# From polygons to AI

Defensive patterns from a five-city PPGIS pipeline

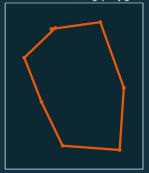

**Mahdi Rahimi** · University of Groningen

FOSS4GNL 2026

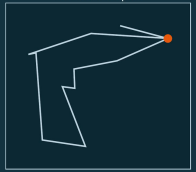
Groningen · Ankara · Münster · Vienna · Lisbon. Real submissions:

**From our real data. Every one passed a check. Every one is wrong.**

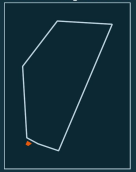
a self-intersecting polygon



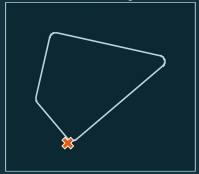
an antenna spike



a 322 km² 'neighbourhood'



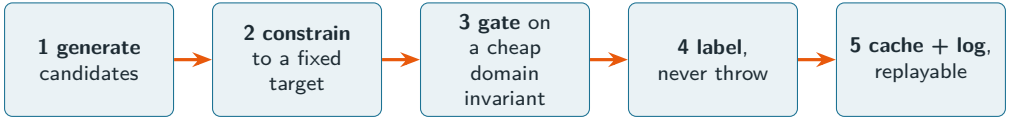
a home outside its neighbourhood



**Citizens drew these. We owe science clean datasets.**

# One architecture, used three times

This is not a survey talk. Inside the pipeline there is **one defensive chain**, instantiated three times: polygons, routes, AI. Every link is FOSS.




POLYGONS five scored repairs, the home must stay inside

ROUTES three engines, one schema, routed $\geq$ straight line

AI cache $\rightarrow$ lookup $\rightarrow$ model, a 0.95 floor, a receipt per decision

*Everything after this slide is proof. At the end you get the whole table to take home.*

# One survey, three accessibilities

1. **Objective**, what network + land use make reachable
2. **Perceived**, how reachable it *feels* (Likert batteries)
3. **Actualized**, where people actually go

**The research object is the mismatch** between the three.

**WP5 / PPGIS via Maptionnaire:** respondents *draw their perceived neighbourhood freehand* and pin home, work, study, shopping, leisure, healthcare.

**5**
cities

**≈4,800**
respondents

**70+**
survey items

**5**
languages

Groningen · Münster · Ankara · Vienna · Lisbon

# The artefact, and why it needs defending

**63,442** lines of R

**422** top-level functions

*One monolithic script, **on purpose**: co-researchers read it end to end.*

FOSS below the survey layer: `R+sf` (GEOS) · `ORS/OSRM/r5r` · GeoPackage out, GeoJSON in · `Mistral` (open-weights path ready)

**Every row was made by a human**

- citizens draw **bow-ties, spikes, holes**; homes outside their own boundary
- free text in **five languages**, sometimes in one column
- **seven EPSGs declared** across five cities; four exercised so far
- a full run costs **2 h 25 m** ⇒ **five caches** ⇒ **cache invalidation is a first-class problem**

Three innovations, one per failure family, one chain under all of them.

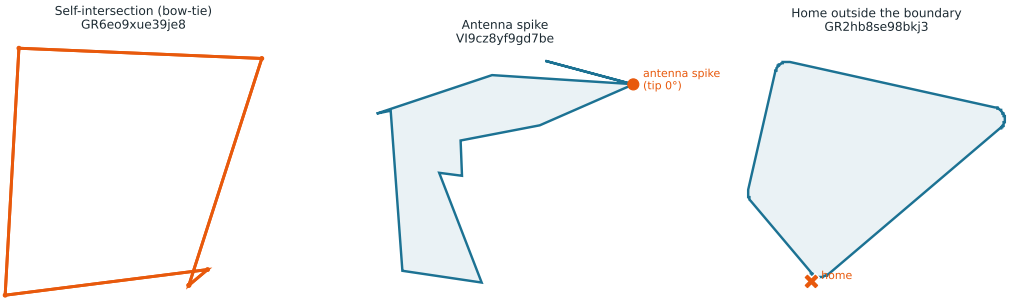
POLYGONS → ROUTES → AI  INNOVATION 1 · citizen drawings into analytic geometry

# Geometry repair for freehand citizen drawings

**1,028** polygons drawn (Groningen)

**944** ship as valid analytic geometry

# The defect catalogue, these are *real* drawings, not diagrams




*0 of 3,821 drawings had an interior ring — nobody drew a hole on purpose (fact-check confirmed)*

real corners **60–120°**; spike tips **<30°** ⇒ detector at **25°/120°/50 m** · homes **200–500 m** outside = drawing errors · compactness floor **0.01**

*Defensive code starts with looking at your data, so we did.*

# The algorithm: one FOSS tool per drawing problem

| Drawing problem | What the step does | FOSS tool (why) |
|---|---|---|
| self-intersection (bow-tie) | score four candidate fixes, keep the least-destructive *valid* + *simple* one | `sf` → GEOS: `st_make_valid`, `st_convex_hull`, concave hull, vertex reorder |
| antenna spike | prune sharp-tip vertices, then morphological *open* | custom R (angle test) + `sf::st_buffer(`$\pm d$`)` |
| ragged edge, slivers | a zero-width buffer re-nodes the boundary | `sf::st_buffer(0)` (GEOS) |
| interior hole | drop interior rings | `nngeo::st_remove_holes` |
| home outside boundary | containment test, 200 m tolerance | `sf::st_contains` |
| degenerate ribbon | compactness floor 0.01 (Polsby–Popper) | `sf::st_area` + custom |

**Why `sf`?** It is the R binding to **GEOS / GDAL / PROJ**, the same C++ engine under PostGIS and QGIS. We write custom R *only* where no primitive exists (spike detection, candidate scoring). **14 stages**, every threshold in config.

# Repair, tiered: the least-destructive fix wins

Five candidates per broken polygon, `make_valid`, concave hull, convex hull, vertex reorder (ccw/cw), each **scored**:

```
if (!st_is_valid(cand) || !st_is_simple(cand)) next  # make_valid's MULTIPOLYGON dies here
if (st_geometry_type(cand) != "POLYGON")       next
if (!check_home_inside(cand))                  next  # a fix that loses the home is no fix
score <- compactness * 0.6 + area_ratio * 0.4        # least-destructive wins
```

then the mechanical stages, **12 repair steps** in all:

buffer(0) → **antenna-spike removal** (25°/120°/50 m, config) → morphological open → close → multipart reconnect → de-holing → convex hull as last resort

Nobody's neighbourhood gets convex-hulled if a gentler repair survives scoring.

The new part is the **choosing**, not the repairing: the candidates come from GEOS, the scored selection is ours.

# Then two gates, and never an exception

```r
check_geom <- sf::st_buffer(cleaned_geom, dist = cfg$neigh_home_inside_buffer_m) # 200 m
is_inside  <- tryCatch(sf::st_contains(check_geom, home_sfc, sparse = FALSE)[1,1],
                       error = function(e) FALSE)
if (!isTRUE(is_inside))
    return(list(idx = idx, wkt_cleaned = NA_character_, status = "HOME_OUTSIDE"))
```

Gate 2: compactness $\geq 0.01$ (Polsby–Popper $4\pi A/P^2$) $\Rightarrow$ `DEGENERATE`.

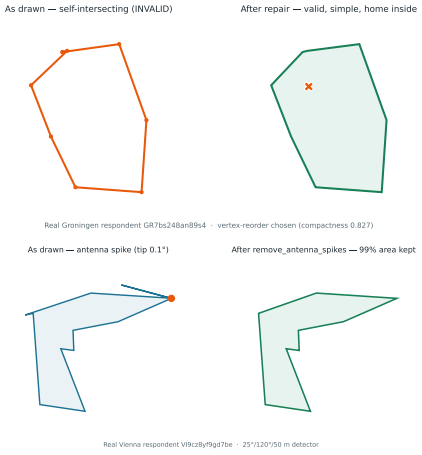
Every row returns `WKT_cleaned` + one of **8 status codes**; downstream *branches*:

`VALID` `HOME_OUTSIDE` `DEGENERATE` `INVALID_WKT`

`CLEANING_FAILED` `NO_GEOMETRY` `NO_HOME_LOCATION` `WKT_CONVERSION_FAILED`

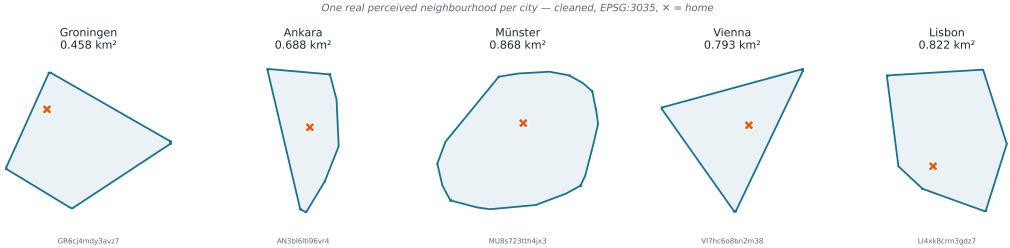
**One invariant, two jobs:** it filters repair candidates · it gates the output (200 m). And a failed row is a **finding**, not a crash.

# Repair, seen: two real fixes




Two real respondents. **Top**, a Groningen **bow-tie** reordered into a valid, simple polygon, footprint preserved, home inside. **Bottom**, a Vienna **antenna spike** removed keeping **99%** of the area. Least-destructive fix wins, no exception thrown.

# Five cities, one cleaner, every context, the same defenses




One real *cleaned* perceived neighbourhood from each city, all EPSG:3035, × = home. One pipeline absorbs five survey dialects, five languages, seven declared CRSes, and every one of these came out of the same fourteen-stage cleaner.

***Takeaway*** *The defenses aren't Groningen-specific, the same code cleans Ankara, Münster, Vienna and Lisbon drawings untouched.*

# What is new here, and what you should not rebuild

We audited our own novelty claims before this talk. The honest split:

| New, claimed narrowly | Standard, use the existing tool |
|---|---|
| **scored, least-destructive selection** among repair candidates (nearest prior art: minimal-change repair in spatial databases, Rodriguez, Bertossi & Caniupán 2013) | single-shot validity repair: GEOS/JTS `make_valid` (structure method), `prepair`, CGAL 2D Polygon Repair |
| **a semantic invariant driving repair choice, gating, and cache validation** (no published geometry tool found that does this) | spike/sliver cleanup: line simplification (Douglas–Peucker, Visvalingam) + morphology (Matheron–Serra); GRASS `v.clean`, mapshaper |
| **per-mode centricity contrast** on routed distances (extends Hasanzadeh 2019, next section) | multi-engine dispatch: `routingpy` (Python), `stplanr` + `r5r` (R) |

*Calibrated claims are the FOSS way: the four novelty-audit reports ship in the talk repo.*

# Steal the invariant, not the cleaner

Any domain with a **cheap, checkable invariant** can run the same chain. Ours is "the home stays inside." Yours already exists:

| **Feature** | **Invariant (all O(1) point-in-polygon or overlap)** |
|---|---|
| cadastral parcel | contains its **address point** |
| building | lies within its **plot** |
| administrative unit | contains its **seat / capital** |
| watershed | contains its **outlet (pour point)** |
| lake / water body | contains its **label point** |
| census tract | contains its **internal point** (guaranteed by design) |
| ML building footprint | keeps $\geq$ X% **IoU with its detection mask** |

Each one can do all three jobs: **filter** candidate repairs, **gate** outputs, **validate** caches.

Packaging: a thin layer over `sf`/GEOS. An R package or a QGIS Processing plugin.

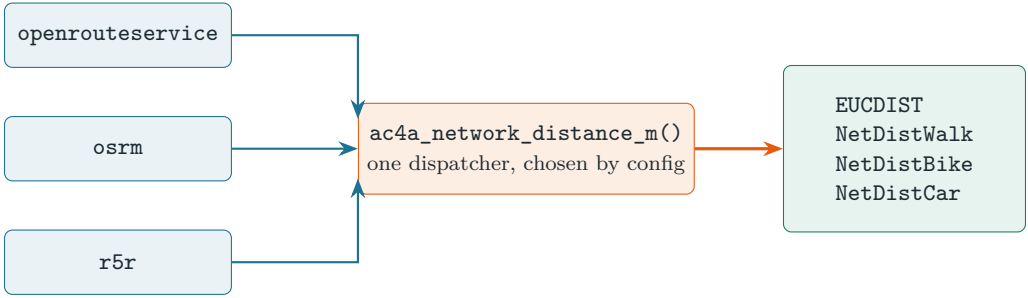
POLYGONS → ROUTES → AI  INNOVATION 2 · one schema over N engines, and a per-mode method

# Routing at survey scale, without trusting the network

≈9,800 pinned home→activity trips

≈29,000 routed legs, two networks live, three queued

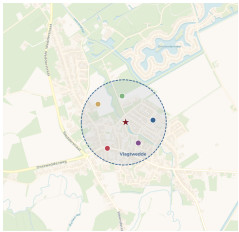
# Three engines, one schema



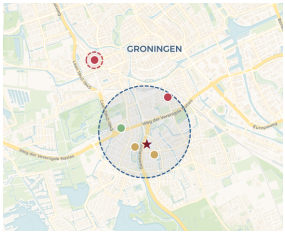

- Downstream code **cannot tell which engine ran**, switching engines touches zero analysis code
- Starting fresh? `routingpy` (Python) or `stplanr`+`r5r` (R) give you this seam off the shelf; ours is the same pattern, R-native, credit where due
- `EUCDIST` always computed alongside, routed $<$ straight-line should fail *loudly*

# Dual-mode activity spaces: single-linkage, pure R

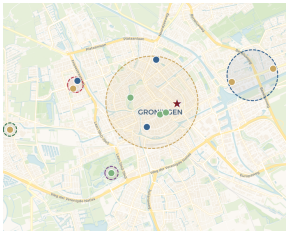
**Per respondent, per mode (walk + bike): single-linkage clustering over routed distances, custom BFS, zero new dependencies. Join threshold = 15 min at door-to-door speed (walk 4.8 km/h → 1.2 km, bike 12.4 km/h → 3.1 km); cluster count ⇒ the three types, on real respondents:**



**Monocentric**, single core



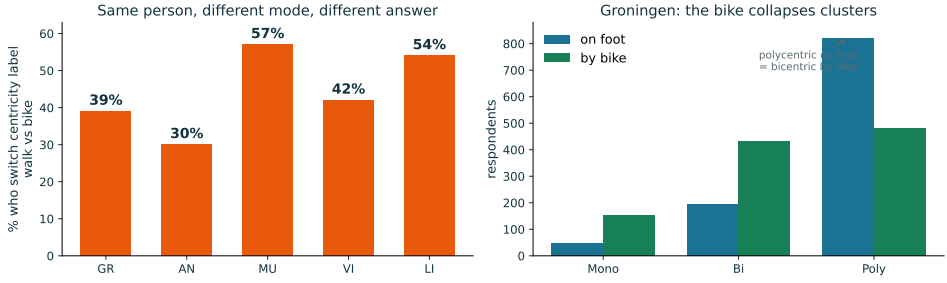

**Bicentric**, two centers




**Polycentric**, dispersed centers

The construct and labels are **Hasanzadeh's** (Travel Behaviour and Society, 2019; K-means, straight-line, one mode-blind label). Ours: **routed** distances, **15-minute budgets**, one label **per mode**.

# The finding, seen: same person, two activity spaces




Per-mode activity-space centricity, five cities, real respondents

Per-mode centricity, **real respondents**, five cities. **Left**: share who *change label* walk vs bike (**30–57%**, Münster highest). **Right**: on a bike Groningen slides toward mono/bi, the bike **glues clusters into one**, always the same direction.

Owed caveat: single-linkage chains, and chaining pushes the same direction; threshold and linkage sensitivity runs are queued before the paper.

# Routing at survey scale: the honest fix is to run it locally

WAR STORY 17 May: a public routing API rejected **1,482 of 1,622** calls, **91% HTTP 429** in one afternoon. We survived with a ladder, **pace** to the quota · **rotate** keys · **degrade** to Euclidean, *labelled*. But that is symptom management.

**The real fix: stop trusting a remote API. Run routing locally.** `r5r` runs the R5 engine *on your own machine*, against a dated OSM extract you downloaded. **No rate limits**, nothing leaves the machine, and **reproducible by construction**, you pin the snapshot and the engine version.

***Takeaway*** *Pace, rotate and degrade only keep a flaky remote API alive. Running one engine locally makes the rate limit disappear, and pins reproducibility for free.*

POLYGONS → ROUTES → AI  INNOVATION 3 · reproducible AI attribute-coding

# An AI step you can defend in peer review

free text in NL DE TR PT EN out: the survey's own codes, nothing else

*The pattern matters more than the model.*

# Three tiers before the model speaks

1 · CACHE → 2 · LOOKUP TABLE → 3 · MISTRAL API

The system prompt, verbatim (trimmed):

```
CORE RULES:
- You ONLY classify into the predefined
  categories. You NEVER invent new ones.
- When unsure, report no match rather
  than a wrong match.
- You respond ONLY with the requested JSON.
```

Confidence calibration, from the prompt:

| | |
|---|---|
| `huisarts` → GP | 0.98 |
| `library` → Study? Leisure? | 0.85 |
| `going outside` | 0.55 |
| gibberish / "nothing" | 0.00 |
| **applied only if** | $\geq$ **0.95** |

below the floor: the original text **stands**

# Tier 2, concretely: the deterministic lookup table

No model, no variance, known phrasings resolve by regex first:

| pattern (verbatim) | maps to | context |
|---|---|---|
| `^hbo$` | Bachelor's or equivalent | Dutch education |
| `bakfiets\|cargo.?bike` | Bike / bicycle | travel mode |
| `regiotaxi\|deeltaxi\|wmo.?vervoer` | On-request PT | travel mode |
| `^chorprobe$` | Social clubs (416) | activity subtype |

**324 patterns** resolve deterministically, no variance, no tokens spent.
Only text that survives *both* the cache and this table ever reaches the API.

# Tier 3, the model: a real request, verbatim from the code

```
SYSTEM  (verbatim)
You are a multilingual classification assistant for PPGIS survey data
about urban accessibility, travel behavior, and mobility.
CORE RULES:
- You ONLY classify into the predefined categories provided.
  You NEVER invent new categories.
- You are conservative: when unsure, report no match rather than a wrong match.
- You respond ONLY with the requested JSON format. [... no fences ...]

USER  (verbatim; the coded list and 4 injected blocks shown as [...])
TASK: ...originally classified as "Other activities" but the respondent
wrote in the "Other" field: "Sanquin". Re-categorize, or set false.
AVAILABLE ACTIVITY TYPES AND SUBTYPES:      [... 6 types, 40 subtypes ...]
  ## Healthcare (ActivityType code: 5)
     - General medical practitioner (GP) (ActivitySubtype code: 510)
     - Dental care practitioner (ActivitySubtype code: 511)   [...+38 more...]
[... language hints . confidence scale . gibberish rules . reasoning steps ...]
OUTPUT FORMAT - respond ONLY with this JSON:
{ "reasoning": "...", "should_reclassify": true or false,
  "new_activity_subtype_code": <integer code> or null, "confidence": 0.0-1.0 }
EXAMPLE  Input "tandarts" -> {..., "new_activity_subtype_code": 511, "confidence": 0.98}
```

The system message is the leash; the user message carries the surveys own codes, a strict JSON contract, and two worked examples baked in.

# Tier 3 with receipts: the audit log, real rows

`ai_reclassifications_20260525_195237.csv`, **380 logged decisions** from the production run:

| original_text (lang) | matched_value | code | conf |
|---|---|---|---|
| Winkelen (NL) | Other goods shopping | 311 | 1.000 |
| Kroeg (NL) | Eating and drinking | 417 | 1.000 |
| hond uitlaten (NL) | Leisure in public parks | 413 | 1.000 |
| Koor (NL) | Social clubs / community | 416 | 1.000 |

**And the failures that set policy** (documented in-code): `Therapy` (Healthcare) → Shopping at **conf 0.98**; `Sanquin` (blood bank) → Sports at **0.95** ⇒ cross-activity AI now restricted to type-6 "Other" only.

**Confidence is triage, not truth**: calibrate the floor on a labelled gold set; the log doubles as the agreement dataset (model as one coder, human as the other, Cohen's kappa as the gate).

# Determinism you can actually get, and a model you can swap

The AI cache keys on the **normalized text itself**, so the same input always replays the same answer:

```
text_norm <- tolower(trimws(as.character(input_text)))   # the cache key IS the content
```

- **temperature 0 is not determinism** (GPU batch order breaks bit-identity); **caching on content is**, the boring fix that actually works
- re-runs **replay the cache**, not the model ⇒ deterministic *and* cheap
- the harness is model-agnostic: a local **Mistral 7B / Llama** is a one-function swap, the fully-FOSS endgame

# Beyond the survey: the same AI step, for GIS at large

**One operation:** messy, multilingual free text → a **controlled vocabulary**, cached, confidence-floored, logged, replayable. Never let a model free-generate into a codelist: **retrieve candidates, constrain, select-or-abstain**.

### ATTRIBUTES, production-ready

messy categoricals (OSM tags, land use) → **CORINE / HILUCS**. Measured: LLM-embedded OSM tags lifted building-function F1 by **12.5%** avg, six cities (peer-reviewed, 2024).

### METADATA, promising

abstracts + keywords → **ISO 19115 / INSPIRE / GEMET**. Evidence: LLMs fill fields well but code weakly against big vocabularies (EuroVoc F1 ≈0.43–0.55 vs BERT 0.73), so the **floor + curator stay**.

### CROSS-LINGUAL, promising

tongues → one code (real): `huisarts`, `Hausarzt`, `aile hekimi`, `médico` → **GP**. Set **per-language floors**.

### GAZETTEER, promising

free-text places → gazetteer IDs. **Ground it** (GeoNames/OSM retrieve-then-rank); ungrounded LLM geocoding invents places.

# The chain, closed: steal a row, or a single cell

| | 1 generate | 2 constrain | 3 invariant gate | 4 label | 5 cache + log |
|---|---|---|---|---|---|
| POLYGONS | five scored repair candidates | valid, simple, single POLYGON | home inside (200 m) | 8 status codes, rejects ship | cleaning cache, replayable |
| ROUTES | three engines, one dispatcher | one 4-column schema | routed $\geq$ straight line | DEGRADE flag per row | distance cache, replayable |
| AI | cache $\rightarrow$ table $\rightarrow$ model | the survey's own codes | confidence $\geq 0.95$ | below floor, text stands | content key, audit CSV |

***Takeaway*** *Generate · constrain · gate on a cheap invariant · label, never throw · cache with receipts. That is the whole reproducibility contract, and every cell is FOSS.*

# Three innovations, one chain

Take these home:

1 `POLYGONS` **Repair + gate at the boundary** , scored candidates, status codes

2 `ROUTES` **One schema over N engines** , pace, rotate, degrade, and make the fallback **confess**; pin what routes

3 `AI` **AI with receipts** , adopt the harness first and the model last; constrained, floored, logged

*The through-line: a reproducibility contract for mixing AI with deterministic geoprocessing. `sf` stays exact, the model stays on a leash, the whole run replays.*