Locating the entrance to the garden of forking paths

Thinking about the wider context of robust data-driven (social) science

Rachel Franklin

Newcastle University | The Alan Turing Institute

The big TMCF picture from my biased perspective

Two main thoughts (+ an addendum, for a treat)

1. **Humans** all the way down

How research domains and questions are framed

- What does it mean to be *poor*?

- How do we define sustainability or social mobility?

- What makes a migrant?

Data, data, data

Take migration, for example

- Conceptualising migration with all sorts of temporal and spatial assumptions

- How we put numbers to concepts

- Big implications for how we visualise migration flows

- Big implications for policy-relevant analysis

Gets at important *human* linkages between concept and data

Still not at the entrance to the garden of forking paths

What about data generation? Humans are (of course) there too

Age *heaping*

Age-Sex Pyramid for the United States
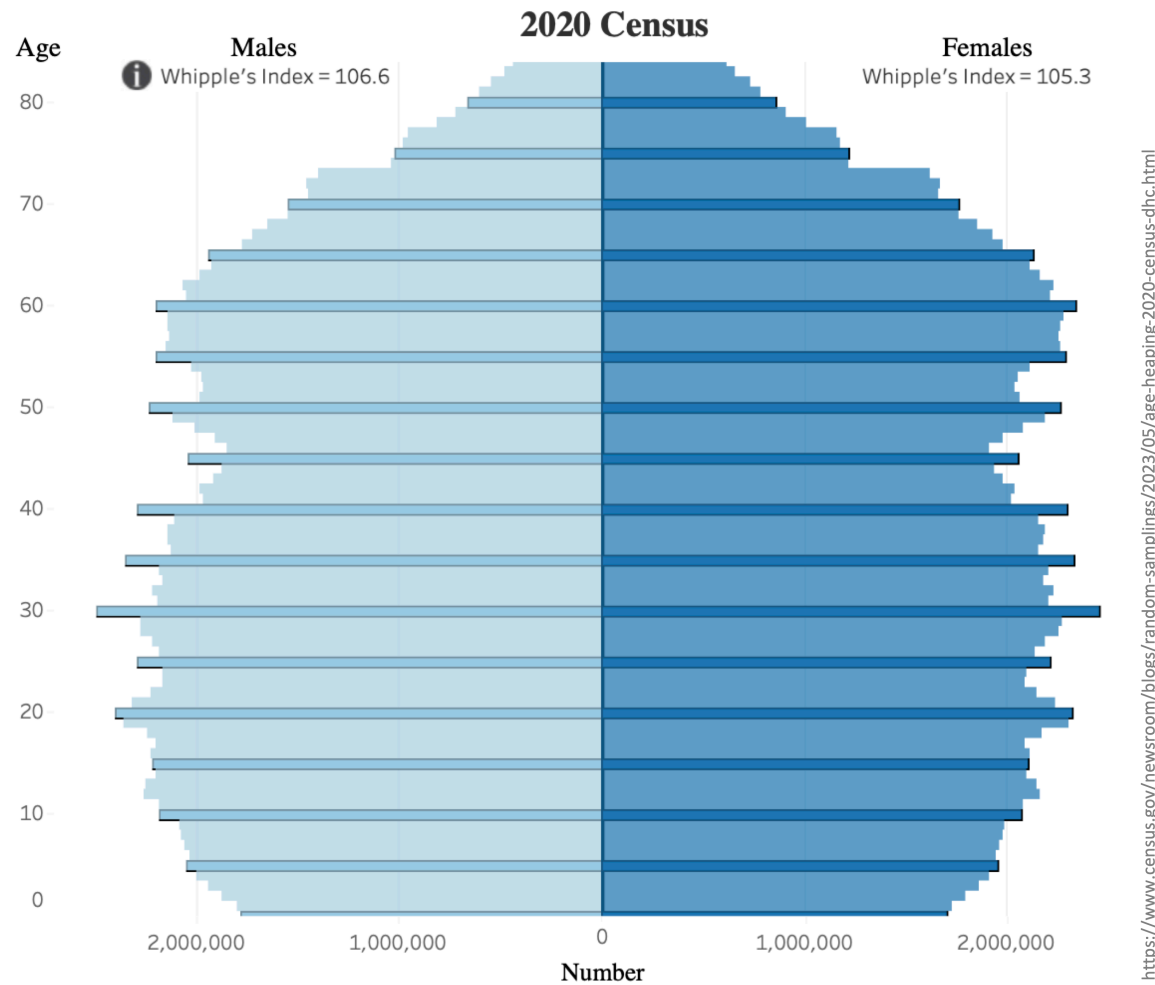(Aged 0 to 84. Outlined bars represent ages ending in "0" or "5")

2020 Census

Press play to watch the pyramid change

2020 Census

Age

Males
Whipple's Index = 106.6

Females
Whipple's Index = 105.3

80

70

60

50

40

30

20

10

0

2,000,000   1,000,000   0   1,000,000   2,000,000

Number

https://www.census.gov/newsroom/blogs/random-samplings/2023/05/age-heaping-2020-census-dhc.html

United States® Census Bureau
U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

Source: 1980 Census Summary File 2C (SF2C), 1990 Census Summary File 2C (SF2C), Census 2000 Summary File 1 (SF1), 2010 Census Summary File 1 (SF1), 2020 Census Demographic and Housing Characteristics File (DHC)

Wrangling, cleaning, choosing models, interpreting results…

Essentially: it's *all* human in the loop

2. Research robustness whack-a-mole

- Theory

- Models and metrics

- Peer review

- "Route maps"

- Pre-registration

- Open science and reproducibility

Tools that render explicit inference *legible* are good!

Bonus addendum: where does this leave us?

Attending to risks

- Hard-wiring some forking paths into commonly used tools or packages

- Aiding and abetting conflation of analytical complexity with quality

Lots of opportunites too though

- Analytical and pedagogical

- Narrative, procedural, and tool-based

Interactive data analysis support

- Toolkits

- But also wider embrace of a cross-validation (and counter-factual) mindset

- And (social science) theory!

- And replication

Big TMCF picture [reprise]

- Knowledge production depends on repetition
  - So tools should embed openness and flexibility alongside inference assistance
- Thinking about meta-mechanisms for better research
- Beware the quick fix

Thanks!!