Monday March 14th

UN POPULATION DATA DISAGGREGATION BY BROAD AGE GROUPS AND SEX

..

- Preliminary
- Brouillon-2020.xlsx
- End of 2020 conflict stock by age.xlsx
- Geoentities _FR_ES_AR_18052021.xlsx
- Population_Age_Sex_2020.xlsx
- README.md
- age_sex_population_example.ipynb
- gen_population_age_sex.py

Disaggregation of UN Population Data by broad age groups and sex. Example application to 2020 Conflict stock data.

Files:

- `Preliminary/` : folder containing files needed to obtain `Brouillon-2020.xlsx`.
- `Brouillon-2020.xlsx` : author-made file, contains (among other intermediary results) the result (before final processing) of the disaggregation. The process of how this file was obtained is documented here.
- `End of 2020 conflict stock by age.xlsx` : example Conflict stock data.
- `Geoentities _FR_ES_AR_18052021.xlsx` : reference file for geoentities.
- `Population_Age_Sex_2020.xlsx` : result of the disaggregation by broad age groups and sex.
- `age_sex_population_example.ipynb` : tutorial on how to apply the disaggregated data to the 2020 Conflict stock data.
- `gen_population_age_sex.py` : script to generate `Population_Age_Sex_2020.xlsx` from `Brouillon-2020.xlsx`.

Of interest for *you*

To apply the disaggregated population data to stock data : `age_sex_population_example.ipynb`.

Additional resources

The thought process behind the realization of this task can be found here.

PRELIMINARY - obtaining the data

Goal : percentage of population by broad age groups and sex (per 100 total population).

Note: the challenge of the disaggregation comes from the fact that "age" and "broad age groups" do not refer to the same age categories. We are interested in the following age categories for the GRID report: '0-4','0-17','5-14','15-17','15-24','25-64','65+','0+'.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Population by age and sex (thousands)													
2	ISO 3166-1 numeric code	Location	Time	Sex	Note	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44
3	900	World												
4	1828	Sustainable Development Goal (SDG) regions												
5	947	Sub-Saharan Africa	2020	Female		84 214	75 838	67 249	58 264	49 854	42 478	36 391	30 740	25 006
6	947	Sub-Saharan Africa	2020	Male		86 609	77 738	68 696	59 221	50 313	42 543	36 119	30 167	24 457
7	910	Eastern Africa	2020	Female		33 606	30 743	27 862	24 645	21 315	17 918	15 021	12 511	10 012
8	910	Eastern Africa	2020	Male		34 362	31 328	28 299	24 870	21 278	17 646	14 592	11 984	9 624
9	108	Burundi	2020	Female		1 016	903	752	608	524	498	442	335	227
10	108	Burundi	2020	Male		1 037	916	757	609	519	489	433	329	219
11	174	Comoros	2020	Female		61	56	50	44	39	35	31	27	21
12	174	Comoros	2020	Male		63	58	52	46	40	36	32	27	21
13	262	Djibouti	2020	Female		49	48	40	41	41	40	41	37	31
14	262	Djibouti	2020	Male		50	50	49	50	49	47	45	39	35
15	232	Eritrea	2020	Female		242	233	238	183	142	143	134	96	72
16	232	Eritrea	2020	Male		253	244	248	190	143	144	135	96	69
17	231	Ethiopia	2020	Female		8 271	7 526	6 854	6 433	5 816	4 802	3 758	3 183	2 488
18	231	Ethiopia	2020	Male		8 520	7 721	6 999	6 543	5 931	4 890	3 761	3 091	2 446

- **PopulationAgeSex.xlsx** - population by age and sex (thousands)

Step 1.0 : combine the columns of PopulationAgeSex into broad age groups (instead of years). For most broad age groups this process is straightforward, as the sum of non-overlapping and continuous age groups. The difficulty comes from the '15-17' column, for which we need to do some processing (= result of step 1.3), as the UN Data spans the '15-19' age category.

Step 1.1 : sum all columns of PopulationAgeSex to get the total population per country by sex (note: in this by-age disaggregation, age categories don't overlap so we can safely add them). Call the result TotalPopulationAgeSex.

	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI
1			Population by broad age groups and sex (thousands)							
2	100+	Total	0-4	0-17	5-14	15-17	15-24	25-64	65+	0+
3										
4										
5	1	548 210	84 214	263 483	143 087		108 118	194 371	18 420	548 210
6	0	546 155	86 609	269 635	146 434		109 534	189 114	14 464	546 155
7	1	224 366	33 606	107 468	58 605		45 960	78 780	7 415	224 366
8	0	221 042	34 362	109 241	59 627		46 148	75 270	5 635	221 042
9	0	5 990	1 016	3 048	1 655		1 132	2 030	157	5 990
10	0	5 901	1 037	3 088	1 673		1 128	1 937	126	5 901
11	0	430	61	194	106		83	166	14	430
12	0	437	63	201	110		86	166	12	437
13	0	467	49	161	88		82	225	23	467
14	0	520	50	179	99		99	249	23	520
15	0	1 769	242	830	471		325	641	90	1 769
16	0	1 777	253	866	492		333	629	70	1 777
17	0	57 447	8 271	26 557	14 380		12 249	20 351	2 196	57 447
18	0	57 517	8 520	27 209	14 720		12 474	19 933	1 870	57 517

	A	B	C	D	E	F	G	H	I	J	K
1	Percentage of female population by broad age group (per 100 female total population)										
2	ISO 3166-1 numeric code	Location	Time	Sex	Note	0-4	0-17	5-14	15-17	15-24	25-64
3	900	World	2020	Female		8.5	29.5	16.3	4.6	15.1	49.7
4	1828	Sustainable Development Goal (SDG) regions									
5	947	Sub-Saharan Africa	2020	Female		15.4	48.0	26.1	6.6	19.7	35.5
6	910	Eastern Africa	2020	Female		15.0	47.9	26.1	6.8	20.5	35.1
7	108	Burundi	2020	Female		17.0	50.9	27.6	6.3	18.9	33.9
8	174	Comoros	2020	Female		14.1	44.9	24.6	6.3	19.3	38.6
9	262	Djibouti	2020	Female		10.4	34.5	18.8	5.2	17.5	48.0
10	232	Eritrea	2020	Female		13.7	46.9	26.6	6.6	18.4	36.2
11	231	Ethiopia	2020	Female		14.4	46.2	25.0	6.8	21.3	35.4
12	404	Kenya	2020	Female		12.9	44.8	25.1	6.8	20.7	38.4
13	450	Madagascar	2020	Female		14.6	46.2	25.0	6.6	20.5	36.6
14	454	Malawi	2020	Female		14.9	49.0	27.1	7.0	20.8	34.0
15	480	Mauritius	2020	Female	1	4.9	20.5	11.4	4.3	14.4	55.3
16	175	Mayotte	2020	Female	2	12.8	44.4	25.0	6.6	19.4	38.4
17	508	Mozambique	2020	Female		16.0	49.6	26.8	6.9	20.3	33.5
18	638	Réunion	2020	Female	2	6.9	25.9	14.4	4.6	14.3	50.7

- **PercFemalePop.xlsx** - percentage of female population by broad age groups (per 100 female total population)
- **PercMalePop.xlsx** - percentage of male population by broad age groups (per 100 male total population)

Step 1.2 : for each country, multiply the Male (resp. Female) TotalPopulationAgeSex by the 15-17 PercMalePop (resp. PercFemalePop) column to get the male (resp. female) population in thousands in the 15-17 age group. Call the result TotalPopMale1517 (resp. TotalPopFemale1517).

	M	N
1	Female population (thousands)	
2	**Total**	**15-17**
3		
4		
5	548 210	36181.86
6	224 366	15256.888
7	5 990	377.37
8	430	27.09
9	467	24.284
10	1 769	116.754
11	57 447	3906.396
12	27 051	1839.468
13	13 877	915.882
14	9 696	678.72
15	643	27.649
16	139	9.174
17	16 068	1108.692
18	462	21.252

Step 2 : now that we have the population by broad age groups and sex (thousands) for all age groups, for each country we divide each age/sex-disaggregated value by the total population in the country (summing the TotalPopulationAgeSex for female and male), and multiply by 100 to get the result in percentage. We finally obtain the Percentage of the population by broad age groups and sex (per 100 total population).

	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ
1	Population by broad age groups and sex (thousands)								Percentage of the population by broad age groups and sex (per 100 total population)							
2	**0-4**	**0-17**	**5-14**	**15-17**	**15-24**	**25-64**	**65+**	**0+**	**0-4**	**0-17**	**5-14**	**15-17**	**15-24**	**25-64**	**65+**	**0+**
3																
4																
5	84 214	263 483	143 087	36181.86	108 118	194 371	18 420	548 210	7.6952388	24.076324	13.074888	3.3061967	9.8795192	17.761076	1.6831679	50.09389
6	86 609	269 635	146 434	36592.385	109 534	189 114	14 464	546 155	7.9140872	24.638524	13.380728	3.3437094	10.008909	17.280706	1.3216797	49.90611
7	33 606	107 468	58 605	15256.888	45 960	78 780	7 415	224 366	7.5449925	24.127965	13.157599	3.4253736	10.318629	17.687154	1.6647658	50.373141
8	34 362	109 241	59 627	15251.898	46 148	75 270	5 635	221 042	7.7147245	24.52603	13.387052	3.4242533	10.360838	16.899113	1.2651322	49.626859
9	1 016	3 048	1 655	377.37	1 132	2 030	157	5 990	8.5442772	25.635943	13.918089	3.1735767	9.5198049	17.071735	1.3203263	50.374233
10	1 037	3 088	1 673	377.664	1 128	1 937	126	5 901	8.7208813	25.966395	14.069464	3.1760491	9.486166	16.289631	1.0596249	49.625767
11	61	194	106	27.09	83	166	14	430	7.0357555	22.38639	12.226067	3.1245675	9.5732411	19.146482	1.6147636	49.596309
12	63	201	110	27.968	86	166	12	437	7.266436	23.1797	12.687428	3.2258362	9.9192618	19.146482	1.384083	50.403691
13	49	161	88	24.284	82	225	23	467	4.964539	16.340831	8.9159068	2.460385	8.3080041	22.796353	2.3302938	47.315096
14	50	179	99	29.64	99	249	23	520	5.0658561	18.099291	10.030395	3.0030395	10.030395	25.227964	2.3302938	52.684904
15	242	830	471	116.754	325	641	90	1 769	6.8245911	23.399718	13.282572	3.292555	9.1652566	18.076706	2.5380711	49.887197
16	253	866	492	120.836	333	629	70	1 777	7.1347998	24.417259	13.874788	3.4076706	9.3908629	17.738297	1.9740553	50.112803
17	8 271	26 557	14 380	3906.396	12 249	20 351	2 196	57 447	7.1944261	23.100619	12.508263	3.3979298	10.65464	17.702063	1.9101632	49.969556
18	8 520	27 209	14 720	3968.673	12 474	19 933	1 870	57 517	7.4110156	23.667124	12.804008	3.4521007	10.850353	17.338471	1.6265962	50.030444

Study of the percentage of the total population per broad year group

Available file: PercTotPop.xlsx

In this sanity check, I compare the UN-provided data on the percentage of the total population per broad year group with the same metric but calculated from the previously obtained data. The results do not show a statistically significant difference between the two sources of data.

For more details, see file Brouillon-2020.xlsx, sheet UNResult, columns M:AG.

	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Percentage of the total population by broad age groups (per 100 total population) - calculated							Percentage of the total population by broad age groups (per 100 total population) - UN						
2	**0-4**	**0-17**	**5-14**	**15-17**	**15-24**	**25-64**	**65+**	**0-4**	**0-17**	**5-14**	**15-17**	**15-24**	**25-64**	**65+**
3	14.571238	49.125515	27.254277	7.3	21.980167	33.55084	2.643477367	14.6	49.2	27.3	7.3	22	33.5	2.6
4								14.6	49.2	27.3	7.3	22	33.5	2.6
5	5.8006252	21.26662	11.462313	4.0036818	14.831539	53.178187	14.72733588	5.8	21.2	11.5	4	14.9	53.2	14.7
6								5.8	21.2	11.5	4	14.9	53.2	14.7
7	11.496271	34.833437	19.286643	4.0505234	13.478072	48.997697	6.741316792	11.5	34.9	19.3	4.1	13.5	49	6.7
8								11.5	34.9	19.3	4.1	13.5	49	6.7

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
1	Absolute difference							Outliers detection						
2	**0-4**	**0-17**	**5-14**	**15-17**	**15-24**	**25-64**	**65+**	**0-4**	**0-17**	**5-14**	**15-17**	**15-24**	**25-64**	**65+**
3	0.03	0.07	0.05	0	0.02	0.05	0.04	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4								FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	0	0.07	0.04	0	0.07	0.02	0.03	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6								FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	0	0.07	0.01	0.05	0.02	0	0.04	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8								FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

EXAMPLE - applying the data
Jupyter notebook

Example: applying the projection reference data to 2020 Conflict data.

Loading reference data

We first load the reference data, downloadable here.

In [4]:
```
final_df = pd.read_excel('Population_Age_Sex_2020.xlsx')
```

In [5]:
```
display(final_df)
```

	iso3	idmc_short_name	GRID_geographical_group	sex	0-4	0-17	5-14	15-17	15-24	25-64	65+	0+
0	AFG	Afghanistan	South Asia	Female	7.103221	23.969642	13.312439	3.553982	10.689513	16.164003	1.415506	48.684684
1	AFG	Afghanistan	South Asia	Male	7.468016	25.155873	13.941838	3.746018	11.290654	17.386837	1.227971	51.315316
2	AFG	Afghanistan	South Asia	Total	14.571238	49.125515	27.254277	7.300000	21.980167	33.550840	2.643477	100.000000
3	ALB	Albania	Europe and Central Asia	Female	2.778743	10.027579	5.383814	1.865023	7.085794	26.154915	7.676276	49.079542
4	ALB	Albania	Europe and Central Asia	Male	3.021883	11.239041	6.078499	2.138659	7.745745	27.023272	7.051059	50.920458
...	...	...	...	...	...	...	...	...	...	...	...	...
595	ZMB	Zambia	Sub-Saharan Africa	Female	7.925801	25.392760	13.882391	3.584567	10.547789	16.830768	1.300114	50.486863
596	ZMB	Zambia	Sub-Saharan Africa	Male	8.099875	25.775684	14.110863	3.564946	10.493391	15.982157	0.826851	49.513137
597	ZWE	Zimbabwe	Sub-Saharan Africa	Male	7.096731	24.477600	13.944572	3.436298	10.016144	15.585901	1.083008	47.726355
598	ZWE	Zimbabwe	Sub-Saharan Africa	Female	7.016010	24.264193	13.850397	3.397787	10.278488	19.198170	1.930580	52.273645
599	ZWE	Zimbabwe	Sub-Saharan Africa	Total	14.112740	48.741793	27.794968	6.834084	20.294632	34.784071	3.013588	100.000000

600 rows × 12 columns

Description: Percentage of the population by broad age groups and sex (per 100 total population).

Loading and pre-processing the example data

In this example, we focus on 2020 Conflict stock data, downloadable here.

Please note that the following code is based on this specific file. Other examples will need a different tailoring of the code. To apply this code to another file, you will also need to change the following variables: `example_filename`, `sheet_name`, `skiprows`, `countries_col`, and `stock_col`.

In [6]:
```
# name of the example file
example_filename = 'End of 2020 conflict stock by age.xlsx'
stock_df = pd.read_excel(example_filename, sheet_name='IDP Data')

# name of the column in the .xlsx file that contains the countries isocodes
countries_col = 'iso3'

# name of the column in the .xlsx file that contains the stock data
stock_col = 'All IDPs end 2020'
```

In [7]:
```
display(stock_df[['iso3', countries_col, stock_col]].head())
```

	iso3	iso3	All IDPs end 2020
0	AFG	AFG	3546858
1	AZE	AZE	735455
2	BGD	BGD	426763
3	BEN	BEN	3494
4	BIH	BIH	98574

Performing the projection

In [33]:
```
# Countries in the stock data
countries = [country.strip() for country in stock_df[countries_col].values]

# Stock values
stocks = stock_df[stock_col].values

# Mapping from country (idmc_short_name) to stock
mapping = dict(zip(countries, stocks))
```

In [34]:
```
def rounding(x):
        """

        Rounding function.

        The following are the rules that define all rounding carried out for IDMC's publications:
        - If the number is less than 100: report the number itself (i.e. 17 people)
        - If the number is between 100 and 999 - round to nearest 10 (i.e. 240 people)
        - If the number is between 1,000 and 9,999 - round to nearest 100 (i.e. 2,300 people )
        - If the number is over 10,000 - round to nearest 1,000 (i.e. 347,000 people)
        - If we want to report a figure rounding to the nearest million (mainly for aggregate figures) - round to the nearest 100,000 so that

        """
        if 0 <= x < 100 : return int(x)
        if 100 <= x < 1000 : return round(x / 10) * 10
        if 1000 <= x < 10000 : return round(x / 100) * 100
        if 10000 <= x < 1000000 : return round(x / 1000) * 1000
        if 1000000 <= x : return round(x / 100000) * 100000
```

Populating the dataframes that will respectively contain the stock disaggregated by broad age groups and sex (`proj_df`), and the rounded figures (`rounded_df`).

In [36]:
```
proj_df = final_df[final_df['idmc_short_name'].isin(countries)]
rounded_df = proj_df.copy()
for index, row in proj_df.iterrows():
        country = row['idmc_short_name']
        proj_df.loc[index, age_groups] = proj_df.loc[index, age_groups].multiply(mapping[country] / 100)
        rounded_df.loc[index, age_groups] = proj_df.loc[index, age_groups].apply(rounding)

rounded_df[age_groups] = rounded_df[age_groups].astype(int)
```

In [13]:
```
display(rounded_df)
```

	iso3	idmc_short_name	GRID_geographical_group	sex	0-4	0-17	5-14	15-17	15-24	25-64	65+	0+
0	AFG	Afghanistan	South Asia	Female	252000	850000	472000	126000	379000	573000	50000	1700000
1	AFG	Afghanistan	South Asia	Male	265000	892000	494000	133000	400000	617000	44000	1800000
2	AFG	Afghanistan	South Asia	Total	517000	1700000	967000	259000	780000	1200000	94000	3500000
30	AZE	Azerbaijan	Europe and Central Asia	Total	60000	200000	113000	28000	98000	415000	50000	735000
31	AZE	Azerbaijan	Europe and Central Asia	Female	28000	94000	53000	13000	46000	212000	29000	368000
...	...	...	...	...	...	...	...	...	...	...	...	...
559	UKR	Ukraine	Europe and Central Asia	Male	18000	71000	42000	11000	36000	203000	42000	340000
560	UKR	Ukraine	Europe and Central Asia	Female	17000	67000	40000	9900	34000	221000	83000	394000
591	YEM	Yemen	Middle East and North Africa	Female	246000	809000	446000	117000	368000	686000	58000	1800000
592	YEM	Yemen	Middle East and North Africa	Male	256000	841000	464000	121000	379000	684000	48000	1800000
593	YEM	Yemen	Middle East and North Africa	Total	502000	1600000	910000	238000	747000	1400000	107000	3600000

159 rows × 12 columns

Description: Rounded 2020 Conflict stock by broad age groups and sex.