XXX *

Riya Chanduka *Clemson University*
Dennis Hammerschmidt *dhammers@mail.uni-mannheim.de*
Soyeon Jin *Springfield University*
Pauline Kleinschlömer

*Final paper submitted as part of the course 'Big Data and Immigration Research' in Spring 2019.

Contents

Introduction

The number of migrants, especially forced migrants are increasing at a rapid phase worldwide. According to the 2018 report from the International Organization for Migration (IOM), 68.5 million individuals are forced to migrate in 2017 because of persecution, conflict, generalized violence, or human rights violations. The issue gets notably salient in 2015 following so-called 'refugee crisis' due to the Syrian civil war. Meanwhile, on 2nd September 2015, the picture of drowned 3-year-old Syrian boy, Aylan Kurdi in the Mediterranean Sea alarmed the world the seriousness of the issue and the difficulties which migrants face during their journey. However, despite its political saliency, there are still a lot of data gaps in migration research. IOM pointed out the issue in its 2018 report, insisting the necessity to fill out data gaps in the topics such as irregular migration, missing migrants and migration flows. "Missing Migrants Project" by International Organization for Migration (IOM) began to fill the gap in quantitative data concerning migrants' journey after the tragic accident at October 2013, when two ships have wrecked near the Italian island of Lampedusa which caused the death of at least 368 individuals. Missing migrants project tracks the missing of refugees and asylum-seekers during their migration journey combining different data sources. By providing information about missing migrants, missing migrants project not only notify the world with the precarious situation migrants face but also point out the possible improvement in regulatory or administrative system to ensure safe migration journey. This achievement is important to tackle the United Nations' Sustainable Development Goals (SDGs) (Global Migration Data Portal, n.d.). Migration is a cross-cutting issue that is relevant to all SDGs. The Agenda's core principle is "leaving no one behind". Especially Goal 10.7 underlines an "orderly, safe, regular and responsible migration and mobility of people". Therefore, it is mandatory to understand patterns of migrants' journeys in order to achieve this goal by 2030. The number of missing migrants can serve as a helpful indicator. However, the data of the project should be seen as approximations because the true number of missing migrants is unknown. It is not possible to report all deaths and disappearances because many bodies will never be found or identified. IOM has recognized the importance of adopting the big data in migration research to fill the gap in quantitative data (IOM, 2017a). With acknowledgement of necessity to implement big data, Missing migrants project combine various sources and type of data to track the missing migrants information, from national authorities to interviews of migrants. To dedicate to the missing migrants project aim, this report tries to explain the present of migration route and the dangerousness of different routes in the Mediterranean area, taking full advantage of the data offered by missing migrants project. Since migration is a continuous process that never stops, it is necessary to keep a constant eye on the changing patterns of migrants' travel. Therefore, this report analyses different routes of migrants in the Mediterranean Sea and compares changes over time. Also, this report tries to figure out possible factors affecting the journey to prevent further incident. To achieve this aim, this report begins by explaining the various migration routes, especially focusing on three different routes in the Mediterranean area. Then this report proceeds to explain the missing migrants project and the uniqueness of the data, along with the necessity to use big data for migration research. After that, death and missing of migrants on those routes are presented, suggesting the possible cause of dangerousness of the routes. Lastly, the report discusses the limitation of the report and suggest a possible improvement of the report by combining the data with weather data.

Theoretical Background

Data and Methods

The present report uses the data provided by the 'Missing Migrants Project', conducted by the International Organization for Migration (IOM). The missing migrants project has begun after witnessing the tragic incident at October 2013 near the Italian island of Lampedusa, where two shipwrecks led to the deaths of at least 368 migrant individuals. In response to the incident and the total rising number of migrants, IOM launched the missing migrants project to report the present migration. The aim of the project is to improve the precarious situation migrants encounter during their journey. Missing migrants project provides information about the missings and death of the migrants at the external borders or during their journey due to transportation accidents, shipwrecks, violent attacks, or medical complications. Missing migrants project;s data and content is freely available at the project's website. All the analysis is done with the software R and Python. The missing migrants project has its strength in the fact that they gather information from diverse sources, employing not only traditional media but also social media to find data. The information sources range from official records to media reports, NGOS, and surveys and interviews of migrants. To provide latest data, missing migrants provide the number of arrivals and crossings for the current year every Monday and Thursday. To be specific, in the Mediterranean region, national authorities deliver the information to IOM field missions. At landing points in Italy and Greece, IOM and other organizations which receive survivors obtain the data. Also, IOM cooperates with UNHCR, the United Nation Refugee Agency, at the Mediterranean region to validate the data on missing migrants. On the other hand, on the United States and Mexico border, U.S county medical examiners, coroners, sheriff's officers and media reports covering the death on the Mexican side of the border provide and accumulate the data. Lastly, in Africa, media, NGOs, such as Regional Mixed Migration Secretariat and International Red Cross or Red Crescent gather the data concerning the issue. The missing migrants project covers various information about migrants' missings and death. It includes the region of incident, reported date, reported year, reported month, number dead, number missing, total dead and missing, number of survivors, number of females, males, and children, age, country of origin, region of origin, cause of death, location description, location coordinates, migration routes, UNSD geographical grouping, source quality, and further comments. The number of deaths indicates the total number of people confirmed dead in an incident. The people who are presumed to be dead due to an incident such as shipwrecks are left to be blank. Meanwhile, the number of missing migrants covers the shipwreck in general. It is recorded by subtracting the number of bodies recovered and the number of survivors from the total original number of migrants on the boat. The information also relies on a report by surviving migrants or witness. Where there are no reported missing migrants, it is left blank. The age of decedents is occasionally substituted with estimated age range. If it is not reported, it is left blank. Region of origin of the decedent's are at times recorded as 'Presumed' or (P) and if it is unknown, 'unknown' is recorded. When the cause of death could not be identified, the reason for the missing of identification is recorded (e.g., Unknown - Skeletal remains only). Migration routes show the migration route where the incident took place. With the help of this variable it is possible to examine changing patterns of the different routes mentioned above in the Mediterranean Sea. Lastly, source quality describes the quality of the incident's information with the 1-5 level. Level 1 incidents are based solely on a media source, level 2 incidents are based on uncorroborated eyewitness or data from survey respondents, and level 3 are based on information from more than one media reports. On the other hand, level 4 incidents have to be based on information from at least one NGO, IGO, or another humanitarian actor with direct knowledge

of the incident. Level 5 incident is the information from official sources such as coroners, medical examiners, or government officials from multiple humanitarian actors. Hence, the methodologies seeking for maximal accuracy and timeliness are employed. However, it is important to keep in mind that the data can only be seen as approximations due to the difficulty in obtaining these data. Nonetheless, the project's data can serve as a good starting point to analyse the scale and trends of the routes that migrants take.

Setup and basic data prepration

We're using the Missing Migrants dataset from 2014 to 2019 (up to the most recent dataset available). In our case, this is the one from May XX.

First descriptive: Where did people go missing the most?

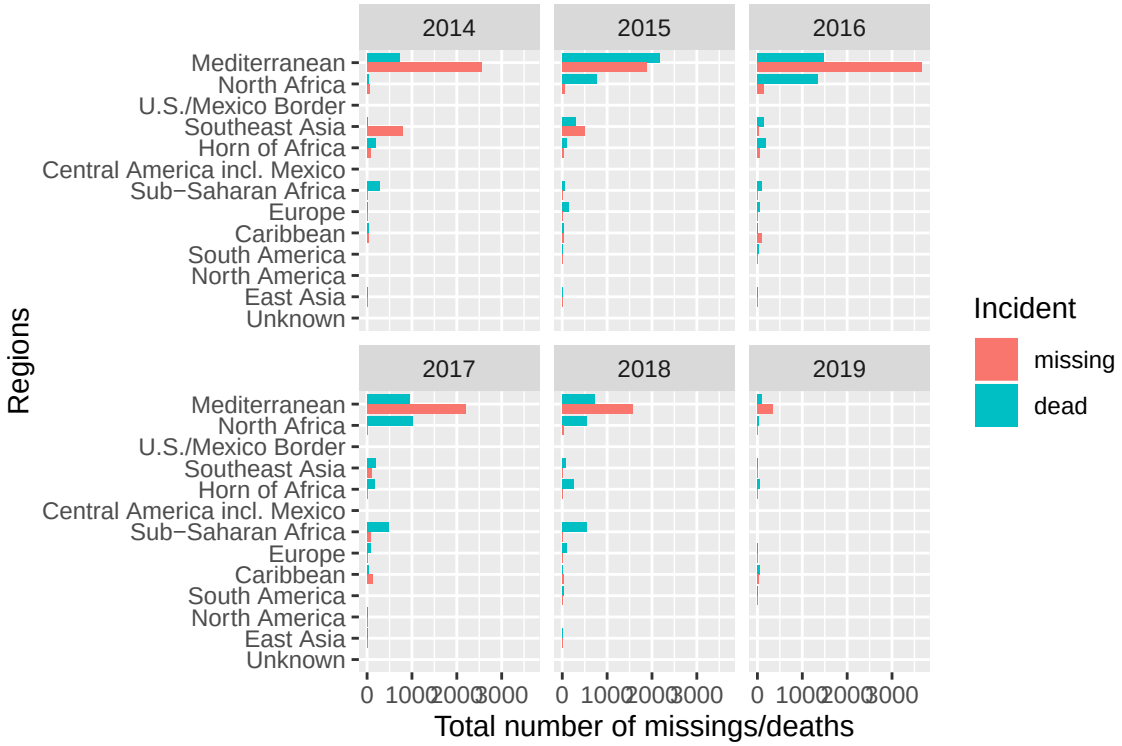
```
# create regions data that includes the sum of missings and dead migrants per year for each regi
regions <- missing %>%
  group_by(region, year) %>%
  dplyr::summarise(sum(est_miss), sum(num_dead))

regions <- as.data.frame(regions)
colnames(regions) <- c("region", "year", "missing", "dead")
regions <- melt(regions, id.vars = c("region", "year"))

# define position ordering for graph
positions <-
  c(
    "Unknown",
    "East Asia",
    "North America",
    "South America",
    "Caribbean",
    "Europe",
    "Sub-Saharan Africa",
    "Central America incl. Mexico",
    "Horn of Africa",
    "Southeast Asia",
    "U.S./Mexico Border",
    "North Africa",
    "Mediterranean"
  )

# plot the total number of missings and dead across regions and years
incident_region <- ggplot(regions) +
  geom_bar(aes(x = region, y = value, fill = variable),
           stat = 'identity',
           position = 'dodge') +
  coord_flip() +
```

```
  labs(x = "Regions", y = "Total number of missings/deaths") +
  scale_x_discrete(limits = positions) +
  labs(fill = "Incident") +
  facet_wrap(vars(year))
# uncomment for html output
incident_region
```




```
# use ggplotly for interactive exploration
# ggplotly(incident_region)
```

As we see and suspected before, most incidences happen at the Mediterranean Sea.

```
# subset only for the Mediterranean Sea
missing_medsea <- subset(missing, region == "Mediterranean")

# create regions data only for Mediterranean Sea that also includes the number of people survivi
region_med <- missing_medsea %>%
  group_by(route, year) %>%
  dplyr::summarise(sum(est_miss), sum(num_dead), sum(num_surv))

region_med <- as.data.frame(region_med)
colnames(region_med) <-
  c("route", "year", "missing", "dead", "survived")
region_med <- melt(region_med, id.vars = c("route", "year"))

# reduce the name of the routes to their direction (since we're only in the Mediterranean Sea)
```

```r
route_short <-
  c(
    "Central Mediterranean" = "Central" ,
    "Western Mediterranean" = "Western",
    "Eastern Mediterranean" = "Eastern"
  )
region_med$route <- as.character(route_short[region_med$route])

# define position ordering for graph
positions2 <- c("Central", "Western", "Eastern")

# plot the total number of missings and dead across regions and years
surv_or_not <- ggplot(region_med) +
  geom_bar(aes(x = route, y = value, fill = variable),
           stat = 'identity',
           position = 'fill') +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Regions", y = "Percentage of incidents") +
  scale_x_discrete(limits = positions2) +
  labs(fill = "Incident") +
  facet_wrap(vars(year))
# uncomment for html output
surv_or_not
```

```r
# use ggplotly for visualization and round the percentage of stacked bar chart by 2 digits
# with_options(list(digits = 1), ggplotly(surv_or_not))
```

```r
monthly <- missing_medsea %>%
  group_by(year, month) %>%
  dplyr::summarise(sum(total_dead_missing), sum(num_surv))

monthly <- as.data.frame(monthly)
colnames(monthly) <- c("year", "month", "dead/missing", "survived")
monthly <- melt(monthly, id.vars = c("year", "month"))

# make the name of the routes shorter
mon_abb <-
  c(
    "01" = "Jan" ,
    "02" = "Feb",
    "03" = "Mar",
    "04" = "Apr",
    "05" = "May",
    "06" = "Jun",
    "07" = "Jul",
    "08" = "Aug",
    "09" = "Sep",
    "10" = "Oct",
    "11" = "Nov",
    "12" = "Dec"
  )
monthly$month <- as.character(mon_abb[monthly$month])

monthly$month <-
  factor(
    monthly$month,
    levels = c(
      "Jan",
      "Feb",
      "Mar",
      "Apr",
      "May",
      "Jun",
      "Jul",
      "Aug",
      "Sep",
      "Oct",
      "Nov",
      "Dec"
    )
  )
```

```
# plot the total number of missings and dead across regions and years
by_month <- ggplot(monthly) +
  geom_bar(aes(x = year, y = value, fill = variable),
           stat = 'identity',
           position = 'dodge') +
  #coord_flip() +
  labs(x = "Months", y = "Total number of incidents") +
  labs(fill = "Incident") +
  theme(axis.text.x = element_text(angle = 45)) +
  facet_wrap(vars(month), nrow = 2, ncol = 6)
# uncomment for html output
by_month
```

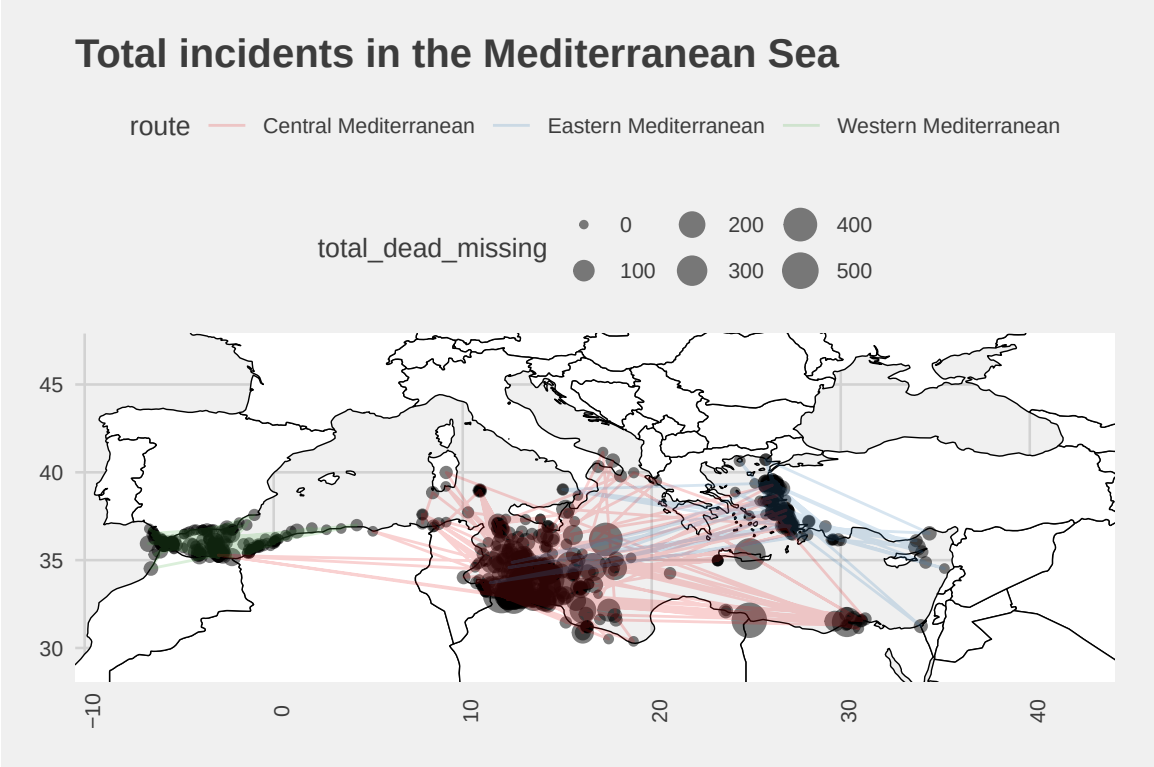

```
# for animated output
# ggplotly(by_month)
```

```
# intiatlize a simple map
worldMap <- fortify(map_data("world"), region = "mediterranean")

# define axis, shape and visual display of the map
map <- ggplot() +
  geom_map(
    data = worldMap,
    map = worldMap,
    aes(
```

```
      x = long,
      y = lat,
      map_id = region,
      group = group
    ),
    fill = "white",
    color = "black",
    size = 0.25
  )

# plot the map with total number of dead and missing migrants including the three routes in the
med_map <-
  map + geom_point(aes(x = lon, y = lat, size = total_dead_missing),
                   alpha = 0.5,
                   data = missing_medsea) +
  geom_path(aes(lon, lat, col = route), data = missing_medsea, alpha =
              0.2) +
  theme_fivethirtyeight(base_size = 10, base_family = "sans") +
  scale_color_brewer(palette = 'Set1') +
  theme(axis.text.x = element_text(size = 8, angle = 90),
        legend.position = 'top') +
  xlab('') + ylab('') +
  ggtitle('Total incidents in the Mediterranean Sea') + ylim(c(29, 47)) +
  xlim(c(-8, 42))
med_map
```



```r
# optionally: use ggplotly for visualization. Caution: takes some time to run!
# ggplotly(med_map)
```

```r
# If not available, load the transformr package directly for github using the following command(
# devtools::install_github("thomasp85/transformr") # load if needed
# library(transformr)

# create an animated map for the total number of dead and missing migrants as well as the three
mapanimated <-
  map + geom_point(aes(x = lon, y = lat, size = total_dead_missing),
                   alpha = 0.5,
                   data = missing_medsea) +
  geom_path(aes(lon, lat, col = route), data = missing_medsea, alpha =
              0.3) +
  theme_fivethirtyeight(base_size = 10, base_family = "sans") +
  scale_color_brewer(palette = 'Set1') +
  theme(axis.text.x = element_text(size = 8, angle = 90),
        legend.position = 'top') +
  xlab('') + ylab('') +
  ggtitle('Total incidents in the Mediterranean Sea') + ylim(c(29, 47)) +
  xlim(c(-8, 42)) +
  labs(title = 'Year: {frame_time}') +
  transition_time(year)
# uncomment for html output
# mapanimated

# optional: save the animation as .gif
#anim_save("mapanimated.gif", animation = last_animation())
```