Two Approaches for Register-based Population Size Estimation: Individual-level and Population-level

Lucy Brown[1]: lyb3@kent.ac.uk

Eleni Matechou[2] Bruno Santos[3] Eleonora Mussino[4]

1 University of Kent, UK
2 Queen Mary University of London, UK
3 Universidade de Lisboa, Portugal
4 Stockholm University, Sweden

Administrative Register Data

- Provided by a national statistics agency (e.g. Statistics Sweden – SCB, or Statistics Norway - SSB)

- All foreign-born residents who first entered the country as adults during the study years (e.g. for Sweden 2003-2016, or for Norway 2007-2023)

An arbitrary number of registers

register1	register2	register3
1	0	0
1	0	1
0	1	0
1	0	1
0	0	1
0	0	0
0	1	1

Emigration, Immigration and Death records

death	immig	emig
0	1	0
0	0	0
0	0	0
0	1	1
1	0	0
0	1	0
0	0	1

	id	year	age	cob	sex
1	XXX1	2007	28	France	1
2	XXX1	2008	29	France	1
3	XXX1	2009	30	France	1
4	XXX1	2012	33	France	1
5	XXX1	2013	34	France	1
6	XXX2	2008	67	India	0
7	XXX2	2009	68	India	0

Covariates (treated as categorical)

- **Age**
- **Country of birth**
- **Sex**
- **Time since first entry**

Capture-Recapture & HMM Formulation

Capture Recapture (CR) Models[1]:

1. An initial capture is made
2. Unmarked individuals are marked in some unique way (tags, rings etc)
3. Individuals are released back into the population
4. Subsequent captures are made, allowing individuals to be tracked over time

[1] Kenneth H. Pollock. Capture-recapture models. Journal of the American Statistical Association, 95(449):293–296, 2000. ISSN 01621459, 1537274X. URLhttp://www.jstor.org/stable/2669550

- CR models are commonly used in ecology to estimate population size for wild animals
- An individual's true state is unobservable (latent)
- Hidden Markov model (HMM) formulation marginalises over the latent state to calculate the marginal likelihood
- This iterative method is very efficient and allows latent states to be inferred

Probability of history at the previous time point

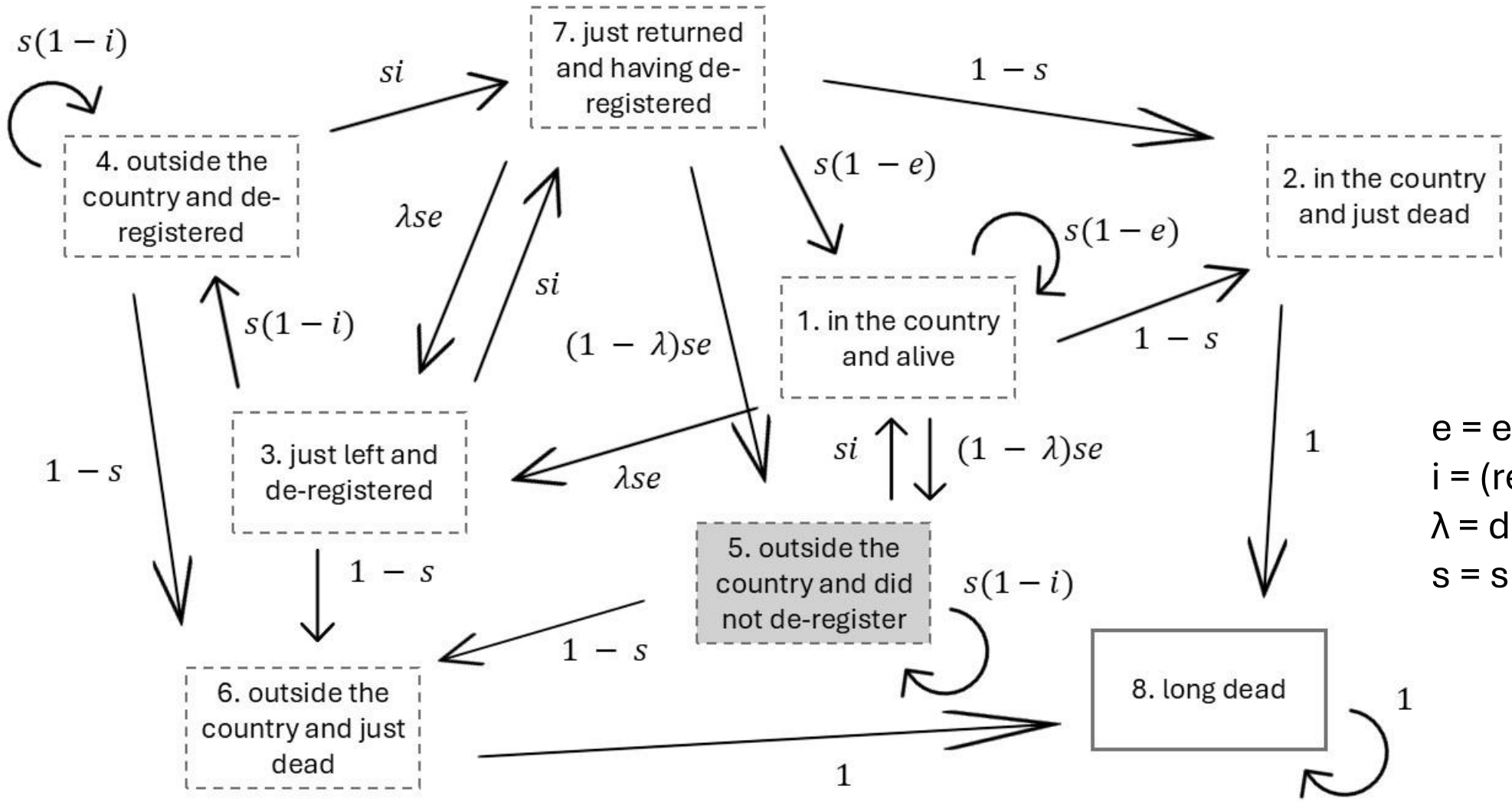
Probability of transitioning to each new state given the state at $t-1$

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\,\boldsymbol{\Gamma}_{t-1}\,\boldsymbol{P}(\boldsymbol{y}_t)$$

$$\boldsymbol{\alpha}_1 = \boldsymbol{\delta}\,\boldsymbol{P}(\boldsymbol{y}_1)$$

$$\Rightarrow \quad \boldsymbol{L}_T = \boldsymbol{\alpha}_T \mathbf{1}'$$

Observation probabilities dependent on the current state



e = emigration
i = (re-)immigration
λ = de-registering
s = survival

Model Specification

[2] A. Agresti. An introduction to categorical data analysis. John Wiley & Sons, 2nd edition, 2007.

1

- Parameters in the model matrices are specified using logistic regression and a multicategory logit model[2], allowing incorporation of individual characteristics and interactions:

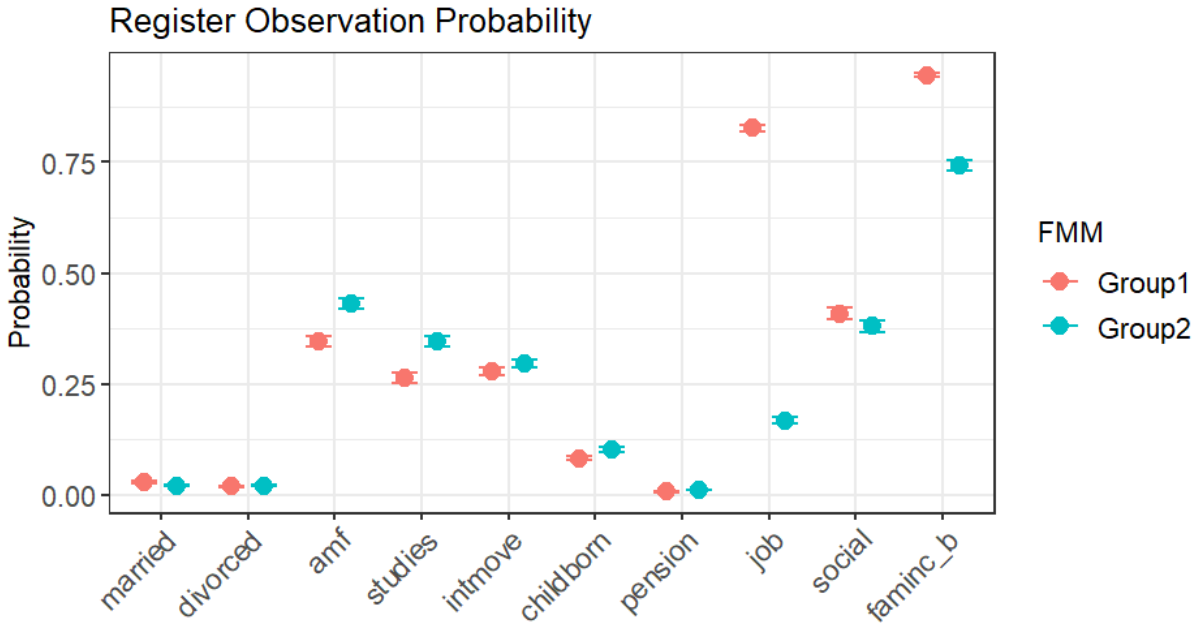
$$\text{logit}(\theta) = \beta_0 + \underbrace{\beta_1 C_1 + \beta_2 C_2 + \ \ldots}_{\text{Individual- and time-dependent covariates C}}$$

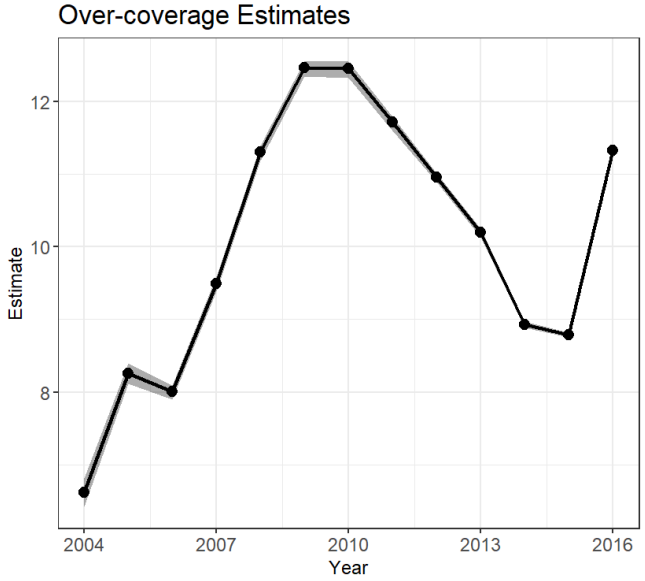
$$p_j^{(g)} = \frac{\exp(\gamma^{(g)} X_{j*})}{\sum_h \exp(\gamma^{(g)} X_{h*})} \qquad X = \begin{bmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 & 1 & \cdots & 1 \\ 1 & \cdots & 1 & 1 & \cdots & 0 & 1 & \cdots & 0 \\ & \vdots & & & \vdots & & & \vdots & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}$$

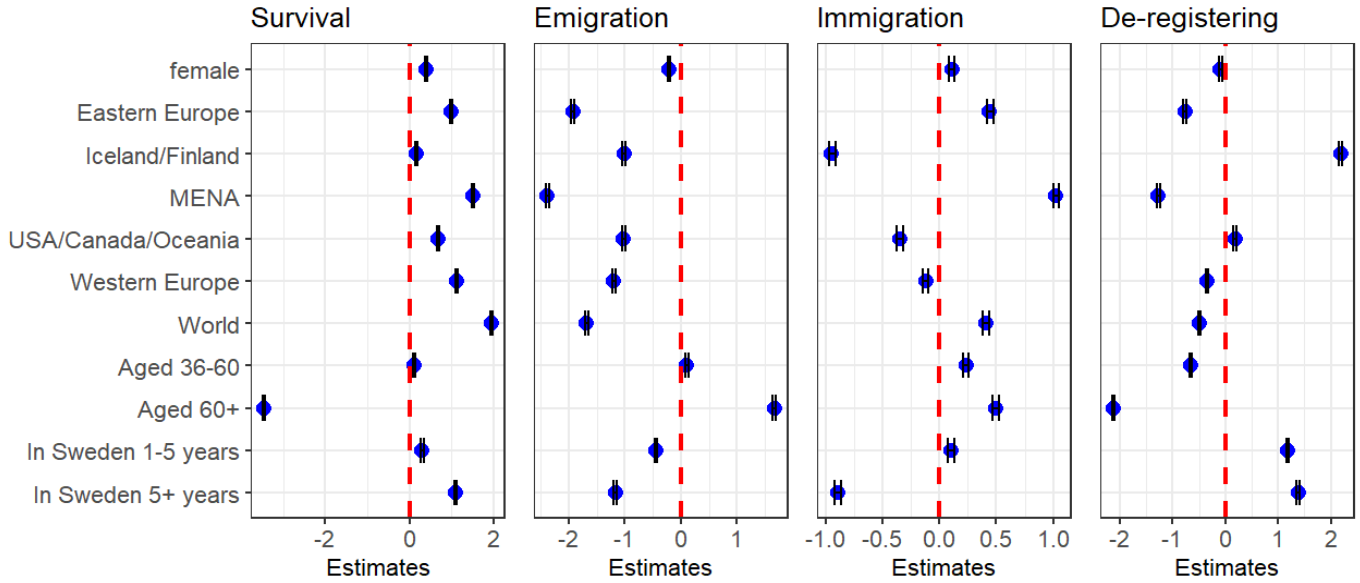
Registers Covariates Interactions

- Finite mixture model incorporates individual heterogeneity in the observation process for $g = \{1,2\}$
- Uncertain sightings (false positives) acknowledge that individuals observed on certain registers may not necessarily be present in the country, e.g. the family income register such that <u>when outside the country and we do not know</u>:

P(observed on family income register only) = ϵ P(unobserved) = $1 - \epsilon$







Baseline:

Survival:	0.996 (0.995, 0.996)
Emigration:	0.481 (0.475, 0.487)
Re-immigration:	0.084 (0.082, 0.086)
De-registration:	0.529 (0.522, 0.536)

1

Benefits:

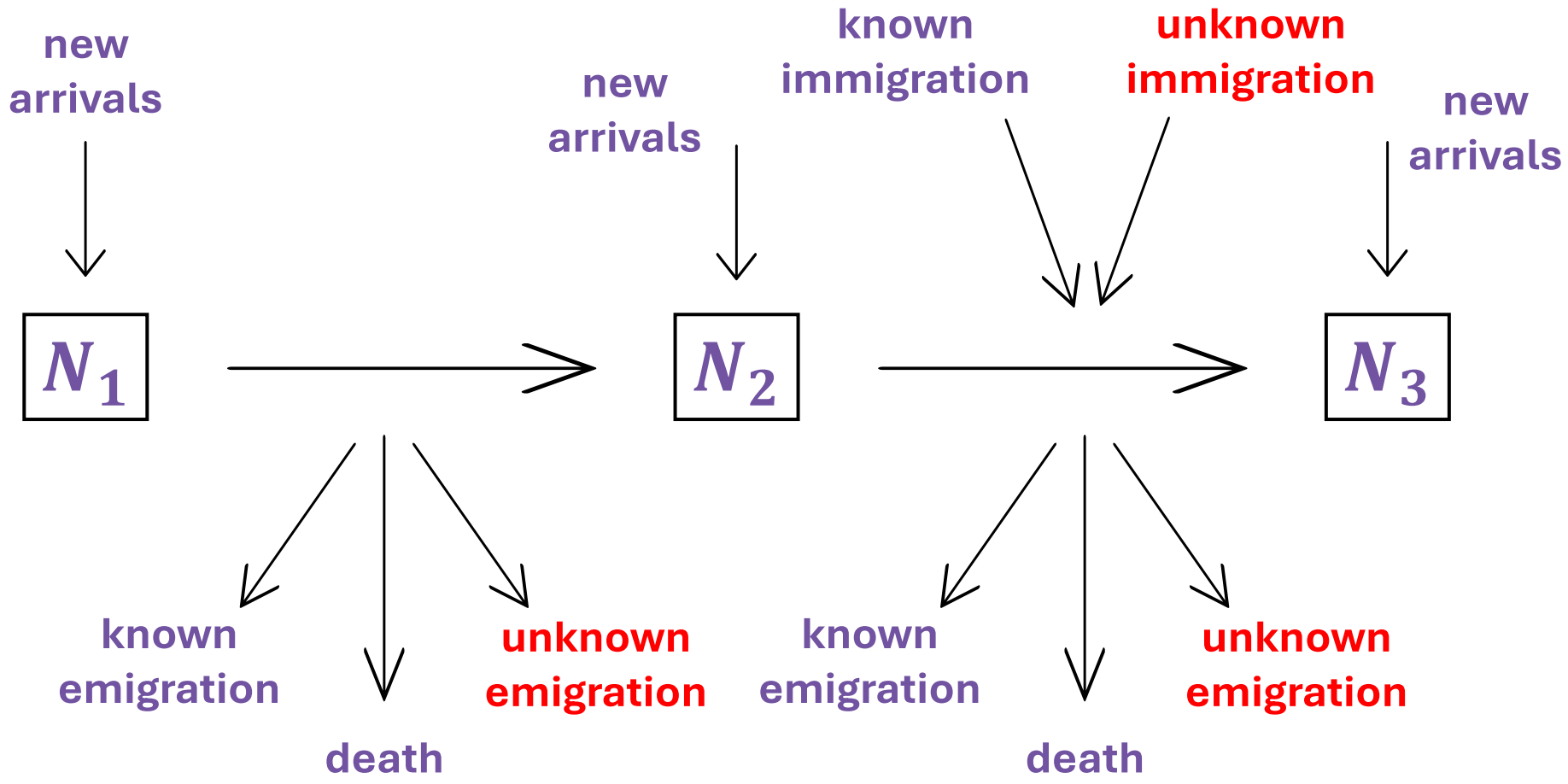
- Incredibly detailed results for each group of individuals
- Able to follow individuals every year they are present during the study, determining which state they are in (using the Viterbi Algorithm)
- Improvement on previous models

Limitations:

- We had near unlimited access to a high-performance computing cluster BIANCA at Uppsala University, allowing huge computational capacity and effectively perfect parallelisation. Realistically, access to similar resources is limited

Propose a population-level model

Overview of Population-level Model



Latent Class Models[3]

[3] Mariano Porcu and Francesca Giambona. Introduction to latent class analysis with applications. The Journal of Early Adolescence, 37(1):129–158, 2017. doi: 10.1177/0272431616648452.

2

- By design assume that multiple latent subpopulations (G) exist with differing observation probabilities
  - Accounts for individual heterogeneity, like FMMs
- Individual response patterns (register observation combinations) y_i for K registers are expressed as a mixture over G latent classes, with class specific mixing proportion π_g

$$P(Y_i = y_i) = \sum_{g=1}^{G} \pi_g \prod_{k=1}^{K} P(Y_{ik} = y_{ik} \,|C_i = g)$$

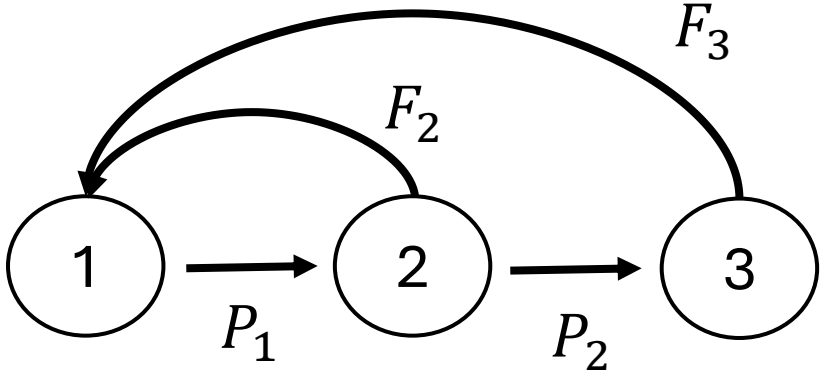
- Class specific response probabilities $P(Y_{ik} = y_{ik} \,|C_i = g)$ are specified using the same multicategory logit model as before, allowing incorporation of individual covariates and/or interactions

Matrix Population Models[4]

[4] Hal Caswell. Matrix population models: construction, analysis, and interpretation. Sinauer Associates, Inc, 2nd edition, 2001.

2

- A class of population models that utilise matrix algebra
- Applications in estimation and projection of populations in both ecology and demography
- Similar to HMM formulation of CR, but on a population-level with counts in each state



$$\underline{\mathrm{n}}(t+1) = A\underline{\mathrm{n}}(t)$$

Projection Matrix $\mathrm{A} = \begin{bmatrix} 0 & \mathrm{F}_2 & \mathrm{F}_3 \\ \mathrm{P}_1 & 0 & 0 \\ 0 & \mathrm{P}_2 & 0 \end{bmatrix}$

- Counts in each stage (geographical location and/or covariate stage) are observed, allowing transition probabilities to be estimated
- Covariate combinations are considered separately and transition between stages allowed

Current Work

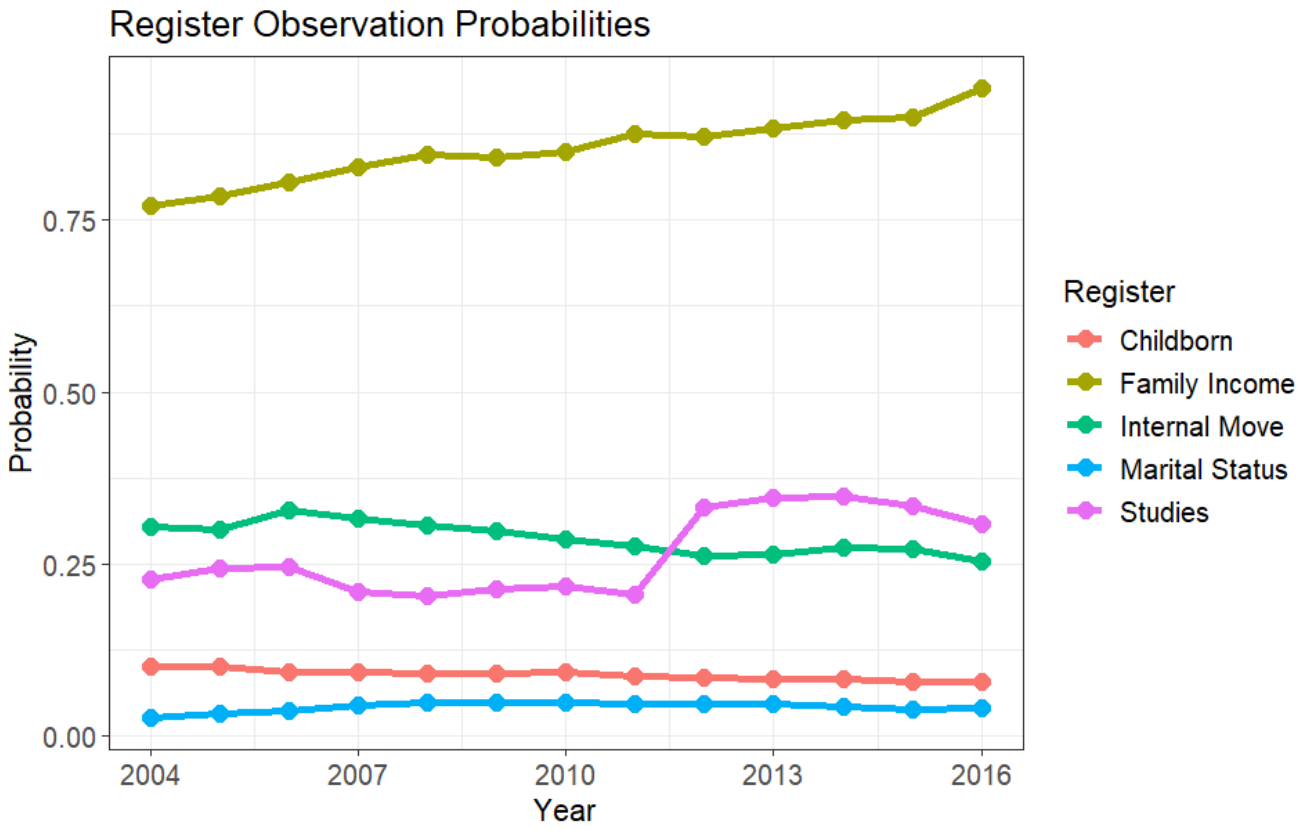
2

- Currently can fit LCM to each year then plug estimates into the MPM
  - LCM: fitted to each covariate group separately, estimates population size, allowing an estimate of the number of overcovered individuals each year
  - MPM: fitted to all years to estimate life event probabilities (survival, emigration, re-immigration and de-registration) for each covariate group, using states
    1. Present in the country
    2. Dead
    3. Outside the country and de-registered
    4. Overcovered (outside country but didn't de-register)

- Goal: combine models and optimise jointly

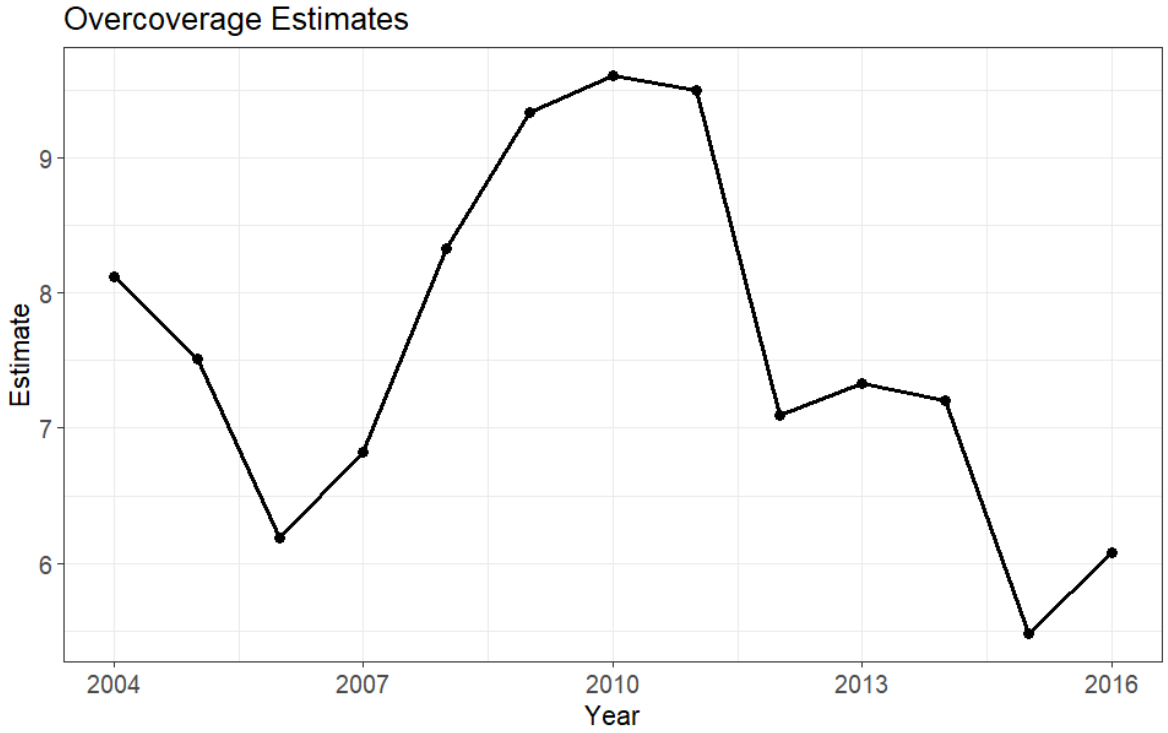
Thank you for your attention!

Email: lyb3@kent.ac.uk



Life Event Probabilities:

- Survival: 0.99862
- Emigration: 0.11450
- Re-immigration: 0.23231
- De-registration: 0.79870



Note: these are only preliminary estimates

- Different registers are used
- No covariates are incorporated
- Only 1 latent class is considered (no FMM equivalent)
- Uncertain sightings are not considered