Creating a Synthetic version of the Labour Force Survey

LFS & APS user conference 27th November 2019

Iain Dove, Office for National Statistics iain.dove@ons.gov.uk

Defining synthetic data

- Synthetic data can be a very broad term

Definition from the US Census Bureau:

"Synthetic data are microdata records created to improve data utility while preventing disclosure of confidential respondent information. Synthetic data is created by statistically modelling original data and then using those models to generate new data values that reproduce the original data's statistical properties. Users are unable to identify the information of the entities that provided the original data."

- In this case we mean record-level data that is not real, that has been produced from statistical models

Why synthesize data?

LFS levels of access

1. Open/teaching data - most basic version ~15 basic variables

Very low disclosure risk, available to anyone

---

2. Safeguarded (EUL) - more detailed version, ~700 variables

Some disclosure risk so only available to researchers under agreed conditions

---

3. Secure version - full detail, ~1100 variables

Potentially very disclosive so only accessible within a secure site (Secure Research Service)

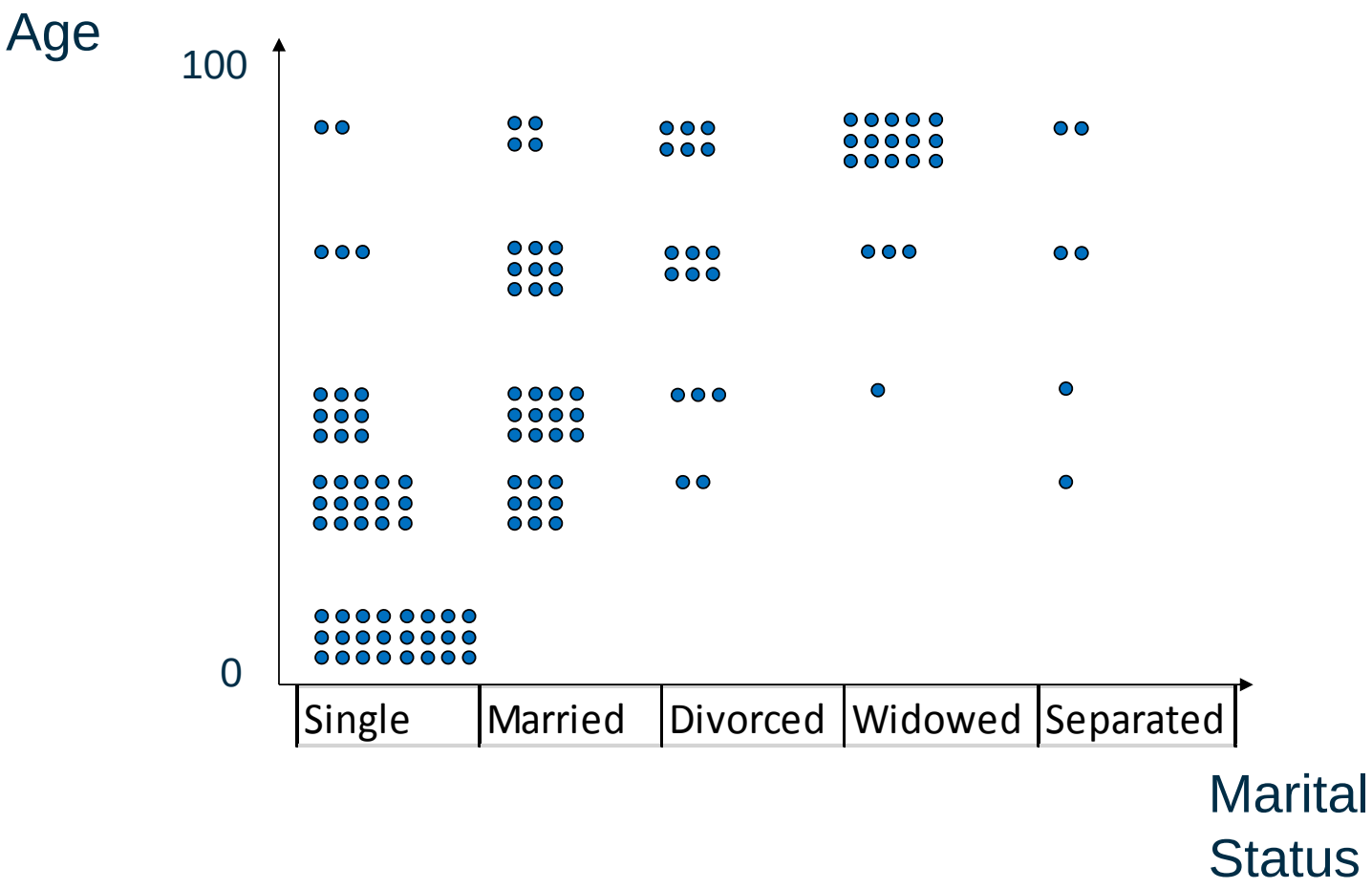
Synthetic data could help users prepare for the real data

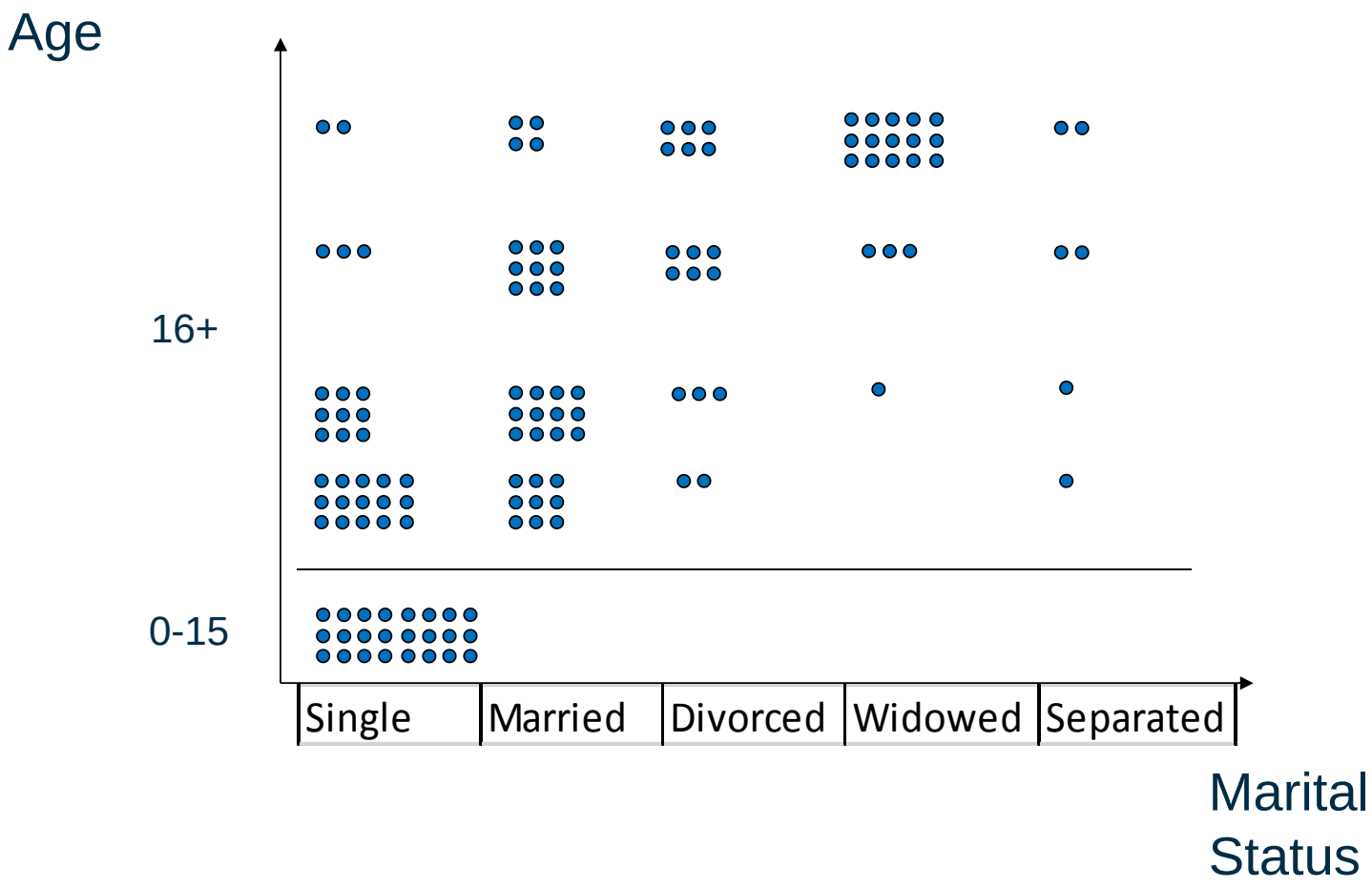
- Ideally the synthetic data would give the same results as the real data, but this will be *very* difficult to achieve!

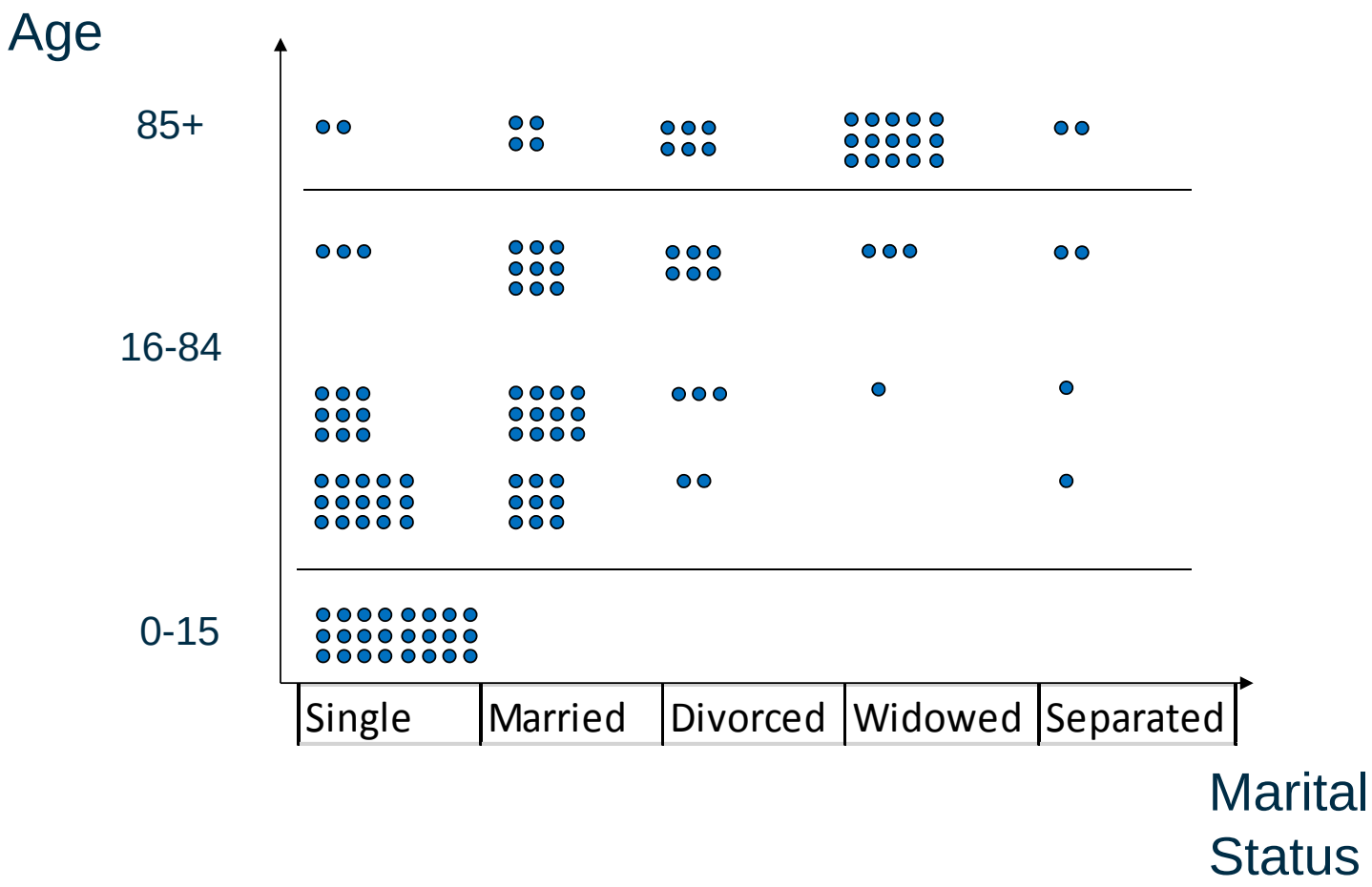
- Instead, a user could write and test code using the synthetic version (which would normally be done within the secure site), then get the final results from the secure version

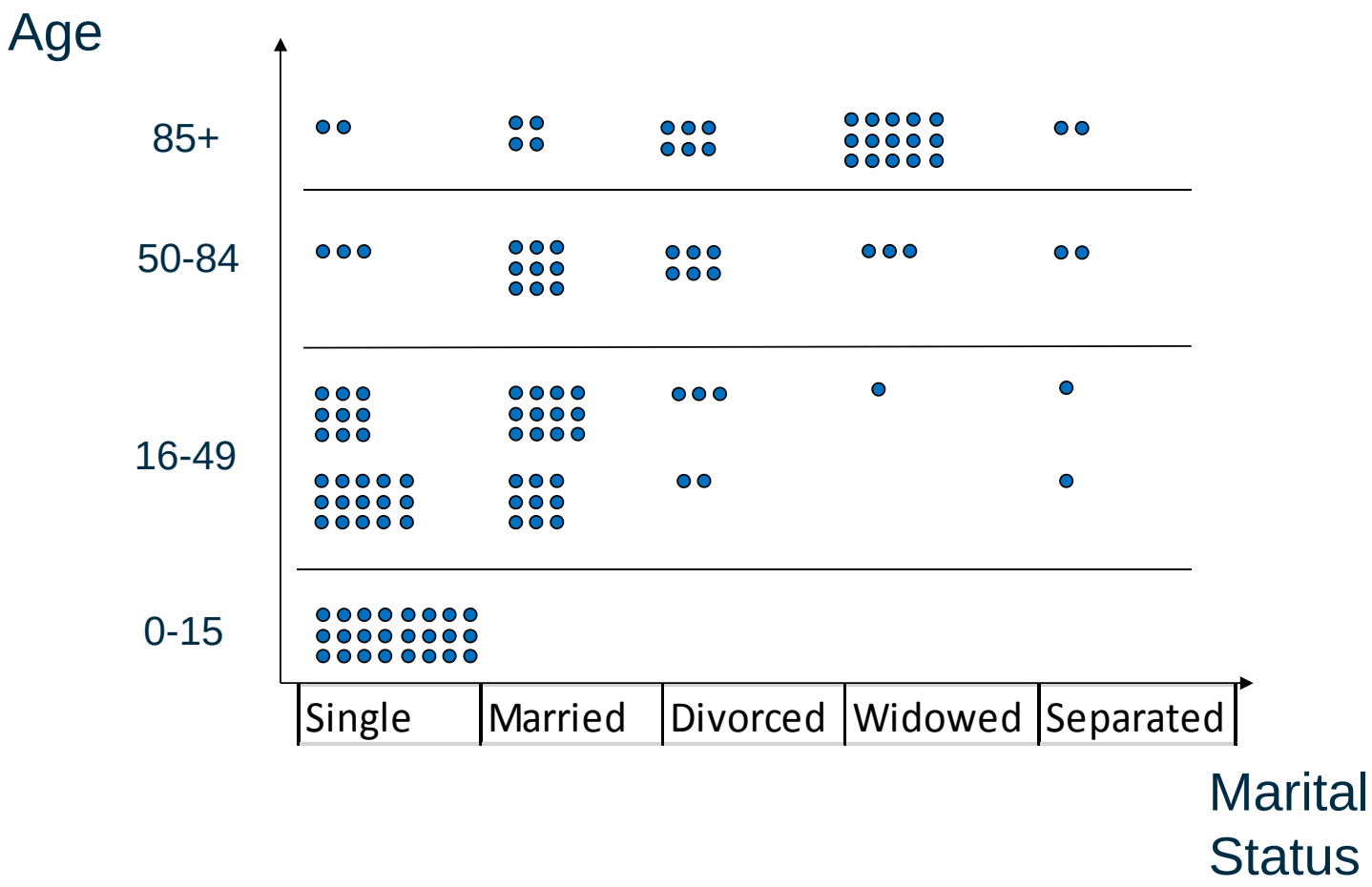
Synthesis methods

- Based on the EUL information, the extra variables are synthesised from statistical models/conditional probabilities
- Most variables are synthesised using classification trees in the ‘Synthpop’ R package
- Given age, sex, random generation of most likely _

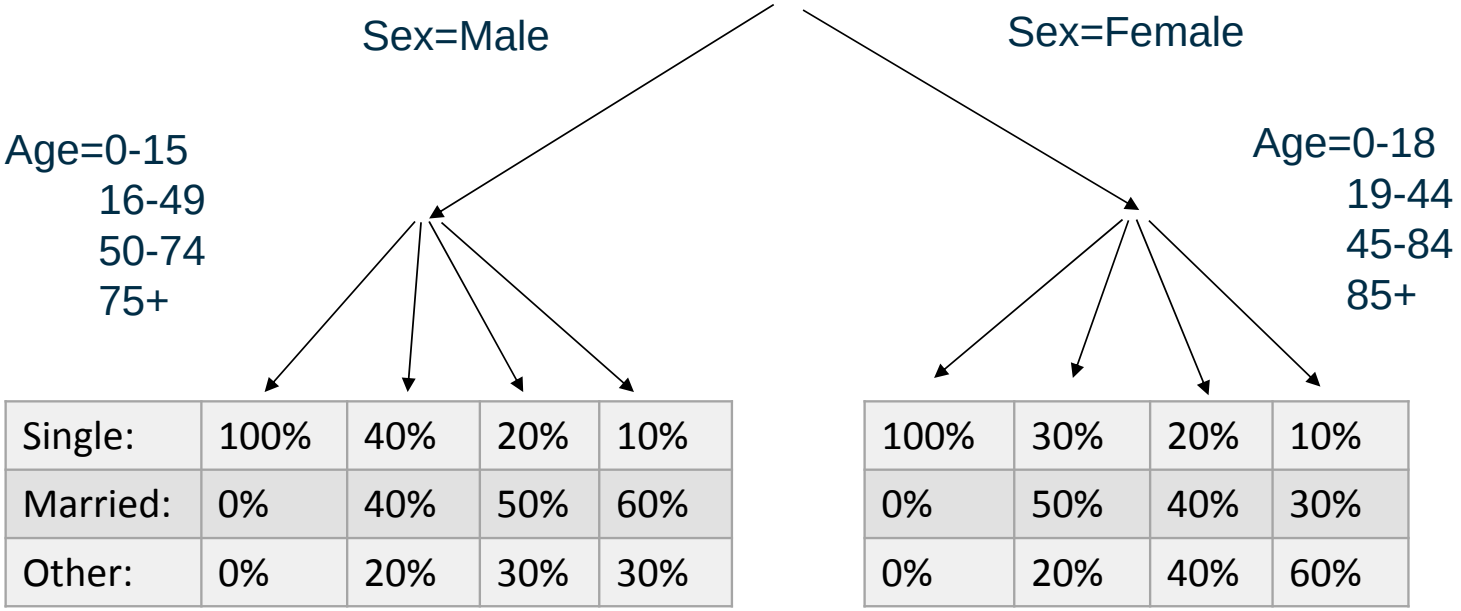

Age
100
0
Single
Married
Divorced
Widowed
Separated
Marital Status


Age
16+
0-15
Single
Married
Divorced
Widowed
Separated
Marital Status


Age
85+
16-84
0-15
Single
Married
Divorced
Widowed
Separated
Marital Status


Age
85+
50-84
16-49
0-15
Single
Married
Divorced
Widowed
Separated
Marital Status

Classification Trees



Single:	100%	40%	20%	10%
Married:	0%	40%	50%	60%
Other:	0%	20%	30%	30%

100%	30%	20%	10%
0%	50%	40%	30%
0%	20%	40%	60%

- The same process is performed with multiple predictor variables, which can be visualised as this tree structure

Synthesis methods - Occupation, Industry, Country of Birth

- Given 3 digit SIC/SOC, very limited plausible options of 4/5 digit SIC/SOC
- So generate these at random from all plausible entries:
- E.g. Basic SOC = 114, so detailed SOC = 114**1**
  114**2**
  114**3**

Conclusion

- We are producing a synthetic version of secure LFS data (not all variables). This will not give the same results as the real data, but it will be non-disclosive
- We hope this data will be helpful in preparing to use the full data at a secure site. Our main goal is to make it easier to work with data in the SRS, by saving time, travel, and accommodation costs.

Feedback

If you're interested in using synthetic data like this, we'd love to hear from you:

- Would synthetic data be useful to you? Why/why not?
- What properties would you want the synthetic data to preserve?
- Which variables do you use, that are only in the secure version?

- For more information, read our paper here (search ONS methodology working series paper 16)
- Email iain.dove@ons.gov.uk with any questions/comments