
  theme(
    legend.position="bottom",
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  )
```



```
pred_rf <- predict(Japan_data.rf, Japan_data)
print(tail(pred_rf))
```

```
##      860      861      862      863      864      865
## 27.02545 34.55483 34.28767 27.91272 28.62552       NA
```

Outcome and Biases

Outcome

New confirmed cases of covid-19 has the most influence to new death cases in Japan, while fully and partially vaccinated numbers are followed. The more new case increase, the more death increase. Vaccine has affected death case increment but not to new confirmed cases.

Possible biases

One possible bias is country bias. I use Japan as a sample in this report. If I need to take care of other areas, I should've applied this random forest to another countries. One way is to use 40 countries as training set, and different 10 countries as test set.

Another bias is inconsistent variable value. I need to compare "apple to apple" for the consistency. Vaccine might be from different maker, while inspection methods might vary on different time span. If I get a bit more detailed data, the outcome will change.