

To,
The Editors
Data Science

Subject: Submission of Manuscript: Heaped Age Adaptive Model (HAAM)
for Mitigating Age Heaping in Demographic Data

Dear Editors,

I am pleased to submit the manuscript entitled “**Heaped Age Adaptive Model (HAAM) for Mitigating Age Heaping in Demographic Data**” for consideration for publication in *Data Science* (SAGE). This paper proposes a novel, data-driven statistical framework for detecting, diagnosing, and correcting age heaping in demographic and survey data, a long-standing data quality challenge with direct implications for data science, population analytics, and applied statistical modelling.

Age heaping—systematic misreporting of ages due to digit preference or culturally salient thresholds—remains pervasive across historical and contemporary datasets. While classical indices (e.g., Whipple’s or Myers’ indices) provide descriptive summaries, they are limited to digit-based heaping and do not offer a principled mechanism for correction or for identifying irregular, age-specific heaping patterns. From a data science perspective, this represents a problem of latent structure inference under noise and misclassification.

The core contribution of this work is the **Heaped Age Adaptive Model (HAAM)**, a penalized Expectation–Maximization framework that jointly estimates (i) a smooth latent “true” age distribution and (ii) age-specific misreporting behavior. Methodologically, HAAM integrates:

- a **Poisson-robust smoothing component (PRISMA)** to recover a demographically plausible latent signal from spiky count data,
- a **flexible misreporting kernel** that models attraction toward specific reported ages with distance-dependent decay, and
- an **ℓ_1 sparsity penalty** that adaptively identifies a small subset of genuinely heaped ages without imposing any a priori assumptions about terminal digits or fixed heaping rules.

This combination allows the model to uncover both conventional digit-preference heaping and irregular, institutionally or culturally driven age preferences. Through controlled simulation experiments, the paper demonstrates that HAAM accurately recovers the underlying age distribution, removes artificial spikes, and correctly identifies focal heaped ages—outperforming traditional graduation and digit-based correction methods while also providing interpretable diagnostic outputs.

I believe this manuscript is particularly well aligned with *Data Science Journal* because it addresses a general data quality problem using modern statistical learning principles: latent-variable modeling, regularization, robust smoothing, and interpretable inference. Beyond demography, the framework is applicable to any domain involving discrete count data with systematic reporting bias, including epidemiology, public health surveillance, historical data

analysis, and survey methodology. The paper emphasizes reproducibility and transparency, and the full codebase for simulations and estimation will be made publicly available upon publication.

The manuscript is original, has not been published previously, and is not under consideration elsewhere. I confirm that there are no conflicts of interest and no external funding sources associated with this work.

Thank you for considering this submission. I would be grateful for the opportunity to have the manuscript reviewed and am happy to respond to any questions or suggestions from the editors and reviewers.

Sincerely,

Santosh Kudtarkar

Centre for Mathematical Modelling

FLAME University, Pune, India

Email: sant@flame.edu.in