
Heaped Age Adaptive Model (HAAM) for Mitigating Age Heaping in Demographic Data

Journal Title
XX(X):1–28
©The Author(s) 2025
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/

SAGE

Santosh Kudtarkar¹

Abstract

Age heaping—systematic misreporting of ages due to digit preference or culturally salient ages—introduces distortions into demographic and survey data. Existing methods primarily address heaping at terminal digits (0 or 5) and are not designed to detect irregular or age-specific heaping patterns. We introduce the Heaped Age Adaptive Model (HAAM), a penalized EM framework that jointly estimates a smooth latent true age distribution and age-specific misreporting behavior. HAAM integrates PRISMA, a Poisson-robust smoothing procedure, with a flexible misreport kernel whose parameters (α, β) govern the attractiveness and locality of reported ages. An ℓ_1 sparsity penalty enables the model to adaptively identify a small set of genuinely heaped ages without imposing any predetermined heaping structure. Across simulation studies containing both digit-based and irregular heaped ages, HAAM closely recovers the true age distribution, removes artificial spikes, and correctly locates the ages that attract disproportionate reports. Compared with classical digit-heaping indices and graduation techniques, HAAM provides both superior correction and richer diagnostic insight. This model offers a general, data-driven tool for mitigating age heaping in demographic, epidemiological, and historical datasets.

Keywords

Age Heaping; EM Algorithm; Poisson Estimator; Smoothing

¹Centre for Mathematical Modelling, FLAME University Pune, India

Corresponding author:

Santosh Kudtarkar, Centre for Mathematical Modelling, FLAME University, Pune-412115, India
Email: sant@flame.edu.in

Introduction

Censuses and large-scale surveys are the foundation of demographic and social science research, and the accuracy of reported ages is central to the validity of population estimates (Manual 1983; Moultrie et al. 2013). Yet age misreporting remains pervasive. Such misreporting can be unintentional, when respondents lack precise knowledge of their birth date, or intentional, when individuals deliberately round or exaggerate ages due to cultural, social, or institutional incentives (Moultrie et al. 2013). For example, individuals below the legal age of marriage or military enlistment may overstate their age, while older adults may inflate their ages to qualify for retirement benefits. These errors distort population structures, leading to systematic biases that complicate demographic analyses.

One of the most studied manifestations of age misreporting is age heaping, a phenomenon in which certain ages occur disproportionately relative to their neighbours (Spoorenberg and Dutreuilh 2007). The most common form is digit preference, where ages ending in particular digits (notably 0 and 5) are over-reported (Whipple 1919; Myers 1954; Bachi 1951). Age heaping has long served both as an indicator of data quality and as a proxy for numeracy, since populations with poor educational attainment often exhibit strong preferences for round numbers (Manual 1983; A'Hearn et al. 2009). For example, historical European data show heavy digit preference in pre-industrial societies (A'Hearn et al. 2009; Mikołaj et al. 2018), while recent African census data reveal persistent heaping on terminal digits despite improvements in literacy (Ferber and Baten 2025). Similarly, Latin American censuses have documented spikes at ages 30, 40, and 50, reflecting strong digit preferences in age reporting (Ramachandran 1965; Shryock and Siegel 1971). This phenomenon is not restricted to poorer countries: census data in the United States also show a preference for ages ending in 0/5 (National Academies of Sciences, Engineering, and Medicine 2023; Zelnik 1961).

However, digit preference does not capture all age misreporting patterns. In many populations, specific ages attract disproportionate reporting even when they do not end in 0 or 5. This phenomenon—sometimes termed heaped age preference—reflects cultural or institutional landmarks rather than mere digit bias. For instance, the 1960 Philippines census recorded spikes at age 18 (legal adulthood) and age 12 (school entry), in addition to the conventional heaping at multiples of five (Myers et al. 1940; Manual 1983). Chinese census data show heaping not only at 0s and 5s but also at retirement thresholds (see (Gu and Feng 2019) and references therein), while South Asian data have documented anomalies around culturally significant ages such as 60 or 65, associated with seniority and pension eligibility (Spoorenberg and Dutreuilh 2007). These heaped ages often arise from cultural norms, institutional thresholds, or policy incentives, and they cannot be fully captured by indices designed only for digit-based heaping.

Over the past century, demographers have developed a wide variety of indices to quantify age heaping. Whipple's index (Whipple 1919) measures excess reporting of ages ending in 0 or 5 within the 23–62 range, with a value of 100 indicating perfect reporting and higher scores reflecting increasing heaping. Myers' blended index (Myers et al. 1940) extends this approach by measuring preference across all ten terminal digits. Further refinements include Bachi's index (Bachi 1951), Carrier's index (Carrier

1959), and Ramachandran’s index (Ramachandran 1965), while the United Nations Age–Sex Accuracy Index (United Nations 1952; Manual 1983) integrates age heaping with sex ratio distortions. More recent innovations include Spoorenberg’s (Spoorenberg and Dutreuilh 2007) modification of Whipple’s index to evaluate arbitrary digits and Noubissi’s (Noubissi 1992) generalized digit-specific index. Collectively, these tools provide useful diagnostics for data quality assessment.

Yet these measures share important limitations. First, they are largely descriptive, offering indices of overall heaping severity rather than identifying which specific ages are disproportionately reported. Second, they are typically constrained to digit-based heaping, overlooking age preferences tied to institutional or cultural thresholds. Hence, they are structurally ill-equipped to detect isolated or irregular heaping. Whipple’s index is by definition blind to heaps at non-standard digits (such as a *coming-of-age* spike at 21). Similarly, Myers’ blended index assumes that a preference for a digit (e.g., “0”) is uniform across the lifespan. Consequently, a large, isolated institutional heap at age 50 is conceptually flattened by averaging it with accurate reporting at ages 30, 80, or 90. This averaging dilutes the signal of specific heaped ages, treating distinct institutional shocks as mere digit preference. Finally, these indices do not provide a framework for correcting distorted distributions but only for evaluating their quality (Shryock and Siegel 1971; Office 2008).

This motivates the need for a more adaptive, model-based approach. The Heaped Age Adaptive Model (HAAM) introduced in this study builds on the need for flexible methods that allow the data itself to reveal the heaped ages, rather than assuming *a priori* that only digits ending in 0 or 5 are affected. By systematically identifying and adjusting for both digit-based and age-specific heaping, HAAM provides a unified framework that enhances demographic data quality assessment and correction. This innovation extends the utility of traditional indices while addressing their blind spots, offering a more robust tool for modern demographic analysis.

In this work, we introduce a novel model-based approach that simultaneously smooths the age distribution and infers the pattern of age misreporting (both digit-based and age-specific). HAAM extends the penalized composite link model (PCLM) framework for demographic data ungrouping (Camarda et al. 2007, 2008) by incorporating an explicit misreporting model for age heaping and embedding it within a Poisson-robust smoothing procedure (PRISMA) for the latent true age distribution. HAAM has several advantages: (1) it produces a corrected age distribution (at single-year intervals) that is smooth and demographically plausible by penalizing irregular fluctuations; (2) it estimates age-specific heaping propensities, allowing identification of which ages were overstated or understated; and (3) it imposes an ℓ_1 sparsity penalty on these propensities, effectively picking out a sparse set of heaped ages that concentrate the heaping, thereby capturing age preference patterns beyond terminal-digit rules. To our knowledge, this is the first approach to use a sparse regularization technique to adaptively discover heaped ages, rather than assuming them *a priori*. By using an Expectation–Maximization (EM) algorithm, our method treats true ages as latent variables and alternates between inferring the true age distribution and updating the misreporting parameters (Dempster et al. 1977; McLachlan and Krishnan 2008). The underlying assumption—common to many heaping adjustments—is that the

true age distribution varies smoothly with age, and that deviations from smoothness in the observed data are attributable to heaping errors.

The paper is organized as follows. In the Methods section, we formulate HAAM mathematically, describing the misreporting model, the penalized likelihood, and the EM algorithm for parameter estimation. We also discuss how the PRISMA smoother, within the PCLM framework, is used to enforce smoothness on the age distribution at each M-step. In the Results section, we apply HAAM to simulated data with known heaping patterns, demonstrating its ability to recover the true distribution and detect heaped ages. We compare these results with traditional methods and highlight how HAAM automatically identifies non-digit-based heaping (age-specific preferences). We conclude by discussing the model’s assumptions, possible extensions (such as two-dimensional age–sex heaping models), and potential applications in historical demography and survey analytics. Mathematical details of the derivation of the optimization steps are provided in the Appendices.

Materials and Methods

Model Framework

We now give provide a detailed explanation of the model. Our goal is to estimate the true age frequency distribution $\theta = (\theta_0, \theta_1, \dots, \theta_K)$ (where K is the maximum age, e.g. 100) from the observed heaped counts $\mathbf{y} = (y_0, y_1, \dots, y_K)$. We treat the true ages of individuals as latent (unobserved) and model the reporting process that produces \mathbf{y} from θ . Specifically, we assume that each individual of true age a reports an age k according to some probability $P(k | a)$. Then the expected observed counts can be written as a linear transformation of the true counts:

$$\mu_k = \sum_{a=0}^K P(k | a) \theta_a, \quad \text{for each reported age } k \quad (1)$$

where μ_k is the expected number of observations reported at age k . In matrix–vector form, $\boldsymbol{\mu} = P \boldsymbol{\theta}$, where P is a $(K + 1) \times (K + 1)$ reporting matrix with entries $P_{k,a} = P(k | a)$. In our model, we posit a parametric form for $P(k | a)$ that captures age heaping behaviour.

We assume that an individual of true age a either reports their age correctly or misreports it, with a bias that decays with distance from the true age. Let α_k be a parameter representing the baseline log-attractiveness of the reported age k , and let β_k be a parameter controlling how quickly the probability of misreporting drops off as the distance $|k - a|$ between true and reported age increases (a larger β_k means the attractiveness of reporting k is concentrated in the immediate neighbourhood of the true age a , whereas a smaller β_k allows misreports from further away). We define an attractiveness function for reporting k when the true age is a :

$$A_{k,a} = e^{(\alpha_k - \beta_k |k-a|)}, \quad \text{for all ages } k \text{ and each true age } a. \quad (2)$$

Here $A_{k,a}$ can be thought of as an unnormalised weight for someone of actual age a to report age k . The form of $A_{k,a}$ implies that, for a given true age a , reporting age k

becomes exponentially less likely as the distance $|k - a|$ increases, consistent with the idea that misreporting typically involves rounding to nearby ages rather than to ages far away. The intercept-like term α_k raises or lowers the overall attractiveness of age k .

To ensure that probabilities sum to 1 for each true age a , we include a baseline probability of reporting correctly versus incorrectly. For each age a , there are $K + 2$ competing channels: a baseline “tell the truth” channel of unit weight and $K + 1$ “heap to k ” channels with weight $A_{k,a}$. Normalising them leads to the following definition of the misreporting probability:

$$P(k | a) = \frac{\delta_{k,a} + A_{k,a}}{1 + \sum_j A_{j,a}}, \quad (3)$$

where $\delta_{k,a}$ is the Kronecker delta. Here $\delta_{k,a}$ represents the weight of truthful reporting: even if all attraction terms are tiny, individuals can still report their exact age through this unit-mass channel. The term $A_{k,a}$ represents the weight of the heaping channel whose destination age is k . The parameter α_k raises or lowers the overall tendency for all ages to be reported as k ; larger α_k makes k a stronger contender for heaping. The parameter β_k controls locality: larger β_k concentrates misreports near k (steeper decay with distance), whereas smaller β_k permits longer jumps. When $k = a$, $A_{k,a} = e^{\alpha_a}$, which implies that the true age can be arrived at via both the truthful channel and the “heap to a ” channel, whose contributions simply add. The normalising term in Eq. (3) has two parts: the 1 is the unit weight of the truthful channel, and $S_a = \sum_j A_{j,a}$ can be interpreted as the total pull exerted by all ages j on someone whose true age is a .

The misreport model above is quite general and can represent both digit preference and heaped-age preference patterns. For example, consider a true age $a = 49$ in our simulated scenario that will be introduced in a later section. Empirically, we made age 49 prone to misreport as 50. In our model, this corresponds to α_{49} being relatively large (making misreporting likely) and β_{49} being moderate so that $A_{50,49} = \exp(\alpha_{49} - \beta_{49}|50 - 49|) = \exp(\alpha_{49} - \beta_{49})$ is one of the largest misreport weights. Ages $k = 48$ or 50 (distance 1) would have the largest $A_{k,49}$; ages at distance 2 (47 or 51) would have $A_{k,49} = \exp(\alpha_{49} - 2\beta_{49})$, which, if β_{49} is not too small, will be substantially lower than at distance 1. Thus, the model inherently centres misreports around the true age. A heaped-age preference (e.g. a tendency for ages 49, 51, 52, etc. to all report 50) emerges not from a single age’s parameters but from a combination: if multiple adjacent ages all have increased α_a and thereby tend to shift their reports towards a common age (50), that age 50 will appear as a heaped reported age with an excess of reports. Our model captures this as several ages (48, 49, 51, 52 in this case) having elevated misreport propensities α_a . One can interpret a “heaped age” in the data as a reported age with an unusually large influx of misreports from neighbouring actual ages. We show how the model can identify such heaped ages after fitting, by examining the fitted parameters and the distribution of misreport probabilities.

We now derive the expression for the complete-data likelihood. Suppose we treat the true ages of all individuals as known for the moment. The data would then consist of counts $W_{k,a}$ representing the number of individuals who are of actual age a and reported age k . (In practice, we only observe the sums $y_k = \sum_a W_{k,a}$ for each reported age k , while the

allocation among actual ages is latent). Given our model, the chance for an actual- a person to be recorded in cell (k, a) is $P(k | a)$, so the complete-data log-likelihood, assuming Poisson–multinomial sampling for counts, can be written as (see Appendix S1):

$$\ell_{\text{complete}} = \sum_{a=0}^K \sum_{k=0}^K W_{k,a} \log P(k | a) + \text{constant} \quad (4)$$

subject to the constraint that $\sum_k W_{k,a} = \theta_a$ (the total actual count of age a). Substituting the expression for $P(k | a)$, this becomes

$$\begin{aligned} \ell_{\text{complete}} &= \sum_{a,k} W_{k,a} \log \frac{\delta_{k,a} + A_{k,a}}{1 + S_a} \\ &= \sum_a W_{a,a} \log(1 + A_{a,a}) - \sum_a \hat{\theta}_a \log(1 + S_a), \\ &\quad + \sum_{a,k: k \neq a} W_{k,a} (\alpha_k - \beta_k |k - a|) \end{aligned} \quad (5)$$

where we have used $\sum_k W_{k,a} = \hat{\theta}_a$, which denotes the estimated count of true age a . This is the starting expression for the objective function that we will use.

Estimating True Age Counts (PRISMA) and Regularization

Fitting the model involves estimating the true age counts θ_a and the heaping parameters α_a and β_a for all ages. This is a high-dimensional problem: for $K + 1$ age categories, there are $K + 1$ θ parameters and $2(K + 1)$ misreport parameters, minus some constraints (one degree of freedom is lost because the total population is fixed by $\sum_a \theta_a = \sum_k y_k$, and extremely large α, β could be unidentifiable without regularization). To obtain meaningful and stable estimates, we incorporate penalty terms that reflect our prior assumptions:

1. The true age distribution θ_a should vary smoothly with a (because, barring heaping, population age structures are usually smooth functions of age) (Cleveland 1979);
2. Most ages should not be heavily prone to misreporting—we expect only a subset of ages to be heaped points, so we impose a sparsity-inducing penalty on α_a to shrink most of them towards 0;
3. We add a weak ridge penalty on all α_a and β_a to aid numerical stability and identifiability.

Within HAAM, smoothing of the latent true age distribution is carried out by PRISMA, a Poisson-robust iterative smoother. PRISMA is an integral component of the model: at each EM iteration it produces a stable, demographically plausible estimate of θ , so that the misreporting parameters α, β are used to capture genuine heaping rather than random noise or small-scale irregularities.

Conceptually, PRISMA (Poisson-Robust Iterative Smoother with count-preserving Adaptation) de-spikes and smooths discrete count distributions (e.g. single-year ages)

while preserving totals and realistic tail behaviour. It combines local averaging with robust down-weighting of upward outliers so that artificial heaping spikes contribute little to the estimate, yet genuine features (including multiple peaks) are retained. The smoother uses edge-aware triangular kernels whose effective window size automatically shrinks near the boundaries, avoiding reflection or padding beyond the observed support and thereby maintaining tail shape (Fan and Gijbels 1996; Abramson 1982; Jones 1993). The local bandwidth is adjusted according to a simple curvature measure, narrowing in high-curvature regions and widening in flat regions, so that true peaks and troughs are not oversmoothed.

In settings where boundary ages must be preserved exactly (e.g. when early ages are anchored by high-quality vital registration), PRISMA can optionally hard lock the first and/or last few ages. In this mode, smoothing is applied only to the interior ages, with radii computed relative to the unlocked interior so that borrowing never crosses the lock boundary. After smoothing, a multiplicative rescaling restores the exact interior mass, leaving locked ages unchanged.

A final, global rescale enforces exact equality between the smoothed and observed totals, ensuring comparability in downstream analyses. In practice, PRISMA is fast (its complexity is linear in the number of ages times a small window size) and governed by a small set of intuitive tuning parameters (base half-window, number of robust iterations, robustness constant, and optional hard-lock lengths). Its main design features—one-sided robustness, edge-aware kernels, adaptive bandwidth, interior-only finishing, and mass preservation—draw on classical ideas in robust and nonparametric smoothing (Cleveland 1979; Fan and Gijbels 1996; Abramson 1982; Boyd and Vandenberghe 2004; Benjamini and Hochberg 1995), but their synthesis for spike-resistant, de-heaping-specific smoothing is particular to this work. The mathematical details and full algorithm are provided in Appendix S3.

Sparsity penalty for α (ℓ_1 norm). To encourage only a few ages to have large misreport tendencies, we add a penalty on the ℓ_1 norm of the α vector. A smooth approximation is used to avoid non-differentiability at 0. Specifically, we penalise $\sum_a \sqrt{\alpha_a^2 + \delta}$, where δ is a small constant (e.g. 10^{-6}) to make the function differentiable at $\alpha_a = 0$. This behaves like an ℓ_1 (absolute value) penalty on α_a , inducing sparsity (many α_a driven to near 0) (Cleveland 1979; McLachlan and Krishnan 2008). The tuning parameter for this penalty is $\gamma \geq 0$. In our experiments we set $\gamma = 0.1$; this value was chosen to sufficiently shrink minor heaping propensities while allowing a few α_a to remain nonzero for heaped ages. Ages that do not strongly improve the likelihood by having $\alpha_a > 0$ are therefore shrunk towards zero, leaving only ages that truly contribute to explaining heaping to stand out with large α_a . This implements our prior belief that age heaping is concentrated at relatively few ages.

Ridge penalty for α and β (ℓ_2 norm). Additionally, a small ridge (quadratic) penalty $\frac{\rho}{2} \sum_a (\alpha_a^2 + \beta_a^2)$ is included to discourage extreme values and ensure numerical identifiability. Here ρ is a very small number (we used $\rho = 10^{-4}$). This penalty is mainly for regularization of the optimisation and does not materially bias the estimates given its small magnitude.

Combining sparsity and ridge penalties, the final penalised log-likelihood objective function $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is given by

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \ell_{\text{complete}} - \gamma \sum_a \sqrt{\alpha_a^2 + \delta} - \frac{\rho}{2} \sum_a (\alpha_a^2 + \beta_a^2) \quad (6)$$

where ℓ_{complete} is the complete-data log-likelihood given in Eq. (5). Because the observed-data likelihood does not have a closed form due to the latent variables $W_{k,a}$, we maximize \mathcal{L} via the Expectation–Maximization algorithm, which iteratively handles the latent assignments $W_{k,a}$.

EM Algorithm for HAAM

We now describe the Heaped Age Adaptive Model (HAAM) estimation procedure as an EM algorithm. The EM algorithm is well suited to problems with latent data; here $W_{k,a}$ (the mapping of observed counts to true ages) is the latent allocation. EM alternates between computing the expected values of the latent counts (E-step) and maximizing the penalised expected complete-data log-likelihood (M-step).

E-step. In iteration t , given current parameter estimates $\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}$, we compute the expected fraction of each observed count y_k that originated from actual age a . Essentially, we calculate the conditional probability $\Pr(\text{actual} = a \mid \text{observed} = k)$ for each (k, a) pair and then multiply by y_k . The expected $\widehat{W}_{k,a}^{(t)}$ is the expected count of observations reported as k that truly belong to age a . Using Bayes' rule:

$$\widehat{W}_{k,a}^{(t)} = y_k \frac{\hat{\theta}_a P^{(t)}(k \mid a)}{\sum_{a'} \hat{\theta}_{a'} P^{(t)}(k \mid a')}$$

for all k, a . Here $P^{(t)}(k \mid a)$ is computed from the current $\alpha_a^{(t)}, \beta_a^{(t)}$ via the formula in the previous section. In matrix form, let $P^{(t)}$ be the reporting matrix at iteration t and $\boldsymbol{\mu}^{(t)} = P^{(t)} \hat{\boldsymbol{\theta}}$ the current fitted values for \mathbf{y} . Then the above can be written as

$$\widehat{W}_{k,a}^{(t)} = P_{k,a}^{(t)} \hat{\theta}_a \frac{y_k}{\mu_k^{(t)}}.$$

In implementation, we form the entire matrix $W^{(t)}$ of size $(K+1) \times (K+1)$. Each column a of \widehat{W} (varying k) represents the probabilistic redistribution of the observed counts across possible actual ages for those who are truly age a . Summing over observed k in that column yields $\sum_k \widehat{W}_{k,a} = \hat{\theta}_a \sum_k P_{k,a}^{(t)} \frac{y_k}{\mu_k^{(t)}} = \hat{\theta}_a \sum_k \frac{\mu_k^{(t)}}{\mu_k^{(t)}} = \hat{\theta}_a$, so the column sums of \widehat{W} are constrained to $\hat{\boldsymbol{\theta}}$. Likewise, summing row k of \widehat{W} gives $\sum_a \widehat{W}_{k,a} = y_k$, so the row sums equal the observed counts. Thus \widehat{W} provides a “soft assignment” of each observed count to actual ages. Here, note that $\hat{\theta}_a$ is estimated by the PRISMA algorithm before the start of the E-step.

M-step. In the M-step, we maximize \mathcal{L} in Eq. (6) with respect to α_a and β_a for each age a , including their penalty terms. This can be done by taking partial derivatives or using a numerical optimiser. We chose to use numerical optimization (the L-BFGS-B algorithm, which allows simple bounds) to update all α_a and β_a jointly, as the objective is smooth and concave (the expected log-likelihood is concave in these parameters given W). For a given $\hat{\theta}$ and $\widehat{W}^{(t)}$ at iteration t , we maximise

$$\mathcal{L}^{(t)} = \ell_{\text{complete}}^{(t)} - \gamma \sum_a \sqrt{(\alpha_a^{(t)})^2 + \delta} - \frac{\rho}{2} \sum_a ((\alpha_a^{(t)})^2 + (\beta_a^{(t)})^2). \quad (7)$$

where the superscript denotes the " t "th iteration. To perform this optimization efficiently, we derive the gradient with respect to $\alpha_m^{(t)}$ and $\beta_m^{(t)}$ which are given by,

$$\frac{\partial \mathcal{L}^{(t)}}{\partial \alpha_m^{(t)}} = \sum_a \widehat{W}_{m,a}^{(t)} - \frac{\widehat{W}_{m,m}^{(t)}}{1 + A_{m,m}^{(t)}} - \sum_a \hat{\theta}_a \frac{A_{m,a}^{(t)}}{1 + S_a^{(t)}} - \gamma \frac{\alpha_m^{(t)}}{\sqrt{(\alpha_m^{(t)})^2 + \delta}} - \rho \alpha_m^{(t)} \quad (8)$$

$$\frac{\partial \mathcal{L}^{(t)}}{\partial \beta_m^{(t)}} = - \sum_a |m - a| \widehat{W}_{m,a}^{(t)} + \sum_a \hat{\theta}_a \frac{|m - a| A_{m,a}^{(t)}}{1 + S_a^{(t)}} - \rho \beta_m^{(t)}. \quad (9)$$

We use these gradients in a limited-memory BFGS optimiser (Fletcher 2000) with bound constraints $\alpha_a^{(t)} \geq 0$, $\beta_a^{(t)} \geq 0$ (the bounds are naturally handled by L-BFGS-B). The result of this procedure yields the updated $\alpha^{(t+1)}$ and $\beta^{(t+1)}$. In practice, because of the concavity of \mathcal{L} , this optimisation converges rapidly (typically a few dozen iterations of L-BFGS-B). If the optimiser does not fully converge due to strict tolerances, we still take the parameters it returns (since EM does not require a global maximum at each M-step, only an increase in likelihood).

After the two-part M-step, we have new estimates $(\alpha^{(t+1)}, \beta^{(t+1)})$. One EM iteration is thus completed.

In summary, the pseudocode version of the algorithm is given in Algorithm 1.

Algorithm 1 HAAM EM Algorithm

- 1: **Initialize:** Set $\theta_a^{(0)} = y_a + 0.5$ for all a (adding a small constant to avoid zeros). Set $\alpha_a^{(0)} = \epsilon$ and $\beta_a^{(0)} = \epsilon$ for all a , with $\epsilon = 10^{-2}$.
- 2: **PRISMA:** Use the PRISMA algorithm to obtain $\hat{\theta}$.
- 3: **for** $t = 0, 1, 2, \dots$ until convergence **do**
- 4: **E-step:** Compute

$$\widehat{W}_{k,a}^{(t)} = y_k \cdot \frac{\hat{\theta}_a P^{(t)}(k | a)}{\mu_k^{(t)}}, \quad \mu^{(t)} = P^{(t)} \hat{\theta}. \quad (10)$$

(This gives the expected allocation of reported age- k counts to estimated true age a .)

- 5: **M-step: Update α, β :** With $\hat{\theta}$ fixed, update $(\alpha^{(t)}, \beta^{(t)})$ by maximizing

$$\mathcal{L}^{(t)} = \ell_{\text{complete}}^{(t)} - \gamma \sum_a \sqrt{(\alpha_a^{(t)})^2 + \delta} - \frac{\rho}{2} \sum_a ((\alpha_a^{(t)})^2 + (\beta_a^{(t)})^2) \quad (11)$$

using $\widehat{W}^{(t)}$ in $\ell_{\text{complete}}^{(t)}$. Optimisation is performed numerically (e.g. L-BFGS-B) with $\alpha^{(t)}, \beta^{(t)} \geq 0$.

- 6: **Check convergence:** If

$$\max \{ \|\alpha^{(t+1)} - \alpha^{(t)}\|_{\infty}, \|\beta^{(t+1)} - \beta^{(t)}\|_{\infty} \} < \text{tol},$$

or the relative log-likelihood change $< \text{tol}$, then **stop**. Otherwise, set $t := t + 1$ and repeat.

- 7: **end for**

Convergence criteria. We monitor the maximum change in parameter values and the increase in log-likelihood. Specifically, we compute the maximum absolute difference in any parameter (α_a or β_a) between iterations. If this is below a tolerance (we used $\text{tol} = 10^{-4}$), or if the relative increase in the penalised log-likelihood is below a threshold, the algorithm is deemed converged. We also set a cap on the number of iterations (e.g. 1000) to prevent infinite loops.

Identifying Heaped Ages Post-Estimation

After convergence, we obtain estimates $\hat{\alpha}$ and $\hat{\beta}$. The smoothed $\hat{\theta}$ is our best estimate of the true age distribution, corrected for heaping. The estimated parameters α_a and β_a describe the misreporting pattern for each true age a . However, it is useful to summarise which ages were the major ages of heaping—that is, which ages attracted an unusually large share of misreported individuals (regardless of whether those ages themselves are heavily misreported). A “heaped age” in terms of reported data is one with an observed excess even after smoothing the distribution.

We define a heaped-age score for each age x (which can be considered either as a true or reported age, since in our formulation ages are indexed consistently) using a simple geometric criterion.

Spike in the observed distribution. We examine the second difference of the fitted observed distribution (the model’s fitted μ_k). If μ_x is much higher than the local trend, that indicates a spike at x . We compute a discrete second difference

$$D_x^{(2)} = \mu_{x-1} - 2\mu_x + \mu_{x+1}$$

(with appropriate handling at boundaries). If μ_x is a local maximum, this second difference will be negative (concave down). Define a heaped score $S_x = -D_x^{(2)}$ as a measure of “peakness” at age x (so S_x is high if x is a spike). Ages with high S_x correspond to spikes in the fitted distribution. Thus, an age x that is itself a heaped reported age would have a high spike score S_x (because many neighbours report as x), while its neighbours $x - 1$ and $x + 1$ would tend to have lower S_{x-1} or S_{x+1} (meaning they “prefer” x over themselves). We then sort the values of S_x in decreasing order to identify the top- p heaped ages. Alternatively, one can define a cutoff value α^* and classify ages with $S_x > \alpha^*$ as heaped ages, since $S_x \approx 0$ for most non-heaped ages.

Results

Overall model fit on simulated data

We applied HAAM to the simulated dataset described in Appendix S4, which exhibits both conventional digit-preference heaping at terminal digits 0 and 5 and additional heaped-age spikes at ages 21, 40, 50, 60, and 70 (Figure 1). The algorithm converged in 7 iterations for a tolerance of 10^{-5} , satisfying the convergence criterion.

The estimated true age distribution $\hat{\theta}_a$ from HAAM is plotted in Figure 2, alongside the fitted distribution, the observed (heaped) distribution, and true values (without heaping) distribution. The HAAM-corrected distribution is almost indistinguishable from the true distribution, indicating that the model successfully recovers the underlying population by removing artificial spikes.

Goodness of fit

To quantify the goodness of fit, we use the Normalised Root Mean Squared Error (NRMSE), defined as

$$\text{NRMSE} = \frac{\text{RMSE}}{\theta_{\max} - \theta_{\min}} \quad \text{where} \quad \text{RMSE} = \sqrt{\frac{1}{N} \sum_i (\theta_i - \hat{\theta}_i)^2} \quad (12)$$

where N is the number of age groups and $\theta_{\max}, \theta_{\min}$ are the maximum and minimum true counts across ages. For the fitted model, the NRMSE is less than 1%, meaning that the typical deviation between θ_i and $\hat{\theta}_i$ is small relative to the overall range of the true distribution.

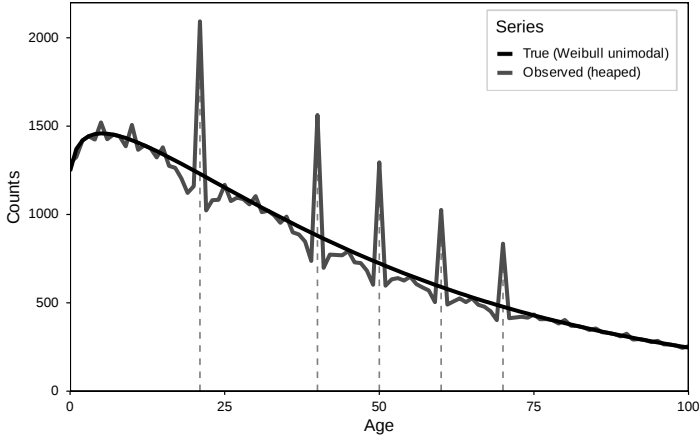


Fig 1. Simulated age distribution. Baseline counts follow a Weibull-shaped age profile with added random perturbations at ages ending in 0 or 5 and strong heaped-age spikes at 21, 40, 50, 60, and 70.

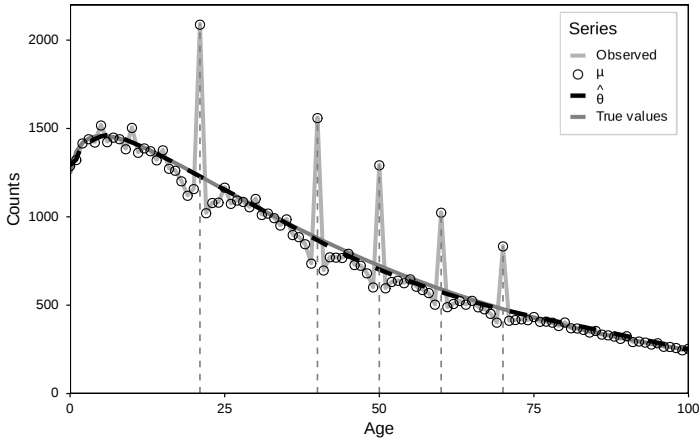


Fig 2. True, observed, and HAAM-fitted age distributions. The true distribution (smooth curve) is distorted by heaping in the observed counts (step curve), while the HAAM-corrected distribution $\hat{\theta}$ (green dotted line) closely recovers the truth. The fitted observed counts $\hat{\mu}$ from the EM algorithm are also shown.

Most of the remaining minor discrepancy occurs around age 0, where the corrected distribution slightly overshoots the true count. At the lower age boundary, the smoother has only right-side neighbours—typically larger counts on an increasing curve—so the local average is pulled upward, producing classic positive boundary bias. PRISMA’s curvature-triggered bandwidth shrinkage preserves the steep rise at young ages but further concentrates the averaging on these larger neighbours, reinforcing this effect. Because the

final mass-preserving rescale is global, it cannot remove such local edge bias. A small boundary correction (e.g. a minimum tail radius, gentler robust weights near the boundary, or a local-linear edge fit) eliminates this overshoot while leaving the interior behaviour unchanged. Overall, the corrected counts match the true counts extremely well across all ages.

Interpreting the heaping parameters

We now examine the estimated heaping parameters. The fitted values of α_k and β_k provide insight into how the model explains the observed heaping.

Figure 3 plots the estimated $\hat{\alpha}_k$ for each age k . As expected, there are clear spikes at the heaped ages, superimposed on a smoothly varying background that declines at high ages.

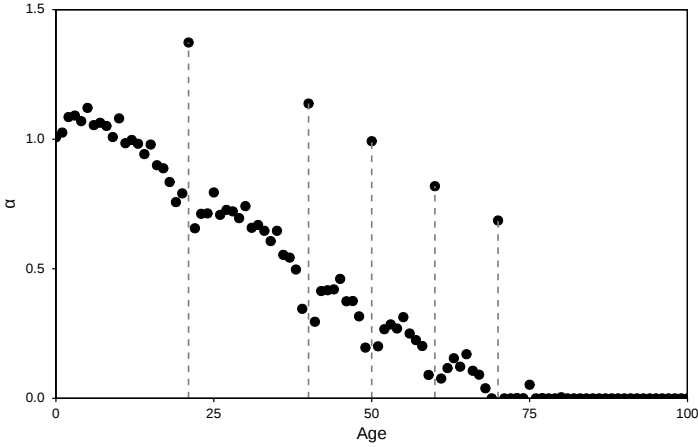


Fig 3. Estimated baseline attractiveness parameters $\hat{\alpha}_k$ for each reported age k . Local peaks occur at heaped ages (21, 40, 50, 60, 70), with a smoothly varying background that tapers off at older ages.

Although Eq. (6) includes an ℓ_1 -type penalty on α_k , in our empirical fits most α_k remain strictly positive. Two modelling details explain this: (i) the penalty is implemented as $\sqrt{\alpha_k^2 + \delta}$ (with small $\delta > 0$), which is a smooth approximation to $|\alpha_k|$ and therefore does not induce exact zeros; and (ii) the non-negativity constraint $\alpha_k \geq 0$, together with the column normalisation in $P(k | a)$, encourages a positive baseline level of attractiveness across ages. Consequently, α should be read not as a sparse selector but as a salience field over labels: focal ages are identified by relative elevations of α_k above a smooth background trend rather than by zeros elsewhere. In summary, α is best interpreted as a dense attractiveness field with local peaks. The peak heights modulate $P(k | k)$, while β controls how quickly this influence decays with distance from k . Figure 4 shows the estimated $\hat{\beta}_k$, which governs the locality of attraction around each age. For ages with strong heaping (e.g. 21, 40, 50 and 70), $\hat{\alpha}_k$ is elevated and $\hat{\beta}_k$ tends to be moderate, producing bright but finite-width attraction bands in $P(k | a)$. Ages that do not act as

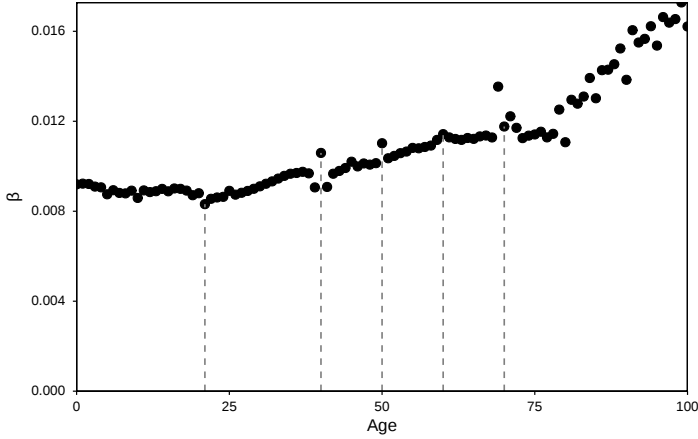


Fig 4. Estimated locality parameters $\hat{\beta}_k$ for each reported age k . Larger β_k values indicate more localized heaping (rapid decay of attraction with distance), whereas smaller β_k values allow misreports from a wider band of true ages.

heaping focal points have lower $\hat{\alpha}_k$ and/or larger $\hat{\beta}_k$, implying weaker or more localized attraction.

Reporting matrix heatmap

The estimated misreporting probabilities $P(k | a)$ are displayed in Figure 5 as a heatmap. The bright diagonal reflects correct reporting,

$$P(a | a) = \frac{1 + e^{\alpha_a}}{1 + \sum_{j=0}^K e^{\alpha_j - \beta_j |a-j|}},$$

which typically exceeds any single off-diagonal entry in the same column. Its level varies smoothly with age via the denominator (competition from other attractive labels). The bright horizontal bands at specific rows ($k = 21, 40, 50, 60, 70$) indicate that a range of true ages is pulled toward the same reported label k . Band intensity is governed by α_k (baseline attractiveness) and band *thickness* by β_k (locality/decay):

$$P(k | a) \propto e^{\alpha_k - \beta_k |a-k|}.$$

Larger α_k makes the band brighter; larger β_k makes it narrower. An interpretable width is the half-life distance $d_{1/2}(k) = \ln(2)/\beta_k$, the distance at which the attraction to label k halves.

The spread of each column away from the diagonal visualises misreporting uncertainty for that true age: tight columns imply precise reporting, whereas columns intersected by strong horizontal bands indicate substantial heaping pressure toward those focal labels. Because $P(k|a)$ is not symmetric (rows and columns encode different conditionals), the

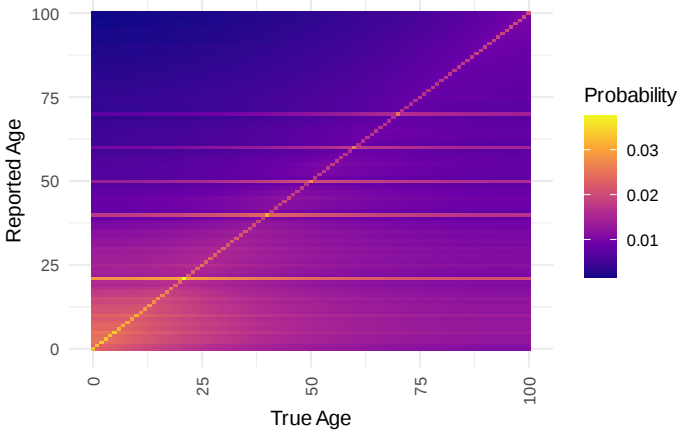


Fig 5. Heatmap of fitted misreporting probabilities $P(k | a)$. Columns correspond to true ages a , rows to reported ages k . The bright diagonal represents correct reporting, while horizontal bands at focal labels (e.g. 21, 40, 50, 60, 70) indicate attraction from a range of true ages. Band intensity is controlled by α_k and band thickness by β_k .

heatmap exhibits a diagonal ridge plus a small number of horizontal filaments. This “ridge plus bands” geometry is characteristic of focal attraction: *height* is set by α , and *width* by β .

Identifying heaped ages

Using the Spiked heaped-age scoring method described in the Materials and Methods section, the model flagged ages 21, 40, 50, 60 and 70 as the top 5 heaped ages by a wide margin. Figure 6 The score for age 70 which is ranked fifth was roughly 3.5 times higher than that of the sixth-ranked age (10, which has a digit-preference spike but far smaller than the spike at 70). This is consistent with the known simulation truth. The high scores reflect both the strong local peak and the fact that neighbouring ages (e.g. 49 and 51 for the heaped age at 50) exhibit low scores, indicating that they “prefer” reporting the focal age rather than themselves as seen in the negative extreme values around heaped ages. In a real dataset, such ages would warrant special attention or external validation (for instance, investigating whether a particular age corresponds to a pension eligibility threshold or an enumeration artefact).

Comparison with traditional methods

We next compare HAAM with more conventional approaches to heaping correction. A typical strategy would redistribute counts at ages ending in 0 or 5, possibly in combination with a graduation method such as the Karup–King approach (Carrier 1959). Such a method would smooth out the spikes at 21, 40, 50, 60 and 70 to some extent but would treat them similarly to ages 25, 30, 35, 45, 65 etc. In our simulated data, however, ages the identified

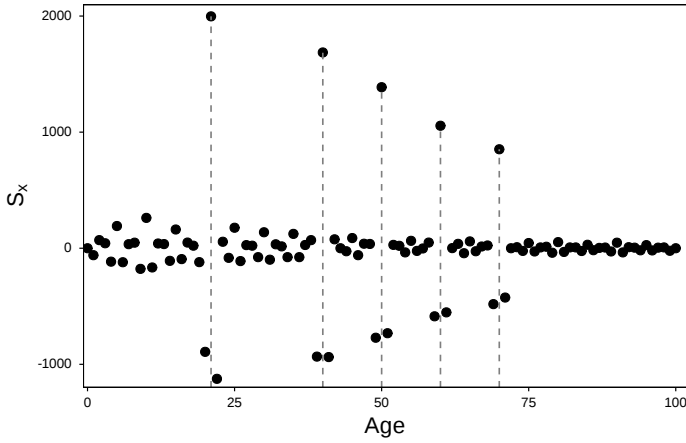


Fig 6. Curvature-based heaped-age spike scores S_x by age, showing pronounced peaks at the simulated heaped ages (21, 40, 50, 60, 70).

heaped ages have much larger spikes (due to accumulation from multiple neighbours) than, say, age 45. A uniform smoothing of all 0/5 ages therefore underestimates how much of the spike at the identified heaped ages is artifactual.

Indeed, when we applied a simple moving-average smoother targeting all ages ending in 0 or 5 equally, the correction at the identified heaped ages was insufficient: the “corrected” distribution still had visible bumps at these ages. In contrast, HAAM recognised via the data likelihood that 21, 40, 50, 60 and 70 were exceptional (requiring much larger α_k and substantial inflow from neighbouring ages to explain the observed spikes) and therefore reallocated substantially more individuals away from these focal ages. The result was a more accurate reconstruction of the true distribution.

Robustness checks

We performed additional robustness checks by varying the random seed of the simulation and altering the magnitude of misreport probabilities. In all cases, HAAM accurately identified the strongest heaped ages and yielded a corrected distribution close to the truth.

One informative scenario was obtained by reducing the sparsity penalty to $\gamma = 0$ (no ℓ_1 penalty on α). In this case, the model tended to spread small misreport propensities across many ages, achieving a similar overall fit but without clearly separating heaped from non-heaped ages. The corrected distribution remained accurate, but the interpretation of heaped ages became harder, as many ages had small but nonzero α_a . This underscores the importance of the sparsity penalty in clearly isolating heaped ages.

Conversely, when we increased γ by an order of magnitude, the model underfit: the penalty forced almost all α_a towards zero and attributed nearly all variation to the smooth θ , failing to fully remove the spikes. Thus, a moderate value of γ (we found 0.1 in these

scale units to be reasonable) is necessary to balance fit and sparsity, yielding both a good reconstruction of the true age distribution and a clear identification of focal heaped ages.

Discussion

We have presented the Heaped Age Adaptive Model (HAAM), a new method for adjusting age distributions for heaping by jointly estimating a smooth true distribution and age-specific misreporting parameters. Our simulation study demonstrated that HAAM can effectively detect and correct both digit-based and heaped-age heaping. In this section, we discuss the model’s advantages, interpretability, assumptions and limitations, and its relationship to existing approaches.

Advantages over traditional techniques. Unlike post hoc graduation methods that apply generic smoothing to an observed distribution, HAAM is a model-based approach that explicitly exploits information in age-heaping patterns to guide redistribution. Traditional methods (e.g. Kannisto’s method (Kannisto et al. 1999), Strong’s method (ARRIAGA et al. 1968), Karup–King (Carrier 1959) interpolators) smooth out fluctuations but do not model the mechanism of heaping. Consequently, they may oversmooth in some regions and undersmooth in others, especially when heaping is highly uneven or concentrated at particular ages.

HAAM, by contrast, directly models the misreporting mechanism via the kernel $P(k | a)$. This allows it to allocate the excess at heaped ages to the most plausible source ages (typically neighbouring ages). For example, in the case of a spike at age 70, HAAM reallocates those counts primarily to ages 69 and 71, rather than spreading them across a wide window. A generic smoother might distribute them more broadly or symmetrically, potentially distorting the local shape of the distribution, for instance around ages 65–75. The penalised likelihood framework also opens the door to uncertainty quantification for the estimated true counts (e.g. via the observed information matrix or bootstrap), which is usually absent in deterministic graduation methods.

Interpretability of heaped ages. An important outcome of our model is the ability to identify which ages are over-reported and to quantify their pull on neighbouring ages. This has direct demographic interpretations. For instance, in some historical data age 100 is often a heaped age due to prestige in reporting longevity; HAAM would be able to detect such a pattern if ages in the 90s had a tendency to report as 100. In demographic data-quality assessment, knowing the specific heaped ages can guide targeted cleaning, imputation, or sensitivity analyses.

The model can also shed light on cultural or institutional factors. For example, one might discover that in a given census ages ending in 8 attract excess reports (as sometimes hypothesised in East Asian contexts where the number 8 is considered auspicious, though empirical evidence is mixed). Traditional indices such as Myers’ blended index (Myers et al. 1940) can indicate a preference for digit 8 if present, but HAAM can go further by quantifying how many individuals shifted to ages ending in 8 and from which true ages they came.

Assumptions and limitations. HAAM assumes that the true age distribution is smooth and that irregularities are primarily the result of misreporting. This assumption can be violated if there are genuine demographic shocks (e.g. wars, pandemics, or migration events) that create real discontinuities or troughs in the age structure. In such cases, there is a risk that the model may attribute a true irregularity to heaping (or conversely, interpret heaping as a real demographic feature). One way to mitigate this is to incorporate external information or constraints—for example, fixing $\alpha_a = 0$ for ages where heaping is unlikely (such as young ages reported by parents, or ages with strong civil registration), or constraining selected age ranges based on prior demographic knowledge.

In its current form, the model treats each age’s misreporting parameters (α_a, β_a) as independent across ages, except for the global ℓ_1 and ridge penalties. We do not impose smoothness on α_a or β_a across age. It could be reasonable, however, to assume that misreporting tendencies change gradually with age, apart from isolated heaped points. A mild smoothness penalty on α_a (e.g. an ℓ_2 penalty on differences $\alpha_{a+1} - \alpha_a$) might improve stability in very sparse data. We did not implement this here in order to preserve the flexibility to allow sharp, isolated spikes in α_a at specific ages.

Another limitation is the assumption of independence of individual age reports (essentially a Poisson or multinomial sampling model). In reality, ages often come in household clusters, and certain survey contexts may induce correlated reporting errors. Incorporating such clustering is beyond the scope of the present work but could be addressed by a hierarchical extension of HAAM (for example, modelling heaping at the household or interviewer level in surveys where one person reports ages for multiple household members).

Relationship to PCLM-based approaches. Our work was inspired by the Penalized Composite Link Model (PCLM) approach developed by Camarda and co-authors (Camarda et al. 2007, 2008). In their application to age heaping, they construct misclassification matrices that encode specific digit-preference patterns (for example, ages ending in 9 or 1 moving to the nearest 0, and similar rules for ages near 5). The structure of $P(k | a)$ is thus specified a priori through known link matrices, and a penalised likelihood is solved for the latent distribution and a small number of pattern parameters. They demonstrate, for instance on the 1960 Philippines census, that this approach can accurately recover known misreport patterns.

HAAM generalises this idea by not imposing any particular heaping pattern in advance: in principle, every age can have its own pattern, and the ℓ_1 penalty then suppresses those that are not needed. In this sense, HAAM allows the data to reveal whether it conforms to a simple digit-preference pattern or whether additional, age-specific heaping is present. In a dataset with pure digit preference and no other heaped ages, we would expect HAAM to recover a pattern similar to that obtained using Camarda’s PCLM framework, but in a more data-driven manner. The trade-off is a larger parameter space, which necessitates stronger regularization to avoid overfitting.

Implications and future directions. In conclusion, the Heaped Age Adaptive Model provides a flexible and powerful framework for addressing age heaping. By integrating smoothing penalties and adaptive sparsity, it avoids overfitting while allowing the data

to reveal irregular heaping patterns. In our simulations, HAAM was able to correct even extreme heaped-age spikes, recovering true distributions that standard methods struggled to reconstruct. We anticipate that HAAM can be a valuable tool both for re-estimating demographic indicators from biased data and for studying the mechanisms of age misreporting themselves.

As data quality continues to be recognised as a central issue—even in modern censuses, where some heaping has been documented in recent enumerations (Bureau 2023)—methods such as HAAM can help practitioners make better use of imperfect data. Future work could apply HAAM to real datasets from different regions and time periods, such as historical European censuses with well-known heaping (A’Hearn et al. 2009; Mikołaj et al. 2018) or contemporary surveys in South Asia and Africa (Spoorenberg and Dutreuilh 2007; Ferber and Baten 2025), to catalogue heaped-age preferences and improve demographic estimates. Extensions to two-dimensional settings (e.g. joint age–sex heaping models) and to hierarchical structures (e.g. household- or cluster-level heaping) also represent promising directions for further research.

Declaration of conflicting interests

The author has declared that no competing interests exist.

Funding

The author received no specific funding for this work.

Data Availability

All data used in this study are simulated. No individual-level or confidential empirical data were analysed. The code used to generate the simulated datasets and to implement the Heaped Age Adaptive Model (HAAM), including the PRISMA smoother and all simulation workflows, will be made available in a public repository (Jupyter Notebook format) upon publication of this article.

Supplemental material

S1 Appendix: From Poisson Counts to the Complete-Data Log-Likelihood

Setup. Let $a, k \in \{0, 1, \dots, K\}$ index true and reported ages, respectively. The latent (true) counts are $\theta_a \geq 0$, and the observed (reported) counts are $y_k \in \mathbb{N}$. We model

$$y_k \sim \text{Poisson}(\mu_k), \quad \mu = P\theta, \quad \mu_k = \sum_{a=0}^K \theta_a P(k | a), \quad (13)$$

where $P(k | a)$ is a column-stochastic reporting matrix which obeys the property $\sum_{k=0}^K P(k | a) = 1$ for each a .

Latent allocations. We will now introduce latent cell counts $W_{k,a} \in \mathbb{N}_0$ which represents the number of true age- a individuals whose age is reported k . Conditional on parameters (θ, P) and assuming independence across cells, we have,

$$W_{k,a} \sim \text{Poisson}(\theta_a P(k | a)) \quad (14)$$

independently over (k, a) with $y_k = \sum_{a=0}^K W_{k,a}$. This augmentation turns the problem into a product of Poisson terms over (k, a) .

The Poissonian-Multinomial joint pmf of $\{W_{k,a}\}$ is given by,

$$L(\theta, P; W) = \prod_{a=0}^K \prod_{k=0}^K \frac{(\theta_a P(k | a))^{W_{k,a}}}{W_{k,a}!} e^{-\theta_a P(k | a)}. \quad (15)$$

Taking the logarithm and defining $\ell_{\text{complete}} = \log L(\theta, P; W)$. We drop the additive constant $-\sum_{k,a} \log(W_{k,a}!)$ (irrelevant for maximization),

$$\ell_{\text{complete}} = \sum_{a,k} W_{k,a} \log \theta_a + \sum_{a,k} W_{k,a} \log P(k | a) - \sum_{a,k} \theta_a P(k | a). \quad (16)$$

Using $\sum_k P(k | a) = 1$, the last term simplifies to $-\sum_a \theta_a$. Hence, expressing this with respect to (P, θ) ,

$$\ell_{\text{complete}} = \sum_{a,k} W_{k,a} \log \theta_a + \sum_{a,k} W_{k,a} \log P(k | a) - \sum_a \theta_a. \quad (17)$$

Moreover, $\sum_k W_{k,a} = \theta_a$ (the latent total at true age a), so

$$\sum_{a,k} W_{k,a} \log \theta_a = \sum_a \left(\sum_k W_{k,a} \right) \log \theta_a = \sum_a \theta_a \log \theta_a. \quad (18)$$

Combining (17) and (18), a convenient form is

$$\ell_{\text{complete}} = \sum_a (\theta_a \log \theta_a - \theta_a) + \sum_{a,k} W_{k,a} \log P(k | a). \quad (19)$$

In EM, the (α, β) -update only depends on the second term $\sum_{a,k} W_{k,a} \log P(k | a)$, while the (θ) -update uses the first plus any smoothness penalty, leading to the expression,

$$\ell_{\text{complete}} = \sum_{a,k} W_{k,a} \log P(k | a) + \text{constant} \quad (20)$$

Using the "attractiveness" given by,

$$A_{k,a} = \exp(\alpha_k - \beta_k |k - a|), \quad S_a = \sum_k A_{k,a}, \quad (21)$$

and substituting the normalized misreport probabilities

$$P(k | a) = \frac{\delta_{k,a} + A_{k,a}}{1 + S_a}, \quad (22)$$

into (20) gives

$$\begin{aligned}
 \sum_{a,k} W_{k,a} \log P(k | a) &= \sum_{a,k} W_{k,a} \log \frac{\delta_{k,a} + A_{k,a}}{1 + S_a} \\
 &= \sum_{a,k} W_{k,a} \log(\delta_{k,a} + A_{k,a}) - \sum_{a,k} W_{k,a} \log(1 + S_a) \\
 &= \sum_a W_{a,a} \log(1 + A_{a,a}) - \sum_a \theta_a \log(1 + S_a) \\
 &\quad + \sum_{a,k:k \neq a} W_{k,a} (\alpha_k - \beta_k |k - a|)
 \end{aligned} \tag{23}$$

where we have used $\sum_k W_{k,a} = \theta_a$ in the last step. This is the final expression of $\ell_{complete}$ given in (5).

S2 Appendix: E-step expectation via Poisson-multinomial conditioning For each reported age $k \in \{0, \dots, K\}$ and true age $a \in \{0, \dots, K\}$, let the latent cell counts be,

$$\lambda_{k,a} := \theta_a P(k | a) \tag{24}$$

For a fixed k and conditioning on $Y_k = y_k$, the vector $(W_{k,0}, \dots, W_{k,K})$ is multinomial given by,

$$(W_{k,0}, \dots, W_{k,K}) | Y_k = y_k \sim \text{Mult}(y_k; p_{k,0}, \dots, p_{k,K}) \tag{25}$$

where

$$p_{k,a} := \frac{\lambda_{k,a}}{\mu_k} = \frac{\theta_a P(k | a)}{\mu_k}. \tag{26}$$

For any nonnegative integers w_0, \dots, w_K with $\sum_a w_a = y_k$, using independence property and the Poisson probability mass function gives,

$$\forall a \quad \mathbb{P}\{W_{k,a} = w_a\} = \prod_{a=0}^K \frac{\lambda_{k,a}^{w_a}}{w_a!} e^{-\lambda_{k,a}} = \left(\prod_{a=0}^K \frac{\lambda_{k,a}^{w_a}}{w_a!} \right) e^{-\mu_k}. \tag{27}$$

Also since $y_k = \sum_a W_{k,a} \sim \text{Poisson}(\mu_k)$, it follows that,

$$\mathbb{P}\{Y_k = y_k\} = \frac{\mu_k^{y_k}}{y_k!} e^{-\mu_k}. \tag{28}$$

Therefore, by conditioning on $Y_k = y_k$ (and using $\sum_a w_a = y_k$), we get,

$$\begin{aligned}
 \mathbb{P}\{W_{k,a} = w_a \forall a | Y_k = y_k\} &= \frac{\mathbb{P}\{W_{k,a} = w_a \forall a\}}{\mathbb{P}\{Y_k = y_k\}} \\
 &= \frac{y_k!}{\prod_a w_a!} \prod_{a=0}^K \left(\frac{\lambda_{k,a}}{\mu_k} \right)^{w_a},
 \end{aligned} \tag{29}$$

which is the multinomial probability mass function with y_k trials and cell probabilities $p_{k,a} = \lambda_{k,a} / \mu_k$.

The expectation value of the conditional is then given by,

$$\mathbb{E}[W_{k,a} \mid Y_k = y_k] = \widehat{W}_{k,a} = y_k p_{k,a} = y_k \frac{\theta_a P(k \mid a)}{\mu_k}. \quad (30)$$

which is the first step of the EM algorithm.

S3 Appendix: Poisson Robust Iterative Smoother with Mass Preserving Adaptation (PRISMA)

Data, notation, and objective Let $y_a \in \mathbb{R}_{\geq 0}$ denote observed counts at integer support points $a = 0, 1, \dots, A$ (e.g., single-year ages). Write vectors $\mathbf{y} = (y_0, \dots, y_A)^\top$ and $\hat{\boldsymbol{\theta}} = (\hat{\theta}_0, \dots, \hat{\theta}_A)^\top$ which is an estimate of the smooth true distribution. Our goal is a smooth sequence $\hat{\boldsymbol{\theta}}$ that (i) suppresses upward spikes (age heaping), (ii) preserves tail *shape* by avoiding boundary reflection, (iii) permits general (possibly multi-modal) structure, (iv) admits optional interior-only post-smoothing that leaves tails largely unchanged, and (v) exactly preserves total mass: $\sum_a \hat{\theta}_a = \sum_a y_a$.

As a first step, we will do triangular averaging of the observations y_a which are away from the edge/corner ages using the following procedure: Fix a base interior half-window $r \in \mathbb{N}$. At each location a , use a location-dependent radius that shrinks near boundaries (edge-aware)

$$r_a = \min\{r, a, A - a\}. \quad (31)$$

Let the (discrete) Bartlett/triangular kernel centered at a be

$$k_{a,i} = r_a + 1 - |i - a|, \quad i \in \{a - r_a, \dots, a + r_a\}, \quad (32)$$

and $k_{a,i} = 0$ otherwise. For generic nonnegative observation weights w_i , the edge-aware local average is

$$\tilde{\theta}_a = \frac{\sum_{i=a-r_a}^{a+r_a} k_{a,i} w_i y_i}{\sum_{i=a-r_a}^{a+r_a} k_{a,i} w_i}. \quad (33)$$

Equation (31) is a standard boundary correction: it avoids borrowing mass beyond the observed support and therefore preserves tail shape (Jones 1993; Fan and Gijbels 1996).

Assuming that the true distribution is smooth, the spikes at the heaped ages should be suppressed. To do this, we treat y_a as Poisson-like around θ_a and define Pearson residuals at iteration t and age i as,

$$r_i^{(t)} = \frac{y_i - \mu_i^{(t)}}{\sqrt{\hat{\theta}_i^{(t)} + \varepsilon}}, \quad \varepsilon > 0 \text{ small for stability}. \quad (34)$$

To down-weight upward spikes while leaving troughs comparatively untouched, a *one-sided* Tukey bisquare (redescending) weight ((Cleveland 1979; Huber and Ronchetti 2009)):

$$w_i^{(t)} = \begin{cases} (1 - (r_i^{(t)}/c)^2)^2, & 0 < r_i^{(t)} < c, \\ 0, & r_i^{(t)} \geq c, \\ 1, & r_i^{(t)} \leq 0, \end{cases} \quad (35)$$

with tuning constant $c > 0$. The consequence of defining this way is that upward outliers (potential spikes) are heavily down-weighted; downward residuals are kept (we don't automatically "fill" troughs). The with smaller c suppresses spikes aggressively.

Adaptive bandwidth from local curvature A fixed bandwidth can oversmooth genuine features or undersmooth flat stretches. To preserve genuine peaks/valleys (multi-modal structure), we adapt the local radius using a curvature proxy. Let $\Delta^2 \tilde{\theta}_a^{(t)} := \tilde{\theta}_{a-1}^{(t)} - 2\tilde{\theta}_a^{(t)} + \tilde{\theta}_{a+1}^{(t)}$, with $\Delta^2 \tilde{\theta}_0^{(t)} = \Delta^2 \tilde{\theta}_A^{(t)} = 0$. We define a robust curvature scale $s^{(t)} = \text{median}(|\Delta^2 \tilde{\theta}_a^{(t)}| : 1 \leq a \leq A-1) + \varepsilon$ and

$$\kappa_a^{(t)} = \frac{|\Delta^2 \tilde{\theta}_a^{(t)}|}{s^{(t)}} \quad (36)$$

which is compressed to $[0, 1)$ via,

$$\kappa_a'^{(t)} = \frac{\kappa_a^{(t)}}{1 + \kappa_a^{(t)}} \quad (37)$$

$\kappa_a'^{(t)}$ is then mapped to an integer radius in $[r_{\min}, r_{\max}]$:

$$\hat{r}_a^{(t)} = r_{\max} - (r_{\max} - r_{\min}) \kappa_a'^{(t)}, \quad r_{\min} \leq \hat{r}_a^{(t)} \leq r_{\max}. \quad (38)$$

Thus, high curvature \Rightarrow smaller window; gentle curvature \Rightarrow larger window. Adaptive bandwidths of this type are in the spirit of variable-bandwidth kernel methods (Abramson 1982; Fan and Gijbels 1996).

An iterative update scheme is adapted for estimating the smoother using the following procedure.

Initialize $\tilde{\theta}^{(0)}$ by a single unweighted pass of (33) with $w_i \equiv 1$. For $t = 0, 1, \dots, T-1$:

1. Compute residuals $r_i^{(t)}$ and weights $w_i^{(t)}$ using (35).
2. Choose radii $r_a^{(t)} = \min\{\hat{r}_a^{(t)}, a, A-a\}$, where $\hat{r}_a^{(t)}$ is from (38) (or set $r_a^{(t)} \equiv r$ if adaptation is disabled).
3. Update by the edge-aware weighted average:

$$\tilde{\theta}_a^{(t+1)} = \frac{\sum_{i=a-r_a^{(t)}}^{a+r_a^{(t)}} k_{a,i} w_i^{(t)} y_i}{\sum_{i=a-r_a^{(t)}}^{a+r_a^{(t)}} k_{a,i} w_i^{(t)}}, \quad a = 0, \dots, A. \quad (39)$$

4. Set $\hat{\theta}_a^{(t+1)} = \max\{\tilde{\theta}_a^{(t+1)}, 0\}$.

This delivers a spike-resistant smoother.

Sometimes, the context of the observed data may require that the observations at the tails be preserved. We can do this by optionally choose a tail length $\tau \in \{0, \dots, \lfloor A/2 \rfloor\}$ and define the interior index set $I = \{\tau, \dots, A-\tau\}$. Apply J additional passes of *unweighted* triangular smoothing (33) only on I (with a chosen radius r_{int}), leaving tails $\{0, \dots, \tau-1\}$ and $\{A-\tau+1, \dots, A\}$ untouched. This removes small interior wiggles while broadly preserving tail properties. After all iterations, the mass conservation constraint is imposed by $\sum_{a=0}^A \hat{\theta}_a = \sum_{a=0}^A y_a =: S$ by multiplicative rescaling:

$$\hat{\theta}_a^* = s \hat{\theta}_a, \quad s = \frac{S}{\sum_{i=0}^A \hat{\theta}_i}. \quad (40)$$

This preserves nonnegativity and relative shape.

Hard lock of boundary ages (exact preservation of early counts) For some data it is desirable to keep the first few ages exactly as observed and smooth only the remainder. Let $\ell \in \{0, \dots, A\}$ be the number of left-boundary ages to *hard lock* and define

$$L = \{0, 1, \dots, \ell - 1\}, \quad R = \{\ell, \dots, A\}.$$

The hard lock enforces

$$\hat{\theta}_a = y_a \quad \text{for all } a \in L, \quad (41)$$

and applies the PRISMA update (39) *only* on R (all references to a in (31)–(39) and the curvature adaptation (38) are restricted to $a \in R$, with edge-aware radii computed relative to R so that borrowing never crosses ℓ). To maintain the original mass on the unlocked block, we rescale the smoothed right part by a factor that matches the observed mass on R while leaving the locked block untouched:

$$\begin{aligned} S_R &= \sum_{a \in R} y_a, & \tilde{S}_R &= \sum_{a \in R} \hat{\theta}_a, \\ \hat{\theta}_a &\leftarrow \begin{cases} \hat{\theta}_a, & a \in L, \\ \hat{\theta}_a \cdot \frac{S_R}{\tilde{S}_R}, & a \in R. \end{cases} \end{aligned} \quad (42)$$

Optionally, to avoid a visible kink at the join $a = \ell - 1 \leftrightarrow a = \ell$, we apply a short, one-time *blend* over the first $b \geq 0$ interior points (no effect on L):

$$\hat{\theta}_{\ell+j} \leftarrow (1 - \omega_j) y_{\ell-1} + \omega_j \hat{\theta}_{\ell+j}, \quad (43)$$

where $j = 0, \dots, b - 1$ and $\omega_j = \frac{j}{b-1}$ (linear ramp; $b \geq 2$) or any convex ramp $\omega_j \in [0, 1]$ (e.g., quadratic). When $b = 0$ no blending is performed.

Summary of the PRISMA Algorithm Insert the hard lock into the iterative scheme as:

- (0) **Lock.** Set $\hat{\theta}_a^{(t)} = y_a$ for $a \in L$. Work only on R thereafter.
- (1) to (4) Run steps (35)–(39) on R (edge-aware radii computed with $a \in R$).
- (5) Apply the restricted mass correction (42) on R and optional blend (43).

S4 Appendix: Simulation Design for Age-Heaped Data

The objective of this simulation is to generate synthetic age distributions that exhibit heaped ages and mild terminal-digit preferences while remaining faithful to a smooth, interpretable baseline. This enables stress-testing of age-heaping diagnostics and correction procedures under controlled conditions. Two principles guide the construction:

1. **Probabilistic formulation.** Heaping is modeled as a stochastic reallocation of individuals from their “true” ages to focal ages with distance-dependent attraction, followed by small binomial “jiggles” that induce preference for ages ending in 0 or 5. Probabilistic modeling makes the data-generating process transparent, reproducible, and easy to tune.
2. **Mass preservation.** Every transformation is defined as either a multinomial split (reallocation) or paired binomial transfers. Consequently, the total population size is invariant at each stage, which is crucial for interpretable downstream comparisons between true and observed distributions.

Age Grid, Notation, and Reproducibility Let the age support be the finite grid

$$\mathcal{A} = \{0, 1, \dots, 100\}, \quad K = |\mathcal{A}| = 101, \quad (44)$$

with index mapping $i(a) = a + 1$ for vector access. Denote by $\mathbf{T} = (T_a)_{a \in \mathcal{A}}$ the true baseline counts and by $\mathbf{R} = (R_a)_{a \in \mathcal{A}}$ the reported (observed) counts after heaping and digit jiggles. We fix the seed of the pseudo-random number generator to guarantee exact replicability of every realized dataset.

Baseline: A Smooth Unimodal Distribution We construct a unimodal baseline by discretizing a Weibull density with shape $\kappa > 1$ and scale $\lambda > 0$ on midpoints $x_a = a + 0.5$:

$$g_a \propto f_{\text{Weib}}(x_a; \kappa, \lambda) = \frac{\kappa}{\lambda} \left(\frac{x_a}{\lambda} \right)^{\kappa-1} \exp \left\{ - \left(\frac{x_a}{\lambda} \right)^\kappa \right\}, \quad (45)$$

for $a \in \mathcal{A}$. Normalizing yields a discrete pmf $p_a = g_a / \sum_b g_b$ with $\sum_a p_a = 1$. For a fixed population size N ,

$$T_a = \text{round}(N p_a), \quad a \in \mathcal{A}, \quad (46)$$

with gentle lower bounds at extremes to avoid structural zeros in demonstrations. Although the Weibull is not a standard model for population age structures per se, it is routine in survival/mortality contexts and serves here as a flexible, low-parameter generator of a smooth unimodal baseline. This choice is appropriate because our goal is to isolate and study heaping mechanisms, not to reproduce a full demographic age pyramid.

Parameterization used in examples. Unless stated otherwise, we take $N = 80,000$, $\kappa = 1.1$ and $\lambda = 50$, which produce a gently right-skewed unimodal profile on \mathcal{A} .

Heaped-Age Attraction (Primary Heaping) Let $\mathcal{F} \subset \mathcal{A}$ denote the set of focal ages (e.g., $\{40, 50, 70\}$). For each non-focal age a , define the in-window focal set

$$\mathcal{F}_a = \{f \in \mathcal{F} : 0 < |a - f| \leq w\}, \quad (47)$$

with window radius $w \in \mathbb{N}$ controlling locality. If $\mathcal{F}_a = \emptyset$, all persons at age a remain at a .

When $\mathcal{F}_a \neq \emptyset$, we assign distance-decayed weights

$$w_{a \rightarrow f} = \exp \{ -\delta (|a - f| - 1) \}, \quad \delta > 0, \quad (48)$$

and define the per-age move probability

$$p_{\text{move}}(a) = \min \left\{ \text{cap}, \beta \sum_{f \in \mathcal{F}_a} w_{a \rightarrow f} \right\}. \quad (49)$$

Here $\beta \in (0, 1)$ is a base attraction strength, while $\text{cap} \in (0, 1)$ is a conceptual and technical safeguard:

- Conceptually, cap is the maximum misreporting probability allowed at any age; it encodes that even under strong attraction some respondents state exact ages.
- Technically, it prevents degenerate reallocations (e.g., draining an age class) and ensures probabilities remain valid.

Conditional on moving, the destination focal is chosen proportionally to the weights:

$$\Pr(a \rightarrow f \mid \text{move}) = \frac{w_{a \rightarrow f}}{\sum_{f' \in \mathcal{F}_a} w_{a \rightarrow f'}}. \quad (50)$$

Thus, for the T_a persons at age a , the multinomial allocation is

$$\begin{aligned} & (\text{stay}, \{m_{a \rightarrow f}\}_{f \in \mathcal{F}_a}) \sim \\ & \text{Mult}\left(T_a; \underbrace{1 - p_{\text{move}}(a)}_{\pi_{\text{stay}}(a)}, \underbrace{\left\{p_{\text{move}}(a) \frac{w_{a \rightarrow f}}{\sum_{f' \in \mathcal{F}_a} w_{a \rightarrow f'}}\right\}_{f \in \mathcal{F}_a}}_{\pi_{a \rightarrow f}}\right) \end{aligned} \quad (51)$$

Reported counts after this stage are updated by

$$R_a \leftarrow R_a + \text{stay}, \quad R_f \leftarrow R_f + m_{a \rightarrow f} \text{ for each } f \in \mathcal{F}_a. \quad (52)$$

Because the updates are purely redistributive, $\sum_a R_a = \sum_a T_a = N$.

Tuning parameters (illustrative). $\mathcal{F} = \{40, 50, 70\}$, $w = 7$, $\delta = 0.50$, $\beta = 0.15$, and $\text{cap} = 0.95$. These produce sharp but realistic focal piles without collapsing neighboring ages.

Terminal-Digit “Jiggles” (Secondary Heaping) To superimpose mild heaping toward terminal digits 0 and 5, we apply small, independent binomial transfers along four patterns:

$$9 \rightarrow 10 \quad \text{with prob. } p_{9 \rightarrow 10}, \quad (53)$$

$$1 \rightarrow 0 \quad \text{with prob. } p_{1 \rightarrow 0}, \quad (54)$$

$$4 \rightarrow 5 \quad \text{with prob. } p_{4 \rightarrow 5}, \quad (55)$$

$$6 \rightarrow 5 \quad \text{with prob. } p_{6 \rightarrow 5}. \quad (56)$$

For each eligible age a , draw $M \sim \text{Binomial}(R_a, p)$, then apply

$$R_a \leftarrow R_a - M, \quad R_{a \pm 1} \leftarrow R_{a \pm 1} + M, \quad (57)$$

with boundary checks. Each move preserves mass; cumulatively, $\sum_a R_a = N$ still holds. The probabilities are set small (e.g., $p \in [0.02, 0.03]$) so that these jiggles add gentle 0/5 spikes without dominating the focal-age attraction.

References

- Abramson IS (1982) On bandwidth variation in kernel estimates—a square root law. *Annals of Statistics* 10(4): 1217–1223.
- A’Hearn B, Baten J and Crayen D (2009) Quantifying quantitative literacy: Age heaping and the history of human capital. *The Journal of Economic History* 69(3): 783–808.
- ARRIAGA E et al. (1968) *New life tables for Latin American populations in the nineteenth and twentieth centuries*. University of California, Berkeley.
- Bachi R (1951) The tendency to round off age returns: Measurement and correction. *Bulletin of the International Statistical Institute* 33: 195–199.

- Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57(1): 289–300.
- Boyd S and Vandenberghe L (2004) *Convex Optimization*. Cambridge University Press.
- Bureau UC (2023) Age heaping in the 2020 census (dhc). Blog. URL <https://www.census.gov/newsroom/blogs/random-samplings/2023/05/age-heaping-2020-census-dhc.html>.
- Camarda CG, Eilers PHC and Gampe J (2007) Modeling digit preference by penalized composite link models. In: *Population Association of America Annual Meeting*. Barcelona, Spain, pp. 148–153.
- Camarda CG, Eilers PHC and Gampe J (2008) Modelling general patterns of digit preference. *Statistical Modelling* 8(4): 385–401.
- Carrier NH (1959) A note on the measurement of digital preference in age recordings. *Journal of the Institute of Actuaries* 85(1): 71–85.
- Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74(368): 829–836.
- Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39(1): 1–38. DOI:10.1111/j.2517-6161.1977.tb01600.x.
- Fan J and Gijbels I (1996) *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC.
- Ferber S and Baten J (2025) Age heaping based numeracy estimates in african regions, 1950–1999: New methodological advances and results. *Economic History of Developing Regions* 40(1): 15–48.
- Fletcher R (2000) *Practical methods of optimization*. 2nd ed. edition. John Wiley & Sons, Ltd. ISBN 9781118723203. DOI:10.1002/9781118723203.
- Gu D and Feng Q (2019) Use of the average age ratio method in analyzing age heaping in censuses: The case of china. *International Journal of Population Studies* 5(1): 13–26.
- Huber PJ and Ronchetti EM (2009) *Robust Statistics*. 2 edition. Wiley.
- Jones MC (1993) Simple boundary correction for kernel density estimation. *Statistics and Computing* 3(3): 135–146.
- Kannisto V, Jeune B and Vaupel J (1999) Assessing the information on age at death of old persons in national vital statistics. *Validation of exceptional longevity (Odense Monographs on Population Aging, Vol. 6)* : 235–249.
- Manual X (1983) Indirect techniques for demographic estimation. *New York: United Nations* .
- McLachlan GJ and Krishnan T (2008) *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics, 2 edition. Hoboken, NJ: Wiley-Interscience. ISBN 978-0-471-20170-0. DOI: 10.1002/9780470191613.
- Mikołaj S, Radosław P and Siegfried G (2018) Age heaping patterns in mosaic data. *Historical Methods. A Journal of Quantitative and Interdisciplinary History* 51: 1.
- Moultrie TA, Dorrington RE, Hill AG, Hill K, Timæus IM and Zaba B (2013) *Tools for demographic estimation*. International Union for the Scientific Study of Population.
- Myers RJ (1954) Accuracy of age reporting in the 1950 u.s. census. *Journal of the American Statistical Association* 49(268): 826–831.

- Myers RJ et al. (1940) Errors and bias in the reporting of ages in census data. *Transactions of the Actuarial Society of America* 41(2): 395–415.
- National Academies of Sciences, Engineering, and Medicine (2023) *Assessing the 2020 Census: Final Report*. Washington, DC: The National Academies Press. DOI:10.17226/27150.
- Noumbissi A (1992) Age heaping index for all terminal digits. *Demographic Research Note* .
- Office UNS (2008) *Principles and recommendations for population and housing censuses. Revision. 2*. New York: United Nations, Department of International Economic and Social Affairs.
- Ramachandran K (1965) An index to measure digit preference error in age data. In: *World Population Conference*. pp. 202–203.
- Shryock HS and Siegel JS (1971) *The methods and materials of demography*, volume 1-2. Department of Commerce, Bureau of the Census.
- Spoorenberg T and Dutreuilh C (2007) Quality of age reporting: Extension and application of the modified whipple’s index. *Population* 62(4): 729–741.
- United Nations (1952) *The Accuracy of Age and Sex Statistics*. Manual II, Methods of Appraisal of Quality of Basic Data for Population Statistics.
- Whipple GC (1919) *Vital Statistics: An Introduction to the Science of Demography*. Wiley.
- Zelnik M (1961) Age heaping in the united states census: 1880–1950. *The Milbank Memorial Fund Quarterly* 39(3): 540–573. DOI:10.2307/3348729.