A Latent Class Analysis Approach to Multiple Systems Estimation with Longitudinal Register Data

Lucy Y. Brown Eleni Matechou Bruno Santos Eleonora Mussino

Abstract

Overcoverage occurs when individuals are registered as living in a country but in fact live elsewhere or have passed away and their death was not recorded, leading to serious bias in demographic rates, negatively influencing policymaking and research. We propose an approach that extends pre-existing methods to estimate the true population size and overcoverage using official population registers. The proposed approach builds on Latent Class Models (LCMs) with a Multiple Systems Estimation (MSE) framework, and allows knowledge exchange over multiple time points, providing an efficient model-fitting framework for longitudinal data. The model will be employed on data from Sweden and Norway, providing new insights into population dynamics and individual trajectories.

Background Overcoverage occurs when individuals are registered as living in a country but in fact live elsewhere (imperfect emigration registration) or have passed away and their death was not recorded (imperfect death registration). Overcoverage can lead to serious bias in demographic rates, negatively influencing policymaking and research, Monti et al. [2020].

In Sweden, as well as in many other European countries, all individuals whose actual, or planned, primary residence is within the country for at least one year are required to register with the Swedish Tax Agency, becoming part of the Register of Total Population (RTB), or equivalent for other countries. Upon registration, a personal identification number is assigned, which is necessary for various life activities such as accessing banking and housing; these high incentives to register result in minimal undercoverage (incorrectly excluding individuals from the population). Individuals are equally required to de-register when leaving Sweden, but a combination of lack of knowledge and low incentives mean many individuals do not, resulting in overcoverage (incorrectly including individuals in the population), Andersson et al. [2023].

Recent work in Sweden has focused on estimating the true population size, and in turn overcoverage, by using population registers. A "zero personal income approach" has been suggested in which individuals with no personal income from a variety of sources in a given year are excluded and they are assumed to be outside the country (Aradhya et al. [2017]; Weitoft et al. [1999]). Statistics Sweden have also developed "register trace" approaches, in which individuals are traced across a series of registers and excluded if not present in a given year/series of years (Statistics Sweden [2015]; Statistics Sweden [2018]). Monti et al. [2020] compared these approaches, finding that precision is low with overcoverage likely to be overestimated by all approaches; furthermore, they found that application to other countries may be difficult due to lower-quality registration systems.

Alternatively, Mussino et al. [2023] propose a model based on log-linear models for contingency tables, using the multiple systems estimation (MSE) framework, i.e. multiple incomplete and overlapping registers. The contingency table contains the count of individuals with each possible combination of register observations, then uses a log-linear model to estimate the number of individuals unobserved on all registers considered. While this model allows for the inclusion of individual covariates and interactions between registers, it only considers annual snapshots of the register data and not individuals themselves over time. For this project, we have extended MSE to create a longitudinal approach, following groups of individuals, and hence registers, over different years, as well as allowing for individual heterogeneity in the observation process.

Data In this project we have access to official administrative register data of the Swedish and Norwegian populations, including birth, death and migration registers. We currently focus on data relating to all foreign-born residents who first entered the study country as adults during a specific time period, e.g. $2003 - 2016$, appearing on a number of available, incomplete, overlapping and possibly interacting registers.

Each year an individual is present in the country, we have a record of their observation in a wide range of registers such as change in marital status, birth of a child, internal moves within the country, enrolment in higher education and income from a range of sources (including household income). We also have data relating to their individual covariates, collected at time of registration, specifically sex, which is treated as binary, as well as country of birth, age, time since first entering the country and reason for immigration, which are all treated as categorical.

Model

Individual heterogeneity refers to variation between individuals in a population and could be a result of behavioural differences, physical traits, spatial heterogeneity or temporal heterogeneity. For example, not all individuals in a population have the same probability of being observed; some individuals naturally have a higher observation probability than others. Failure to incorporate individual heterogeneity may lead to bias in

parameter estimates and a misjudgement of the size of effect of individual covariates (inflated type I error rate), Gimenez and Choquet [2010].

To account for individual heterogeneity, we consider latent class models (LCMs) (Porcu and Giambona [2017]; Forcina [2008]). LCMs suppose that there are G latent (unobserved) subgroups in the population, each with differing observation probabilities; for example, group 1 may have a high probability of being observed while group 2 may have a low probability. The model then works to cluster individuals into these subgroups based on observable characteristics, such as their observation/response patterns and individual characteristics (covariates).

In LCMs individual response patterns (register observation combinations) y_i are expressed as a mixture over G latent classes. Each class g has a mixing proportion π_g and class specific response probabilities $P(Y_{ik}|C_i = g)$ for each item/register j. In the most standard model, for each individual response pattern y_i, local independence is generally assumed for responses y_{ik} across items $k = 1, ..., K$, given the latent class g, giving the following response pattern probabilities:

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

LCMs are a well established tool in social sciences where the identification of underlying groups is of interest but difficult to specify, for example, when studying mental health symptoms, Lanza et al. [2013]. Di Cecco et al. [2018] also addresses population size estimation in the presence of overcoverage and multiple lists/registers using LCMs, however, they adopt a different approach. We propose using LCMs to account for individual heterogeneity in the observation process, while Di Cecco et al. [2018] use LCMs to directly model overcoverage by identifying a target population. Many extensions have been made to the LCM literature such as the incorporation of categorical covariates and the relaxation of local independence assumptions.

We model register combination observation probabilities using a multicategory logit model, specifically a baseline-category logit model for nominal responses, Agresti [2007]. We consider R interacting observation registers and C dummy variables to model the effect of categorical covariates, which we treat as additional registers, resulting in a total of $J = [2^R \times$ (number of covariate categories)] categories, or register combinations an individual may fall into. For example, if we include only the sex covariate, we would have $J = 2^R \times 2$ as sex is binary, with observations "observed in all registers and male", "observed in all registers and female" and so on. Individuals can only be observed (1) or not observed (0) in each register/covariate category, thus, we consider the positions of these zeros and ones, as well as the product of each pair of observations to account for all two-way interactions, creating new columns in the design matrix X. Columns in this matrix are grouped into three: the first R columns correspond to R registers, the next C columns correspond to the covariate categories, and the remaining columns correspond to all two-way interactions between registers and between registers and covariates.

$$X = \begin{bmatrix} 1 & ... & 1 & 1 & ... & 1 & 1 & ... & 1 \\ 1 & ... & 1 & 1 & ... & 0 & 1 & ... & 0 \\ & \vdots & & & \vdots & & & \vdots & \\ 0 & ... & 0 & 0 & ... & 0 & 0 & ... & 0 \end{bmatrix} \tag{2}$$

Therefore, we model observation probabilities using this design matrix, alongside a vector of coefficients $\gamma^{(g)}$ for each latent subgroup g, resulting in the following model observation probabilities:

$$p_j^{(g)} = \frac{\exp(\gamma^{(g)} X_{j*})}{\sum_h \exp(\gamma^{(g)} X_{h*})} \qquad j = 1, ..., J; \qquad h = 1, ..., J \tag{3}$$

From the data we are able to obtain an absolute lower and upper bound for the true population size each year using death and migration records; this is illustrated in Figure 1.

When considering an estimate on population size there are two key aspects of our data: (1) we only consider foreign born individuals, and (2) we only consider individuals who enter for the first time during our study. These properties mean we can be certain of the population size in the first year of the study as the population consists only of individuals who first entered that year, i.e. $N_1 = n_1$. Additionally, we have death records d, emigration/de-registration records e_d and re-immigration/re-registration records i_d for each year of the study which can be incorporated as follows to produce an upper bound for the true population size:

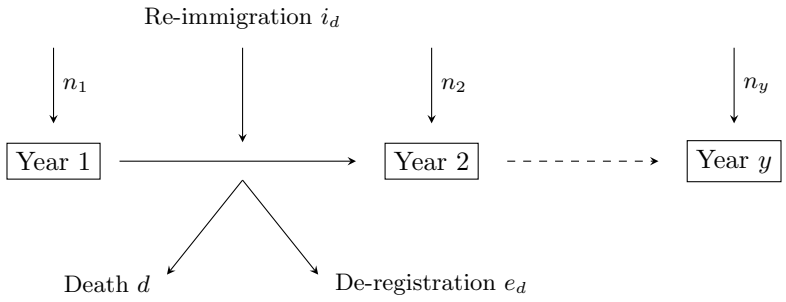

Figure 1: This diagram illustrates the observable processes influencing population size. Each year in the study a known number of individuals enter the country n_i, and a known number of individuals die d, emigrate (having de-registered) e_d and re-immigrate (having previously de-registered) i_d.

$$
\begin{aligned}
N_1 &= n_1 \\
N_2 &= N_1 - d - e_d + n_2 + i_d \\
&\vdots \\
N_y &= N_{y-1} - d - e_d + n_y + i_d
\end{aligned}
$$

In reality, we do not have knowledge of all individuals who leave the country each year which is why we can only determine an upper bound, i.e. we do not know how many individuals leave without de-registering e_o, and in turn how many of these individuals re-enter i_o. Traditionally in MSE, each year is considered individually and an estimate of the number of individuals unobserved in all registers is made. This estimate can then be compared to the number of individuals unobserved in the data and in turn estimate overcoverage. We propose a model that will similarly look at each year of the study separately using a LCM but then share this information over all years in the study and incorporate the idea of transitioning between states over a period of time using, for example, matrix population models, Caswell [2001]. We consider a total of eight latent states an individual can be in at any point in time: (1) in the country and alive, (2) in the country and just dead, (3) just left the country and de-registered, (4) outside the country having de-registered, (5) outside the country but did not de-register, (6) outside the country and just dead, (7) just returned having de-registered, (8) long dead. Considering these states and the idea of "transitioning between them" also allows us to consider and estimate the probability of different life events: survival s, emigration e, re-immigration i and de-registering λ, which can be specified using individual characteristics using logistic regression. This results in a longitudinal MSE approach for open populations, allowing for temporary emigration and individual heterogeneity.

Discussion We propose an efficient approach for estimating population size from incomplete, overlapping registers which utilises LCMs for interacting registers and extends them over time. Our model accounts for temporary emigration and uses multinomial regression to incorporate an arbitrary number of registers, whilst accounting for individual heterogeneity in the observation and latent processes.

Recently, Santos et al. [2024] proposed a capture-recapture (CR) model to estimate overcoverage in Sweden, using register data, however they employed a Bayesian framework where all latent variables are sampled at each MCMC iteration, resulting in a computationally expensive model. Due to this, Santos et al. [2024] used a 5% sample of the total population. Our previous work (currently unpublished) similarly proposed a CR model, however, it was based in a hidden markov model formulation and was able to run for the full dataset as well as incorporate individual heterogeneity. One disadvantage was that as this model works at an individual level it does not run in a reasonably short amount of time to be used by demographers. Therefore, we propose a new model that works at a population level yet still incorporates individual characteristics and multiple time points.

While we do not expect this new model to produce results as detailed or individual-specific as our previous (currently unpublished) version, it operates efficiently at a population level. This balance between detail and runtime makes it practical for demographers to adapt and apply in real-world or official settings for population level analyses.

References

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

Gunnar Andersson, Andrea Monti, and Martin Kolk. Vem bor här?: en eso-rapport om gamla och nya folkräkningar. *Rapp. till Expertgr. Stud. i Offent. Ekon*, 2(2023), 2023.

Siddartha Aradhya, Kirk Scott, and Christopher D Smith. Repeat immigration: A previously unobserved source of heterogeneity? *Scandinavian Journal of Public Health*, 45(17_suppl):25–29, 2017.

Hal Caswell. Matrix population models, 2001.

Davide Di Cecco, Marco Di Zio, Danila Filipponi, and Irene Rocchetti. Population size estimation using multiple incomplete lists with overcoverage. *Journal of Official Statistics*, 34(2):557–572, 2018.

Antonio Forcina. Identifiability of extended latent class models with individual covariates. *Computational Statistics & Data Analysis*, 52(12):5263–5268, 2008.

O Gimenez and R Choquet. Individual heterogeneity in studies on marked animals using numerical integration: capture–recapture mixed models. *Ecology*, 91(4):951–957, 2010.

Stephanie T Lanza, Xianming Tan, and Bethany C Bray. Latent class analysis with distal outcomes: A flexible model-based approach. *Structural equation modeling: a multidisciplinary journal*, 20(1):1–26, 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.

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.

Bruno Santos, Eleonora Mussino, Sven Drefahl, and Eleni Matechou. Using population register data and capture-recapture models to estimate over-coverage in sweden. 10 2024. doi: 10.17045/sthlmuni.27323550.v1.

Statistics Sweden. Övertäckning i registret över totalbefolkningen—en registerstudie [overcoverage in the total population register—a register study]. *Befolkning och Välfärd*, 1, 2015.

Statistics Sweden. The registration bias—a methodological report on the estimation of over-coverage, under-coverage and registration at the wrong address. *Swedish:" Folkbokföringsfelet. En metodrapport om skattning av övertäckning, undertäckning och folkbokförda på fel adress.") Örebro: SVB BV/REG and PMU/MIÖ*, 2018.

G Ringbäck Weitoft, Anders Gullberg, Anders Hjern, and Måns Rosén. Mortality statistics in immigrant research: method for adjusting underestimation of mortality. *International journal of epidemiology*, 28(4): 756–763, 1999.