Statistik över namn efter födelseland 2020

Peter M. Dahlgren

2021-10-12

Sammanfattning

Detta dataset innehåller statistik över namn (tilltalsnamn på kvinnor, tilltalsnamn på män, samt efternamn) efter födelseland. Totalt är det 231 505 namn uppdelade på 202 länder. Datan kommer från SCB:s befolkningsstatistik/namnregister och avser personer folkbokförda i Sverige 31 december 2020. Vissa namn är dock exkluderade på grund av sekretess, såsom namn med färre än fem bärare. I detta dataset hittar du (förutom originaldatan från SCB) även bearbetningar där ISO-kod för varje land har lagts till samt data i så kallat *wide format* och *long format* för att underlätta vidare databehandling. Datan är licensierad med Creative Commons Attribution 4.0 International (CC BY 4.0) och får användas så länge SCB anges som källa.

Översikt

Titel:	Statistik över namn efter födelseland 2020
Engelsk titel:	Statistics on Swedish names by birth country 2020
Ansvarig forskare:	Peter M. Dahlgren, peter.dahlgren@jmg.gu.se (https://peterdahlgren.com/)
Institution:	Institutionen för journalistik, medier och kommunikation (JMG), Göteborgs universitet (https://www.gu.se/jmg)
Finansiär:	Institutet för mediestudier (https://mediestudier.se)
Ämne:	samhällsvetenskap, journalistik, medie- och kommunikationsvetenskap
Nyckelord:	namn, förnamn, tilltalsnamn, efternamn, födelseland, SCB
Geografi:	Sverige
Tidsperiod:	31 december 2020
Observationer:	231 505 namn
Dataformat:	XLSX (Excel), CSV
Datalicens:	Creative Commons Attribution 4.0 International (CC BY 4.0)
DOI:	https://doi.org/10.5878/s91g-y391

Delning och tillgängligörande av data

Licens för datan

Publikationer som använder datan

Dahlgren, P. M. (2021). Svenskar eller utrikesfödda i medierna? – att identifiera födelseland från namn. I: *L. Truedsson (red.) Vitt eller brett? – vilka får ta plats i medier och på redaktioner*. Stockholm: Institutet för mediestudier.

Rekommenderad citering av datan

Som källa ska *SCB* eller *Statistiska centralbyrån* anges. I internationella sammanhang anges *Statistics Sweden* i stället. Förutom källan rekommenderas även följande citering:

APA6:

> Dahlgren, P. M. (2021). Statistik över namn efter födelseland 2020. *Svensk nationell datatjänst*. doi:10.5878/s91g-y391

BibTeX:

```
@misc{dahlgren_namn_2021,
  title = {Statistik över namn efter födelseland 2020},
  url = {https://doi.org/10.5878/s91g-y391},
  abstract = {Detta dataset innehåller statistik över namn (tilltalsnamn
              på kvinnor, tilltalsnamn på män, samt efternamn) efter
              födelseland. Totalt är det 231 505 namn uppdelade på
              202 länder. Datan kommer från SCB:s
              befolkningsstatistik/namnregister och avser personer
              folkbokförda i Sverige 31 december 2020. Vissa namn är
              dock exkluderade på grund av sekretess, såsom namn med
              färre än fem bärare. I detta dataset hittar du (förutom
              originaldatan från SCB) även bearbetningar där ISO-kod
              för varje land har lagts till samt data i så kallat
              wide format och long format för att underlätta vidare
              databehandling. Datan är licensierad med Creative Commons
              Attribution 4.0 International (CC BY 4.0) och får användas
              så länge SCB anges som källa.},
  language = {Svenska},
  publisher = {Svensk nationell datatjänst},
  author = {Dahlgren, Peter M.},
  year = {2021}
}
```

Om namnen

Datan med namn och födelseland är beställd från SCB som beskriver datan på följande vis:

> Bearbetningen innebär framställning av ett Exceldokument innehållande statistik över namn efter födelseland 2020. Exceldokumentet består av tre flikar:
>
> - Tilltalsnamn kvinnor
> - Tilltalsnamn män
> - Efternamn
>
> Uppgifterna hämtas från befolkningsstatistikens namnregister 2020 och avser personer folkbokförda i Sverige 31 december 2020. Personer med skyddad identitet vid tidpunkten för framställningen av registret ingår inte i materialet.
>
> För att inte riskera att röja uppgifter som kan härledas till en enskild redovisas endast namn med minst fem bärare. I de fall ett namn har under fem bärare från ett enskilt födelseland så har uppgiften prickats (primärprickas). För att inte kunna härleda primärprickade uppgifter med hjälp av totalsummor så har även sekundärprickning gjorts i aktuella fall. Sekundärprickning har gjorts i de fall endast ett födelseland primärprickats samt i fall där endast enstaka individer från olika födelseländer primärprickats. Sekundärprickning har gjort av det födelseland med lägst antal personer. Prickade födelseländer (både primär- och sekundärprickade) har summerats och redovisas under ”Ospecificerade länder”. Prickade födelseländer redovisas med tecknet ”..”.
>
> Materialet levereras utan namnsammanslagningar, t.ex. av Clara och Klara eller Gustafsson och Gustavsson, vilket betyder att varje stavning av ett namn redovisas som ett unikt namn.
>
> På SCB:s hemsida https://www.scb.se/contentassets/5a07e6b5601f49ffbb1f31a14d0ad59f/namn-med-minst-tva-barare-31-december-2020.xlsx[1] redovisas namn och antal bärare. Vid jämförelse av uppgifterna på hemsidan med detta uppdrag så finns det små differenser av antal bärare för ett fåtal namn. Skillnaden beror på historiska skäl i uppbyggnaden av namnstatistikens produktionssystem. Skillnaderna är marginella och produktionssystemet kommer att justeras till nästa års statistikframställan.

Beställningen av data från SCB kostade 7 500 kronor exklusive moms.

[1] Denna fil finns även i detta dataset med namnet `namn-minst-tva-barare-2020.xlsx`.

Vidare bearbetningar

Från Excel-dokumentet som levererades av SCB har sedan ett antal bearbetningar gjorts med syfte att underlätta fortsatta analyser:

1. Maskinläsbara CSV-filer har skapats (med UTF-8 teckenkodning).
2. Datan tillhandahålls både i *wide format* och i *long format*.[2]
3. En variabel för landskod har lagts till, som exempelvis `SWE` för Sverige.[3] Landskoden är på tre bokstäver i formatet ISO 3166-1 alpha-3.[4]
4. Tomma celler (som indikerar noll) har ersatts med värdet `0`.
5. Prickade länder med värdet `..` (två punkter) har ersatts med värdet `1`. Detta värde är unikt (eftersom endast namn med 5 bärare eller mer är inkluderade). Det innebär att du enkelt kan använda sök-och-ersätt på alla celler om du vill ändra detta värde.

[2] https://en.wikipedia.org/wiki/Wide_and_narrow_data.

[3] Om du behöver en lista med landsnamn på svenska (`SWE` = Sverige, `DAN` = Danmark etc.) eller engelska (`SWE` = Sweden, `DAN` = Denmark etc.) kan du använda Svensk Text (https://github.com/peterdalle/svensktext) som också finns på SND (https://snd.gu.se/sv/catalogue/study/ext0278).

[4] https://en.wikipedia.org/wiki/ISO_3166-1_alpha-3.

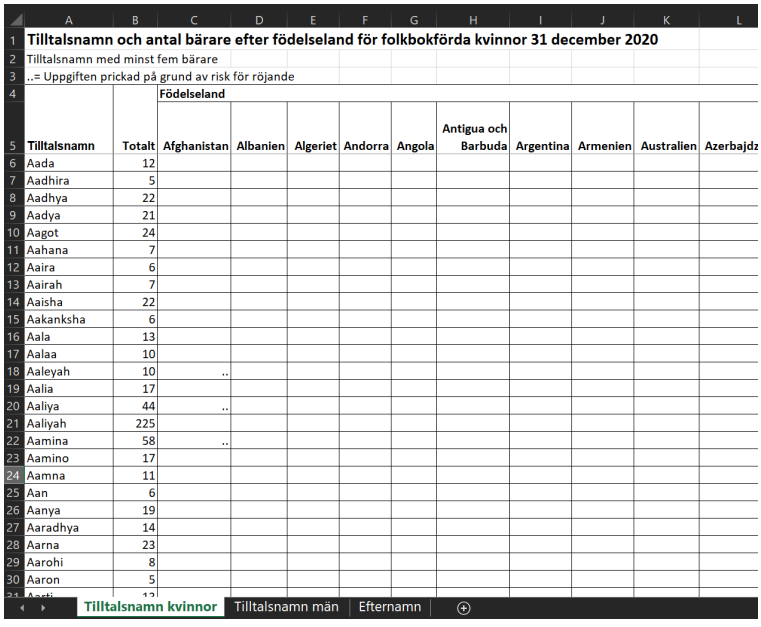
Excel-dokument från SCB

Fil	Filstorlek	Beskrivning	N
`namn-efter-fodelseland-2020.xlsx`	129 MB	Originalfil beställd från SCB med namn efter födseleland 31 december 2020.	231 505
`namn-minst-tva-barare-2020.xlsx`	12 MB	Namn med minst två bärare 31 december 2020 från SCB.[5]	649 289
`SCB villkor.pdf`	0,1 MB	SCB:s allmänna villkor för avtal och överenskommelser vad gäller användandet av datan.	

`namn-efter-fodelseland-2020.xlsx` är den obearbetade originalfilen levererad från SCB. Excel-dokumentet har tre flikar:

- Tilltalsnamn kvinnor (n = 24 787)
- Tilltalsnamn män (n = 21 414)
- Efternamn (n = 185 304)

Datan är i så kallat *wide format* med namn på raderna och länderna i kolumnerna (se figur 1). Totalt är det 202 länder plus kolumnen *Ospecificerade länder* längst till höger.



	A	B	C	D	E	F	G	H	I	J	K	L
1	**Tilltalsnamn och antal bärare efter födelseland för folkbokförda kvinnor 31 december 2020**											
2	Tilltalsnamn med minst fem bärare											
3	..= Uppgiften prickad på grund av risk för röjande											
4			**Födelseland**									
5	**Tilltalsnamn**	**Totalt**	**Afghanistan**	**Albanien**	**Algeriet**	**Andorra**	**Angola**	**Antigua och Barbuda**	**Argentina**	**Armenien**	**Australien**	**Azerbajdz**
6	Aada	12										
7	Aadhira	5										
8	Aadhya	22										
9	Aadya	21										
10	Aagot	24										
11	Aahana	7										
12	Aaira	6										
13	Aairah	7										
14	Aaisha	22										
15	Aakanksha	6										
16	Aala	13										
17	Aalaa	10										
18	Aaleyah	10	..									
19	Aalia	17										
20	Aaliya	44	..									
21	Aaliyah	225										
22	Aamina	58	..									
23	Aamino	17										
24	Aamna	11										
25	Aan	6										
26	Aanya	19										
27	Aaradhya	14										
28	Aarna	23										
29	Aarohi	8										
30	Aaron	5										

Tilltalsnamn kvinnor | Tilltalsnamn män | Efternamn

Figur 1: Utseende och struktur på datan i Excel-dokumentet `namn-efter-fodelseland-2020.xlsx`. Alla tre flikar har samma datastruktur. Ungefär 98% av cellerna är tomma. Celler med `..` (två punkter) är prickade av sekretesskäl. Läs mer under Om namnen.

[5] Filen är nedladdad från internet och nämndes i meddelandet från SCB. Filen inkluderas i detta dataset som en service om webbadressen skulle sluta fungera: https://www.scb.se/contentassets/5a07e6b5601f49ffbb1f31a14d0ad59f/namn-med-minst-tva-barare-31-december-2020.xlsx.

CSV-filer i wide format

Fil	Filstorlek	Beskrivning	N
`lastname_wide.csv`	75 MB	Efternamn	185 304
`men_wide.csv`	9 MB	Tilltalsnamn män	21 414
`women_wide.csv`	10 MB	Tilltalsnamn kvinnor	24 787

Originalfilen har bearbetats till CSV-filer i *wide format* som underlättar maskinläsning. Vid wide format har originalstrukturen bibehållits med namnen på raderna och länderna i kolumnen. Däremot har länderna ersatts med sina respektive ISO-koder för att undvika problemet med mellanslag och svenska tecken i kolumnnamnen.

CSV-filerna i wide format har följande kolumner:

Variabel	Datatyp	Beskrivning
`name`	Text	Namnet ifråga.
`total`	Heltal	Totala antalet bärare av namnet (oavsett födelseland).
`unspecified`	Heltal	Antal bärare av namnet där namnet har prickats av sekretesskäl.
`unknown`	Heltal	Antal bärare av namnet med okänt födelseland.
`afg` ... `tmp`	Heltal	202 stycken kolumner för varje födelseland, där värdet anger antalet bärare av namnet som är födda i landet ifråga. Prickade länder har värdet `1`.[6] Variabelnamnet har tre bokstäver i gemener och representerar landskoden i formatet ISO 3166-1 alpha-3.[7]

Nedan visas de tio första raderna i filen `lastname_wide.csv`:

```
"name","total","unspecified","unknown","afg","alb","dza","and","ag...
"A Anthony",15,15,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,...
"A Deild",13,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0...
"Aab",7,7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,...
"Aabakken",7,7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0...
"Aabel",7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,...
"Aaben",8,8,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,...
"Aaberg",24,6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,...
"Aabi",10,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,...
"Aabol",9,9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,...
```

Eftersom inte alla kolumner får plats har de ersatts med `...` i texten ovan.

Övriga filer som slutar på `_wide.csv` har samma struktur.

[6] Läs varför under Vidare bearbetningar.
[7] https://en.wikipedia.org/wiki/ISO_3166-1_alpha-3.

CSV-filer i long format

Fil	Filstorlek	Beskrivning	N
`lastname_long.csv`	1,2 GB	Efternamn	37 800 840
`men_long.csv`	135 MB	Tilltalsnamn män	4 325 560
`women_long.csv`	156 MB	Tilltalsnamn kvinnor	4 982 186

Originalfilen har bearbetats till CSV-filer i *long format* som underlättar maskinläsning. Vid long format återupprepas namnet på varje rad för respektive land och det totala antalet bärare för namnet ifråga. Dessutom finns en landskod i ISO-format med varje land för att enkelt kunna integreras i andra typer av analyser.

CSV-filerna i long format har följande kolumner:

Variabel	Datatyp	Beskrivning
`name`	Text	Namnet ifråga.
`total`	Heltal	Totala antalet bärare av namnet (oavsett födelseland).
`country`	Text	Födelselandet (på svenska). För okända födelseland används värdet `unknown` och för ospecificerade (prickade) födelseland används `unspecified`.
`persons`	Heltal	Antalet bärare av namnet födda i födelselandet i `country`. Prickade länder har värdet `1`.[8]
`isocode`	Text	Landskod i formatet ISO 3166-1 alpha-3 (tre bokstäver versaler).[9] När `country` är `unknown` eller `unspecified` så är `isocode` en tom sträng, det vill säga `""`.

Nedan visas de tio första raderna i filen `lastname_long.csv`:

```
"name","total","country","persons","isocode"
"A Anthony",15,"Afghanistan",0,"AFG"
"A Anthony",15,"Albanien",0,"ALB"
"A Anthony",15,"Algeriet",0,"DZA"
"A Anthony",15,"Andorra",0,"AND"
"A Anthony",15,"Angola",0,"AGO"
"A Anthony",15,"Antigua och Barbuda",0,"ATG"
"A Anthony",15,"Argentina",0,"ARG"
"A Anthony",15,"Armenien",0,"ARM"
"A Anthony",15,"Australien",0,"AUS"
```

Övriga filer som slutar på `_long.csv` har samma struktur.

[8] Läs varför under Vidare bearbetningar.
[9] https://en.wikipedia.org/wiki/ISO_3166-1_alpha-3.

R-kod för att snabbt komma igång

Koden nedan läser in filen `women_wide.csv`, gör om den till long format och tar sedan fram grundläggande statistik.

```r
library(tidyverse)

wide <- read.csv("women_wide.csv", fileEncoding="UTF-8")

# Convert to long format
long <- wide %>%
  pivot_longer(cols = c(-name), names_to="country", values_to="persons")

# Turn into factors to minimize memory footprint
long$name <- as.factor(long$name)
long$country <- as.factor(long$country)

# Most common names
long %>%
  filter(country == "total") %>%
  group_by(name) %>%
  summarize(n = sum(persons)) %>%
  arrange(desc(n))

# Most common names by birth country
long %>%
  filter(!country %in% c("unknown", "total", "unspecified")) %>%
  group_by(name, country) %>%
  summarize(n = sum(persons)) %>%
  arrange(desc(n))
```