synthpop

From grainy buurt aggregates to a synthetic city

A reproducible, open-source pipeline that upscales public CBS neighbourhood statistics into micro-level synthetic people — linked to wastewater catchments.

OneGov #2 · CBS · ODISSEI · Erasmus MC · Digicampus | Team BringOn · built in Rust

The problem

- Epidemiological models need individual people — privacy means that data can't exist openly.
- All that exists openly are coarse buurt aggregates.
- Wastewater (RWZI) signals lack the demographic context to interpret them.

	Lena — epidemiologist, RIVM	Daan — wastewater analyst, Erasmus MC
Needs	micro-level people for SEIR models	demographic context per RWZI
Wall	buurt aggregates too coarse	“a signal without an address”

How do we generate, from only public data, a micro-level population that is statistically correct, spatially coherent, and feeds both infection models and wastewater surveillance?

Our approach

Like upscaling a grainy photo: more detail out than in, still statistically faithful to the original.

```
CBS buurt aggregates (86165NED)
        |   + national cross-domain seed
        v
   [ IPF fit per buurt ]            exact marginals, preserved associations
        v
   [ reconstruct households + people ]   exact age pool, role constraints
        v
   [ place in real buurt polygon ]   (RD New)        <- Layer 1
        v
   [ point-in-polygon -> RWZI catchment ]            <- Layer 2
        v
   population.csv + catchment_join.csv + quality-report.md
```

One command · fixed seed · fully reproducible · region-agnostic via config.

Method — cross-domain by construction

Iterative Proportional Fitting + synthetic reconstruction

- A national seed joint (household × housing) encodes plausible associations.
- IPF rescales it to each buurt's real CBS marginals, preserving odds ratios (unit-tested).
- Households built to exact CBS counts; sizes calibrated to hit the buurt population exactly.
- Ages drawn from an exact age pool under role rules — children in with-kids households, adults elsewhere.

Result: synthetic age & household marginals equal CBS, while age, household and housing form a realistic whole — not reformatted aggregates.

Layer 2 — a wastewater signal with an address

- Buurt polygons (PDOK CBS kaart) + GWSW catchment polygons, all in RD New (EPSG:28992).
- Each buurt centroid → point-in-polygon → the rioleringsgebied that serves it.
- Households jittered inside their polygon (privacy). Aggregate population + density per catchment.

Catchment (Utrecht)	Synthetic pop	Density /km²
Zuilen / Ondiep	38,235	9,723
Overvecht	32,400	5,480
Baden Powellweg	32,295	10,224

Now a pathogen peak at an RWZI can be read against who lives in the catchment.

Results — gemeente Utrecht

Metric	Value
Synthetic persons / households	376,770 / 194,055 (109 buurten)
Age-band marginal fit (WMAPE)	0.10 %
Household-type marginal fit (WMAPE)	0.000 % (exact)
Housing-type marginal fit (WMAPE)	0.04 %
Children (0-14) in with-kids households	100 %
Buurten → catchment	109 / 109 → 33 catchments
Runtime	seconds, single binary

Generated today from live CBS open data.

Quality & validation — it ships with its own report

`synthpop report` auto-generates a quality report mapped to the judging criteria:

- Marginal fit — MAPE per variable, per buurt.
- Cross-domain (S1) — age × household cross-tab; 100 % role consistency.
- Spatial coherence (S2) — Moran's I across neighbouring buurten.
- Wastewater (S4) — synthetic population & density per catchment.
- Privacy — k-anonymity over the quasi-identifier set, honestly flagged.

Every claim on these slides is regenerated by the tool, not hand-typed.

Privacy & open source

- 100 % synthetic — no real person-level data at any stage; built only from public aggregates.
- Coordinates jittered inside the buurt polygon — never a real address.
- k-anonymity suppression pass: min k=5 over [buurt,age,household,housing], only 0.08% of records generalised.
- Apache-2.0 code · CC BY 4.0 data · one-command reproducible · public sources only.

Honest finding for the challenge owners: the starter catalogue's 85870NED is actually Bruto investeringen (not proximity) and 70262NED land use is gemeente-level — reported so the catalogue can be corrected.

Impact & next steps

Today it gives Lena and Daan a demographically realistic, catchment-linked population they can load in seconds — transferable to any Dutch region by config.

Next

- Validate as input to a working SEIR / infection-and-recovery model (C1).
- Join catchments to RWZI plant capacity (Emissieregistratie).
- Education vs CBS 82275NED; income from a dedicated CBS table.
- Second region (Rotterdam / Den Haag) to demonstrate transferability.

synthpop · syntechd.bringon.io · github.com/BreachWhite/HackAthonnie · thank you