Measuring statistical uncertainty in admin-based population estimates.

Dr Louisa Blackwell and Katy Stokes, Methods Data and Research, Office for National Statistics, United Kingdom

Professor Peter Smith, Southampton Statistical Sciences Research Institute, University of Southampton

Presentation Outline

- Office for National Statistics, UK Population Statistics Transformation Programme
- Measuring uncertainty in ONS mid-year population estimates
- Theoretical basis and underlying assumptions
- Uncertainty measurement methods for SPD (version 2)
- Indicative results
- Further work in progress

ONS Population Statistics Transformation

- We are the UK’s largest independent producer of official statistics and its recognised national statistical institute

- We collect and publish statistics related to the economy, population and society at national, regional and local levels

- We also conduct the census in England and Wales every 10 years

- We are working with the Admin Data Census programme to transform population statistics – with admin data also at the core of population statistics by 2020

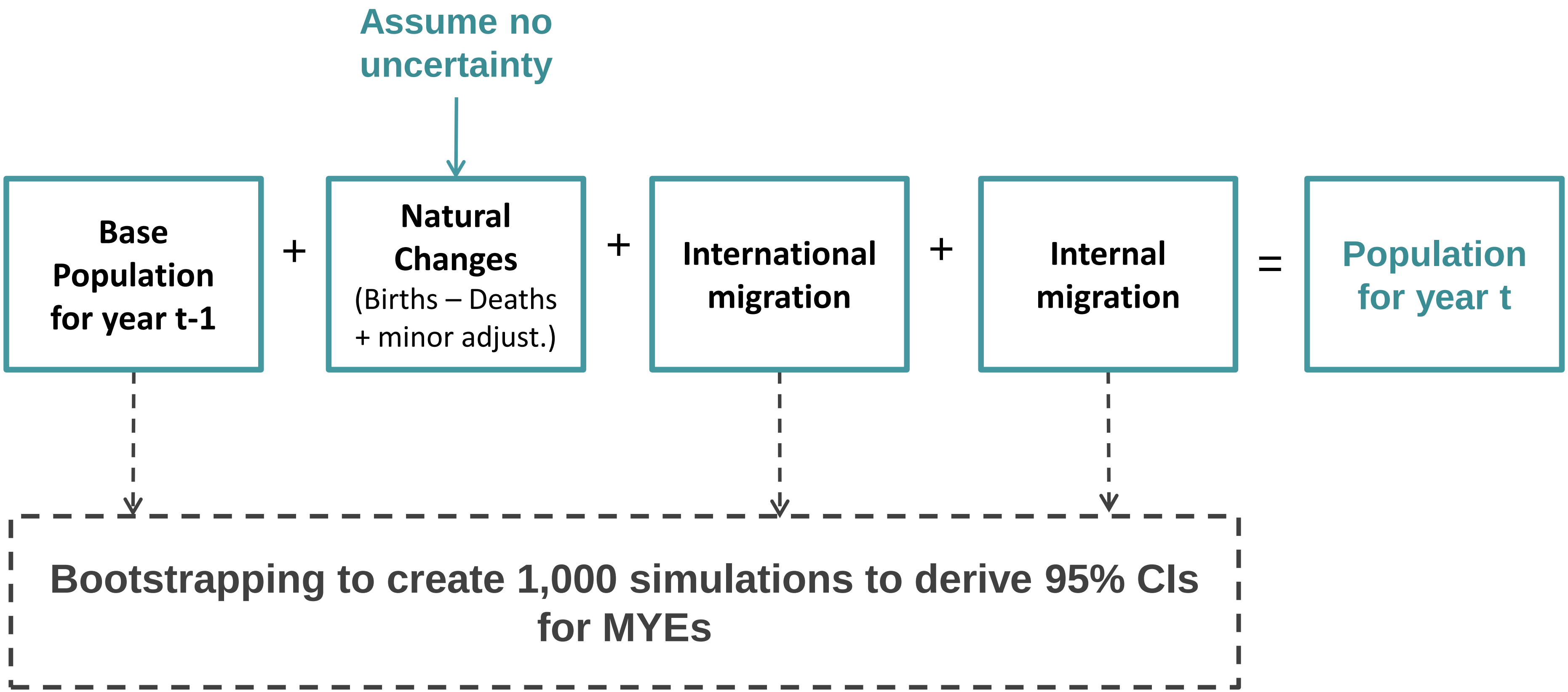
- Our research into statistical uncertainty is helping to guide that programme

MYEs: why measure uncertainty?

- Mid-year estimates (MYEs) are annual estimates UK usual residents (66M in 2017)
- Age, sex, local authority, components of change
- Primary use: Resource allocation for local government, health,
- Secondary use: Policy areas: education planning and provision, ageing and pension planning, housing demand and planning &&&
- Code of Practice for Official Statistics: levels of quality are measured and reported, including main sources of bias and other errors.
- ***Statistical and statutory imperatives***
- ***Definition: The quantification of doubt about an estimate***

Cohort component method for MYE uncertainty

Uncertainty Estimates =



Assume no uncertainty

Base Population for year t-1

+

Natural Changes (Births – Deaths + minor adjust.)

+

International migration

+

Internal migration

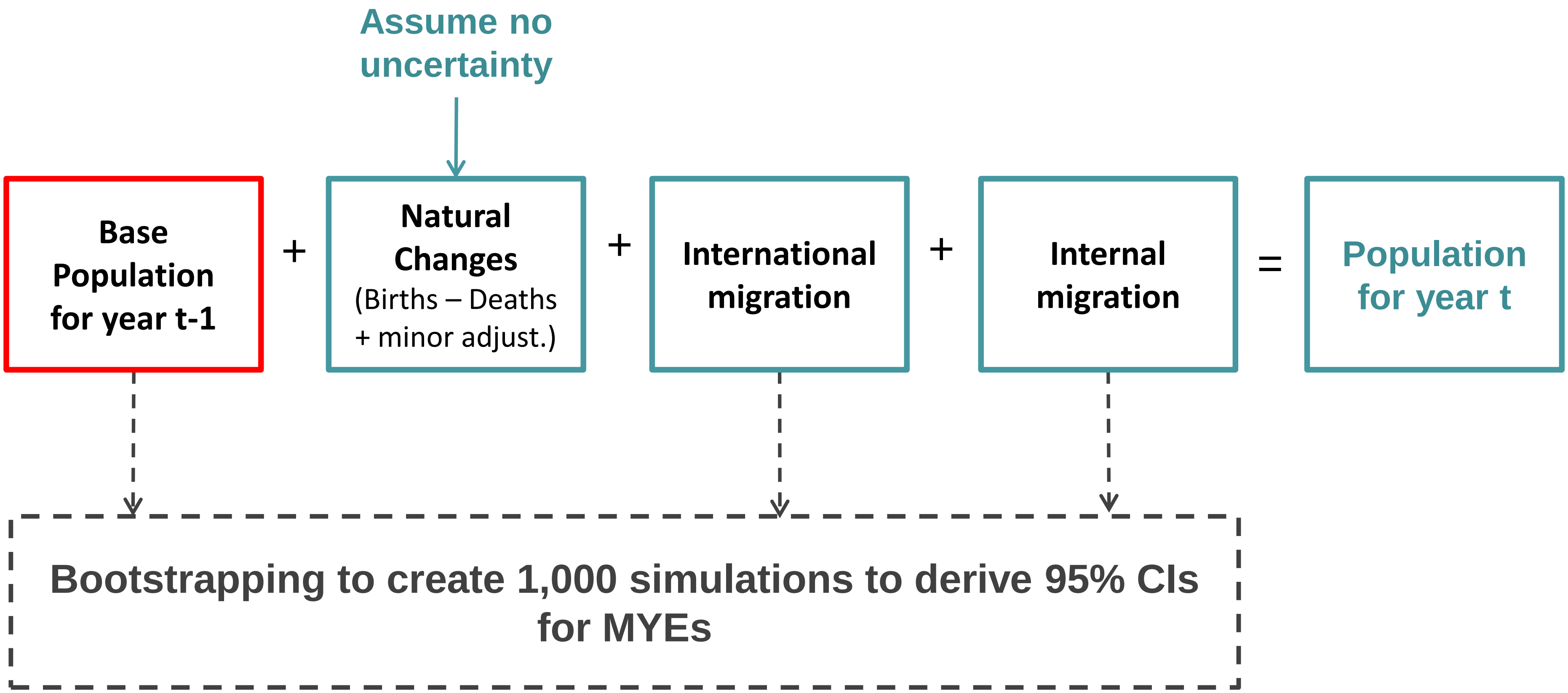
=

Population for year t

Bootstrapping to create 1,000 simulations to derive 95% CIs for MYEs

Cohort component method for MYEs

Uncertainty Estimates =



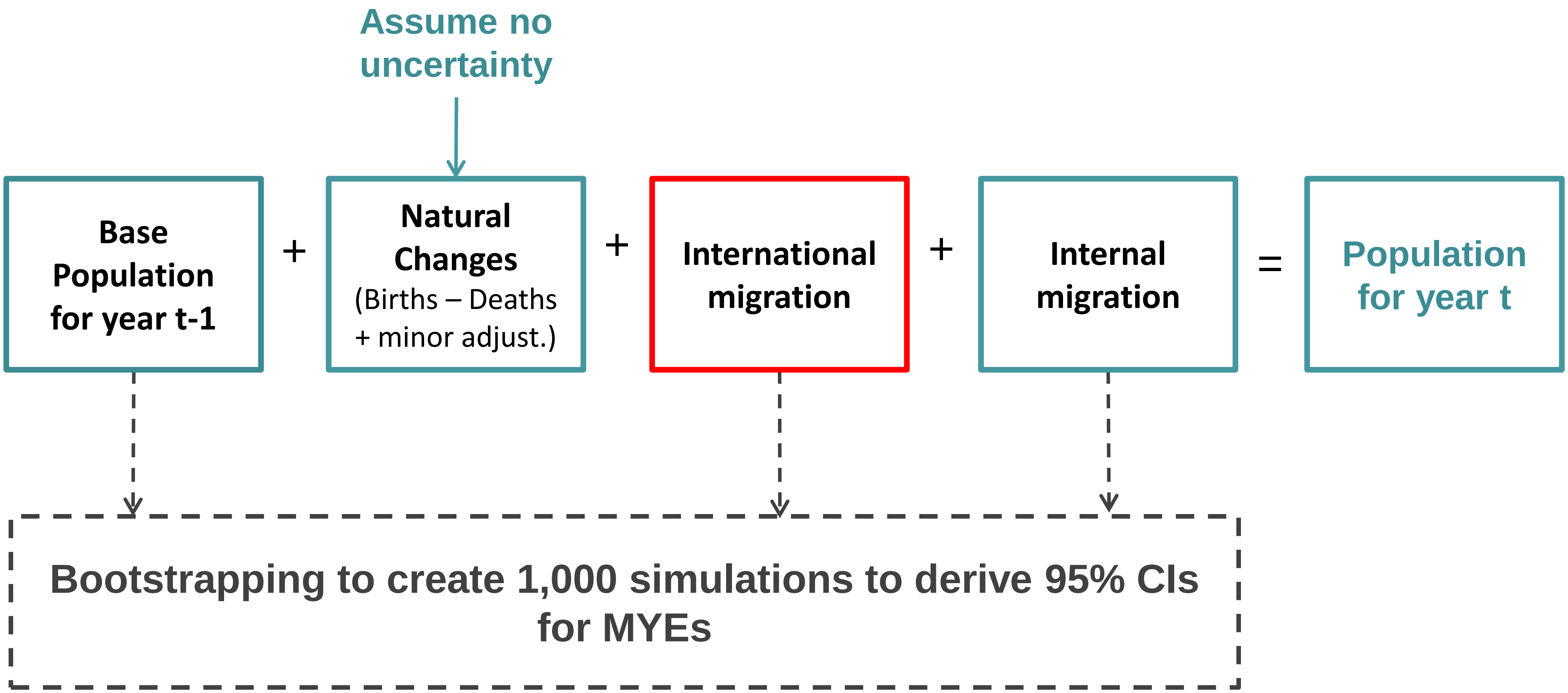
Assume no uncertainty

Base Population for year t-1 + Natural Changes (Births – Deaths + minor adjust.) + International migration + Internal migration = Population for year t

Bootstrapping to create 1,000 simulations to derive 95% CIs for MYEs

Cohort component method for MYEs

Uncertainty Estimates =



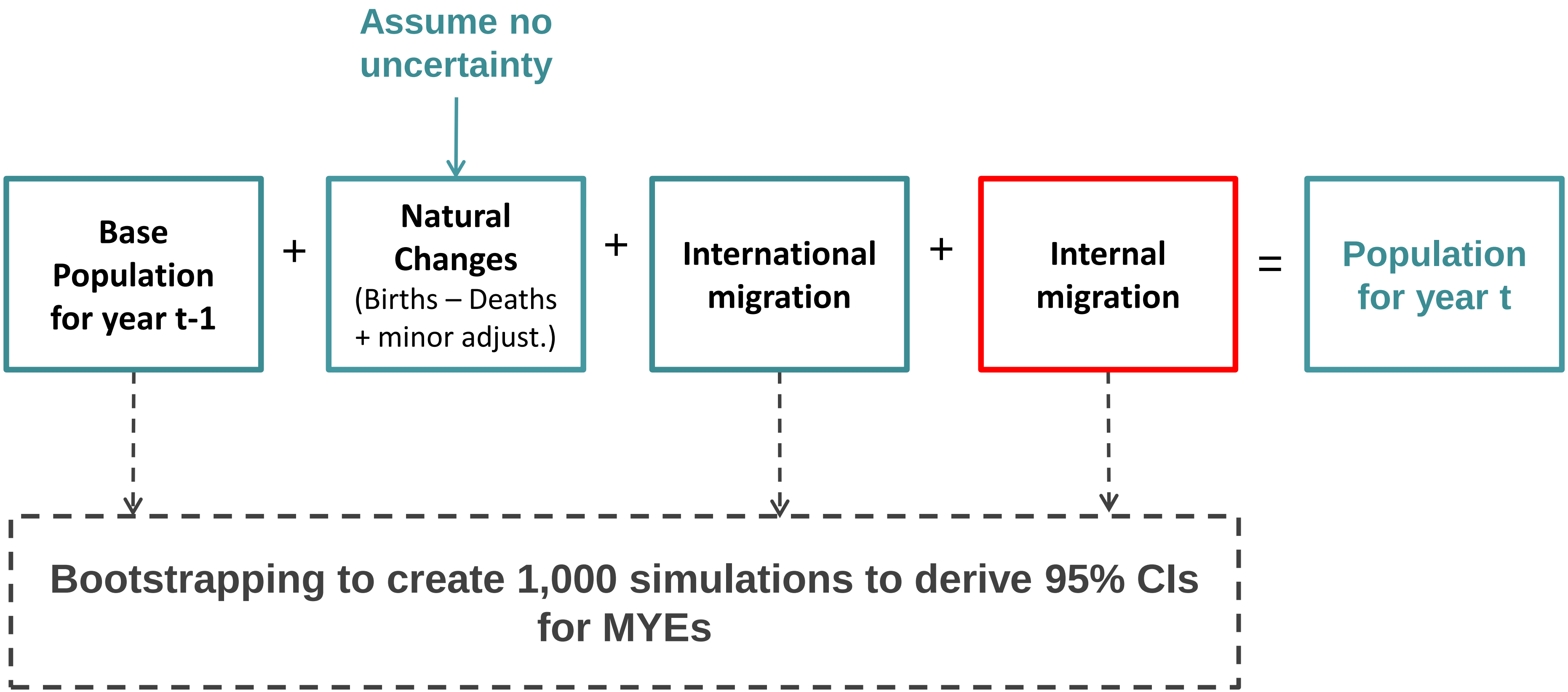
Assume no uncertainty

Base Population for year t-1 + Natural Changes (Births – Deaths + minor adjust.) + International migration + Internal migration = Population for year t

Bootstrapping to create 1,000 simulations to derive 95% CIs for MYEs

Cohort component method for MYEs

Uncertainty Estimates =



Assume no uncertainty

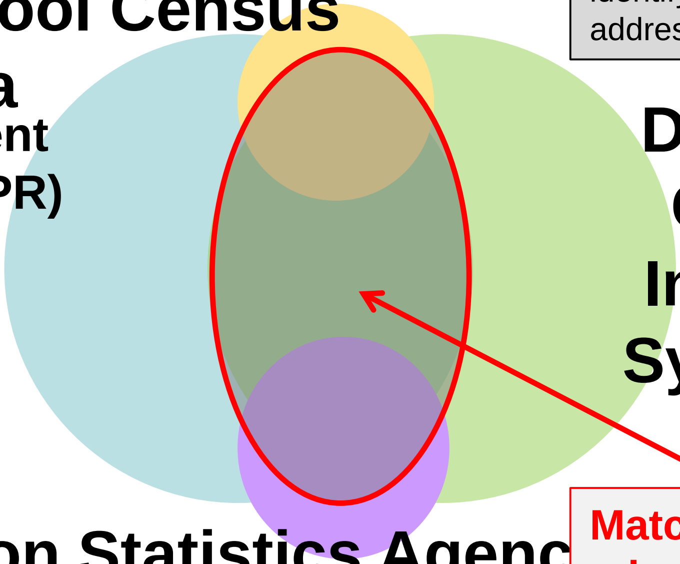
Base Population for year t-1 + **Natural Changes** (Births – Deaths + minor adjust.) + **International migration** + **Internal migration** = **Population for year t**

Bootstrapping to create 1,000 simulations to derive 95% CIs for MYEs

Design of the SPD to produce the population estimates

Statistical Population Dataset – SPD

The SPD estimates have been produced by matching individual records across the administrative data sources. To protect privacy of individuals the process involves replacing identifying fields (names, dates of birth and addresses) by one or more artificial identifiers.



School Census Data

NHS Patient Register (PR)

DWP/HMRC Customer Information System (CIS)

Higher Education Statistics Agenc Data (HESA) - students

Matched records from the various administrative data sources are included in the Statistical Population Dataset (SPD)

Aggregate totals of Armed Forces personnel are added in to the estimates

=

SPD population estimates

A future plan is to use a coverage survey to adjust for biases on the SPD

ADC requirements for SPD uncertainty

ADC requirements:

- Conduct methodological research to develop uncertainty measures by single-year of age, sex and LA for SPDs 2011-2016
- Produce corresponding measures for the mid-year population estimates
- Compare these by LA, age, sex.

Benchmark approach for SPDs

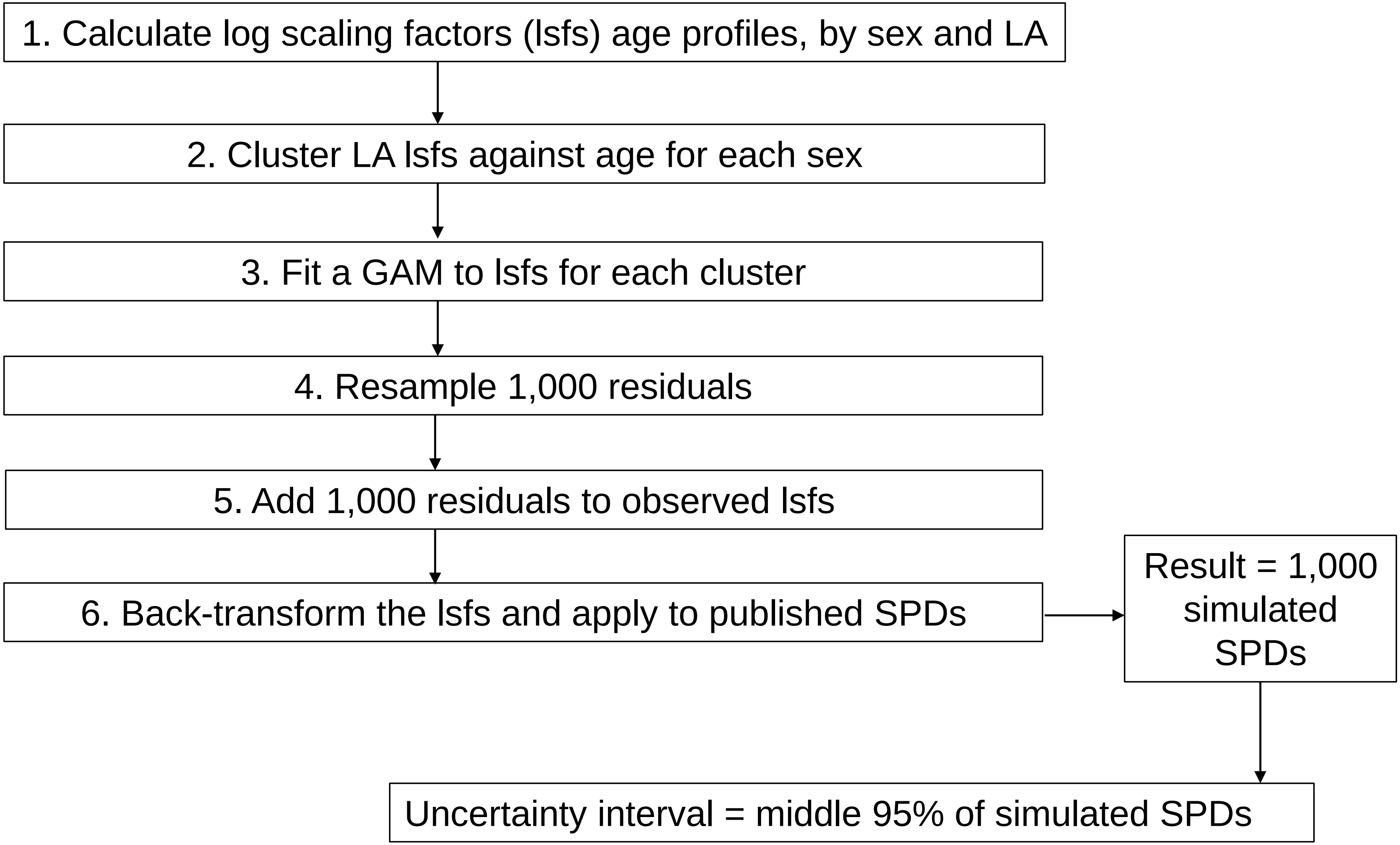
- Uncertainty: Benchmark SPD by comparing against Census

$$lsf = \log\left(\frac{census}{spd}\right)$$

- Ratio provides a measure of how the 2011 SPD differs from the Census

- Assumption: Relationship between the SPD and Census remains constant over time

SPD Uncertainty process



1. Calculate log scaling factors (lsfs) age profiles, by sex and LA

2. Cluster LA lsfs against age for each sex

3. Fit a GAM to lsfs for each cluster

4. Resample 1,000 residuals

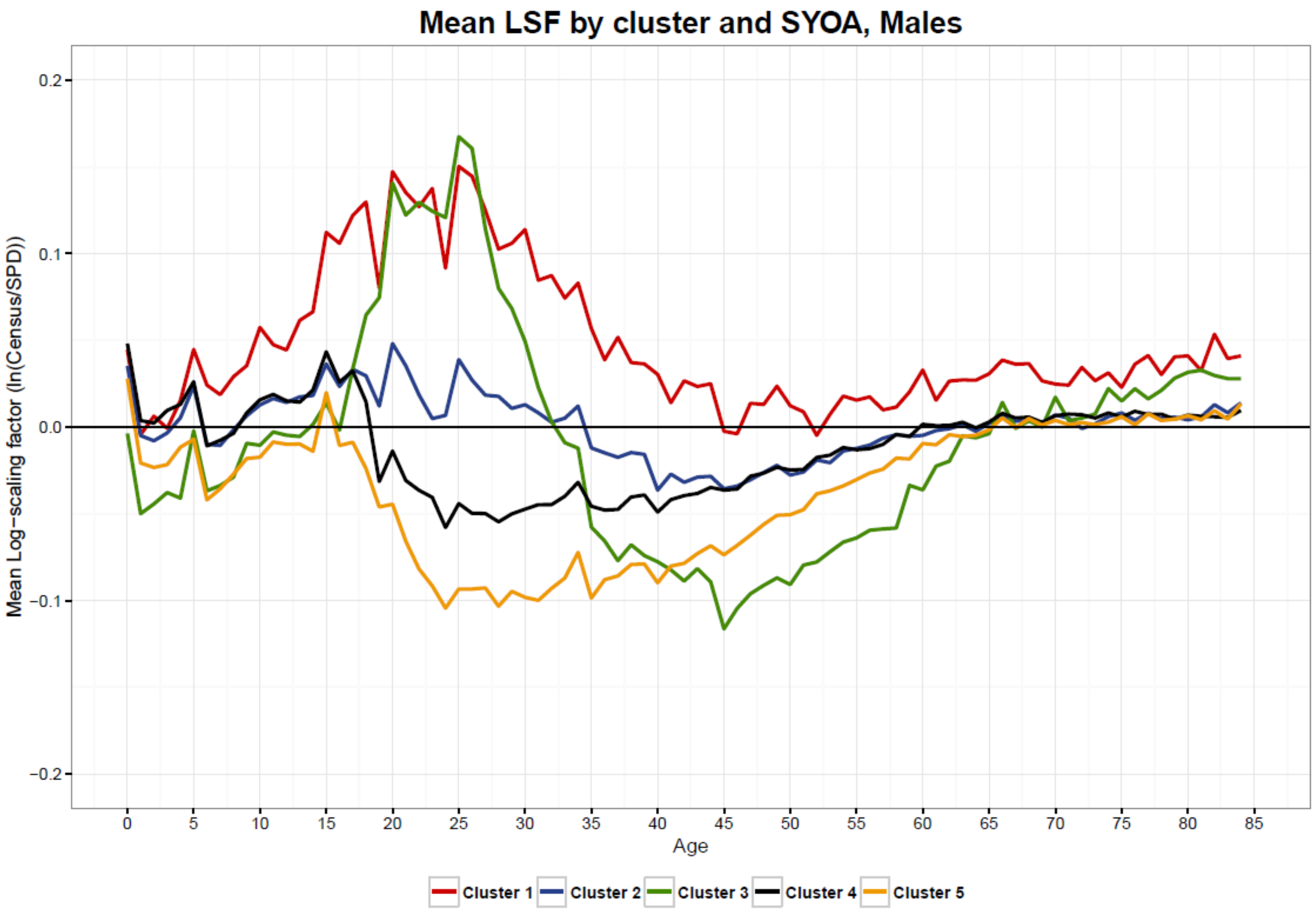
5. Add 1,000 residuals to observed lsfs

6. Back-transform the lsfs and apply to published SPDs

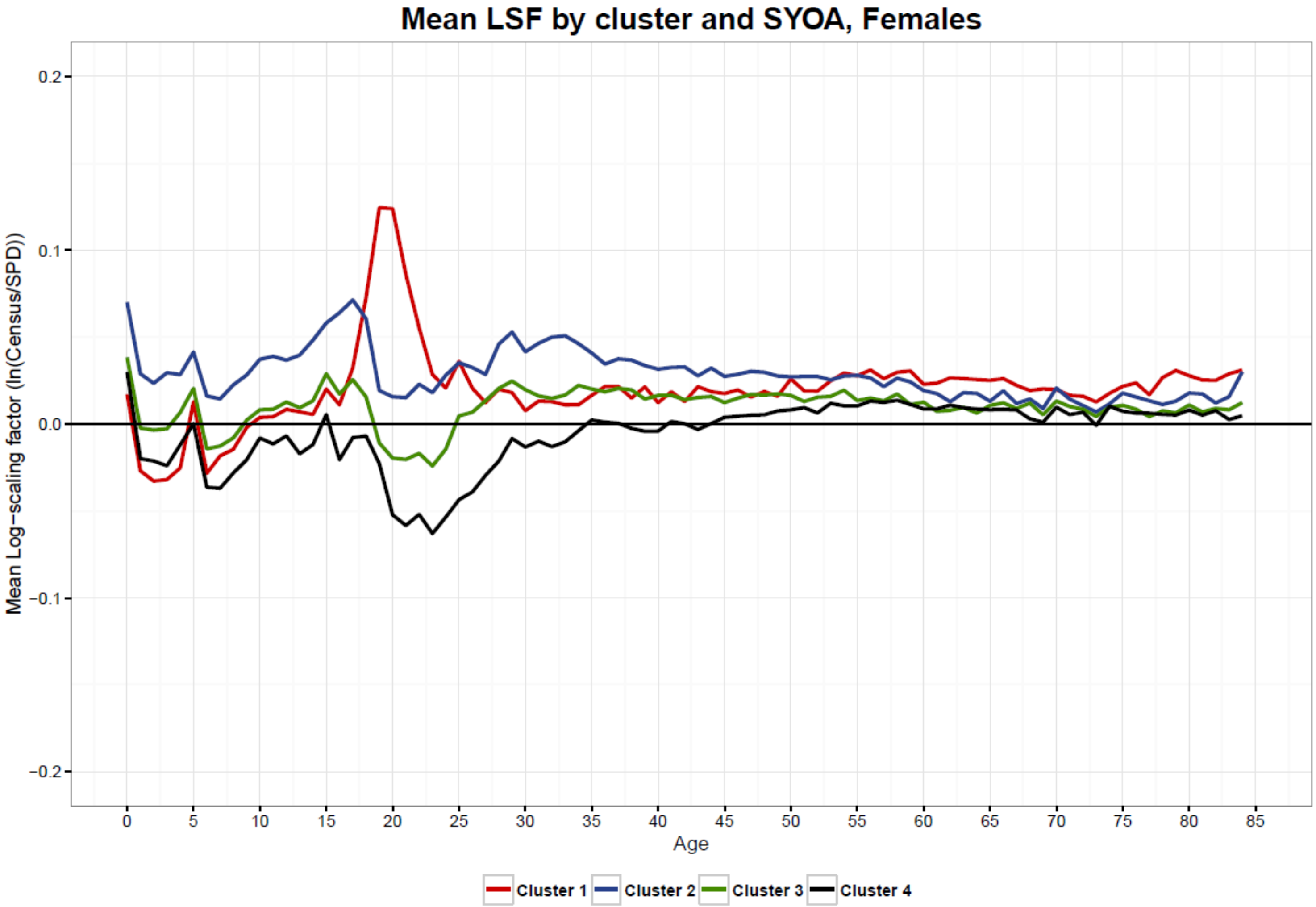
Result = 1,000 simulated SPDs

Uncertainty interval = middle 95% of simulated SPDs

Male clusters



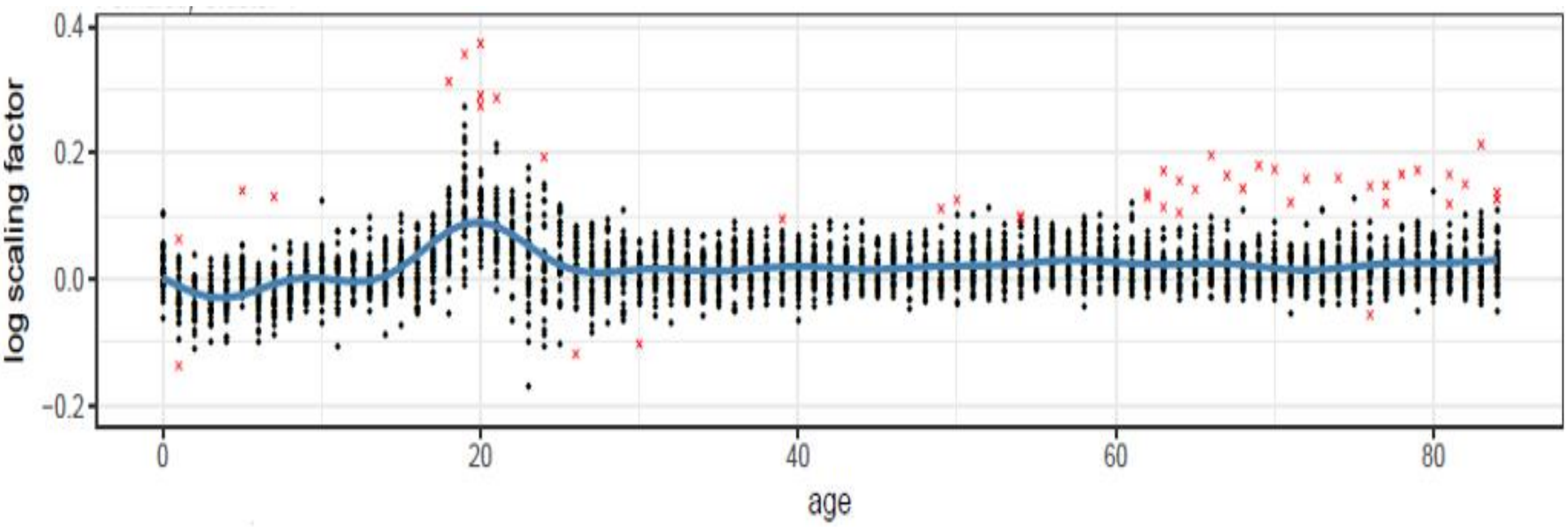
Female clusters



Fit a GAM to each cluster

- Generalised Additive Models (GAM)
- Generates fitted LSFs and corresponding residuals (observed – fitted) for each combination of sex, age and LA

Scatterplot of LSF vs. age for females in cluster 1, with fitted GAM (blue curve).



Resampling residuals

- Define a group as a unique combination of cluster and age *(and sex)*
- Pool the groups of residuals to obtain a better estimate of variability
- Standardise the residuals so that the variance for each group is one:

standardised residual →

$$s = \frac{r}{\sigma}$$

← raw residual

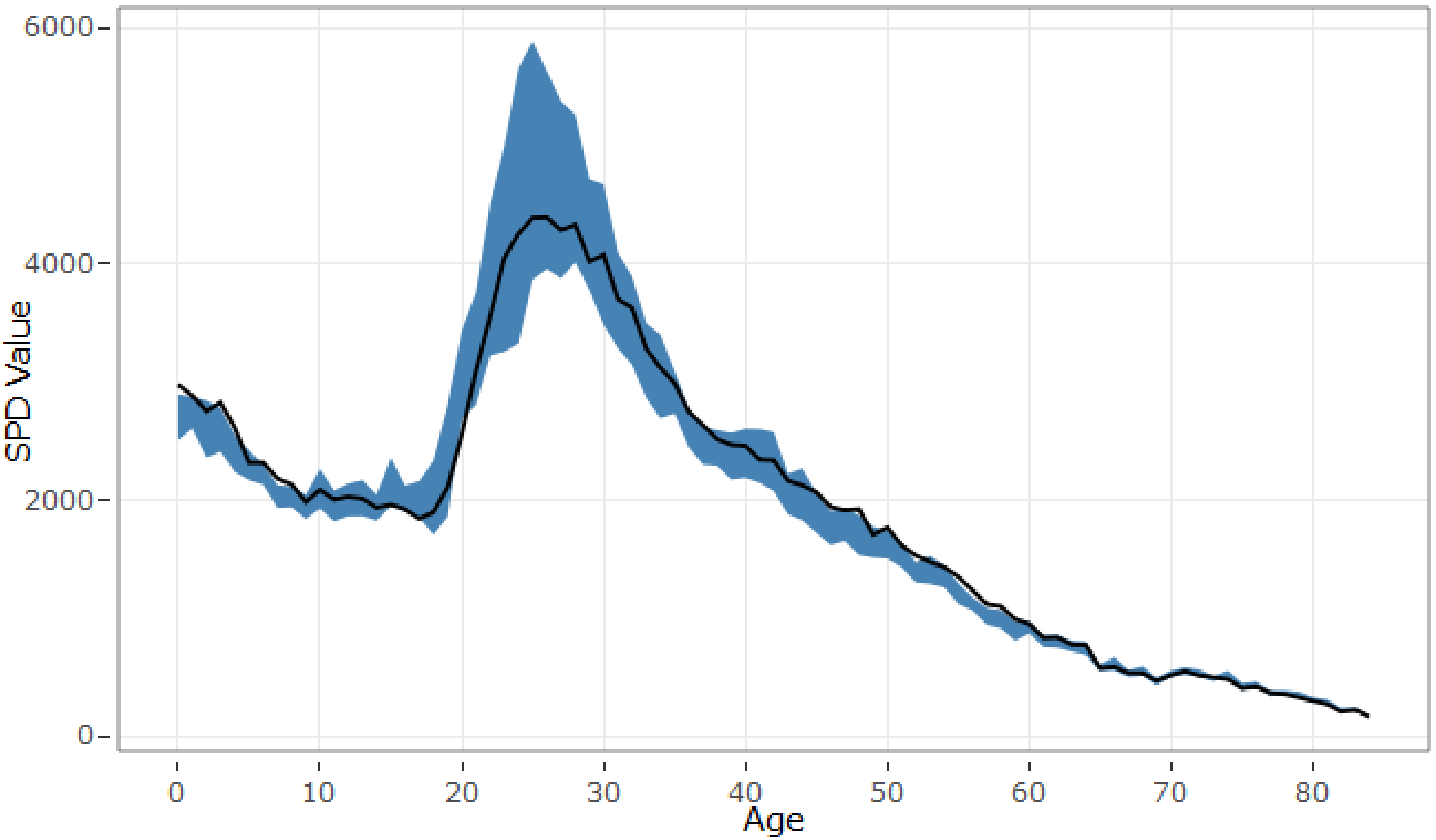
← group standard deviation

- Put all the standardised residuals in one pot
- 1000 residuals are resampled (with replacement)

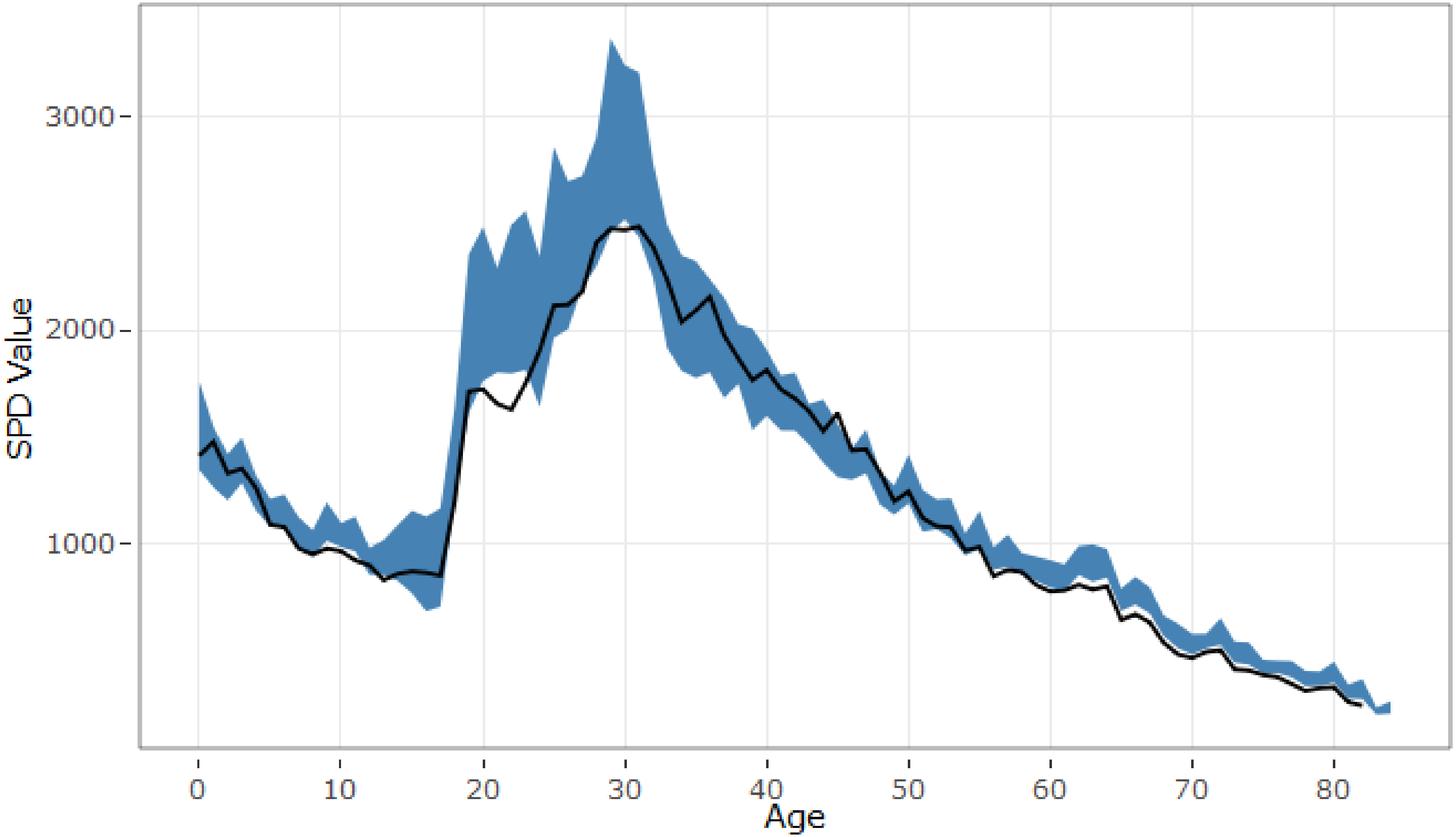
Then:

- **Un-standardize residuals** for each LA by multiplying them by their group standard deviation- then add these to observed LSFs to create 1,000 alternative values

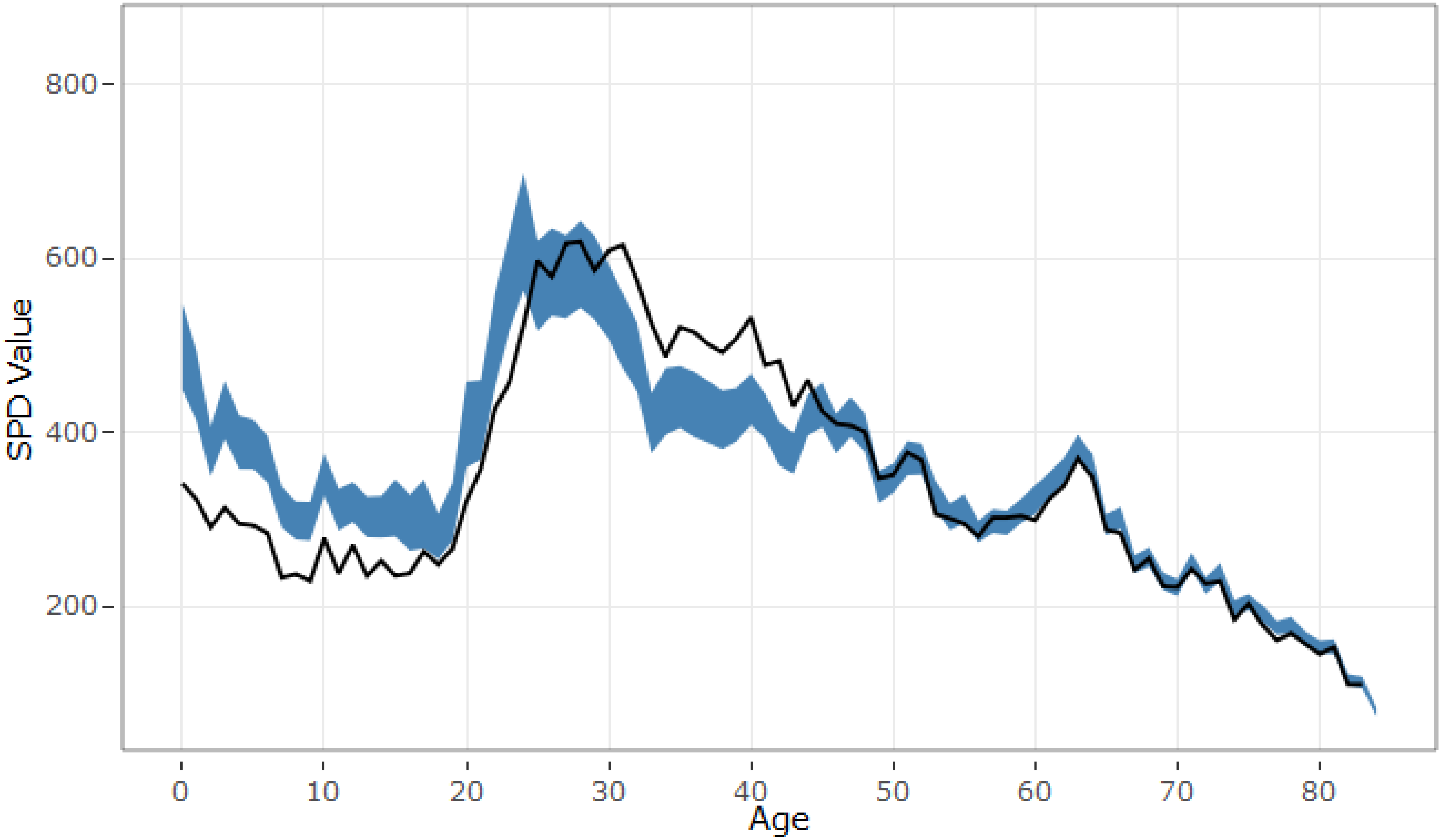
2011 SPDs and uncertainty intervals for males in Newham, by age



2011 SPDs and uncertainty intervals for males in Camden, by age



2011 SPDs and uncertainty intervals for males in Forest Heath, by age



Next steps

- Producing uncertainty measures for MYEs by sex and single year or age within each LA
- Developing new Uncertainty Intervals for SPD3
- Revisiting the assumption that error in the SPD when compared against the Census is unchanged through the intercensal decade
- Considering the impact of change in the key NHS source data

Questions

- Do you have examples of similar (or completely different!) approaches for measuring uncertainty in population estimates?

- We are thinking of using the benchmark approach to assess the behaviour of key admin sources over time. Do you have a use for this, or do you do something different?

- Arguably our ability to cluster LAs based on their log scaling age profiles could suggest that SPD methods could be refined for different types of LA. Your views on taking a stratified approach to the methodology would be welcome.

Benchmark approach for SPDs: Conceptual framework

- Benchmark approach uses the following model:

$$\log(P_{i,j,k}) = \log(SPD_{i,j,k}) + LSF_{i,j,k} + \varepsilon_{i,j,k},$$

- Exponential of benchmark approach:

$$P_{i,j,k} = SPD_{i,j,k} \times \exp(LSF_{i,j,k} + \varepsilon_{i,j,k}).$$

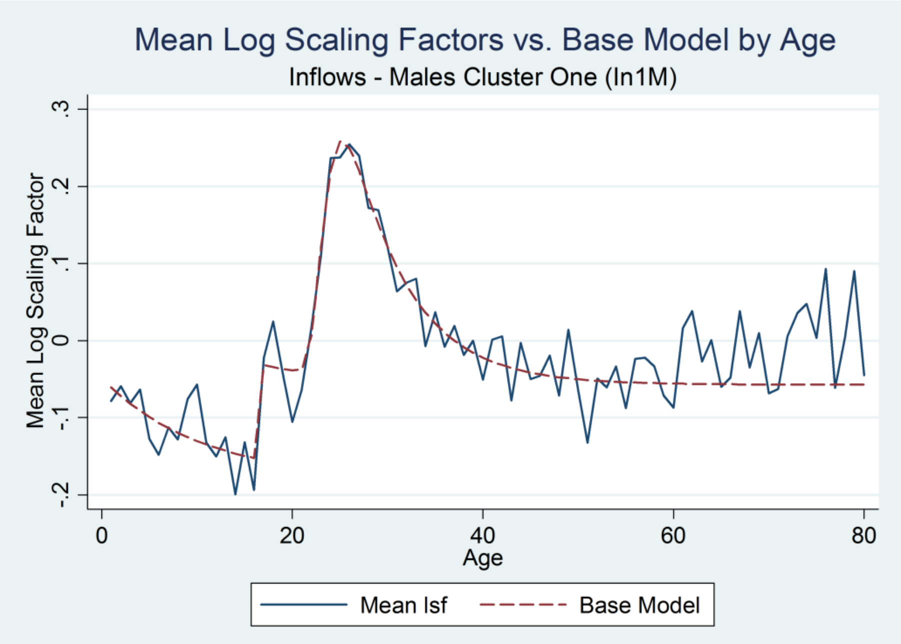
- Uncertainty: Benchmark SPD comparing against Census

$$lsf_{i,j,k} = \log\left(\frac{census_{i,j,k}}{spd_{i,j,k}}\right).$$

Rogers Castro Model

7 parameter Rogers-Castro Migration Curve

$$\underbrace{a_6}_{\text{constant}} - \underbrace{a_7 \sum_{j=0}^{17} I_j(x)}_{\text{step change lowers curve for <18}} + \overbrace{\underbrace{a_0}_{\text{height of childhood curve}} \exp[- \underbrace{a_1}_{\text{rate of descent}} x]}^{\text{childhood}} + \overbrace{\underbrace{a_2}_{\text{height of curve}} \exp\left\{- \underbrace{a_3}_{\text{rate of descent}} (x - \overbrace{a_4}^{\text{position of curve}}) - \exp[- \underbrace{a_5}_{\text{rate of ascent}} (x - a_4)]\right\}}^{\text{labour force}}$$

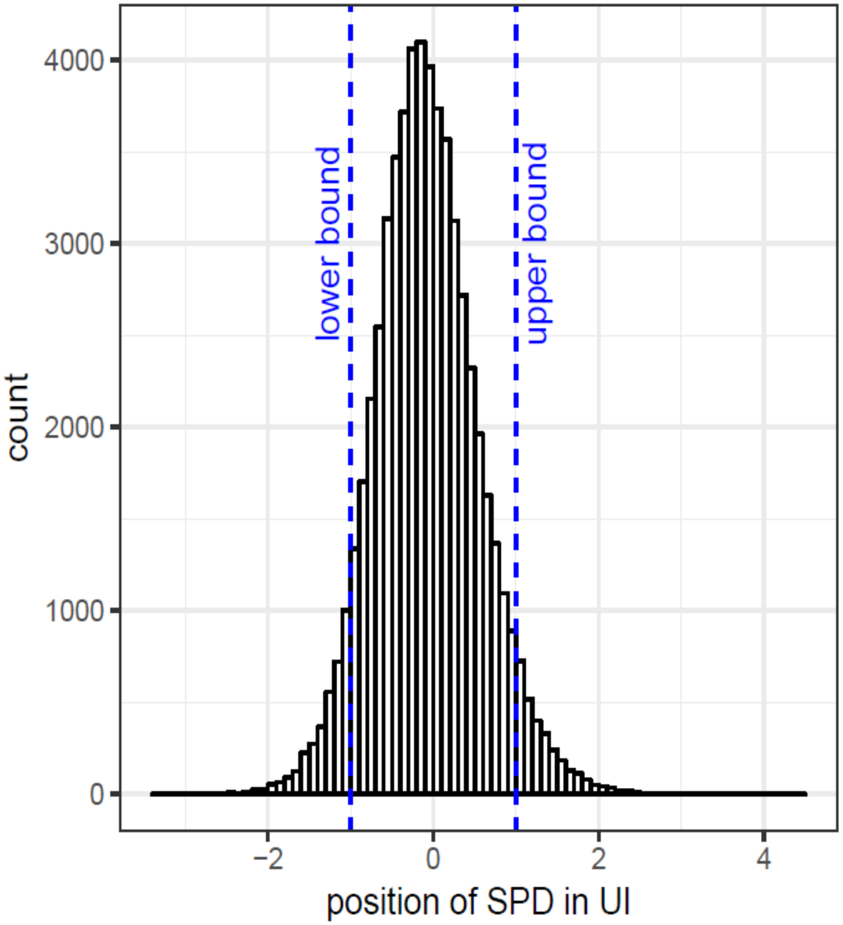


Bootstrapping methods used

Component	Re-sampling method	Type of bootstrap
Census base	Uses published variances Assumes errors are normally distributed.	Parametric
Correction of census base to mid-year	Assumes the same CV as for the 2012 mid-year estimate Assumes errors are normally distributed.	Parametric
International in-migration	IPS: re-sampling with replacement, 1,000 new samples re-run through IPS imputation.	Non-Parametric
	Admin-based allocation to LAs: benchmark approach, comparing each admin source to 2011 Census to derive variances. Assumes errors are log-normally distributed.	Parametric using a benchmark
International out-migration	IPS: re-sampling with replacement, 1,000 new samples re-run through Poisson regression.	Non-parametric
Internal migration	Benchmark approach. Re-samples residuals from the non-linear (Rogers-Castro) regression model of LSFs (Census/ PR moves). Sampled residuals added to predicted values from the model. Model updated with contemporaneous covariates for the inter-censal period	Re-sampling residuals using a benchmark

Uncertainty Interval (UI) calculation

Mid 95% UI



MYE Uncertainty measurement: composite vs benchmark approach

- **‘Uncertainty’ = quantification of doubt about an estimate**

- Sources of error: Internal, international migration and the Census

- Uncertainty in MYE: captured in a composite way using the Cohort Component approach

- But MYE uncertainty also involves using a ‘benchmark approach’