Microdata enrichment for sub-regional analyses: comparing the performance of synthetic population approaches for Emilia-Romagna municipalities

S. Scarpa, M. Baldini, A. Barigazzi, G. Gallo and M. Morciano

---

Discussant: Laura Sigalotti (Bank of Italy)

July 4th, 2025 - 3rd Bank of Italy Workshop on Microsimulation Modelling

Summary - Methods for Spatial Microsimulation

- Ingredients:
  - **sample data** S (individual or household level)
  - area-level **benchmarks** X (census or administrative)
- Goal: to reconstruct the population from which sample data are drawn, within the sub-areas into which it is partitioned
- A priori choices:
  - areas of interest
  - performance measures
  - model or algorithm $\rightarrow$ reweighting methods (IPF, GREG, SA)
- Re-weighting: start with a set of initial weights w_j, $j \in S$, and iteratively update them to obtain a set of weights $w^*_{j,d}$ for each $j \in S$ and each area d, consistent with area-level benchmarks X. The synthetic population is then generated using weights $w^*_{j,d}$, possibly rounded.

Summary – Empirical analysis

- **Data:** 2022 EU-SILC data for Emilia-Romagna (2.937 observations); synthetic populations created for all 330 municipalities (220 for SA).
- **Benchmarks:** 6 constraints (48 attributes), including both univariate (age class, gender, household size) and multivariate ones (e.g., gender by age, education by age and gender), drawn from administrative and census sources.
- **Methods:**
  - Deterministic: IPF, GREG (using different starting weights)
  - Probabilistic: Simulated Annealing (individual-based and household-based variants)
- **Performance measures:** TAE, PSAE, and additional validation exercises on subsamples.
- **Key findings:**
  - *GREG* outperforms IPF in most settings, especially in larger municipalities.
  - *SA* performs better in small and rural areas, with integer weights and household-level coherence.
  - A hybrid GREG–SA strategy could be optimal.

Discussion: Highlights

Key contributions:

- Comprehensive comparison of spatial microsimulation techniques
- Rigorous implementation using relevant data
- Practical guidance on methodological trade-offs
- Discussion relevant for both academics and policymakers

Some comments follow.

Deterministic vs Probabilistic Methods: Questions and Trade-offs

The paper offers a balanced comparison between deterministic (IPF, GREG) and probabilistic (SA) methods. A few comments:

- **Performance gap:** Many studies find that probabilistic methods outperform deterministic ones. Any explanation why this is not fully the case here?
- **Transparency vs Stability:** Probabilistic methods generate multiple synthetic populations — does this provide a more *honest* picture of uncertainty, or does it introduce instability?
- **Hybrid strategy:** The paper suggests using SA in small areas and GREG in larger ones — could this idea be formalized and validated more systematically?

Benchmark Constraints

Constraints at the municipal level are a core component of the reweighting procedure.

A few points could strengthen the methodological transparency:

- It would be helpful to see the full list of **admissible** constraints (at both individual and household levels) from which the final set of P benchmarks is selected.
- More detail on the benchmark selection **algorithm** would enhance replicability.
- A brief discussion on the trade-offs between univariate and bivariate constraints would be valuable—particularly regarding accuracy, convergence, and the risk of overfitting.

Role of DegUrba

The inclusion (or exclusion) of the DegUrba variable — the degree of urbanisation of the municipality (cities, towns and suburbs, rural areas) — appears to have a **significant impact** on the empirical results.

- A brief **theoretical justification** for filtering based on urbanisation level would help clarify the rationale for including this variable. Could other municipal-level characteristics be considered as alternative grouping criteria?
- When used jointly with the **SA** approach, DegUrba does not seem to improve model fit — could you elaborate on the underlying intuition?
- More broadly, does **restricting** the sample to municipalities that are *very similar* risk discarding relevant **variation** in the data?

Further Validation

Section 4.1 presents additional validation to assess the consistency between the synthetic population and real data, focusing on GREG and two key variables: age and income (*).

- While this type of validation requires significant analytical effort, conducting a **systematic extension** to the other two reweighting methods and municipalities could yield valuable insights into the **external validity** of the different approaches.
- A formal comparison between distributions — e.g., using the Kullback–Leibler divergence — could strengthen the analysis.
- (*) Beware that EU-SILC data on income are cross-validated with evidence from MEF and INPS.
- Refer to Figures 8 and 9: not all the commented variables are visible in the plots.

Another possibility for additional validation:

- consider different waves of EU-SILC

Alternative Approaches

- A brief comparison with recent machine learning-based approaches (e.g. random forest reweighting, XGBoost-based synthetic data) could provide useful context on trade-offs between flexibility and interpretability.
- Do the authors view these newer methods as viable alternatives or complementary tools, particularly in household-level modelling?

Concluding thoughts - POV: regional economic analysis

- SM methods can be used to estimate spatial indicators, thus complementing and sometimes encompassing SAE methods.
- SM enable policy evaluations at the small area level by inserting a spatial or geographic dimension into the simulation process.
- **Question for the authors:** Could SAE results be used as an external benchmark to assess the accuracy of reweighting methods in specific domains?

Thank you for the opportunity of discussing this stimulating paper!