# From Fractions to Individuals: Information-Theoretic Integerization for Population Synthesis

Jeppe Rich*[1]

[1]Department of Technology, Management and Economics, Technical University of Denmark, Denmark

## Short summary

Every population synthesis method produces fractional expected counts, but simulation models require whole individuals. This paper treats integerization as an inference problem: among all feasible integer tables, choose the one closest to the fractional baseline in Kullback–Leibler divergence. We show that an optimal solution always assigns each cell either its floor or ceiling value, which reduces the task to a deterministic two-step procedure: a second IPF pass on the residuals followed by largest-remainder allocation. Applied to 98 Danish municipalities with 635,000 joint cells, the method places 98.4% of population mass in the correct cells compared with 92.1% for multinomial sampling, while preserving controlling margins exactly and eliminating run-to-run sampling noise.
**Keywords**: agent-based models; deterministic rounding; entropy minimisation; integerization; population synthesis; reproducibility

## 1 Introduction

Agent-based travel-demand and land-use models need whole individuals. Every person in the simulation must exist as a discrete entity. Yet every population synthesis method produces fractions, not integers.

This is true across paradigms. Classical Iterative Proportional Fitting (IPF) scales a seed table to match official totals and yields fractional cell counts. Bayesian network synthesisers (Farooq et al., 2013) and deep generative models, including variational autoencoders, generative adversarial networks, and diffusion models (Borysov et al., 2019; Zhu et al., 2024; Garrido et al., 2020), output per-cell probabilities. Household-formation models (Aemmer & MacKenzie, 2022) also produce fractional composition counts before any rounding is done. In every case the last step is the same: turn a table of fractional expected counts into a table of whole people. We call this step *integerization*.

That last step is often treated as a minor technicality. It is not. Once fractional counts are converted into integers, they become the actual synthetic persons and households used by downstream simulation models. The integerization rule therefore decides which rare groups survive, which cells are emptied, and how faithfully the final population reflects the fitted joint distribution. If this step is handled badly, much of the value created upstream by careful fitting is lost at the point where fractions become actual agents.

The problem is becoming more pressing. Zone systems are getting finer, and demographic classification is getting richer by crossing more dimensions of age, income, education, labour-market status, and household type at the same time. Together these trends multiply the number of cells in the population table while the total population stays the same. A cell that once held ten people may now hold 0.3. As this shift continues, integerization stops being a minor rounding detail and becomes the main factor deciding who appears in the synthetic population.

This matters most for the groups analysts typically care about when studying distributional effects. Low-frequency demographic combinations are exactly the ones most at risk of being rounded away. What looks like harmless residual noise at the table level can become a systematic loss of vulnerable or unusual households in the realized population.

Two methods currently dominate. *Naive rounding* rounds each cell on its own and usually breaks the margin constraints. *Stochastic sampling* draws integers from the normalised fractional distribution. Sampling is not reproducible: running the same pipeline twice gives different populations. It also suppresses rare groups systematically, because any cell with an expected count below one is likely to be set to zero. Fixing this through Monte Carlo replication is expensive.

The problem is made worse by how much effort has gone into improving the upstream synthesis step. A deep generative model that captures the joint distribution of age, income, and household type offers little benefit if the rounding step then deletes rare cells or injects random noise. Choupani & Mamdoohi (2016) flag integerization as a critical open problem. It has remained largely unsolved since Lovelace & Ballas (2013) and Barthelemy & Toint (2013).

This paper treats integerization as a statistical inference problem under Kullback–Leibler (KL) divergence. Given any fractional population table, from IPF, a generative model, or anything else, the method finds the integer table that is closest to it while respecting demographic margins. The result is deterministic, fast, and works as a drop-in post-processing step for any synthesis pipeline.

## 2 Methodology

### *Problem formulation*

Let $\mathcal{K}$ index the cells of a population table, for example Zone × Age × Gender × Income × Labour-market status. Any synthesis method delivers fractional expected counts $x = (x_k)_{k\in\mathcal{K}}$. For IPF these come from scaling iterations. For a generative model, $x_k = N\,\hat{p}_k$, where $\hat{p}_k$ is the predicted cell probability and $N$ the target population. Either way, the counts satisfy margin constraints

$$\sum_{k\in\mathcal{K}(s)} x_k = m_s, \quad \forall\, s \in \mathcal{S}, \tag{1}$$

where $\mathcal{K}(s)$ is the set of cells contributing to margin $s$ and $N = \sum_k x_k$.

We want integer counts $\hat{x} = (\hat{x}_k)$ that meet integer margin targets $b_s$ and stay as close as possible to $x$. We measure closeness with KL divergence. Let $p_k = x_k/N$ and $q_k = \hat{x}_k/N$. The problem is

$$\min_{\hat{x}\in\mathbb{Z}_{\geq 0}^{|\mathcal{K}|}} D_{\mathrm{KL}}(q\,\|\,p) \quad \text{s.t.} \quad \sum_{k\in\mathcal{K}(s)} \hat{x}_k = b_s, \quad \forall\, s. \tag{2}$$

The KL-minimising table is also the maximum-likelihood integer population under the multinomial distribution implied by $p$.

This gives integerization a clear statistical meaning. Rather than asking how to round each cell independently, we ask which feasible integer table is most consistent with the fractional population as a whole.

### *Floor–ceiling property*

Equation (2) looks hard: it is a discrete optimisation over potentially millions of variables. The key insight is that the solution space is much smaller than it appears.

**Theorem (Floor–ceiling optimality).** *Let $x^*$ minimise $D_{\mathrm{KL}}(q\,\|\,p)$ subject to the integer margin constraints. Then for every cell, $x_k^* \in \{\lfloor x_k\rfloor, \lceil x_k\rceil\}$.*

*Proof sketch.* Suppose $x_k^* \geq \lceil x_k\rceil + 1$ for some cell $k$. Margin balance then forces some other cell $\ell$ to have $x_\ell^* \leq \lfloor x_\ell\rfloor - 1$. Moving one unit from $k$ to $\ell$ strictly reduces KL divergence, because $\varphi(n) = \frac{n}{N}\log(n/(Np_k))$ is strictly convex. This contradicts optimality. The circulation theorem guarantees such a move exists even when cells appear in multiple overlapping margins. A complete proof is provided in Rich (2026). □

In plain terms, the best integer table never puts more than one unit above or below the fractional value in any cell. The search therefore reduces to a binary yes/no decision for each cell: floor or ceiling.

### *Two-step algorithm*

Write each fractional count as a floor plus a binary top-up:

$$\hat{x}_k = \underbrace{\lfloor x_k\rfloor}_{n_k} + \underbrace{z_k}_{\in\{0,1\}}. \tag{3}$$

The floor $n_k$ is fixed. The residual budget for each margin is $\beta_s = b_s - \sum_{k\in\mathcal{K}(s)} n_k$, and we need $\sum_{k\in\mathcal{K}(s)} z_k = \beta_s$ for all $s$.

We allocate the top-ups in two steps.

1. **Second IPF pass.** Take the fractional parts $r_k = x_k - n_k$ as a seed and run IPF with the residual budgets $\beta_s$ as margins. This gives rebalanced residuals $r_k^*$ that are consistent with all constraints simultaneously.
2. **Largest-remainder rounding.** Set $z_k = 1$ for the cells with the largest $r_k^*$ values until all budgets are filled. This is deterministic and corresponds to MAP estimation with a Bernoulli prior, giving an information-theoretic basis for Hamilton's classical apportionment rule.

The whole pipeline runs two IPF passes, both near-linear in the number of cells.

***Handling infeasibility***

Some margin combinations cannot be satisfied exactly because the seed table has structural zeros that block certain cell combinations. We handle this with a three-level priority scheme. *Controlling margins*, typically Zone × Age × Gender, are always met exactly, following the harmonization approach in Rich (2018). *Secondary margins* are projected onto the supported cells and rescaled to remove impossible demands. A final *swap-repair* step shifts $+1/-1$ pairs of cells to reduce remaining secondary-margin errors while leaving controlling margins intact.

## 3 Results and discussion

***Toy example***

We use a simple $2 \times 3$ table crossing Age (Young, Old) with Income (Low, Mid, High) to compare two methods: Hamilton's largest-remainder rule without margin enforcement and the proposed entropy approach.
**Baseline margins.** When the margins already match the fractional row and column sums, both methods give the same result.
**Shifted controls.** Shift one person from Mid to Low income. Hamilton ignores the new margins, so Low is underallocated and Mid is overallocated. The entropy approach meets the shifted margins exactly.
**Small cell at risk.** Set the Young–Low cell to 0.4 persons. Hamilton rounds it to zero and violates the row margin. The entropy approach gives it one person and satisfies all margins. This is the failure mode that grows with finer segmentation: rare cells get deleted by greedy rounding.

***Danish national application***

We apply the method to a national synthesis covering Municipality × Age × Gender × Labour-market status × Family structure × Income across 98 Danish municipalities with 635,000 joint cells. Between 2.2% and 7% of population mass sits in sub-unit fractional cells; the floor step captures the rest.
**Reproducibility.** Running the pipeline twice on the same input gives identical tables (maximum absolute difference $< 10^{-10}$). Sampling variance is completely eliminated.
**Allocation accuracy.** We measure accuracy with the overlap metric

$$\theta = \sum_k \min\{p_k, q_k\} \in [0, 1],$$

where $p$ is the fractional baseline rescaled to shares and $q$ the integer solution rescaled likewise. $\theta = 1$ means every person lands in exactly the right cell.

Table 1: Allocation accuracy and distance from the fractional baseline.

| Method | Mean $\theta$ | Median $\theta$ | L2 distance |
|---|---|---|---|
| Integerised (swap-repair) | 0.984 | 0.988 | 27.9 |
| Integerised (no swap-repair) | 0.977 | 0.982 | 32.0 |
| Multinomial sampling (mean) | 0.921 | 0.927 | 357.2 |
| Multinomial sampling (5%–95%) | — | — | 333–384 |

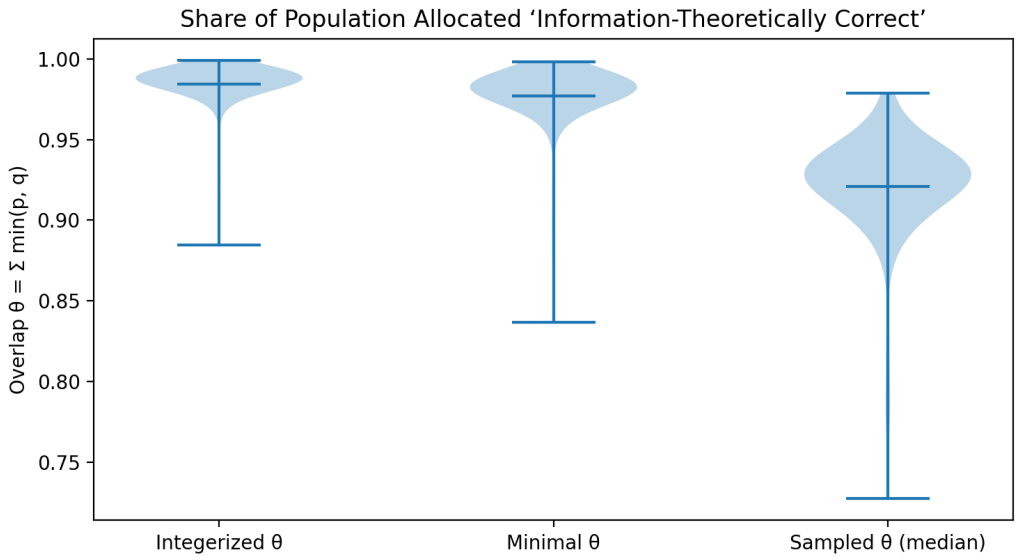



Figure 1: Distribution of overlap accuracy $\theta$ across municipalities. Higher values mean a larger share of population mass is placed in the same cells as the fractional baseline.

The swap-repaired solution places 98.4% of population mass in the correct cells. Stochastic sampling manages 92.1%. The L2 distance from the fractional baseline is 13 times smaller. The deterministic result is a single fixed number, so there is no spread to report.

**Margin satisfaction.** Controlling margins (Municipality × Age × Gender) are met exactly in all 98 municipalities. Secondary margins show mean L1 deviations below 0.5% of municipality totals, with a maximum below 2%. These gaps reflect genuine structural conflicts between non-controlling margins, not algorithmic error.

**Zone-size pattern.** Accuracy varies across municipalities (range 0.968–0.995). Smaller municipalities have fewer people per cell, so each rounding decision carries more weight. Larger municipalities give swap-repair more room to work. Within any municipality, the result is the same every time the algorithm runs.

The biggest practical gain is determinism. Monte Carlo ensembles are no longer needed to average out rounding noise. Uncertainty in downstream models comes from the input margins and the seed table, not from the integerization step itself.

The floor–ceiling theorem gives a formal basis for residual methods used in earlier work (Smith et al., 2017) and proves optimality properties that were previously assumed. The MAP interpretation shows that Hamilton's apportionment rule is a first-order approximation to the full KL objective. It works well when margins are loose and independent. When margins overlap and compete, the second IPF pass resolves conflicts that greedy rounding cannot handle.

This framework slots in after any method that outputs a fractional population table. A diffusion model predicting 3.7 people in a given cell still needs a principled rule for whether to place 3 or 4, consistent across all cells and all margins at once.

The method works best when most cells have fractional parts spread across the $[0, 1]$ range rather than clustered near zero. If nearly every cell has an expected count below 0.1, the table is probably too fine-grained for the available data. That is a modelling problem that affects all integerization methods equally.

## 4 Conclusions

This paper addresses a problem every synthesis pipeline faces but few address well: how to turn fractional cell counts into whole individuals without losing accuracy or reproducibility.

The solution rests on three results. First, the KL-optimal integer table always rounds each cell to its floor or ceiling. Second, a two-step algorithm, a second IPF pass plus largest-remainder rounding, solves those decisions efficiently and deterministically. Third, on a Danish dataset with 635,000 cells, the method places 98.4% of the population in the correct cells against 92.1% for sampling, with controlling margins satisfied exactly.

The method works after any synthesis model. It is fast, reproducible, and removes the need for Monte Carlo replication. More broadly, integerization should be treated as part of the model, not as an afterthought.

## Acknowledgements

The author thanks Gunnar Flötteröd for discussions that helped clarify the floor–ceiling property and the information-theoretic interpretation of the method.

## References

Aemmer, Z., & MacKenzie, D. (2022). Generative population synthesis for joint household and individual characteristics. *Computers, Environment and Urban Systems*, *96*, 101852. doi: 10 .1016/j.compenvurbsys.2022.101852

Barthelemy, J., & Toint, P. L. (2013). Synthetic population generation without a sample. *Transportation Science*, *2*, 266–279. Retrieved from `https://www.jstor.org/stable/43666649`

Borysov, S. S., Rich, J., & Pereira, F. C. (2019). How to generate micro-agents? a deep generative modeling approach to population synthesis. *Transportation Research Part C*, *106*, 73–97. doi: 10.1016/j.trc.2019.07.006

Choupani, A.-A., & Mamdoohi, A. R. (2016). Population synthesis using iterative proportional fitting (ipf): A review and future research. *Transportation Research Procedia*, *17*, 223–233. doi: 10.1016/j.trpro.2016.11.078

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B*, *58*, 243–263. doi: 10.1016/j.trb.2013.09.012

Garrido, S., Borysov, S., Pereira, F., & Rich, J. (2020). Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. *Transportation Research Part C*, *121*, 102787. doi: 10.1016/j.trc.2020.102787

Lovelace, R., & Ballas, D. (2013). Truncate, replicate, sample: A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, *41*, 1–11. doi: 10.1016/j.compenvurbsys.2013.03.004

Rich, J. (2018). Large-scale spatial population synthesis for denmark. *European Transport Research Review*, *10*(2), 21. doi: 10.1186/s12544-018-0336-2

Rich, J. (2026). *From fractions to individuals: Information-theoretic integerization for population synthesis.* (In review Transportion Part B)

Smith, D. M., Lovelace, R., & Birkin, M. (2017). Population synthesis with quasi-random integer sampling. *Journal of Artificial Societies and Social Simulation*, *20*(4), 14. doi: 10.18564/ jasss.3550

Zhu, K., Yin, L., Liu, K., Liu, J., Shi, Y., Li, X., & Zou, H. (2024). Generating synthetic population for simulating the spatiotemporal dynamics of epidemics. *PLOS Computational Biology*, *20*(2), e1011810. doi: 10.1371/journal.pcbi.1011810