Usability of age distributions in the DAS demonstration data

Jan Vink
Cornell Program on Applied Demographics
jkv3@cornell.edu

Applied Demography Conference
February 9, 2023

Assumptions & Postulations

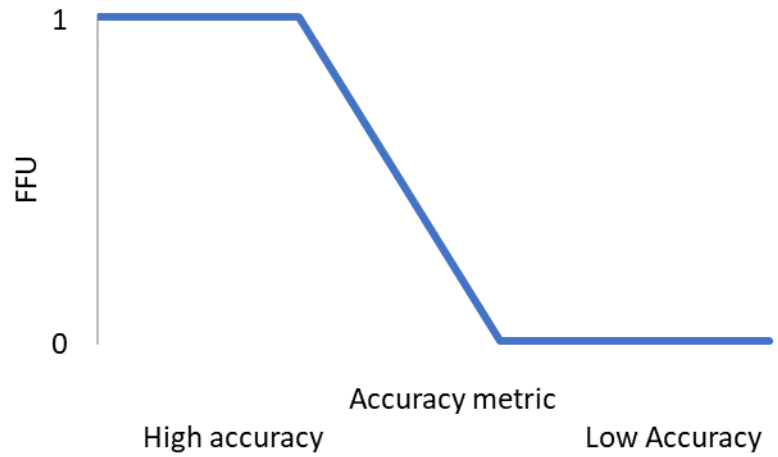
- We can define functions that transform accuracy metrics into Fitness-for-use scores, e.g.

$$\text{FFU} = \begin{cases} 1 \; if \; Abs \; Perc \; Error < 10\% \\ 0 \; if \; Abs \; Perc \; Error \geq 10\% \end{cases}$$

- Fitness-for-use is better scored at each geographic area and aggregated independent from accuracy metrics
- At the local level, fitness-for-use scores can be combined, allowing for scoring multiple accuracy metrics
- Fuzzy membership functions transform multi-valued accuracy metrics into fitness-for-use scores in interval [0, 1]
- Age is measured on an interval scale and usability is impacted differently whether biases are clustered or spread out

Fuzzy membership

- Transforms value X into degree of membership: μ_A: X $\rightarrow$ [0, 1]
- In this paper:
  - Highly accurate $\rightarrow$ FFU = 1
  - Highly inaccurate $\rightarrow$ FFU = 0
  - Linear decreasing membership in between



Aggregating Fuzzy membership

- Often used Fuzzy "and" operators

FFU1	0.2	0.5	0.8	0.2
FFU2	**0.2**	**0.5**	**0.8**	**0.8**
Min	0.2	0.5	0.8	0.2
Product	0.04	0.25	0.64	0.16
Hamacher product	0.11	0.33	0.67	0.19

- Hamacher product: $\frac{FFU1 * FFU2}{FFU1 + FFU2 - FFU1 * FFU2}$

Data used

- August 2022 demonstration data state summary files (New York)
- 2010 SF1 state summary files (New York)
- Different Geographic Summary levels, places split by using PLACECC
  - Unincorporated places farther from “Optimized Spine”
- 3*24=72 age distributions per geography
  - Sex: Total, Male, Female
  - Table PCT12: age by sex is repeated in iterations A through O
    - Total
    - A-G: 6 race alone + TOMR
    - H: Hispanic
    - I-O: Non Hispanic race alone + TOMR
    - All Non Hispanic and Hispanic race alone + TOMR added by subtraction
- Total group size in SF1 is used to create population size bins (>= 500)

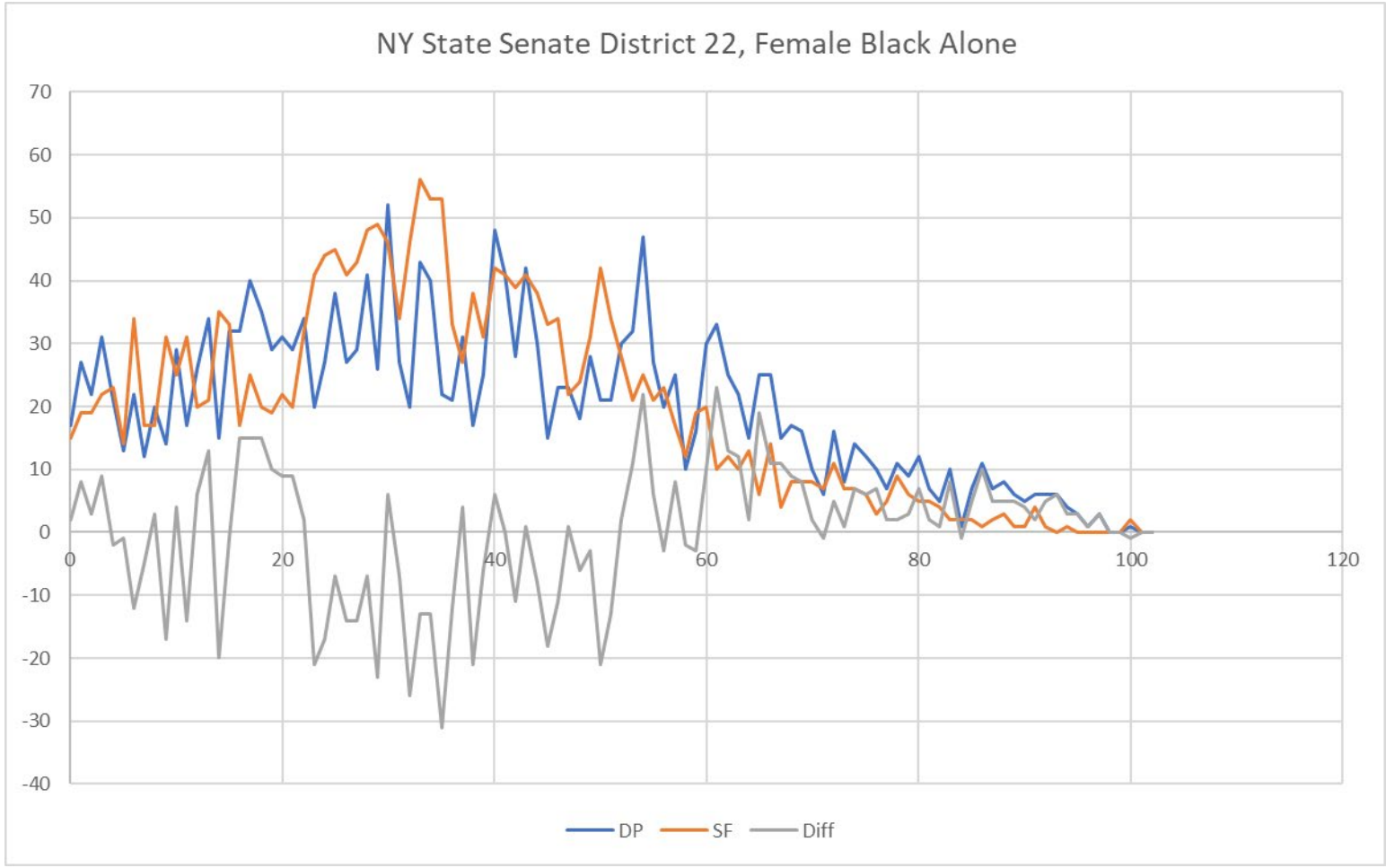
Example Age distribution

Example of DP (August 2022) and SF1 age distributions for the Female Black alone population in NY State Senate District 22

In grey is the difference between those two

Total SF1 population is 2,056

Total DP population is 2,068



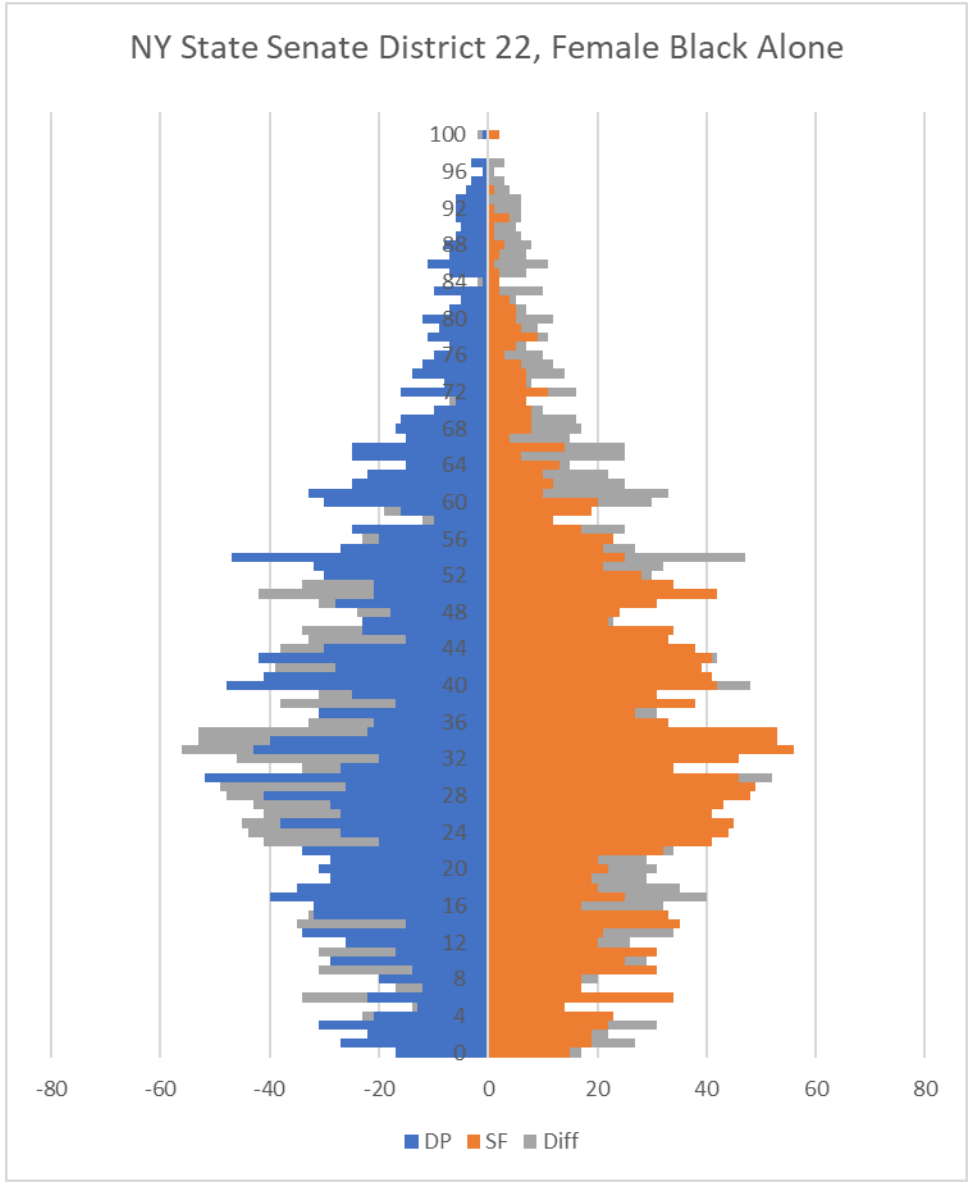
Metric 1: Signal Noise

Grey area makes the 'pyramid' symmetrical and is the absolute error for each age.

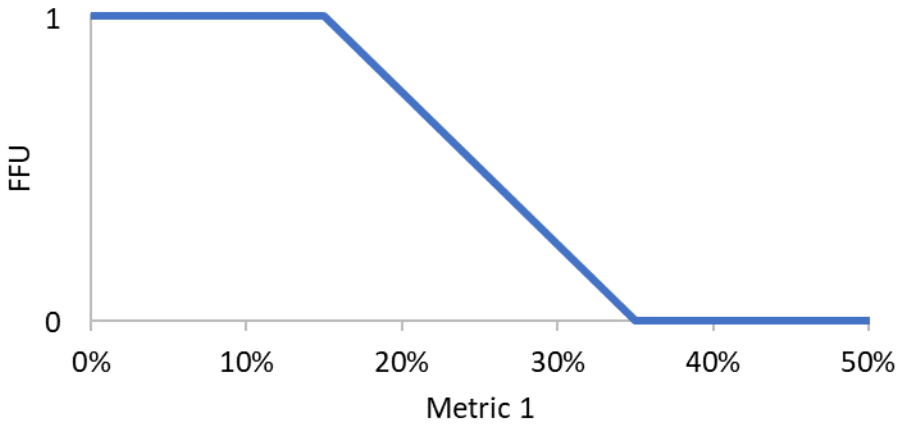
$$Metric_1 = \frac{Total\ Absolute\ Error}{SF1\ population}$$

Metric value in this example =

824 / 2056 = 40%



Metric 1: Signal Noise



	County	MCD	Place inc	Place uninc	Tract	School District	Legislative	ZCTA
500-999	0.004	0.209	0.194	0.050	0.127	0.071	0.018	0.048
1,000-1,999	0.228	0.681	0.676	0.387	0.609	0.491	0.289	0.383
2,000-4,999	0.765	0.968	0.965	0.883	0.941	0.929	0.790	0.843
5,000-10,000	0.995	1.000	1.000	1.000	1.000	1.000	0.994	0.998
10,000+	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Metric 2:
Moving Totals with big errors

Step 1: calculate 5-year moving total

Step 2: Flag ages where:

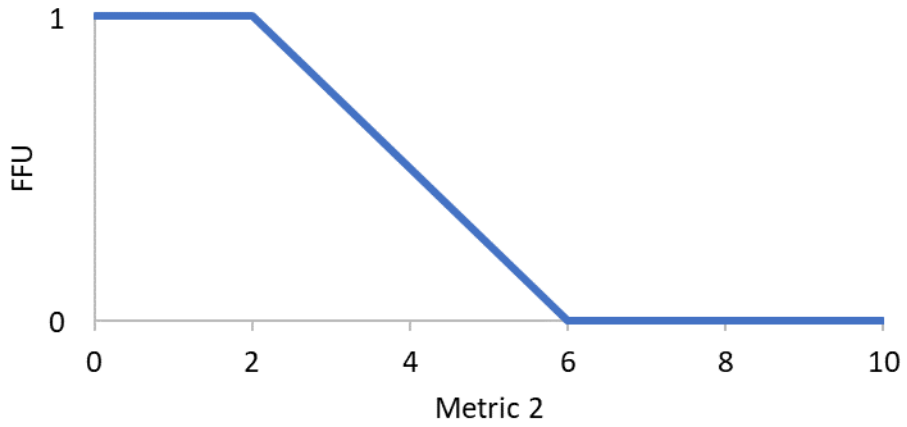
- Absolute difference >= 50 and
- Absolute % difference >= 10%

$Metric_2$ = number of ages with big errors in 5 yr moving totals

NY State Senate District 22, Female Black Alone

	5 yr Moving Total			
Age midpo	Demo	SF1	Abs Diff	Abs % Diff
17	168	114	54	47%
18	167	103	64	62%
19	164	106	58	55%
24	146	203	57	28%
25	141	214	73	34%
26	162	221	59	27%
27	161	226	65	29%
28	175	227	52	23%
30	166	223	57	26%
31	168	231	63	27%
32	182	235	53	23%
33	152	242	90	37%
34	146	241	95	39%
35	157	222	65	29%
36	131	204	73	36%
37	116	182	66	36%
61	126	71	55	77%
62	125	65	60	92%
63	120	51	69	135%
64	112	55	57	104%
65	102	47	55	117%
66	97	45	52	116%
67	98	40	58	145%
N=23				

Metric 2: Big errors in 5yr Moving Average



	County	MCD	Place inc	Place uninc	Tract	School District	Legislative	ZCTA
500-999	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000
1,000-1,999	0.996	1.000	1.000	0.997	1.000	1.000	0.991	0.996
2,000-4,999	0.950	0.999	0.998	0.989	1.000	0.995	0.893	0.953
5,000-10,000	0.828	0.992	0.994	0.973	0.999	0.987	0.706	0.863
10,000+	0.985	0.996	0.993	0.998	1.000	0.993	0.921	0.952

Metric 3: large age group with large % error

Steps:

1. Create cumulative populations over age and cumulative difference.
2. Find ages where cumulative difference is minimal (a1) and maximal (a2)
3. Range of cumulative difference is max – min
4. SF1 population for ages a1 < age <= age2 can be found in cumulative pop chart
   NOTE: SF1 = DP + Range

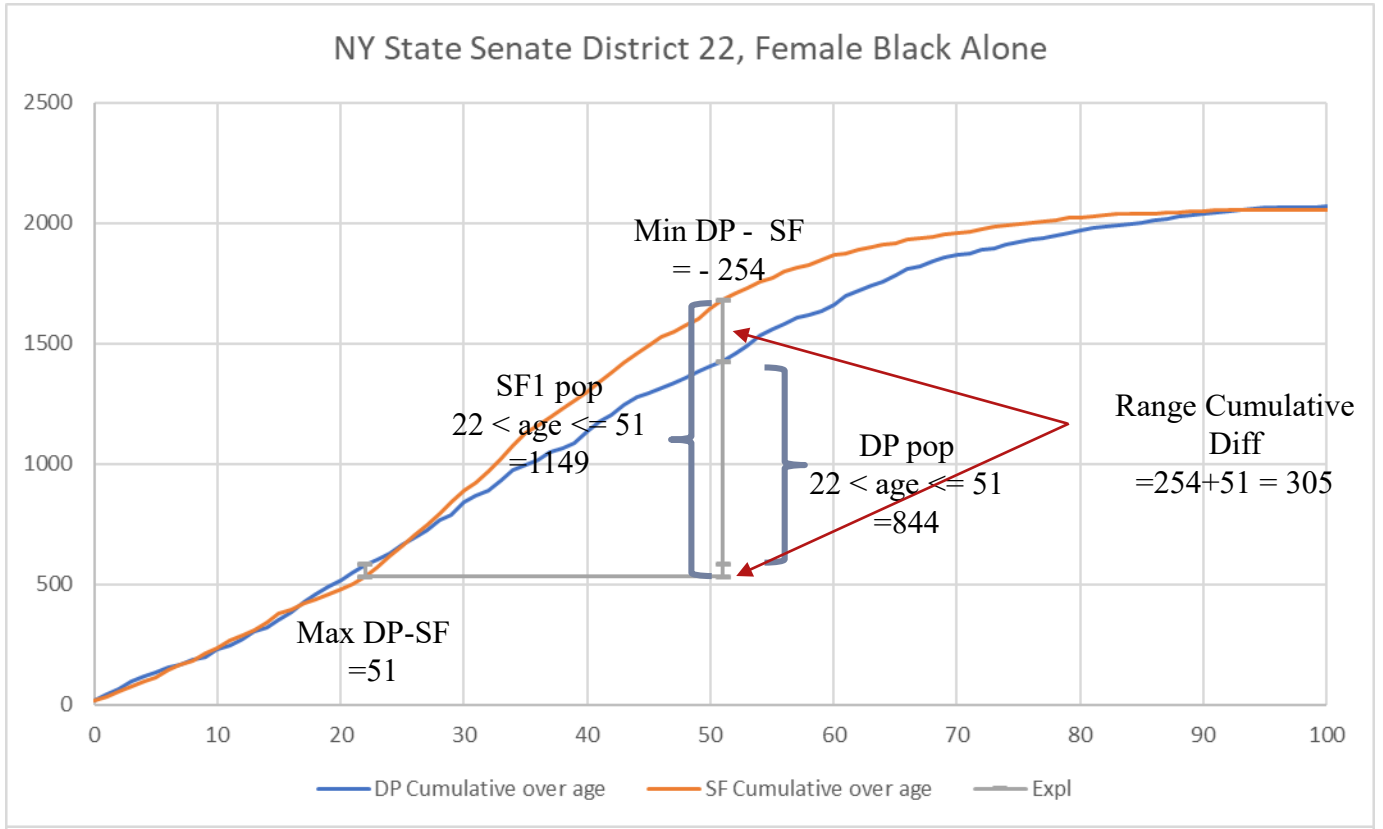
$$Metric_3 = \frac{Range}{SF1_{a1+1,a2}}$$

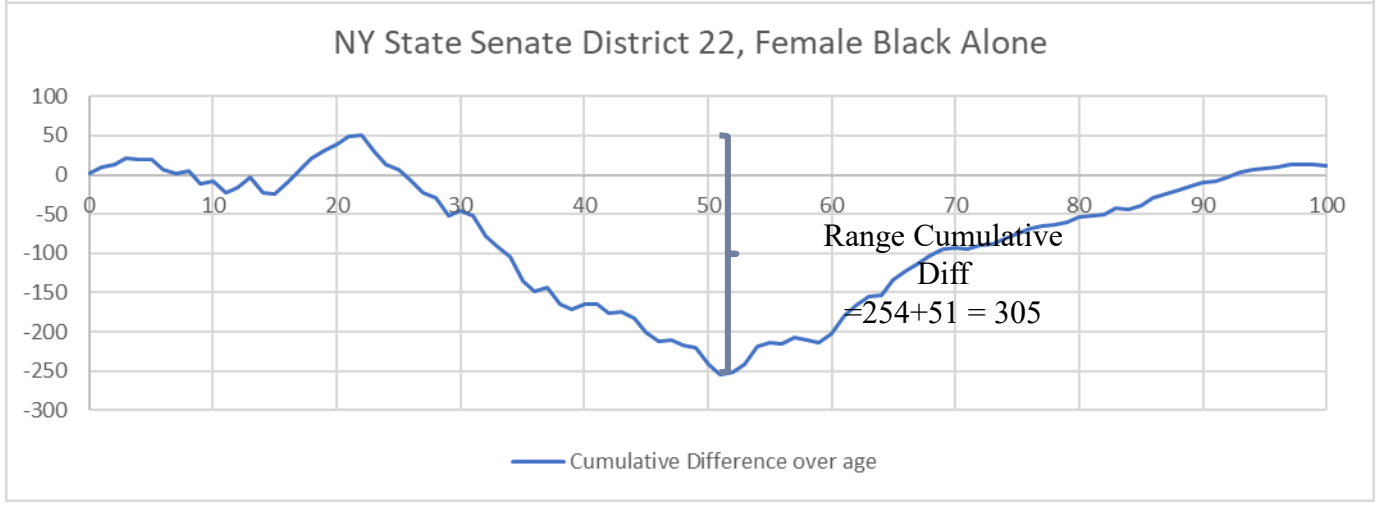
$Metric_3$ is **Absolute Percent Error** for this age range

Metric value in this example =

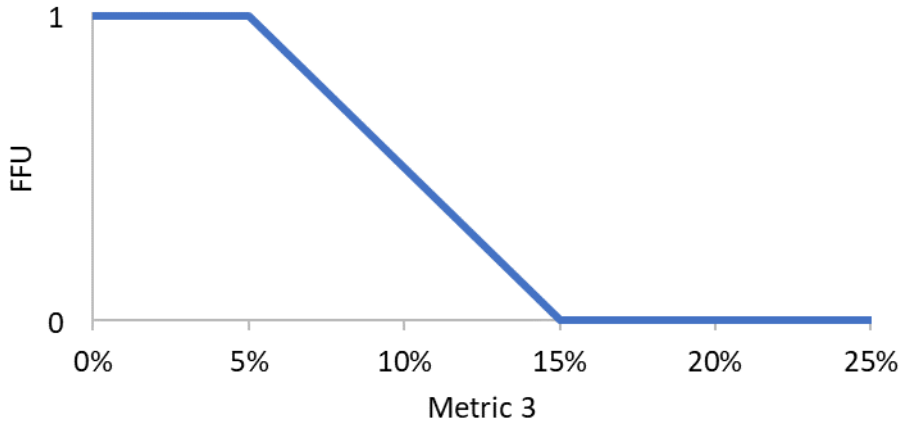
305 / 1149 = 27%

$Metric_3$ is only calculated if age2 – age1 >= 10 and $SF1_{a1+1,a2}$ >= 10% of $SF1_{total}$ to eliminate results from noise and low numbers





Metric 3: % error of large age group with large % error



	County	MCD	Place inc	Place uninc	Tract	School District	Legislative	ZCTA
500-999	0.293	0.615	0.624	0.336	0.538	0.410	0.272	0.344
1,000-1,999	0.533	0.833	0.834	0.608	0.815	0.723	0.518	0.581
2,000-4,999	0.834	0.954	0.953	0.877	0.944	0.924	0.784	0.832
5,000-10,000	0.953	0.992	0.995	0.989	0.990	0.987	0.929	0.960
10,000+	1.000	0.999	0.999	0.999	1.000	0.999	0.994	0.994

Metric 4: Absolute difference in median age

Steps:

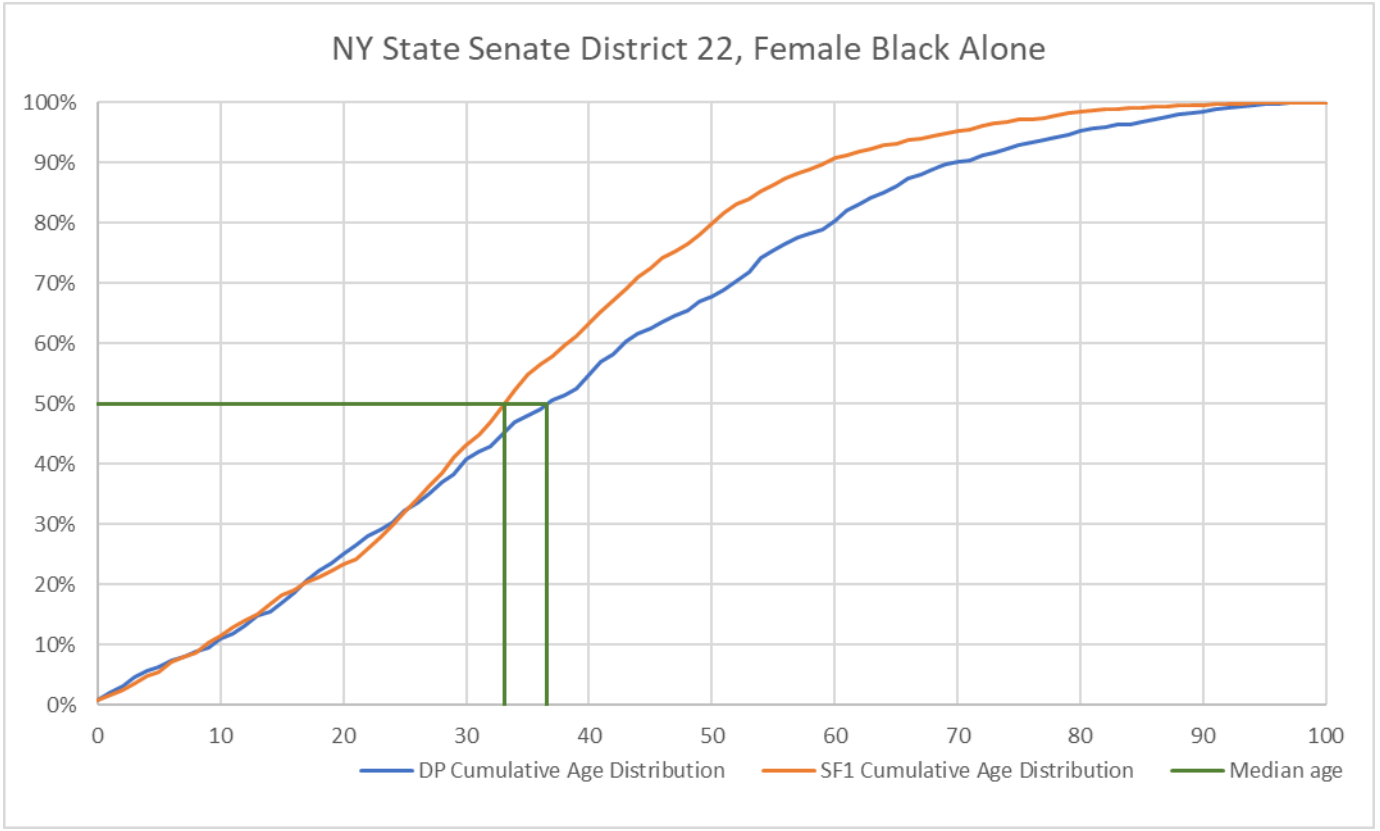
1. Create cumulative age distributions
2. Determine at what age 50% is less than age x (median age)

$$Metric_4 = |Median\ Age_{DP} - Median\ Age_{SF1}|$$

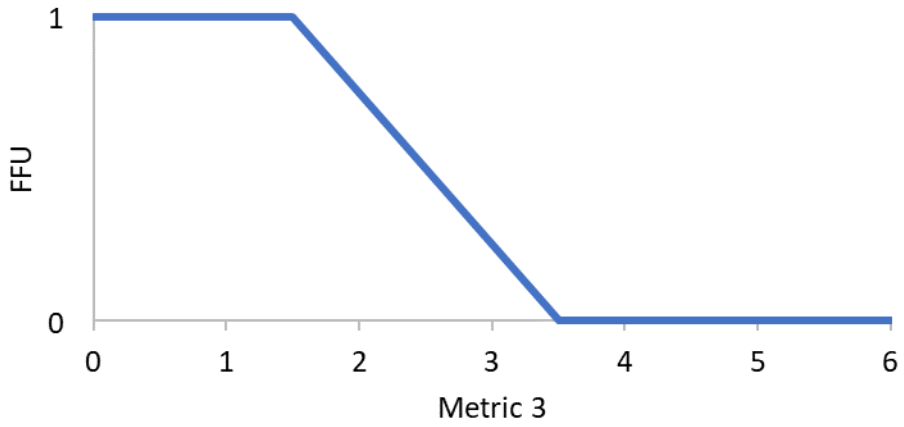
Metric$_4$ is **Absolute Difference in median age**

Metric value in this example =

$$37.6 - 34.1 = 3.5$$



Metric 4: Difference in median age



	County	MCD	Place inc	Place uninc	Tract	School District	Legislative	ZCTA
500-999	0.903	0.975	0.973	0.848	0.967	0.918	0.844	0.867
1,000-1,999	0.959	0.997	0.996	0.954	0.997	0.987	0.942	0.960
2,000-4,999	0.995	1.000	0.999	0.997	1.000	0.999	0.989	0.994
5,000-10,000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000
10,000+	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Aggregated Fitness-For-Use score

Average Fitness-For-Use score

	County	MCD	Place inc	Place uninc	Tract	School District	Legislative	ZCTA
500-999	0.003	0.176	0.164	0.038	0.102	0.051	0.011	0.037
1,000-1,999	0.166	0.604	0.601	0.306	0.534	0.404	0.197	0.284
2,000-4,999	0.648	0.929	0.926	0.795	0.895	0.866	0.604	0.705
5,000-10,000	0.796	0.985	0.989	0.963	0.989	0.975	0.678	0.842
10,000+	0.985	0.996	0.993	0.997	1.000	0.993	0.919	0.949

% of age distributions with Fitness-For-Use score >= 0.5

	County	MCD	Place inc	Place uninc	Tract	School District	Legislative	ZCTA
500-999	0%	10%	9%	1%	3%	1%	0%	2%
1,000-1,999	7%	68%	68%	25%	57%	37%	8%	20%
2,000-4,999	73%	97%	97%	86%	96%	93%	66%	78%
5,000-10,000	82%	99%	99%	97%	100%	99%	71%	87%
10,000+	99%	100%	100%	100%	100%	100%	93%	96%

Fitness-For-Use score (Aug 2022 vs Oct 2019)

Average Fitness-For-Use score August 2022

	County	MCD	Place inc	Place uninc	Tract	School District	Legislative	ZCTA
500-999	0.003	0.176	0.164	0.038	0.102	0.051	0.011	0.037
1,000-1,999	0.166	0.604	0.601	0.306	0.534	0.404	0.197	0.284
2,000-4,999	0.648	0.929	0.926	0.795	0.895	0.866	0.604	0.705
5,000-10,000	0.796	0.985	0.989	0.963	0.989	0.975	0.678	0.842
10,000+	0.985	0.996	0.993	0.997	1.000	0.993	0.919	0.949

Average Fitness-For-Use score October 2019

	County	MCD	Place inc	Place uninc	Tract	School District	Legislative	ZCTA
500-999		0.000	0.000	0.002	0.001	0.000		0.000
1,000-1,999		0.000	0.000	0.003	0.002	0.000		0.006
2,000-4,999	0.000	0.000	0.000	0.006	0.001	0.000		0.005
5,000-10,000	0.000	0.000	0.000	0.004	0.000	0.000		0.001
10,000+	0.211	0.033	0.013	0.000	0.000	0.012	0.049	0.000

Wrap-up

- Fuzzy membership functions are useful to score fitness-for-use
- Fitness-for-use scores allow for considering multiple accuracy metrics at the record level
- There is room to develop more accuracy metrics for variables measured on an interval or ratio scale
- Similar frameworks can be used in estimates evaluation


CORNELL UNIVERSITY
FOUNDED A.D. 1865