A Capture-Recapture Model with Temporary Emigration for Estimating Population Size from Incomplete Registers

Lucy Y. Brown [1] Eleni Matechou [1] Bruno Santos [2]
Eleonora Mussino [2] Ruth King [3] Blanca Sarzo [4]

[1]School of Mathematics, Statistics and Actuarial Science, University of Kent

[2]Department of Sociology, Stockholm University

[3]School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh

[4]Foundation for the Promotion of Health and Biomedical Research of Valencia Region, FISABIO

Motivating Case Study: Sweden

- When people move to Sweden they must register with the Swedish Tax Agency and will receive a personal identification number
- They are equally required to de-register when leaving Sweden for at least one year however incentives and knowledge to do so are low
- This leads to serious bias in population estimates, negatively influencing policy-making and research

Swedish Register Data

- Access to administrative register data of the Swedish population
- Provided by the Swedish National Institute of Statistics (Statistics Sweden - SCB)
- All foreign born residents who first entered Sweden between the years 2003 and 2016 as adults
- We have record of the individual characteristics sex, age, country of birth and time since first entering Sweden
- We have a total of nine incomplete and overlapping registers on which individuals are either observed (1) or not observed (0)

Swedish Data

(This is not real Swedish data)

Nine Registers

emp	stud	intmove	faminc	amf	child
0	0	0	0	1	0
1	0	0	1	0	0
1	0	0	1	0	0
1	0	0	1	0	0
1	0	0	1	0	0
1	0	0	1	0	0
1	0	0	1	0	0

immig	emig	reimmig	marr	div	swecit
1	0	0	0	0	0
0	1	0	0	0	0
1	0	1	0	0	0
0	0	0	1	0	0
0	1	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

Emigration, (Re-)Immigration and Death records

	id	year	age	cob	sex	death
1	XXX1	2003	27	France	1	0
2	XXX1	2004	28	France	1	0
3	XXX1	2007	31	France	1	0
4	XXX1	2008	32	France	1	0
5	XXX1	2009	33	France	1	0
6	XXX2	2014	68	India	0	0
7	XXX2	2015	69	India	0	1

Covariates (treated as categorical):

- **Sex**
- **Country of birth**
- **Age**
- **Time Since first entering Sweden**

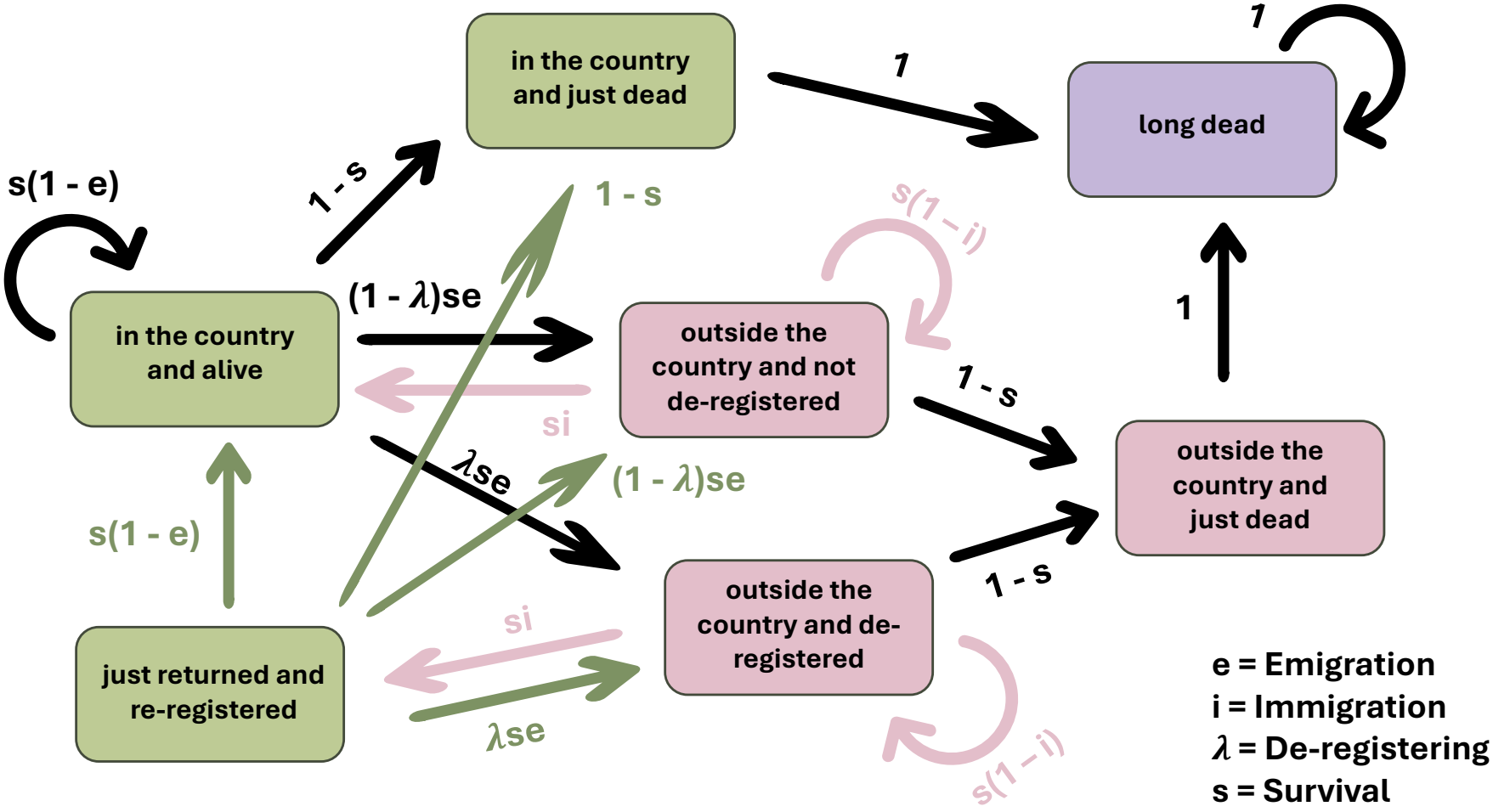
Previous Work

Current approaches include:

- Income-based exclusion method
- Register-trace approaches:
  - Cross-sectional register trace approach
  - Longitudinal approach
- (Supervisors' Work) Multiple Systems Estimation approach, looking at each year separately, Monti et al. [2020], Mussino et al. [2023]

I have used a longitudinal approach, following individuals over a period of time, as well as allowing temporary emigration to be incorporated.

States


in the country and just dead
long dead
1
1
s(1 - e)
1 - s
1 - s
s(1 − i)
in the country and alive
(1 - λ)se
outside the country and not de-registered
1
si
1 - s
λse
(1 - λ)se
outside the country and just dead
s(1 - e)
1 - s
si
outside the country and de-registered
just returned and re-registered
λse
s(1 − i)
e = Emigration
i = Immigration
λ = De-registering
s = Survival

Transition Matrix

$$\Gamma = \begin{bmatrix} s(1-e) & 1-s & \lambda se & (1-\lambda)se & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & s(1-i) & 0 & 1-s & si & 0 \\ si & 0 & 0 & s(1-i) & 1-s & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ s(1-e) & 1-s & \lambda se & (1-\lambda)se & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

e: emigration probability

i: immigration probability

s: survival probability

λ: de-registering probability

Observation Matrix

An individual's observation at some point in time is entirely conditional on the state they are in.

- Observations 1 to 2^L: occurring only when an individual is "in the country and alive", these are **register observations** coming from a multicategory logit model, Agresti [2007]
- Observation 2^L: any individual outside the country or when "long dead" will be **unobserved** with probability 1
- Observation $2^L + 1$: when an individual dies inside the country they will be **recovered dead** with probability 1
- Observations $(2^L + 2)$ to $(2(2^L) + 1)$: an individual re-entering the country (having previously de-registered) will be observed **re-registering** *plus* can appear on some register combination

Observation Matrix

$$\Omega = \begin{bmatrix} p_{111} & \dots & p_{100} & p_{000} & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & p_{111} & \dots & p_{000} \\ 0 & \dots & 0 & 1 & 0 & 0 & \dots & 0 \end{bmatrix}$$

p_{111}: observed in all 3 registers

p_{100}: observed in register 1 only

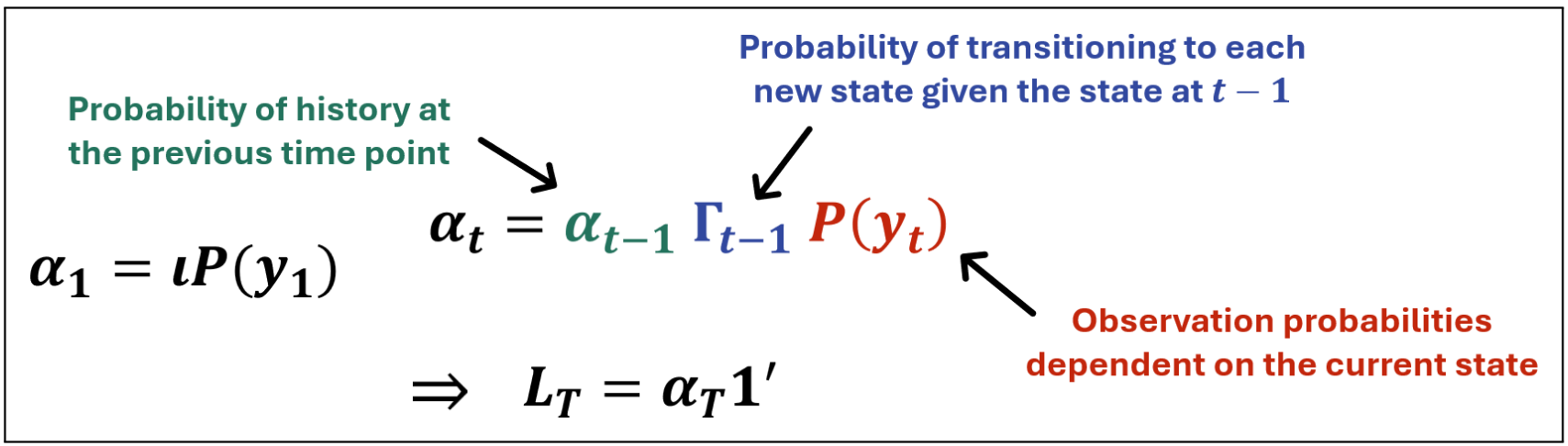
p_{000}: observed in no registers/unobserved, etc

Likelihood Calculation: HMMs

We marginalise over the latent states and obtain a marginal likelihood for each individual j:

$$L_T = \iota P(y_1)\Gamma_1 P(y_2)\Gamma_2 ... \Gamma_{T-1} P(y_T)1'$$

Using the forward algorithm, define $\alpha_t(j) = Pr(y_1, ..., y_t, z_t = j)$.



Model Overview

- Capture-recapture model for multiple incomplete and overlapping observation registers
- Allows for Temporary Emigration
- Uses Hidden Markov Models (HMMs) and the forward algorithm for Likelihood calculation
- Allows for all two-way interactions between registers
- Incorporates individual covariates (in survival and observation probabilities)
- Allows for interactions between registers and individual covariates

Case Study Model Specification

In order to run the model for the **full population** we need to make simplifications. In this case study we will use the following model:

- Three observation registers:
  - Employment
  - Being linked to any household income
  - Enrolment in *Arbetsförmedlingen*, i.e. active unemployment
- Sex and Country of Birth covariates in survival probability
- Sex incorporated into observation probabilities
- Interactions between registers, as well as between sex and registers

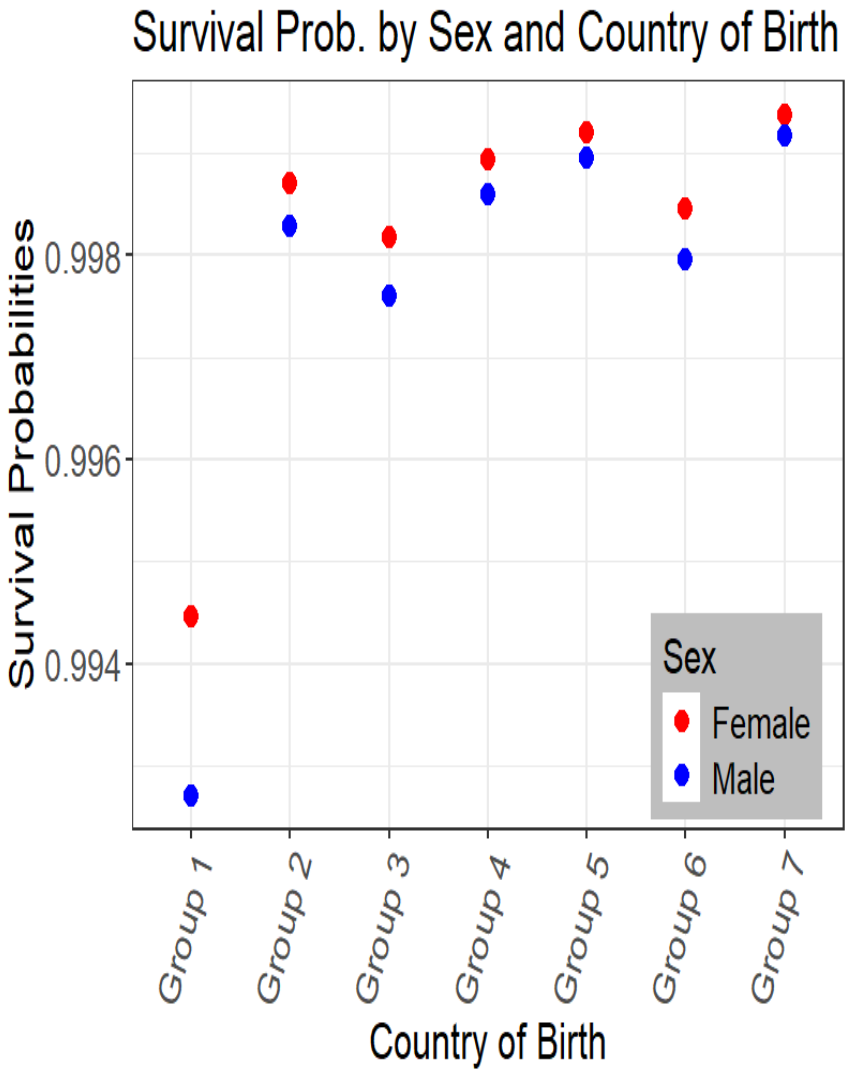
Preliminary Results

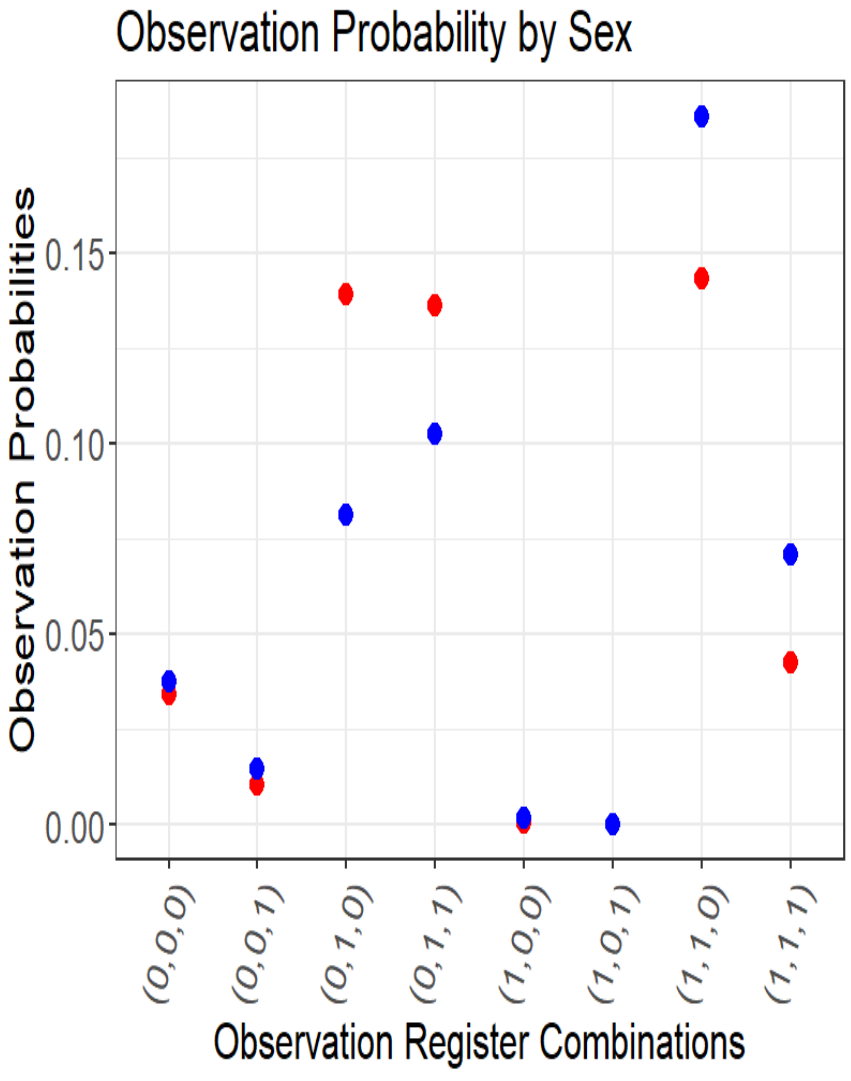
We fit this model using maximum likelihood estimation, using C++, obtaining the following preliminary results:

Probability	Notation	Estimate	95% CI
Emigration	e	0.0684	$(0.0681, 0.0686)$
De-registering	λ	0.3880	$(0.3823, 0.3937)$
Re-immigration	i	0.0296	$(0.0292, 0.0299)$
		Time:	10639.85 sec $\approx$ 3 hours

These estimates are consistent with previous studies' estimates

Preliminary Results





Future Work

- **Next Task:** Experiment with methods to obtain an estimate of population size, e.g. Viterbi Algorithm.
- Currently working on incorporating time-dependent covariates (age and time since first entering Sweden) and running for all nine registers (i.e. the full model)
- Exploring other applications
  - We have access to the equivalent data from Norway
  - Only slight adaptation to the model would be needed
- Considering possible model extensions, such as the incorporation of dependence between family units

Thank you for your attention!

Email: lyb3@kent.ac.uk

References

Alan Agresti. *An introduction to categorical data analysis*. John Wiley & Sons, 2nd edition, 2007.

Olivier Gimenez. *Bayesian analysis of capture-recapture data with hidden Markov models: Theory and case studies in R and NIMBLE*. 2023. URL `https://oliviergimenez.github.io/banana-book/index.html`. Last Updated 2023-08-26.

Jeffrey Lee Laake. Capture-recapture analysis with hidden markov models. In *AFSC Processed Rep. 2013-04*, 2013.

Andrea Monti, Sven Drefahl, Eleonora Mussino, and Juho Härkönen. Over-coverage in population registers leads to bias in demographic estimates. *Population Studies*, 74(3):451–469, 2020.

Eleonora Mussino, Bruno Santos, Andrea Monti, Eleni Matechou, and Sven Drefahl. Multiple systems estimation for studying over-coverage and its heterogeneity in population registers. *Quality & Quantity*, pages 1–24, 2023.