Estimation of stateless populations given a population of unknown citizenship

Contents

1 Overview

This section provides a step-by-step summary of a method to estimate stateless population sizes when there exist data on both confirmed stateless populations and also persons of unknown citizenship. This is quite a common scenario in many countries as people may not have the required documentation to determine their citizenship status. The method discussed here draws on patterns of stateless persons, persons of unknown citizenship, and the broader migrant population, to estimate the likelihood that persons of unknown citizenship are stateless (in addition to those already confirmed to be stateless).

1.1 Case study

The case study discussed is estimating stateless populations by origin in the Netherlands. This method could conceivably be applied to other countries where the minimum data requirements are met (see below).

2 Data requirements

2.1 Ideal requirements

Broadly, the data required to estimate stateless populations using this method are:

1. The number of stateless persons by **age** and **origin country** for a particular year
2. The number of persons with unknown citizenship by **age** and **origin country** for the same year as above
3. The number of persons currently living in the country of interest by **age** and **origin country** for the same year as above
4. Expert opinion or other information source on the probability that a person from a certain origin country in the country of interest is stateless.

Items 1. and 2. are likely to come from official administrative sources. Item 3. is likely to come from a national census or other large-scale survey. Item 4. is likely to be consolidated information from a range of different subject matter experts.

In order for this method to be applied, there needs to be a set of origin countries that have information on **both** confirmed stateless counts and counts of persons with unknown citizenship. It is the comparison of both groups to a baseline of the migrant population in general which is the basis of the method.

2.2 Minimum requirements

Note that point 3. refers to the 'stock' of migrants in a particular country from a particular origin. As stated above, if it is available, this information is likely to be available through censuses or national surveys. However, it is often the case that up-to-date information from censuses or surveys is not readily available. As an alternative, we can draw on two readily-available sets of estimates to apply this method. In particular, we used data from UNDESA on migrant stocks by origin country[1]. These data are broadly available for all countries, although the breakdown by age is not available. As such, an alternative is obtaining information on the age patterns of migrants using data on the recent in-migrant *flows* to a country, which is often readily available through national offices, or through other statistical agencies. For example, migrant flows data is published annually for all European countries by Eurostat[2].

In this case, the data required would be

1. The number of stateless persons by **age** and **origin country** for a particular year
2. The number of persons with unknown citizenship by **age** and **origin country** for the same year as above
3. The number of persons currently living in the country of interest by **origin country** for the same year as above
4. The number of in-migrants to the country of interest by **age** and **origin country** for the same year as above
5. Expert opinion or other information source on the probability that a person from a certain origin country in the country of interest is stateless.

3 Method

3.1 Background and rationale of method

The goal of the method is to estimate the number of persons with unknown citizenship from a particular origin country that are likely to be stateless. These numbers are then added to the number of stateless persons from that origin country to give an updated estimate of the number of stateless persons.

The method draws on patterns of stateless persons, persons of unknown citizenship, and the broader migrant population, to estimate the likelihood that persons of unknown citizenship are stateless. The rationale is that, firstly, if the patterns by age of persons of unknown citizenship is more similar to the patterns by age of stateless persons, compared to the broader migrant population, then there

[1](https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/undesa_pd_2020_ims_stock_by_sex_destination_and_origin.xlsx)

[2](https://ec.europa.eu/eurostat/databrowser/view/migr_imm5prv/default/table?lang=en)

is an increased likelihood that they are stateless. Secondly, the overall sizes of the unknown, stateless and broader migrant population from the same origin country to get an indication of the prevalence — the larger the size of the stateless population compared to the total migrant population, the more likely those with unknown citizenship are likely to be stateless.

In summary, the approach is as follows:

1. We use information about the shape of the age distributions of various migrant groups to calculate a 'similarity' index
2. We also use information about underlying migrant stocks to calculate a 'prevalence' index
3. We then combine these two indexes and convert to a probability of being stateless

There are two possible groups of countries:

A) those that have data on both stateless persons and persons of unknown citizenship
B) those that have data on persons of unknown citizenship only

The method can be applied to contexts where there exist only countries in group A). However, there needs to be at least some countries in this group, that is, we need some countries that have information on both confirmed stateless counts and also counts of persons of unknown citizenship.

3.2 Similarity index

The logic of the first step is that the age distribution of unknown persons is more similar to stateless persons than the general migrant populations, then one could argue they are more likely to be stateless.

We can summarize this idea of 'similarity' by calculating the root mean squared error (RMSE) of the unknowns compared to stateless and unknowns compared to migrants:

$$RMSE = \sqrt{\sum_{i=1}^{G} \frac{(p_i^U - p_i^g)^2}{G}}$$

where i refers to age group (there are G in total); p_i^U refers to the proportion of unknown persons who are in a particular age group i, p_i^g refers to the same proportion in either the stateless group or migrant group.

Steps to calculate the similarity index are as follows:

1. For countries with both stateless and persons of unknown citizenship (group A), calculate the RMSE for stateless compared to unknowns and migrants compared to unknowns, and calculate the ratio of the two RMSEs

2. For countries with observations of persons of unknown citizenship only (group B), calculate the RMSE for migrants compared to unknowns, then calculate the ratio of this RMSE to the maximum RMSE from the countries in step 1.

3.3 Prevalence index

The logic of the second step is that the larger the proportion of known stateless persons of total migrants from that country, the more likely the persons of unknown citizenship are to be stateless.

For countries in Group A we calculate a prevalence index as the number of stateless persons from a particular country of origin divided by the total number of migrants from the same origin living in the country of interest.

For countries in Group B, we cannot calculate prevalence as there is no information on stateless populations.

3.4 Conversion to probability of being stateless

The final step is to convert the indexes into a probability of being stateless.

1. For countries in group A) we consider the product of the two indices. This product is converted to a probability by adding a fraction between 0 and 1 based on likely values from expert opinion.
2. For countries in group B), a likelihood index is calculated by considering the RMSE of unknowns versus the migrant population, divided through by the maximum equivalent RMSE in the group A countries. This is converted to a probability by multiplying by a fraction between 0 and 1 based on likely values from expert opinion.

3.5 Assumptions

There are a number of assumptions underlying this method. The main assumptions are:

- That there is no systematic reason that people have unknown citizenship status that also means they are more likely to be citizens rather than stateless
- That age is an important explanatory variable in distinguishing stateless versus non-stateless migrants
- That the likelihood that migrants from particular origin countries are stateless is well captured by expert opinion

4 Step-by-step example

The following section walks through the proposed method using data on stateless and other migrants in the Netherlands. The data specific to the Netherlands was provided by the Institute on Statelessness and Inclusion.

4.1 Data

The datasets used were as follows:

1. The number of stateless persons by age and origin country for 2019 and 2020 were provided by the Institute on Statelessness and Inclusion. The format of these files is

- every row is a different country of origin
- every column is an age group (0-14 years, 15-24, 25-34, 35-44, 45-54, 55-64, 65+)
- every cell contains the count of stateless persons in that group

2. The number of persons with unknown citizenship by age and origin country for 2019 and 2020 were provided by the Institute on Statelessness and Inclusion. The format of these files is

- every row is a different country of origin
- every column is an age group (0-14 years, 15-24, 25-34, 35-44, 45-54, 55-64, 65+)
- every cell contains the count of persons with unknown citizenship in that group

3. The number of persons currently living in the country of interest by origin country in 2020 were downloaded from UN DESA using this link: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/undesa_pd_2020_ims_stock_by_sex_destination_and_origin.xlsx. The format of these files is

- every row is a different country of origin
- every column is a year
- every cell contains the count of migrants

4. The number of in-migrants to the country of interest by age and origin country for 2020 were download from Eurostat using this link: https://ec.europa.eu/eurostat/databrowser/view/migr_imm5prv/default/table?lang=en. The format of these files is

- every row is a different country of origin, age group and sex category
- every column is a year
- every cell contains the count of in-migrants

4.2 Empirical analysis

All calculations were down in the statistical software package R. The analysis involves a moderate knowledge of statistics, with calculations of root-mean-squared differences, ratios, and probabilities. Each of the four files described above must be read into R, cleaned and manipulated to be in the right format, such the the information from each database can be merged and compared. Calculations are performed and the results are saved and outputted as a CSV file.

4.3 Step 1: Read in and clean each data file

First, the packages used for analysis are loaded in:

```
library(tidyverse)
library(readxl)
library(here)
library(kableExtra)
```

4.3.1 Data on statelessness and unknown citizenship from ISI

We read in the data on statelessness and unknown citizenship for 2020, split the two count types into two data frames, and rename the columns. The data is then changed to be in 'long' format, where every row represents a unique country of origin/ age combination, and there is a single column that has the count of persons. We then add a new columns called `type` (either "unknown" or "stateless") and then recombine the two data frames.

```
d20 <- read_xlsx(here("data/netherlands/Staatlozen_onbekenden_leeftijd_01januari2020.xlsx"), sh

d20_unknown <- d20[4:37,]
colnames(d20_unknown) <- c("origin",
                           "0-14",
                           "15-24",
                           "25-34",
                           "35-44",
                           "45-54",
                           "55-64",
                           "65+",
                           "Total")

d20_unknown <- d20_unknown %>%
```

```
  pivot_longer(-origin, names_to = "age") %>%
  mutate(type = "unknown")

d20_stateless <- d20[45:56,]

colnames(d20_stateless) <- c("origin",
                             "0-14",
                             "15-24",
                             "25-34",
                             "35-44",
                             "45-54",
                             "55-64",
                             "65+",
                             "Total")

d20_stateless <- d20_stateless %>%
  pivot_longer(-origin, names_to = "age") %>%
  mutate(type = "stateless")

d20 <- bind_rows(d20_stateless, d20_unknown)
```

A similar process is carried out for the year 2019:

```
## 2019

d19 <- read_xlsx(here("data/netherlands/Staatlozen_onbekenden_01januari2019.xlsx"), skip = 3)

d19_unknown <- d19[4:37,]
colnames(d19_unknown) <- c("origin",
                           "0-14",
                           "15-24",
                           "25-34",
                           "35-44",
                           "45-54",
                           "55-64",
                           "65+",
                           "Total")

d19_unknown <- d19_unknown %>%
```

```
  pivot_longer(-origin, names_to = "age") %>%
  mutate(type = "unknown")

d19_stateless <- d19[49:66,]

colnames(d19_stateless) <- c("origin",
                             "0-14",
                             "15-24",
                             "25-34",
                             "35-44",
                             "45-54",
                             "55-64",
                             "65+",
                             "Total")

d19_stateless <- d19_stateless %>%
  pivot_longer(-origin, names_to = "age") %>%
  mutate(type = "stateless")

d19 <- bind_rows(d19_stateless, d19_unknown)
```

We then combine the two years into the one data frame, and translate the country names from Dutch.

```
d <- bind_rows(d19 %>% mutate(year=2019), d20 %>% mutate(year=2020)) %>%
  mutate(value = as.numeric(value)) %>%
  mutate(age_group = case_when(
                       age=="0-14"~0,
                         age== "15-24"~15,
                        age==  "25-34"~25,
                         age== "35-44"~35,
                         age== "45-54"~45,
                         age== "55-64"~55,
                          age=="65+"~65,
                          age=="Total"~as.numeric(NA),
                       TRUE~as.numeric(NA)
  ))

country_translations <- tribble(
```

```
~country_dutch, ~country,
"Nederland", "Netherlands",
"Voormalig Sovjet-Unie", "Former Soviet Union",
"Voormalige Sovjet-Unie", "Former Soviet Union",
"Voormalig Joegoslavië", "Former Yugoslavia",
"Irak", "Iraq",
"Somalië", "Somalia",
"Afghanistan", "Afghanistan",
"Syrië", "Syrian Arab Republic",
"Iran", "Iran (Islamic Republic of)",
"China", "China*",
"Angola", "Angola",
"Sierra Leone", "Sierra Leone",
"Egypte", "Egypt",
"Ver. Arabische Emiraten", "United Arab Emirates",
"Indonesië", "Indonesia",
"Eritrea", "Eritrea",
"Soedan", "Sudan",
"Congo (DR)", "Democratic Republic of the Congo",
"Ethiopië", "Ethiopia",
"Guinee", "Guinea",
"Turkije", "Turkey",
"Sri Lanka", "Sri Lanka",
"Burundi", "Burundi",
"Myanmar", "Myanmar",
"Nigeria", "Nigeria",
"Uganda", "Uganda",
"Libanon", "Lebanon",
"Liberia", "Liberia",
"Ivoorkust", "Côte d'Ivoire",
"Libië", "Libya",
"Israël", "Israel",
"Duitsland", "Germany",
"Togo", "Togo",
"Pakistan", "Pakistan",
"Congo", "Congo",
"Saoedi-Arabië", "Saudi Arabia",
"Mongolië", "Mongolia",
"Koeweit", "Kuwait",
```

```
  "Polen", "Poland",
  "Thailand", "Thailand",
  "Rwanda", "Rwanda",
  "Overige landen", "Other countries",
  "Totaal", "Total"
)

d <- d %>%
  left_join(country_translations %>% rename(origin = country_dutch))
```

4.3.2 In-migrants

Secondly, we read in the data on in-migrant flows from Eurostat. The file is read in, and then only the age groups and countries of origin of interest are retained. The age groups in this dataset are in five year age groups (not ten years as is with the ISI data), so we sum up migrant counts by ten year age groups. The data on in-migrants are then combined with the data on stateless and unknown citizenship populations from above.

```
df <- read_tsv(here("data/netherlands/migr_imm5prv.tsv"))
colnames(df)[1] <- "code"



ages_of_interest <- c("_LT15" ,"_GE65",
                      "15-19", "20-24", "25-29",
                      "30-34", "35-39",
                      "40-44", "45-49",
                      "50-54", "55-59",
                      "60-64", "TOTAL")

df <- df %>%
  filter(str_ends(code, "T,NL")) %>%
  select(code:`2019`) %>%
  filter(str_detect(code, "REACH")) %>%
  filter(!str_detect(code, "15-64")
         &!str_detect(code, "10-14")
         &!str_detect(code, "5-9")
         &!str_detect(code, "65-69")
         &!str_detect(code, "70-74")
         &!str_detect(code, "75-79")
```

```
         &!str_detect(code, "80-84")
         &!str_detect(code, "85-89")
         &!str_detect(code, "90-94")
         &!str_detect(code, "GE100")
         &!str_detect(code, "GE85")
         &!str_detect(code, "LT5")) %>%
  separate(code, c("partner", "age_def", "age", "unit", "sex", "country"), "," ) %>%
  select(-unit)

partners_of_interest <- c("SY", "NL", "LY", "IL", "ID", "LB",
                          "AE", "SA", "EG", "IQ", "KW", "SU",
                          "ER", "AF", "TR", "IR", "PL", "NA",
                          "SU","SO", "YU", "CN", "AO","ET",
                          "SL", "SD", "CD", "GN", "LK", "BI",
                          "MM", "NG", "UG", "LR", "CI", "PK",
                          "DE", "CG", "TG", "MN", "TH")

de <- df %>%
  filter(partner %in% partners_of_interest)

de <- de %>%
  filter(age!="UNK", age!="TOTAL") %>%
  mutate(age_group = str_sub(age, 2, 3),
         age_group = as.numeric(ifelse(age_group=="_G", 65, ifelse(age_group=="_L", 0, age_grou
  mutate(age_group_broad = case_when(age_group<15~0,
                                     age_group<25~ 15,
                                     age_group<35~25,
                                     age_group<45~35,
                                     age_group<55~45,
                                     age_group<65~55,
                                     TRUE~65)) %>%
  select(partner, `2020`:age_group_broad) %>%
  rename(age_five = age_group,
         age_group = age_group_broad)

correspond <- bind_cols(origin = unique(d$origin), partner = partners_of_interest)

de <- de %>%
  left_join(correspond) %>%
```

```
  left_join(country_translations %>% rename(origin = country_dutch)) %>%
  select(partner, origin, country, age_five, age_group, `2020`:`2019`) %>%
  pivot_longer(-(partner:age_group), "year")

de <- de %>%
  mutate(value = as.numeric(value), year = as.numeric(year)) %>%
  mutate(type = "migrants")

de_collapsed <- de %>%
  group_by(partner, origin, country, age_group, year, type) %>%
  summarize(value = sum(value))

d <- d %>%
  left_join(correspond)

d_all <- de_collapsed %>%
  bind_rows(d %>% filter(age!="Total")) %>%
  select(-age)

d_all <- d_all %>%
  mutate(origin = ifelse(origin=="Voormalige Sovjet-Unie", "Voormalig Sovjet-Unie", origin))
```

4.3.3 Migrant stocks

The final dataset to read in is the total counts of migrant stocks by country of origin. Here there is minimal cleaning, just renaming column names.

```
du <- read_xlsx(here("data/netherlands/nl_stock.xlsx"))

du <- du %>%
  rename(country = `Region, development group, country or area of origin`) %>%
  right_join(country_translations) %>%
  select(country, country_dutch, `2020`) %>%
  rename(origin = country_dutch)
```

4.4 Step 2: Calculate similarities between age distributions for origin countries that have both stateless and unknown persons

Now we are interested in comparing the shape of the age distributions of migrants in total, stateless persons, and persons of unknown citizenship. First, find all the origin countries that have information on stateless persons, unknown persons and migrants:

```
all_three_sets <- d_all %>%
  group_by(country, year) %>%
  tally() %>%
  filter(year==2020, n ==21) %>%
  filter(country!="Netherlands", country!="Other countries") %>%
  select(country) %>% pull()
```

Then calculate the RMSEs for stateless compared to unknowns, and migrants compared to unknowns, and the ratio of the two RMSEs. The results are shown in Table 1. Countries with a higher ratio have a greater similarity between unknowns and stateless populations.

```
rmse_complete <-  d_all %>%
  group_by(origin, type, year) %>%
  mutate(prop = value/sum(value, na.rm = TRUE)) %>%
  mutate(complete = ifelse(country %in% all_three_sets, 1, 0)) %>%
  filter(year==2020, complete==TRUE) %>%
  select(country, age_group, type, prop) %>%
  pivot_wider(names_from = "type", values_from = "prop") %>%
  group_by(country) %>%
  drop_na() %>%
  summarize(stateless = sqrt(mean((stateless - unknown)^2)),
            migrants = sqrt(mean((migrants - unknown)^2))) %>%
  arrange(stateless) %>%
  mutate(ratio = migrants/stateless) %>%
  arrange(ratio)

rmse_complete %>%
kable(booktabs = TRUE, linesep="", caption = "RMSE for unknown age distribution compared to sta
```

Table 1: RMSE for unknown age distribution compared to stateless and migrants for countries with complete observations, 2020

country	stateless	migrants	ratio
Saudi Arabia	0.0809651	0.0606322	0.7488690
Israel	0.0680249	0.0578980	0.8511291
Syrian Arab Republic	0.0355195	0.0315014	0.8868774
Libya	0.1149196	0.1727553	1.5032711
Iraq	0.0666945	0.1137189	1.7050724
Lebanon	0.0446298	0.0819731	1.8367365

4.5 Step 3: Calculate similarities between age distributions for origin countries that have unknown persons only

For many countries we don't observe any stateless counts. We can still calculate the RMSE between unknowns and migrant populations. The magnitude of these RMSEs can be compared against those countries where we do have complete information to get an idea of the relative similarity.

```
has_unknowns <- d_all %>%
  filter(type== "unknown", year==2020) %>%
  group_by(country) %>%
  select(country) %>%
  slice(1) %>%
  pull()

rmse_unknown <- d_all %>%
  group_by(origin, type, year) %>%
  mutate(prop = value/sum(value, na.rm = TRUE)) %>%
  mutate(complete = ifelse(country %in% all_three_sets, 1, 0),
         has_unk = ifelse(country %in% has_unknowns, 1, 0)) %>%
  filter(year==2020, !complete==TRUE, has_unk==TRUE) %>%
  select(country, age_group, type, prop) %>%
  pivot_wider(names_from = "type", values_from = "prop") %>%
  group_by(country) %>%
  summarize(stateless = sqrt(mean((stateless - unknown)^2, na.rm = TRUE)),
            migrants = sqrt(mean((migrants - unknown)^2, na.rm = TRUE))) %>%
  arrange(stateless, -migrants) %>%
  filter(country!= "Other countries", country!="Former Yugoslavia", is.nan(stateless))  %>%
  select(-stateless)

rmse_unknown %>%
```

Table 2: RMSE for unknown age distribution compared to migrants for countries with no stateless observations

country	migrants
Ethiopia	0.2375831
Togo	0.2187182
Liberia	0.1862965
Sudan	0.1843691
China*	0.1809260
Guinea	0.1786588
Sierra Leone	0.1713322
Burundi	0.1668281
Mongolia	0.1583439
Sri Lanka	0.1337818
Democratic Republic of the Congo	0.1286550
Angola	0.1269378
Côte d'Ivoire	0.1192712
Pakistan	0.1179023
Congo	0.1159192
Myanmar	0.1151933
Eritrea	0.1138321
Uganda	0.1059840
Somalia	0.1034728
Turkey	0.0936685
Afghanistan	0.0903883
Nigeria	0.0813377
Iran (Islamic Republic of)	0.0801461
Germany	0.0712691

```
kable(booktabs = TRUE, linesep="", caption = "RMSE for unknown age distribution compared to mi
```

4.6 Step 4: Calculate the prevelance of stateless persons as compared to migrant stocks

The next step calculates the proportion of all migrants from a particular country of origin who are stateless for countries in group A, that is, countries where we have some counts of stateless persons.

```
stateless_prev <- d_all %>%
  filter(type!="migrants") %>%
  group_by(origin, year, type) %>%
  summarize(value = sum(value, na.rm = TRUE)) %>%
  pivot_wider(names_from = "type", values_from = "value") %>%
  left_join(du %>%  rename(stock = `2020`)) %>%
```

```
  mutate(prop_stateless = stateless/stock,
         prop_unknown = unknown/stock) %>%
  pivot_longer(prop_stateless:prop_unknown) %>%
  select(origin, country, year, name, value) %>%
  mutate(origin = as_factor(origin)) %>%
  filter(!is.na(value), name == "prop_stateless") %>%
  ungroup() %>%
  mutate(country = fct_reorder(country, value))
```

4.7 Step 5: Calculate an overall likelihood index and probability for countries with stateless observations

The next steps take the indexes calculated above and converts them into probabilities of persons with unknown citizenship being stateless. For countries in group A, we first multiply the similarity index and prevalence index together, to produce an overall likelihood index. This is then converted to a probability by adding 0.4. Ideally, this constant would be determined basic on expert opinion about the likely range of values.

```
sim <- d_all %>%
  group_by(origin, type, year) %>%
  mutate(prop = value/sum(value, na.rm = TRUE)) %>%
  mutate(complete = ifelse(country %in% all_three_sets, 1, 0)) %>%
  filter(year==2020, complete==TRUE) %>%
  select(country, age_group, type, prop) %>%
  pivot_wider(names_from = "type", values_from = "prop") %>%
  group_by(country) %>%
  drop_na() %>%
  summarize(stateless = sqrt(mean((stateless - unknown)^2)),
            migrants = sqrt(mean((migrants - unknown)^2))) %>%
  arrange(stateless) %>%
  mutate(ratio = migrants/stateless) %>%
  arrange(ratio)

prev <- d_all %>%
  filter(type!="migrants") %>%
  group_by(origin, year, type) %>%
  summarize(value = sum(value, na.rm = TRUE)) %>%
  pivot_wider(names_from = "type", values_from = "value") %>%
  left_join(du %>%  rename(stock = `2020`)) %>%
```

Table 3: Statelessness likelihood index and converted probabilities

country	similarity	prevalence	index	probability
Libya	1.5032711	0.0767167	0.1153261	0.5153261
Lebanon	1.8367365	0.0330938	0.0607847	0.4607847
Syrian Arab Republic	0.8868774	0.0583080	0.0517120	0.4517120
Israel	0.8511291	0.0349282	0.0297284	0.4297284
Saudi Arabia	0.7488690	0.0344311	0.0257844	0.4257844
Iraq	1.7050724	0.0014809	0.0025250	0.4025250

```
  mutate(prop_stateless = stateless/stock,
         prop_unknown = unknown/stock) %>%
  pivot_longer(prop_stateless:prop_unknown) %>%
  select(origin, country, year, name, value) %>%
  mutate(origin = as_factor(origin)) %>%
  filter(!is.na(value), name == "prop_stateless")


prob_known <- sim %>%
  left_join(prev %>% filter(name=="prop_stateless", year==2020)) %>%
  select(-origin, -name, -stateless, -migrants, -year) %>%
  rename(prevalence = value, similarity = ratio) %>%
  mutate(index = similarity*prevalence) %>%
  arrange(-index) %>%
  mutate(probability = index+0.4)

prob_known %>%
  kableExtra::kable(booktabs = TRUE, linesep="", caption = "Statelessness likelihood index and
```

4.8 Step 6: Calculate an overall likelihood index and probability for countries with unknown observations only

For the group B countries, we create a likelihood index by considering the RMSE of unknowns versus the migrant population, divided through by the maximum equivalent RMSE in the group A countries. This is converted to a probability by multiplying by 0.15. Again, ideally, this constant would be determined basic on expert opinion about the likely range of values.

```
max_mig <- sim %>%
  filter(migrants ==max(migrants)) %>%
```

```
  select(country, migrants)

prob_unk <- d_all %>%
  group_by(origin, type, year) %>%
  mutate(prop = value/sum(value, na.rm = TRUE)) %>%
  mutate(complete = ifelse(country %in% all_three_sets, 1, 0),
         has_unk = ifelse(country %in% has_unknowns, 1, 0)) %>%
  filter(year==2020, !complete==TRUE, has_unk==TRUE) %>%
  select(country, age_group, type, prop) %>%
  pivot_wider(names_from = "type", values_from = "prop") %>%
  group_by(country) %>%
  summarize(stateless = sqrt(mean((stateless - unknown)^2, na.rm = TRUE)),
            migrants = sqrt(mean((migrants - unknown)^2, na.rm = TRUE))) %>%
  arrange(stateless, -migrants) %>%
  filter(country!= "Other countries", country!="Former Yugoslavia", is.nan(stateless))  %>%
  select(-stateless) %>%
  mutate(similarity = migrants/max_mig$migrants) %>%
  select(-migrants) %>%
  arrange(-similarity) %>%
  mutate(probability = 0.15*similarity/max(similarity))

prob_unk %>%
  kable(booktabs = TRUE, linesep="", caption = "Statelessness likelihood index and converted pr
```

4.9 Step 7: Calculate the estimated number of persons with unknown citizenship who are stateless and add to existing populations

The estimated probabilities for countries in groups A and B are then combined. These are used to calculate the estimated number of persons of unknown citizenship. Standard errors are calculated based on assuming a binomial distribution.

```
probs <- prob_known %>%
  select(country, probability) %>%
  bind_rows(prob_unk %>% select(country, probability))

add_stats <- probs %>%
  left_join(d_all %>% filter(type == "unknown", year == 2020)) %>%
  mutate(add_stat = probability*value,
```

Table 4: Statelessness likelihood index and converted probabilities for countries with no observed stateless populations

country	similarity	probability
Ethiopia	1.3752577	0.1500000
Togo	1.2660577	0.1380895
Liberia	1.0783839	0.1176198
Sudan	1.0672271	0.1164030
China*	1.0472961	0.1142291
Guinea	1.0341725	0.1127977
Sierra Leone	0.9917625	0.1081720
Burundi	0.9656904	0.1053283
Mongolia	0.9165792	0.0999717
Sri Lanka	0.7744006	0.0844642
Democratic Republic of the Congo	0.7447236	0.0812274
Angola	0.7347838	0.0801432
Côte d'Ivoire	0.6904056	0.0753029
Pakistan	0.6824813	0.0744385
Congo	0.6710022	0.0731865
Myanmar	0.6668005	0.0727282
Eritrea	0.6589212	0.0718688
Uganda	0.6134918	0.0669138
Somalia	0.5989560	0.0653284
Turkey	0.5422035	0.0591384
Afghanistan	0.5232158	0.0570674
Nigeria	0.4708261	0.0513532
Iran (Islamic Republic of)	0.4639284	0.0506009
Germany	0.4125437	0.0449963

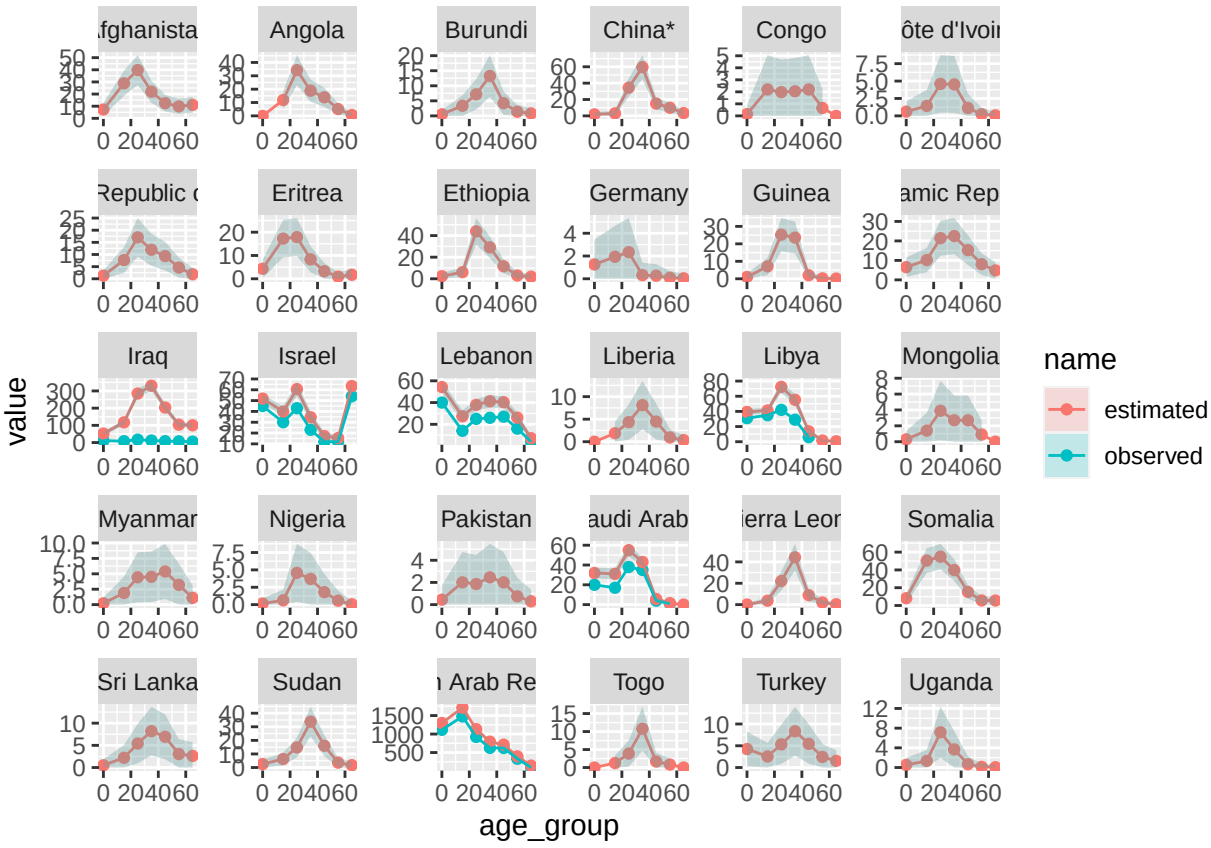
```
        se_stat = sqrt(probability*(1-probability)*value)) %>%
  select(country, age_group, add_stat, se_stat)
```

The additional estimated stateless persons are then added on to the observed values to come up with final estimates.

```
estimates <- add_stats %>%
  left_join(d_all %>% filter(type == "stateless", year == 2020)) %>%
  rowwise() %>%
  mutate(stat_final = sum(add_stat, value, na.rm = TRUE)) %>%
  select(country, age_group, value, stat_final, se_stat) %>%
  rename(observed = value, estimated = stat_final) %>%
  mutate(lower = max(0, estimated - 2*se_stat), upper = estimated + 2*se_stat)
```

Observed and updated estimates are plotted below:

```
estimates %>%
  pivot_longer(observed:estimated) %>%
  mutate(se_stat = ifelse(name=="observed", NA, se_stat)) %>%
  ggplot(aes(age_group, value)) +
  geom_point(aes(color = name)) +
  geom_line(aes(color = name)) + facet_wrap(~country, scales = "free")+
  geom_ribbon(aes(fill = name, ymin = lower, ymax = upper), alpha = 0.2)
```



We can also convert age-specific estimates to total estimates by country of origin.

```
estimates %>%
  group_by(country) %>%
  summarize(estimated = round(sum(estimated))) %>%
  kable(booktabs = TRUE, linesep="", caption = "Estimated stateless populations by origin coun
```

Table 5: Estimated stateless populations by origin country

country	estimated
Afghanistan	131
Angola	85
Burundi	31
China*	127
Congo	9
Côte d'Ivoire	13
Democratic Republic of the Congo	54
Eritrea	54
Ethiopia	98
Germany	6
Guinea	59
Iran (Islamic Republic of)	89
Iraq	1192
Israel	283
Lebanon	234
Liberia	20
Libya	224
Mongolia	12
Myanmar	21
Nigeria	11
Pakistan	10
Saudi Arabia	168
Sierra Leone	81
Somalia	180
Sri Lanka	29
Sudan	79
Syrian Arab Republic	6193
Togo	18
Turkey	30
Uganda	14