GIS/Spatial analytics : quel potentiel pour les données géographiques

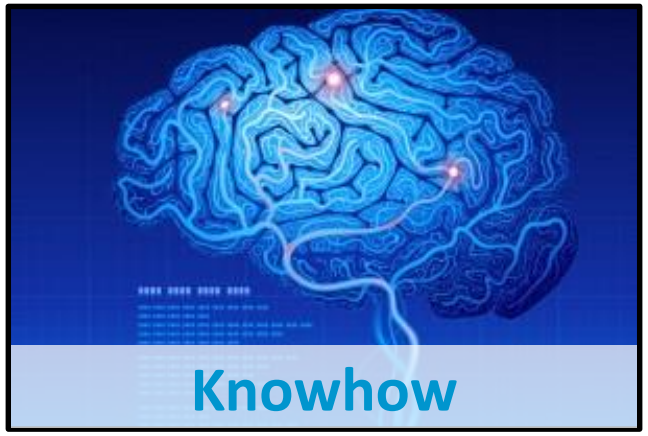
Vandy Berten – Smals Research

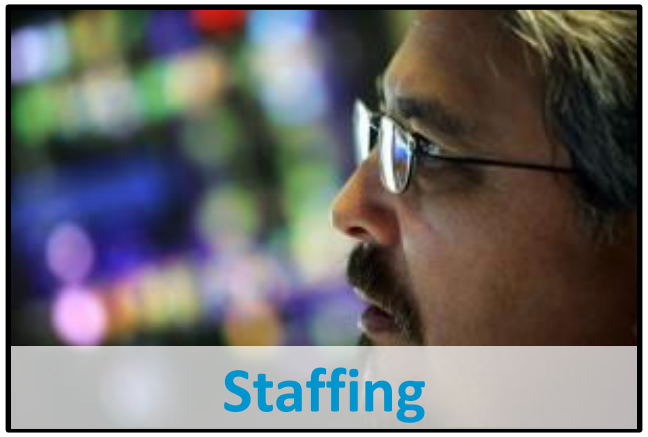
uMons, 23/2/2022

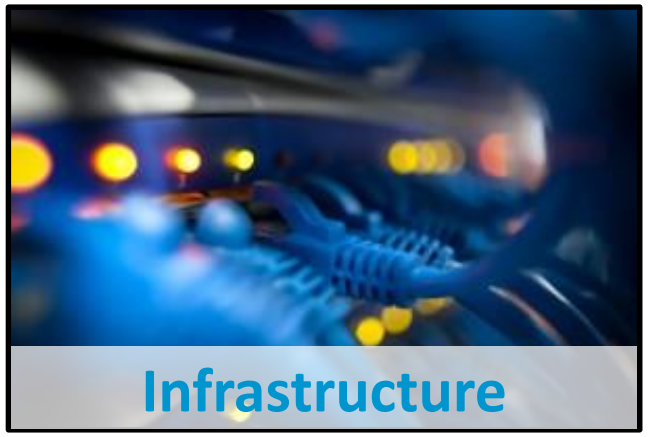
Smals
ICT for society

Smals
ICT for society

SUPPORT FOR E-GOVERNMENT







g-cloud

top EMPLOYER BELGIË BELGIQUE
CERTIFIED EXCELLENCE IN EMPLOYEE CONDITIONS

WWW.SMALS.BE

Smals
ICT for society

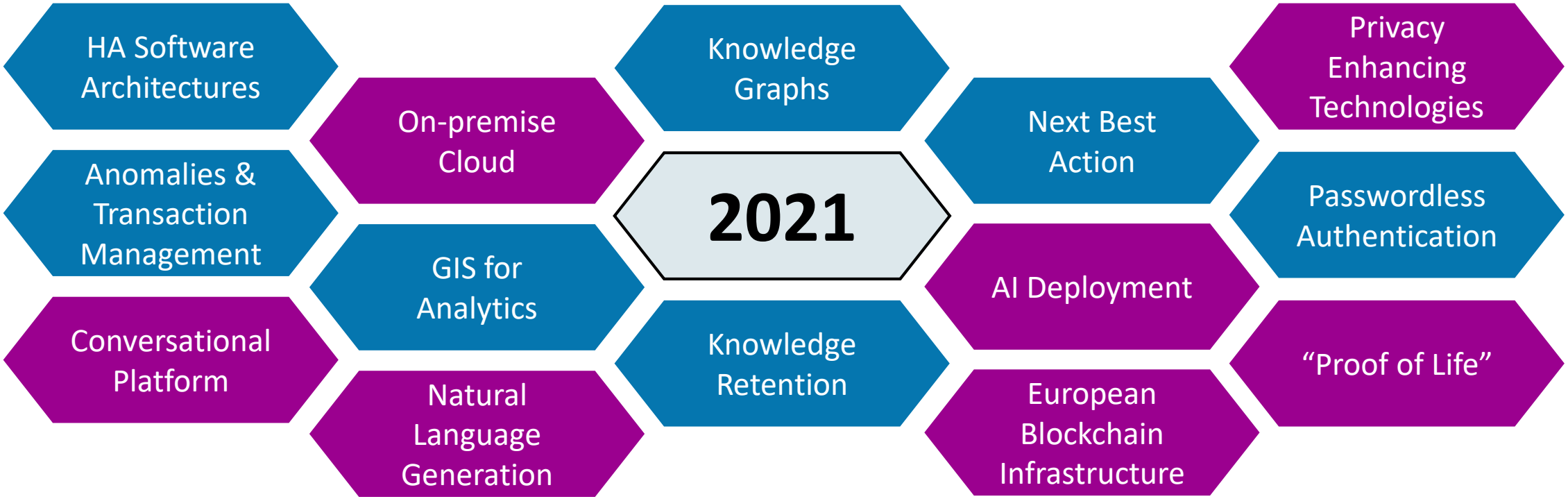
Smals Research 2021

Innovation with new technologies

Consultancy & expertise

Internal & external knowledge transfer

Support for going live



Smals ICT for society

www.smalsresearch.be

Contenu

- Définition
- Géocodage
- Sources de données
- Visualisation
- Jointure spatiale
- Autocorrélation spatiale
- Atteignabilité
- Anomalies CAW

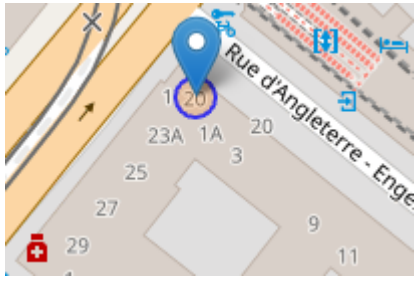
Contexte

- On s’intéresse à l’aspect collecte de données/analytique, pas/peu aux aspects visualisation
- Pas à des modèles épidémiologiques
- Pas à de la cartographie (nature de sols, réseau routier, hydrologie…)

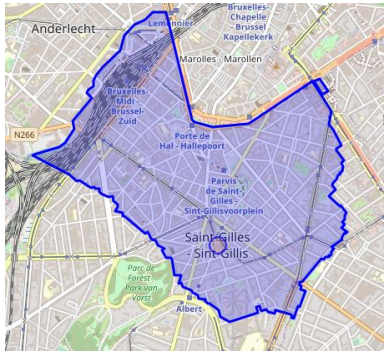
Donnée géographique

- Donnée géographique : donnée que l'on peut placer sur une carte, directement ou indirectement
- Directement :
  - 50.8361263, 4.3382716 (WGS 84 – EPSG:4326)
  - 50°50'10.1"N, 4°20'17.8"E
  - 147859.164113, 169482.701283 (Lambert 1972 – EPSG:31370)
- Indirectement :
  - Avenue Fonsny 20, 1060 Saint-Gilles
  - 1060
  - Saint-Gilles
  - Smals, Saint-Gilles
- Objet spatial/géographique :
  - Point (localisation précise)
  - Ligne (segment de route)
  - Rectangle ("bounding box")
  - Polygone

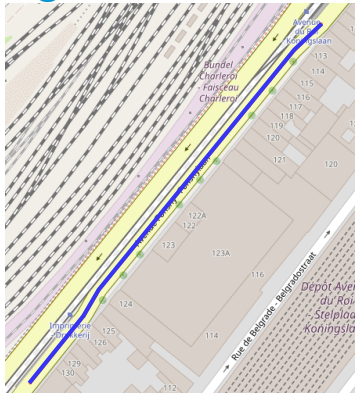
Point



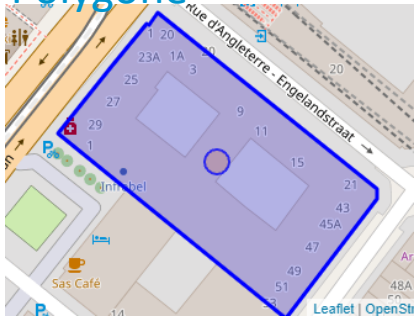
Polygône



Ligne



Polygône



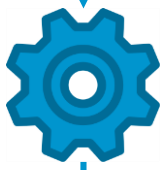
Smals
ICT for society

Géocodage

Geocoding: concept

Pour être exploitable géographiquement, un adresse doit être “géocodée”

“Av. Fonsny 20, 1060 Bruxelles”



```
{“street”  : “Avenue Fonsny”,
 “housenbr”: “20”,
 “zipcode” : “1060”,
 “city”    : “Saint-Gilles”,
 “country” : “Belgique”,
 “location”:[50.8358,4.3361]}
```

Address validation:

- Data quality:
  - Standardization
  - Deduplication
  - Matching
- Auto-completion (type-ahead)
- …

Geocoding:

- Visualization
- Controls
- Spatial (GIS) analytics
- …

Or polygon (point sequence)

Smals ICT for society

Geocoding: tools

- Cloud commercial providers:
  - Here we go
  - Bing maps
  - Google maps
  - TomTom
- Sometimes free in specific conditions

- Mainly one open-source/open-data solution: OpenStreetMap/Nominatim
- Variant (with ElasticSearch): Photon
- On-premise possible (1 Docker cmd line) for a specific region

OpenStreetMap

Smals
ICT for society

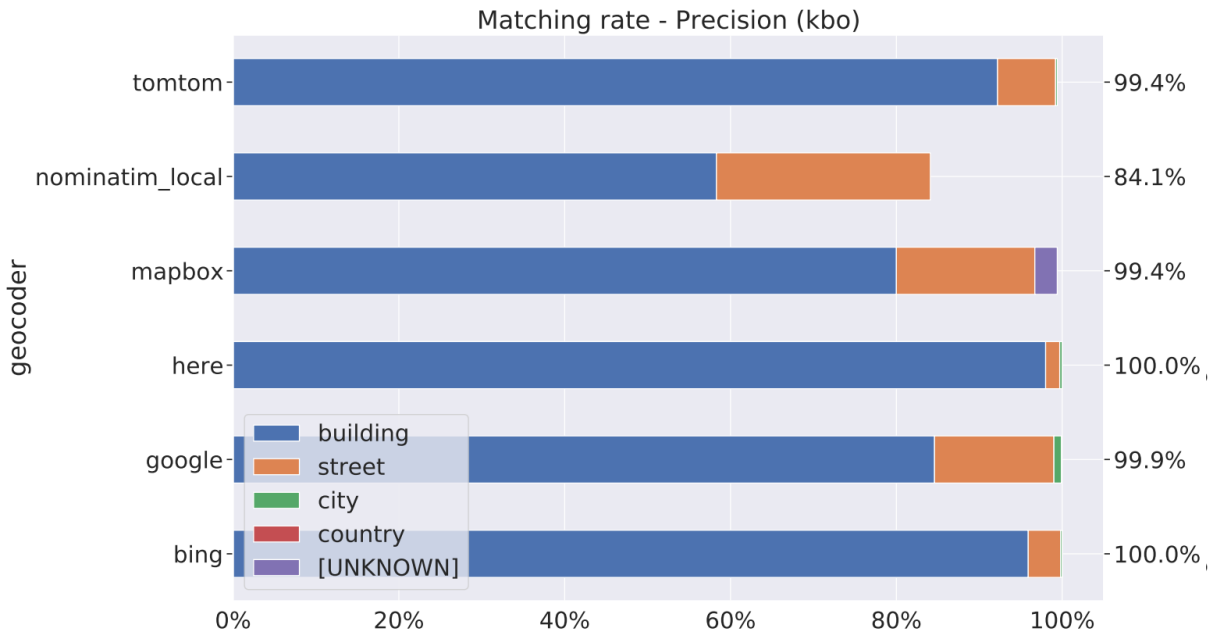
Geocoding: pros/cons

	Cloud solutions	On-premise solutions
Pros	• Easy to implement • No infrastructure cost • High quality	• Confidentiality • OK for large volumes • OSM: no data/soft costs
Cons	• GDPR • Confidentiality • Requires internet access • Cost for large volumes (time, €)	• World coverage unaffordable • Quality? • Infra costs

Smals
ICT for society

Geocoding: quality

Dataset : 1,000 (belgian) addresses from KBO OpenData



- Matching rate: proportion of addresses with a match
- Precision:
  - Building: localizes the building
  - City: localizes the city center
- Main weaknesses of Nominatim/OSM:
  - Not very robust
  - Often street level
    → not a problem in most of our applications
  - Weak output coherence (~10 fields for « street »)
- But: on-premise!

Smals
ICT for society

Nominatim: weak robustness (input)

- In general, does not support “small typos”
- Robustness for house numbers:
  - Avenue Fonsny 20, 1060 Bruxelles → OK
  - Avenue Fonsny 20 bte 2, 1060 Bruxelles → No result
  - Avenue Fonsny 20/2, 1060 Bruxelles → Street level (skip 20/2)
- Robnustness with names:
  - Avenue Fonsny 20, 1060 Bruxell**e** → No result
  - Avenue Fons**ni** 20, 1060 Bruxelles → No result
  - Aven**u** Fonsny 20, 1060 Bruxelles → No result
- With an unknow house number:
  - Rue des Guildes, 22, 1000 Bruxelles → Street level (skip 22)
  - Rue Gray 96, 1040 Etterbeek → No result
  - … but : Rue Gray, 1040 Etterbeek → Street level
- Real life example : « Avenue Fonsny, en face du 20 », « … (entrée rue X) », « Gallerie Y »

Nominatim: weak coherence (output)

- In general : https://nominatim.openstreetmap.org/search.php?q=**Avenue+Fonsny+20,+1060+Bruxelles**&format=jsonv2&addressdetails=1

➔

```
{ place_id: 48588284,
lat: "50.8358216", lon: "4.3386884", (…)
address: {
   house_number: "20",
   road: "Avenue Fonsny",
   town: "Saint-Gilles", (…) }
```

- But “street” in sometimes in another field : “road”, “pedestrian”, ”footway”, “cycleway”, “path”, “address27”, “construction”, “hamlet”, “park”
- “ city” is sometimes in “town”, “village”, “city_district”, “county”, “city”.

Smals ICT for society

Nominatim: observations

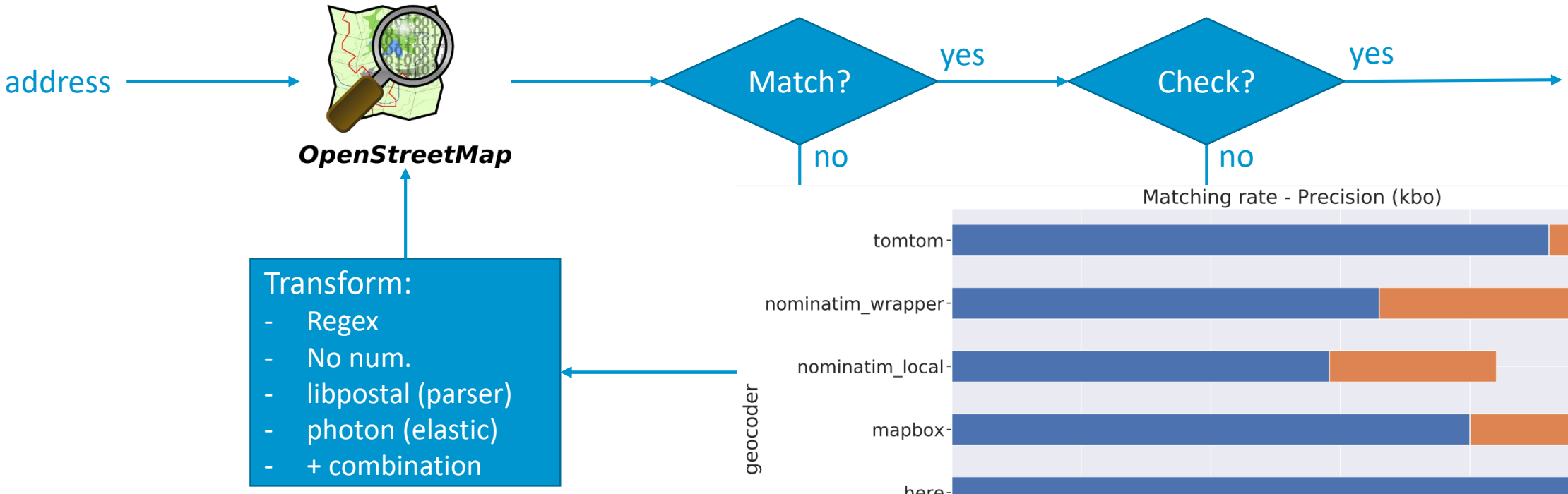
- Often, simple change allows matching :
  - Av. Fonsny → Avenue Fonsny
  - Avenue Fonsny 20 bte 2 → Avenue Fonsny 20
  - Rue Gray 96 → Rue Gray
  - Rue de Namur, Spy (Jemeppe-s-s) → Rue de Namur, Spy
  - Specific to dataset: “Fonsny (Avenue)”, “A.FONSNY” → Avenue Fonsny
- Simple tools allow to “clean” addresses:
  - Regex
  - Libpostal (address parser) : “Smals, Avenue Fonsny 20 bte 5, 1060 Saint-Gilles” →

```
[('smals', 'house'),
 ('avenue fonsny', 'road'),
 ('20', 'house_number'),
 ('bte 5', 'unit'),
 ('1060', 'postcode'),
 ('saint-gilles', 'city')]
```

  - Photon = Nominatim+ElasticSearch

NominatimWrapper

- To address the robustness problem, we build a (REST API) « wrapper » around Nominatim
- Main idea: if Nominatim does not recognize an address, « clean it » beforehand



Transformers performances (KBO – 3M addresses)

Method	Sent	Match	Match rate (%)	Glob MR	Cum MR
orig	3.062.332	2.551.079	83,31	83,31	83,31
regex[init]	348.011	319.321	91,76	10,43	93,73
nonum	184.503	36.653	19,87	1,20	94,93
libpostal+regex[lpost]	69.780	16.881	24,19	0,55	95,48
libpostal+regex[lpost]+nonum	129.906	895	0,69	0,03	95,51
libpostal+regex[lpost]+photon	91.867	86.853	94,54	2,84	98,35
libpostal+regex[lpost]+photon+nonum	5.400	2.450	45,37	0,08	98,43
photon	8.591	6.038	70,28	0,20	98,62
photon+nonum	2.949	454	15,40	0,01	98,64
nostreet	41.708	40.875	98,00	1,33	99,97

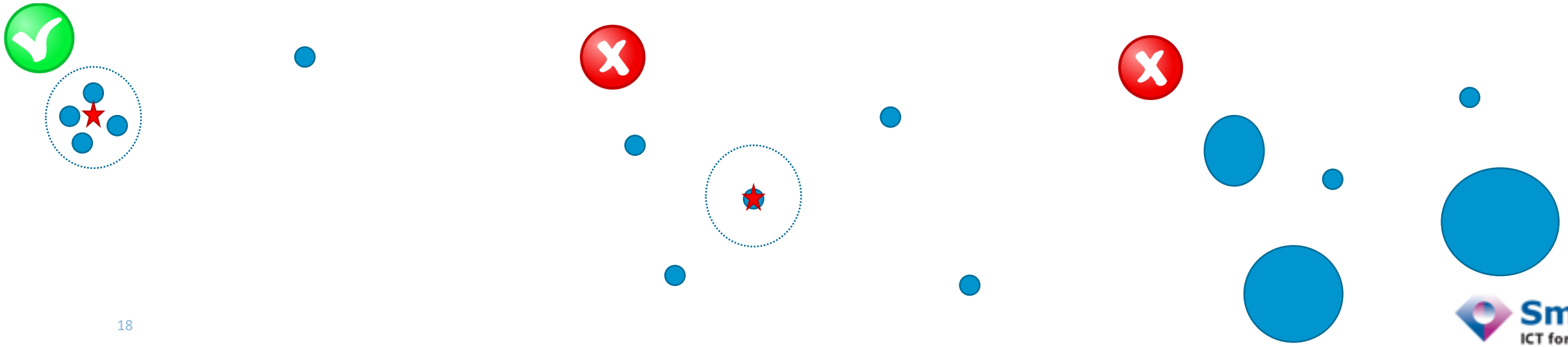
NominatimWrapper: API

- URL: https://github.com/SmalsResearch/NominatimWrapper
- Input:
  - 3 input modes:
    - Non structured: "Av. Fonsny 20, 1060 Bruxelles"
    - Structured: {'street': 'Av. Fonsny 20', 'housenbr': '', 'zipcode': '1060', 'city': 'Bruxelles'}
    - Batch: [{'street': 'Av. Fonsny 20', 'housenbr': '', 'zipcode': '1060', 'city': '…'},
      {'street': 'Quai de Willebroeck', 'housenbr': '38','zipcode': '1000', 'city': '…'}]
  - Options:
    - output mode=geo|short|long
    - check_result=yes|no: checks that the output is not "too far away" from the input
    - with_rejected=yes|no: provides also the rejected results
- Output:
  - Structured address(es): {'street': 'Avenue Fonsny', 'housenbr': '20',
    'zipcode': '1060', 'city': 'Saint-Gilles'}
  - Location: {'lat': '50.8358216', 'lon': '4.3386884'}
  - Level of precision: {'place_rank': 30} (30:building, 26:street, …)
  - Method (transformation): {'method': 'libpostal;regex'}
  - Misc. extra indicators

Smals ICT for society

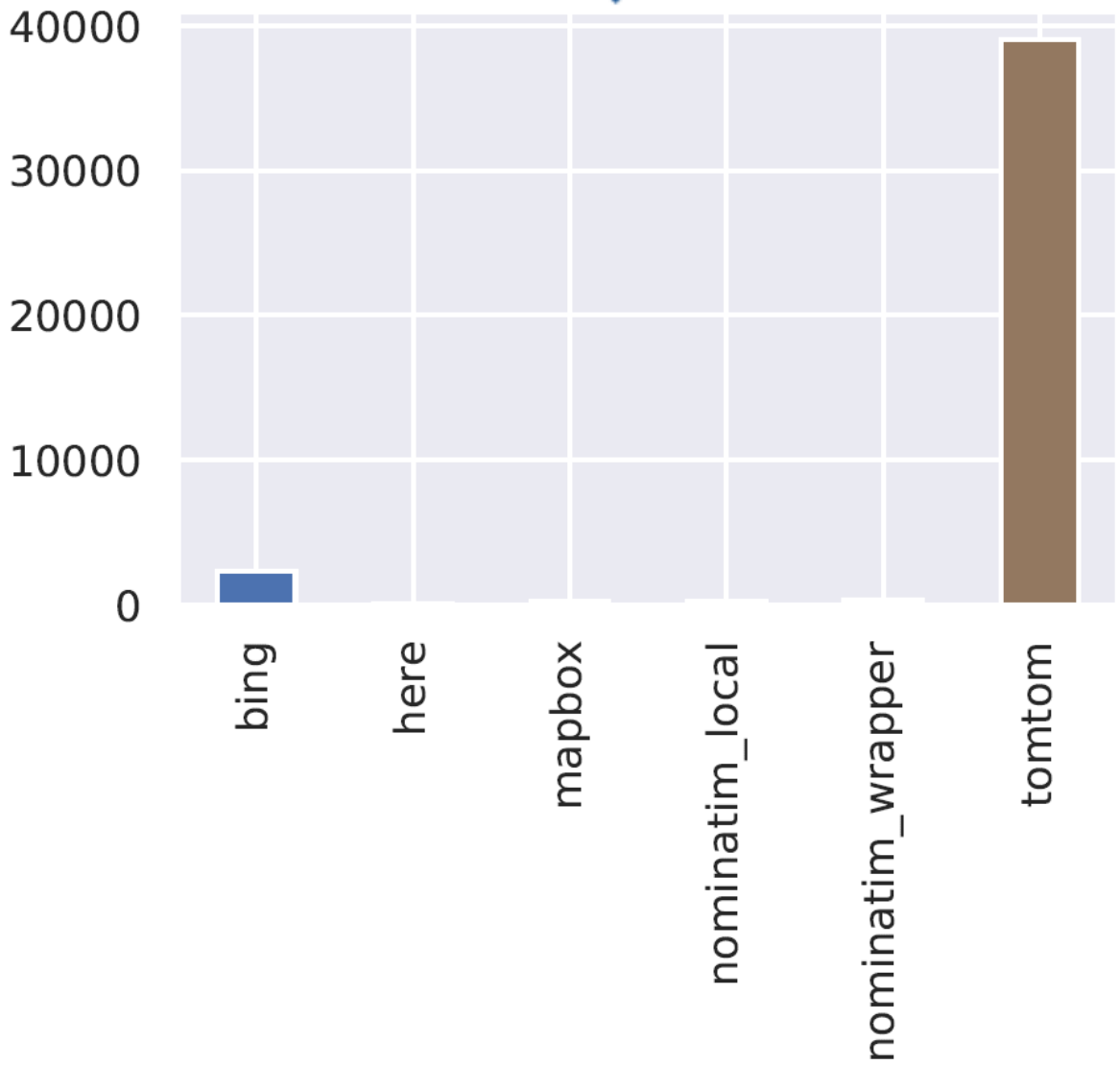
Performances

- Giving an answer does not make this answer to be correct!
- If we don’t have a reference address list with known location, evaluating performance is complex
- Methodology:
  - Collect a (not too clean) address dataset → 1000 addresses from KBO OpenData
  - Send this dataset to several coders → 7: here, bing, mapbox, nominatim (local, BEL), tomtom, google, NW)
  - Use median point for each address as reference, with some conditions :
    - Enough (precise) matches → 950 with ≥ 3 /6 “building” matches
    - Enough answers close to median → 942 with ≥ 2/7 within 100 m
  - Compute misc. stats about « distance to median »



Smals
ICT for society

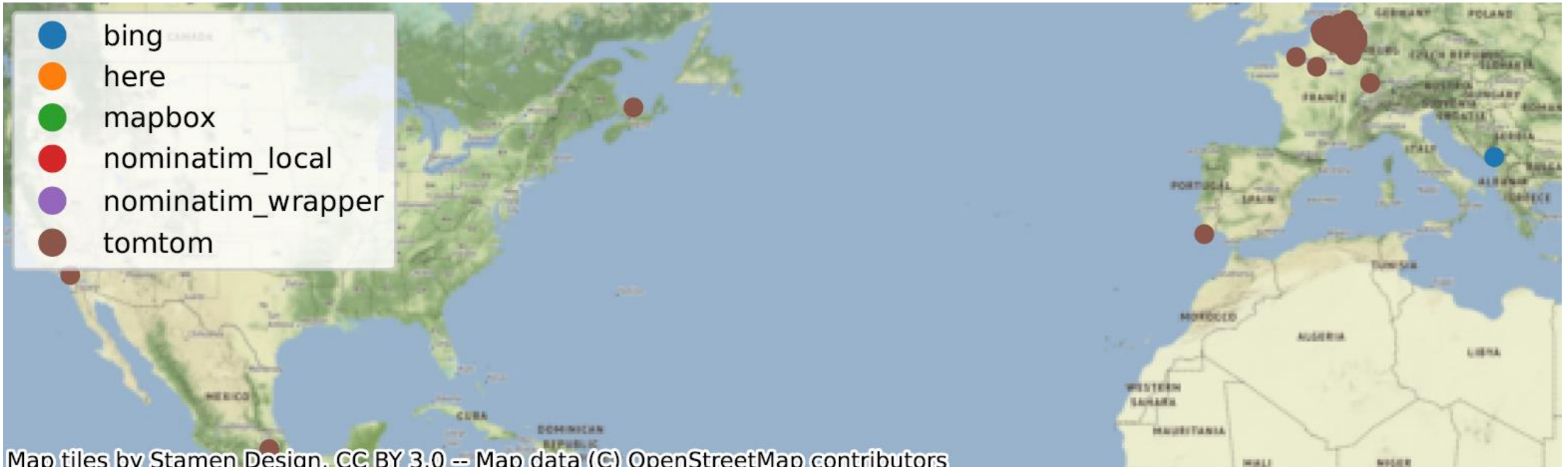
Avg distance to median



Dataset : 1000 addresses from KBO (in Belgium) – a few months ago

Smals
ICT for society

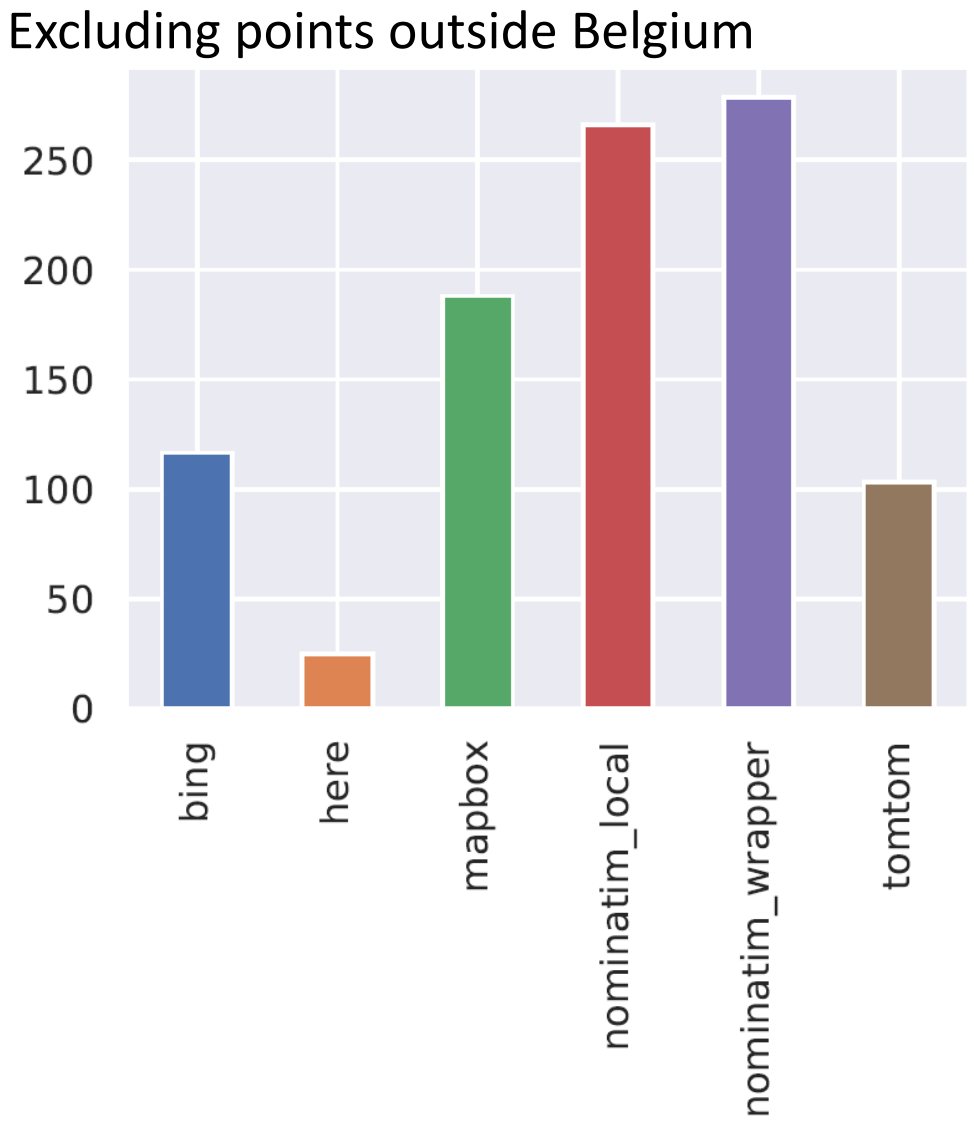
Global view



A few months ago

Smals
ICT for society

Reliable distance to median

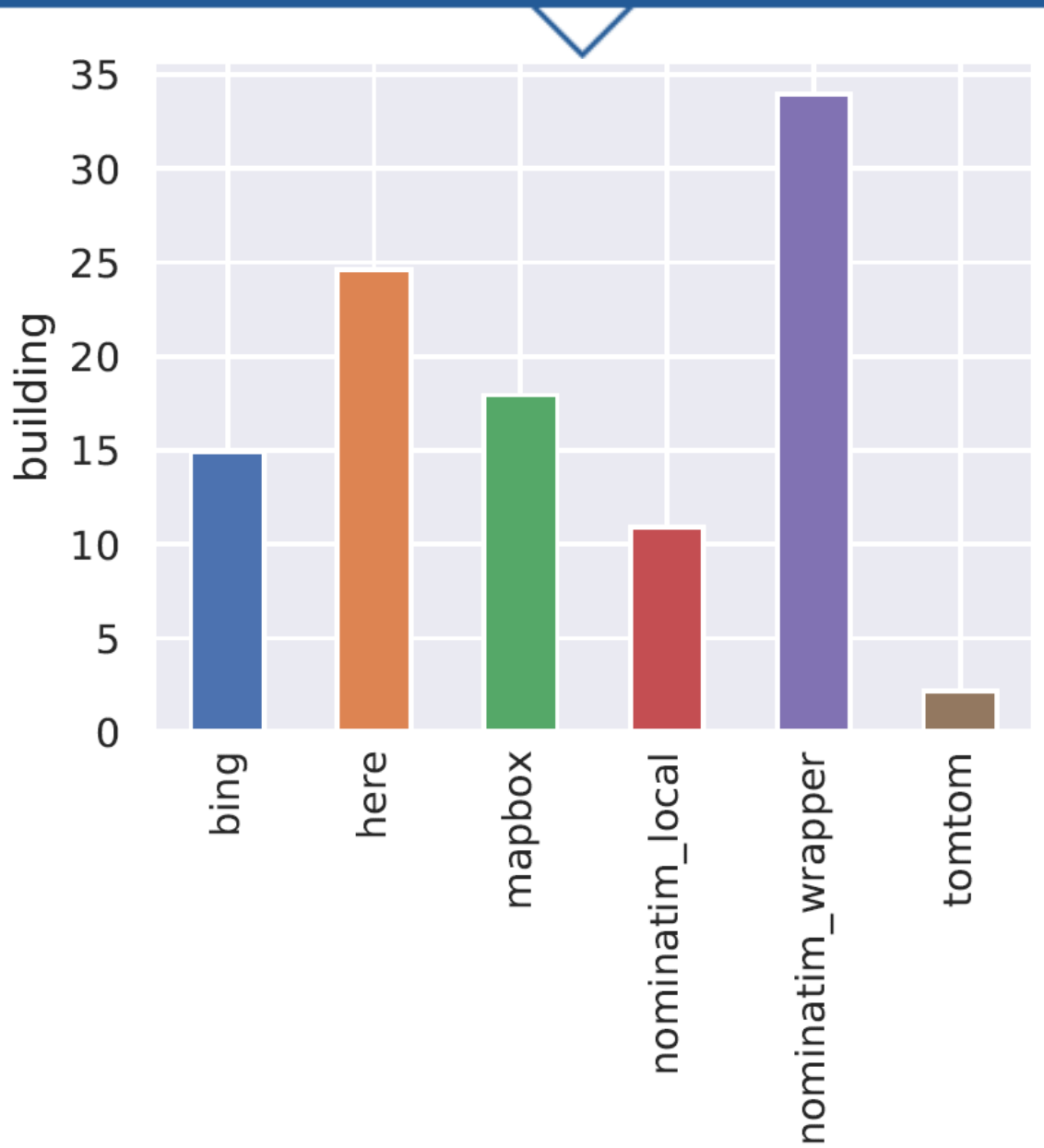


Remind: OSM localizes building only in ~ 65% - only Belgium dataset

A few months ago

Smals
ICT for society

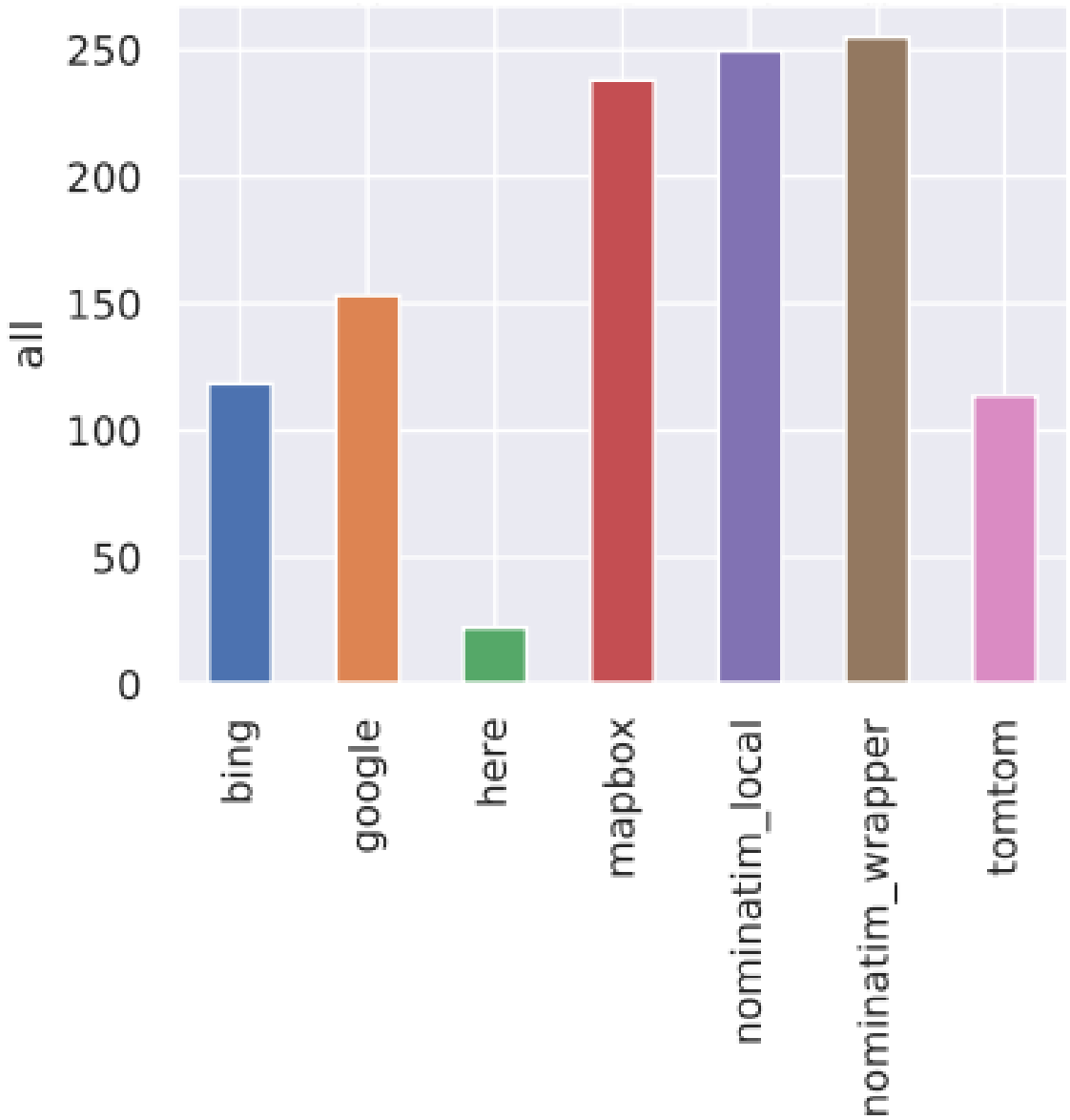
Reliable distance, if “building”



A few months ago

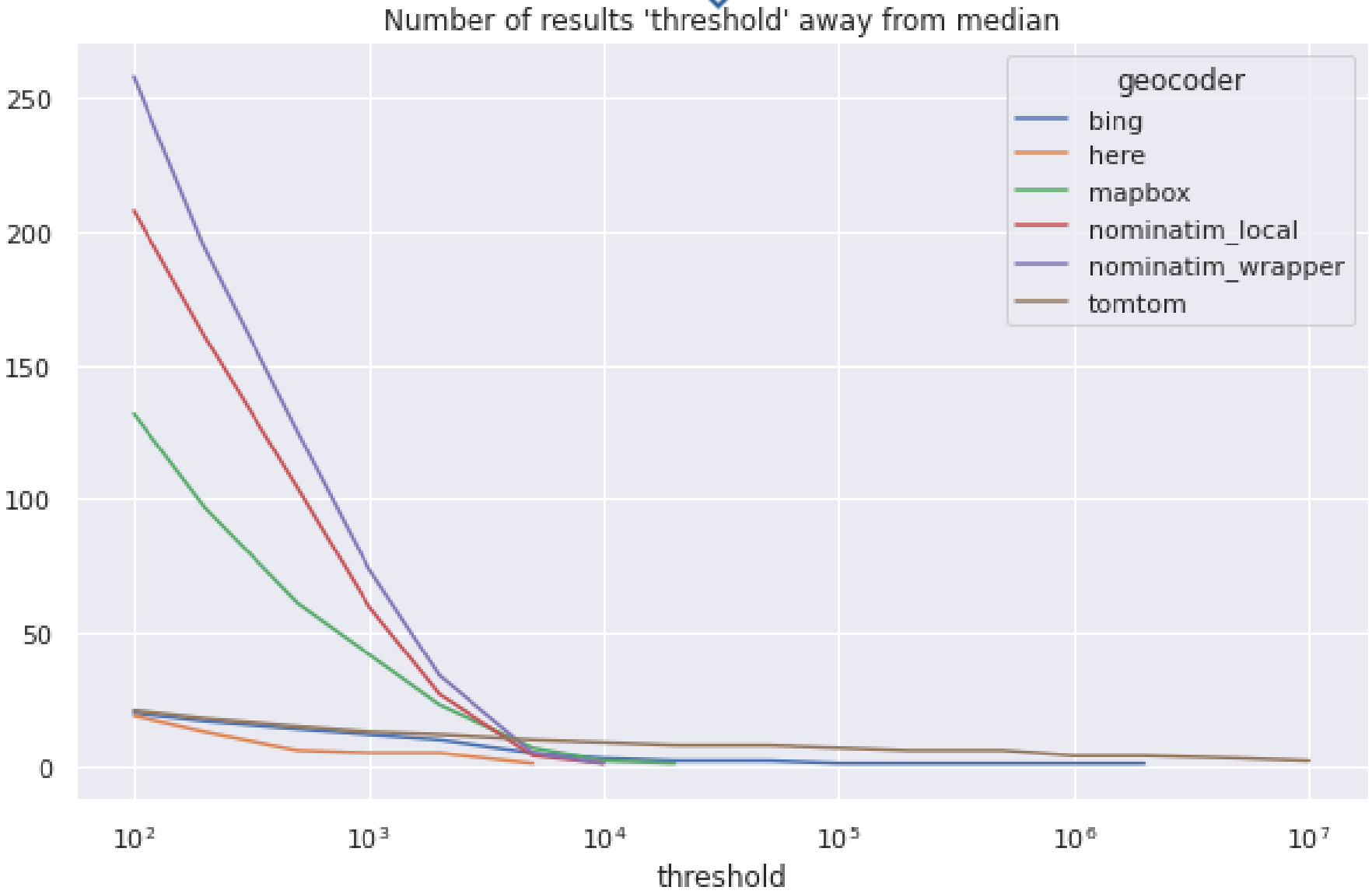
Smals
ICT for society

Avg distance to median - Rerun last week



Smals
ICT for society

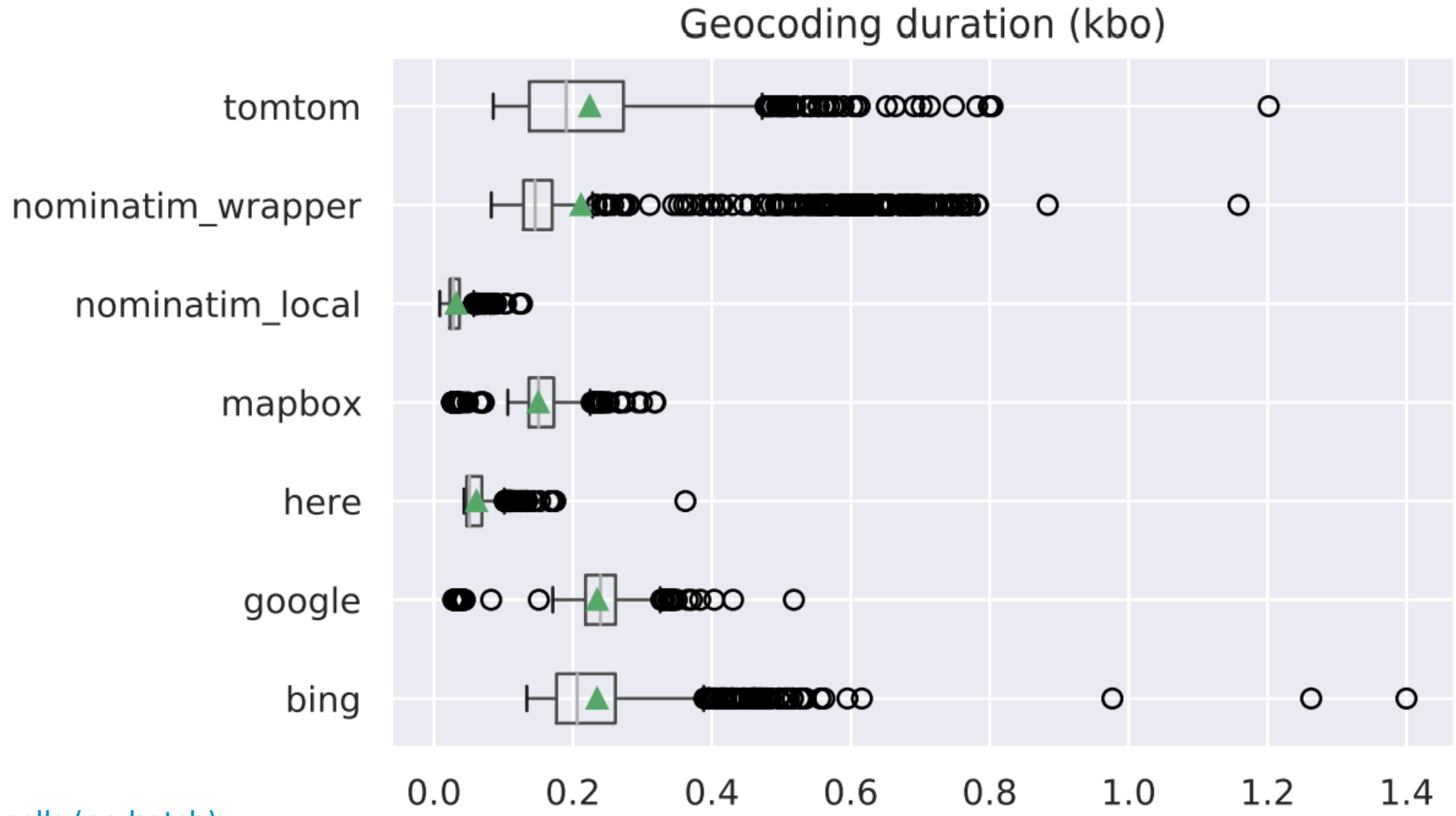
Distance distribution



A few months ago

Smals
ICT for society

Geocoding speed



Sequential calls (no batch)

Smals ICT for society

Data sources

Sources

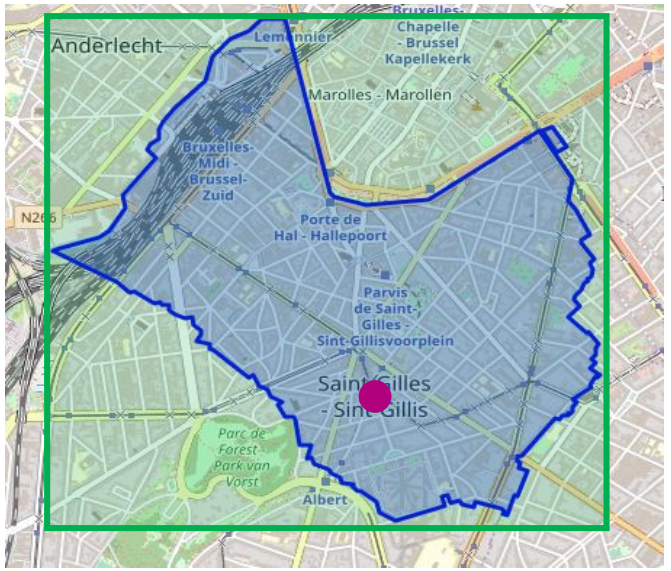
- Principalement deux façons d'obtenir des données géographiques :
  - En enrichissant les données "internes" (adresses, …), via des APIs
  - En utilisant des données authentiques/Open Sources (frontières et hiérarchies d'entités…)
- Les APIs permettent (souvent gratuitement) de collecter une multitude d'information:
  - Géocodage
  - Collecte de POI
  - Courbes isochrones
  - Itinéraires

API – Nominatim (OpenStreetMap)

https://nominatim.openstreetmap.org/search.php?q=**1060+saint-gilles**&format=jsonv2&addressdetails=1&polygon_geojson=1

[{**place_id**: 281729964,
boundingbox:
["50.8200036", "50.8401825", "4.3264847", "4.3588636"],
lat: "50.8249958", **lon**: "4.3454841",
display_name: "Saint-Gilles, Brussels-Capital, 1060, Belgium",
place_rank: 14,
address:
{**town**: "Saint-Gilles",
county: "Brussels-Capital",
region: "Brussels-Capital",
postcode: "1060",
country: "Belgium",
country_code: "be"
}, (...)

geojson:
{**type**: "MultiPolygon",
coordinates:
[[[[4.3264847, 50.8308493], [4.3302516, 50.829586], [4.3310
50.8291262], [4.3312378, 50.8292228],(...)],



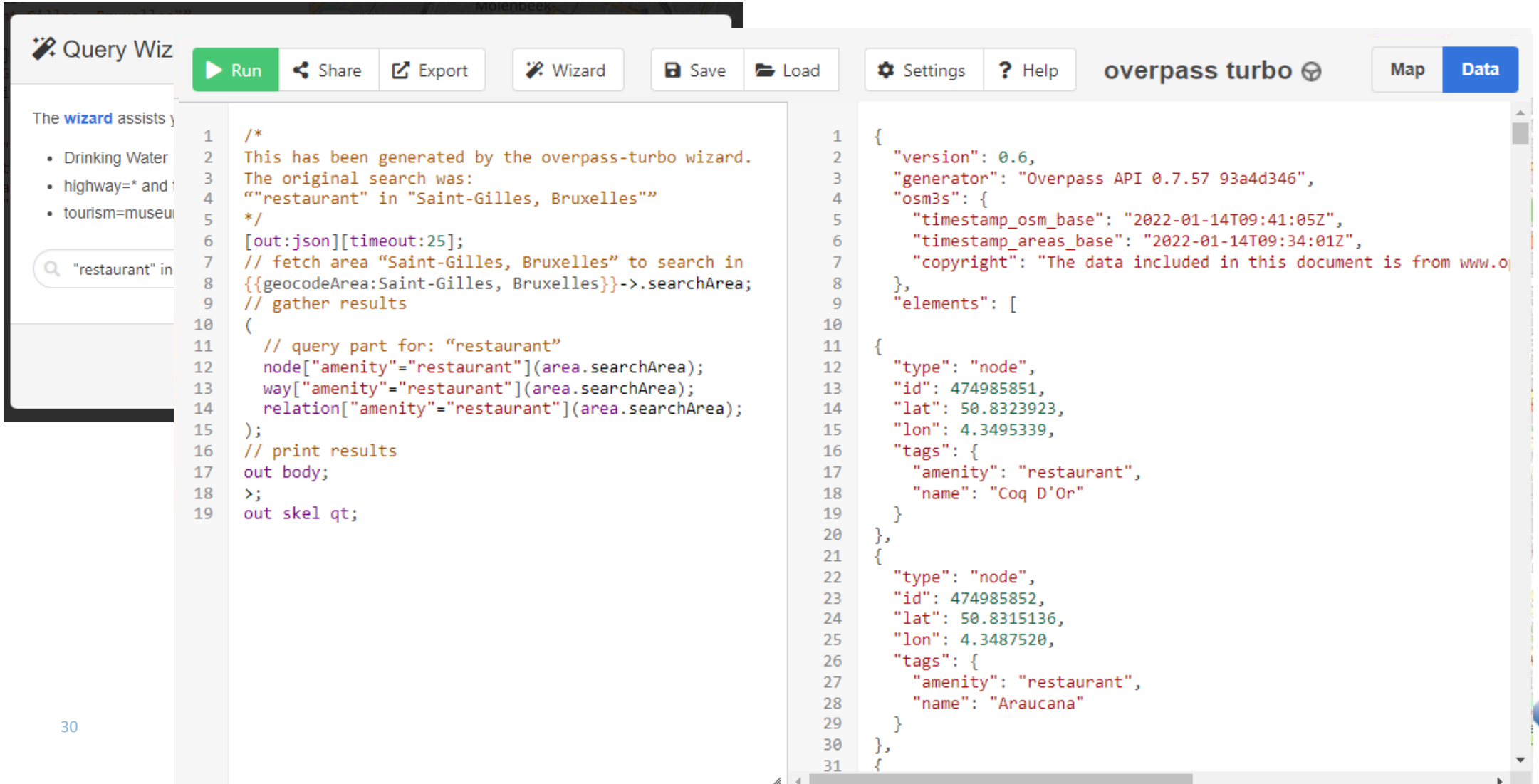
Smals ICT for society

API – Nominatim (OpenStreetMap)

- Base: https://nominatim.openstreetmap.org/(...)
- Adresse exacte :
  (...)search?q=**Avenue+Fonsny+20,+1060+saint-gilles**&format=jsonv2&addressdetails=1
- POI (restaurant, Saint-Gilles):
  (...)search?q=**restaurant,+1060+saint-gilles,+Bruxelles**&format=jsonv2&addressdetails=1
- POI (restaurant, par “bounding box”):
  (...)search?q=**restaurant&viewbox=4.35886,50.84018,4.32648,50.82&bounded=1**&format=jsonv2
- Reverse:
  (...)**reverse?lat=50.83582&lon=4.33868**&format=jsonv2

API : Overpass (OpenStreetMap)

- http://overpass-turbo.eu/



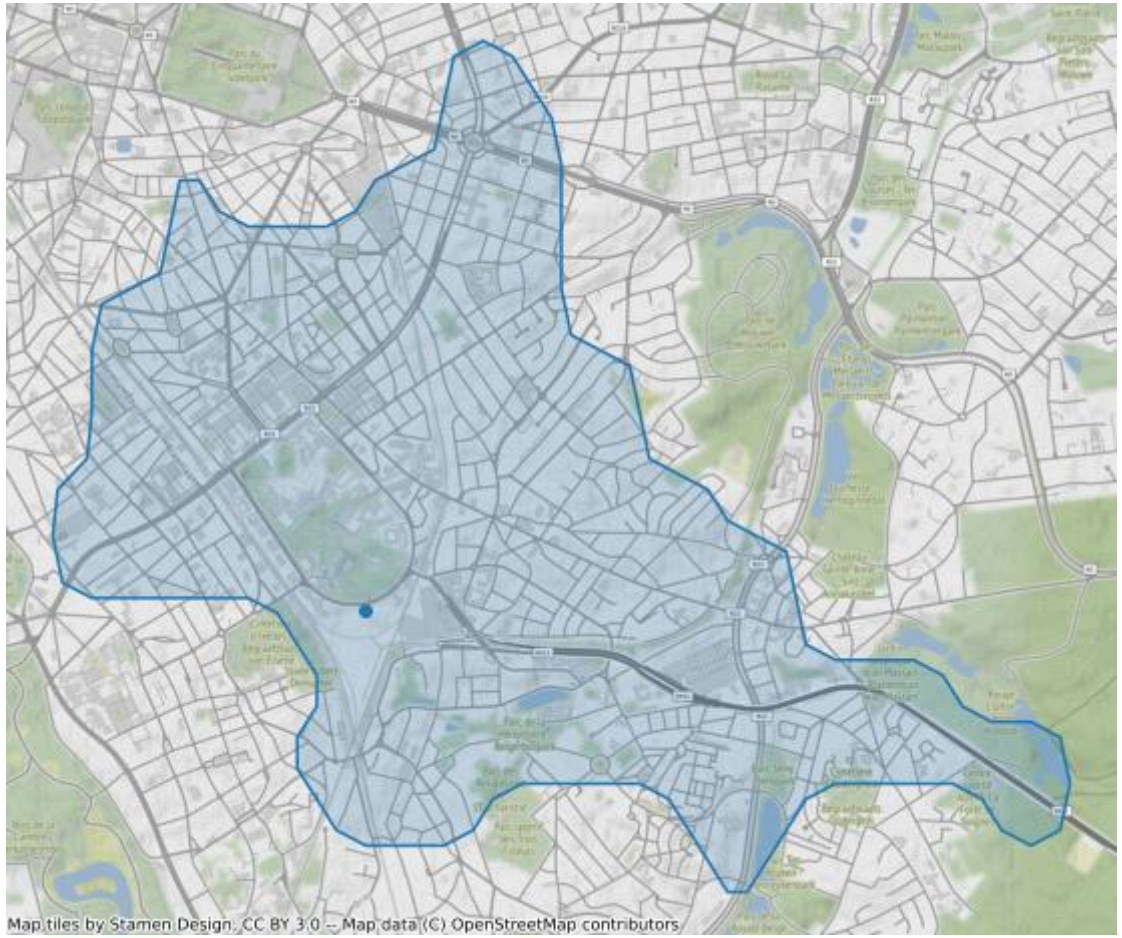
Smals
ICT for society

API : Here (routing)

- Routing:
  - https://router.hereapi.com/v8/routes?transportMode=car&**origin=50.83582,4.33868&destination=50.8249958,4.3454841**&return=summary&apiKey=X(...)zE&departureTime=any
  - **(…) travelSummary**:
    {**duration**: 294,
    length: 1636,
    baseDuration: 294},
- "Full routing": out of scope
- Matrix routing : tous les trajets possibles entre une liste de points
- Public transport routing: https://transit.hereapi.com/v8/routes?origin=50.83582,4.33868&destination=50.8249958,4.3454841&apiKey=X(...)zE

API : Here (isochrone)

- Isoline : ensemble des points accessibles à partir de/vers un point, en un temps (isochrone) ou une distance définie (voiture, camion, piéton)
- Résultat : polygône
- Exemple : zone accessible en 5 minutes à partir de l'hôpital Delta (Auderghem, Bruxelles)



Smals
ICT for society

Sources officielles

- Contours des codes postaux (BPost):
  - https://www.geo.be/catalog/details/9738c7c0-5255-11ea-8895-34e12d0f0423?l=fr
  - https://bgu.bpost.be/assets/9738c7c0-5255-11ea-8895-34e12d0f0423_x-shapefile_3812.zip
- Contours des secteurs statistiques (StatBel):
  - https://statbel.fgov.be/fr/open-data/secteurs-statistiques-2020
  - https://statbel.fgov.be/sites/default/files/files/opendata/Statistische%20sectoren/sh_statbel_statistical_sectors_31370_20200101.shp.zip
- Localisation de chaque adresse (BestAddress) : https://opendata.bosa.be/index.fr.html
- Quartiers Bruxellois : https://monitoringdesquartiers.brussels/
- Structures des codes postaux/localité/communes/province (BPost): https://www.bpost2.be/zipcodes/files/zipcodes_alpha_fr_new.xls

BestAddress: open data

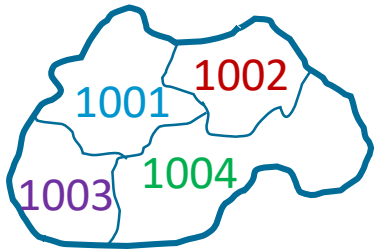
- BestAddress permet de télécharger toutes les addresses belges : https://opendata.bosa.be/
- Un fichier CSV par région :

	C	D	E	G	H	J	L	O	P	Q	R	U
1	EPSG:4326	EPSG:4326	address_i	house_nu	municipal	municipal	postcode	street_id	streetnam	streetname_fr	streetname_nl	
2	50.88995	4.31196	212450	6	21010	Jette	1090	2559		Place Jean-Louis Th	Jean-Louis Thysplaats	
3	50.82095	4.28556	200187	20	21001	Anderlech	1070	3959		Square Marie Curie	Marie Curiesquare	
4	50.79599	4.41243	25899	25	21017	Watermae	1170	4637		Avenue Delleur	Delleurlaan	
5	50.85773	4.30862	19920	31	21012	Molenbee	1080	3811		Rue Joseph Diongre	Joseph Diongrestraat	

- Difficulté : mapping “adresse en input” vs “adress BestAddress”
- À venir: REST API de géocodage (avril 2022?)

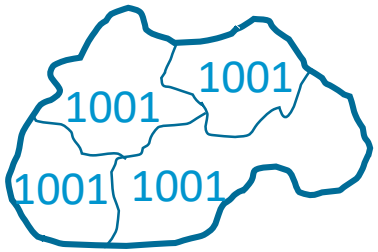
Belgium division schemes – Postal scheme

- Postal schemes (BPost): main commune – localities



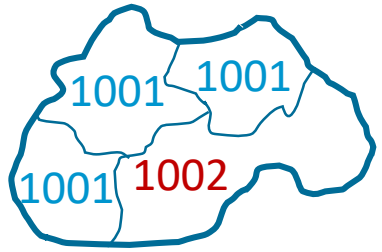
Main commune : Bruxelles(-Ville)

Zipcode	Locality
1000	Bruxelles
1020	Laeken
1120	Neder-Over-Hembeek
1130	Haren



Main commune : Jemeppe-Sur-Sambre

Zipcode	Locality
5190	Jemeppe-sur-Sambre
5190	Spy
5190	Onoz
5190	Balâtre
5190	…



Main commune : Assesse

Zipcode	Locality
5330	Assesse
5330	Maillen
5330	Sart-Bernard
5332	Crupet
5333	…

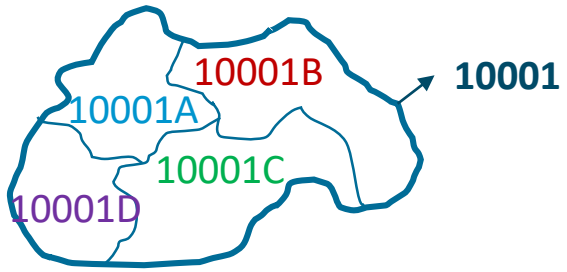
- No (numerical) ID for a commune (only « commune name ») → 581 values
- Locality ID : Zipcode + locality name
- Lots of « special codes » (1110 NATO, 1044 RTBF, …)

https://www.bpost.be/fr/codes-postaux

Smals ICT for society

Belgium division schemes – statistical scheme

- Statistical scheme (Statbel):
  - Main commune (NIS5) → 581 values
  - Sections (NIS6)

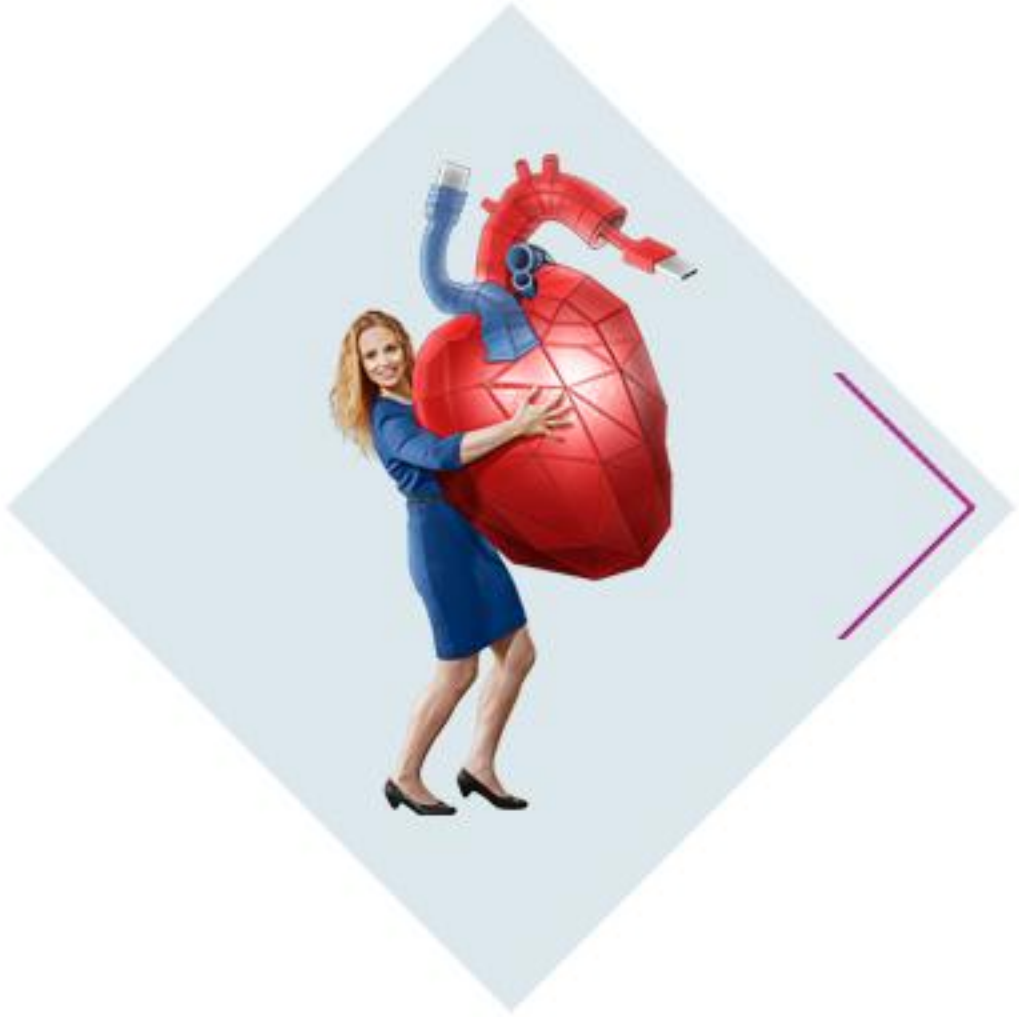


- No mapping Zipcode-NIS5 (or NIS6)
- No special code

Main commune : Bruxelles(-Ville)

NIS5		NIS6
21004	BRUXELLES PENTAGONE	21004A
21004	BRUXELLES-RUE DE LA LOI	21004B
21004	BRUXELLES-LOUISE	21004C
21004	BRUXELLES-CHAUSSEE D'ANVERS	21004D
21004	BRUXELLES-LAEKEN	21004E
21004	BRUXELLES-NEDER-OVERHEEMBEEK	21004F
21004	BRUXELLES-HAREN	21004G

https://statbel.fgov.be/fr/propos-de-statbel/methodologie/classifications/geographie



Visualisation

Visualisation

- Question : y a-t-il une composante/dépendance géographique dans un indicateur ?
- Premier réflex avant d'analyser des données : les visualiser
- Exemple : revenu moyen déclaré par commune :

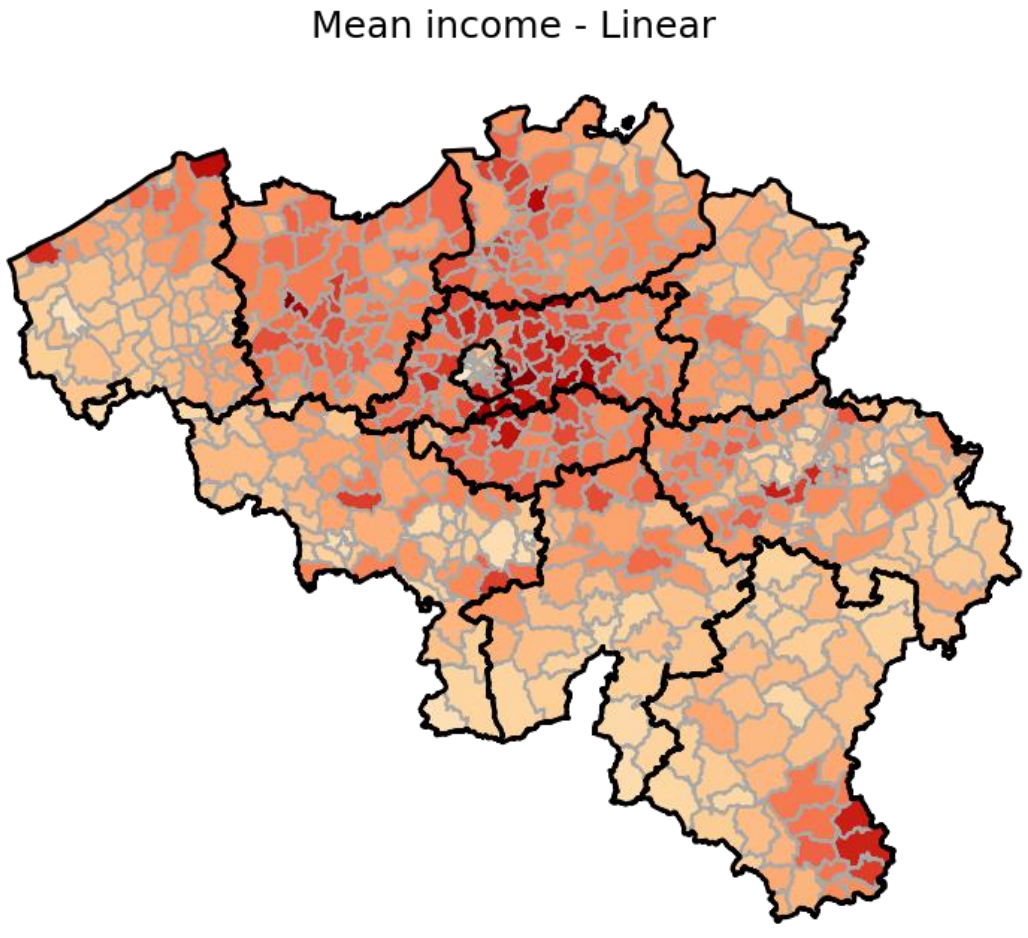
INS	MEAN_INCOME
11001	21 966.0
11002	18 417.0
11004	20 091.0
…	…

INS	GEOM
11001	polygon(pt_1, pt_2,…)
11002	polygon(…)
11004	polygon(…)
…	…

INS	GEOM	MEAN_INCOME
11001	polygon(pt_1, pt_2,…)	21 966.0
11002	polygon(…)	18 417.0
11004	polygon(…)	20 091.0
…	…	…

Smals ICT for society

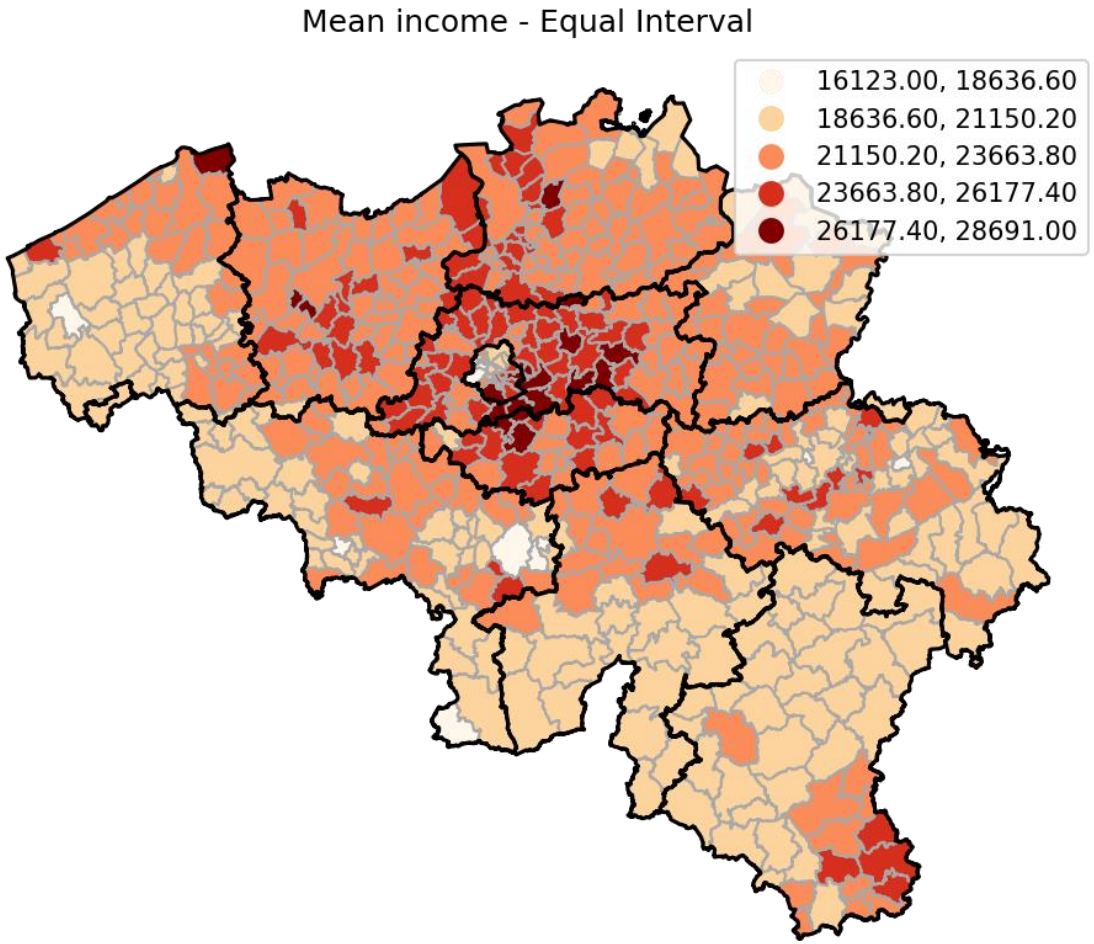
Colorisation linéaire



- Colorisation linéaire entre min et max
- Si quelques outiers très éloignés du reste, peu de distinction
- Petites communes (ex: Bxl) difficile à voir

Smals
ICT for society

Equal intervals

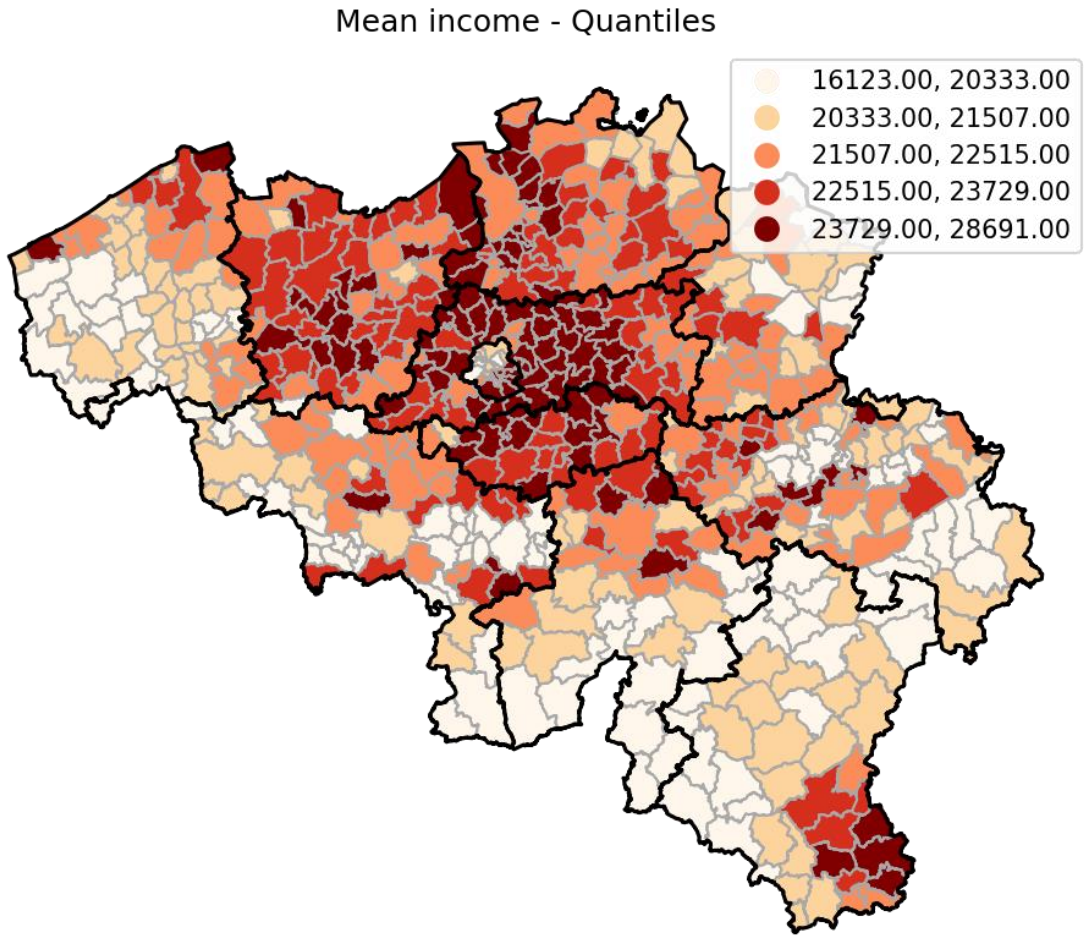


- On découpe l'intervalle [min, max] en K (ici: 5) classes de taille égale
- Rend les couleurs plus distinctes, mais on perd en précision
- Pas de garantie que toutes les classes soient utilisées

From	To	Width	Count
16123,0	18636,6	2513,6	**11**
18636,6	21150,2	2513,6	**190**
21150,2	23663,8	2513,6	**256**
23663,8	26177,4	2513,6	**103**
26177,4	28691,0	2513,6	**21**

Smals ICT for society

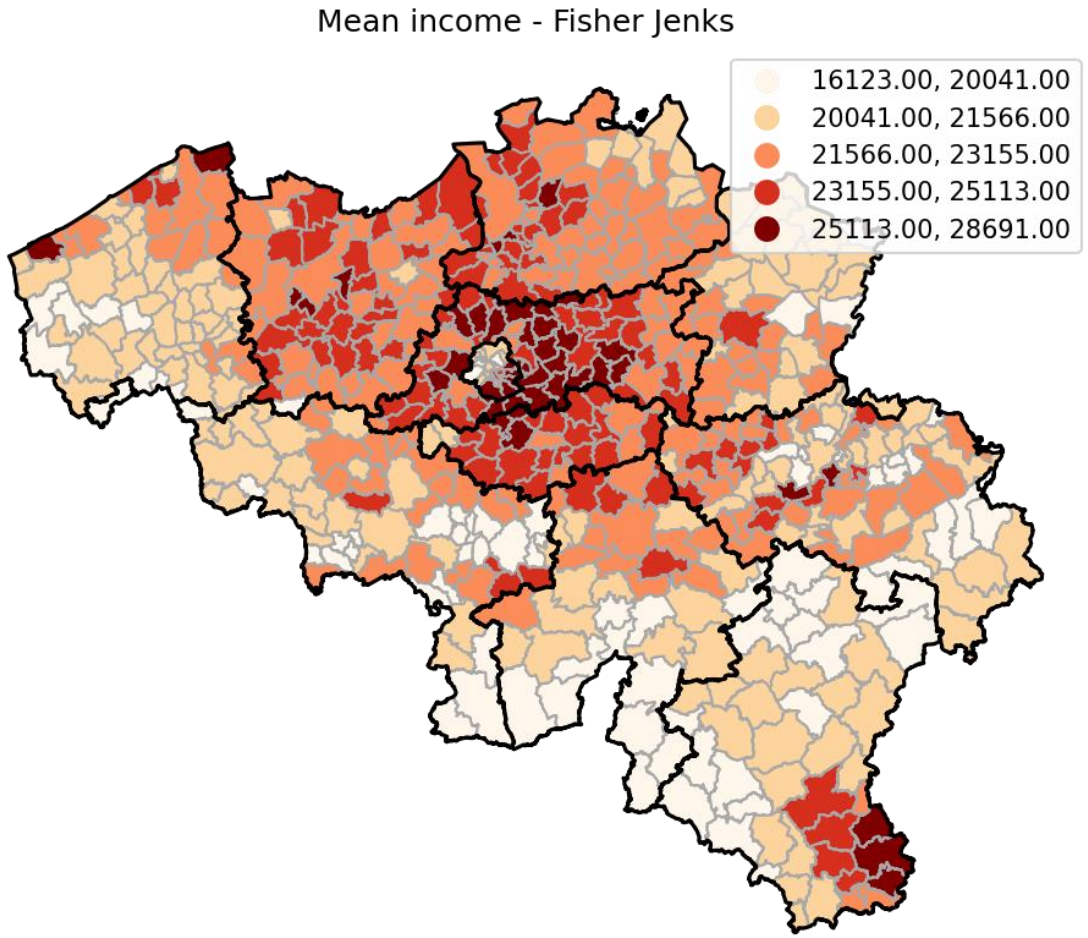
Quantiles



- On découpe les données en K (ici: 5) classes de taille égale (si possible)

From	To	Width	Count
16123	20333	**4210**	117
20333	21507	**1174**	116
21507	22515	**1008**	116
22515	23729	**1214**	116
23729	28691	**4962**	116

Fisher-Jenks

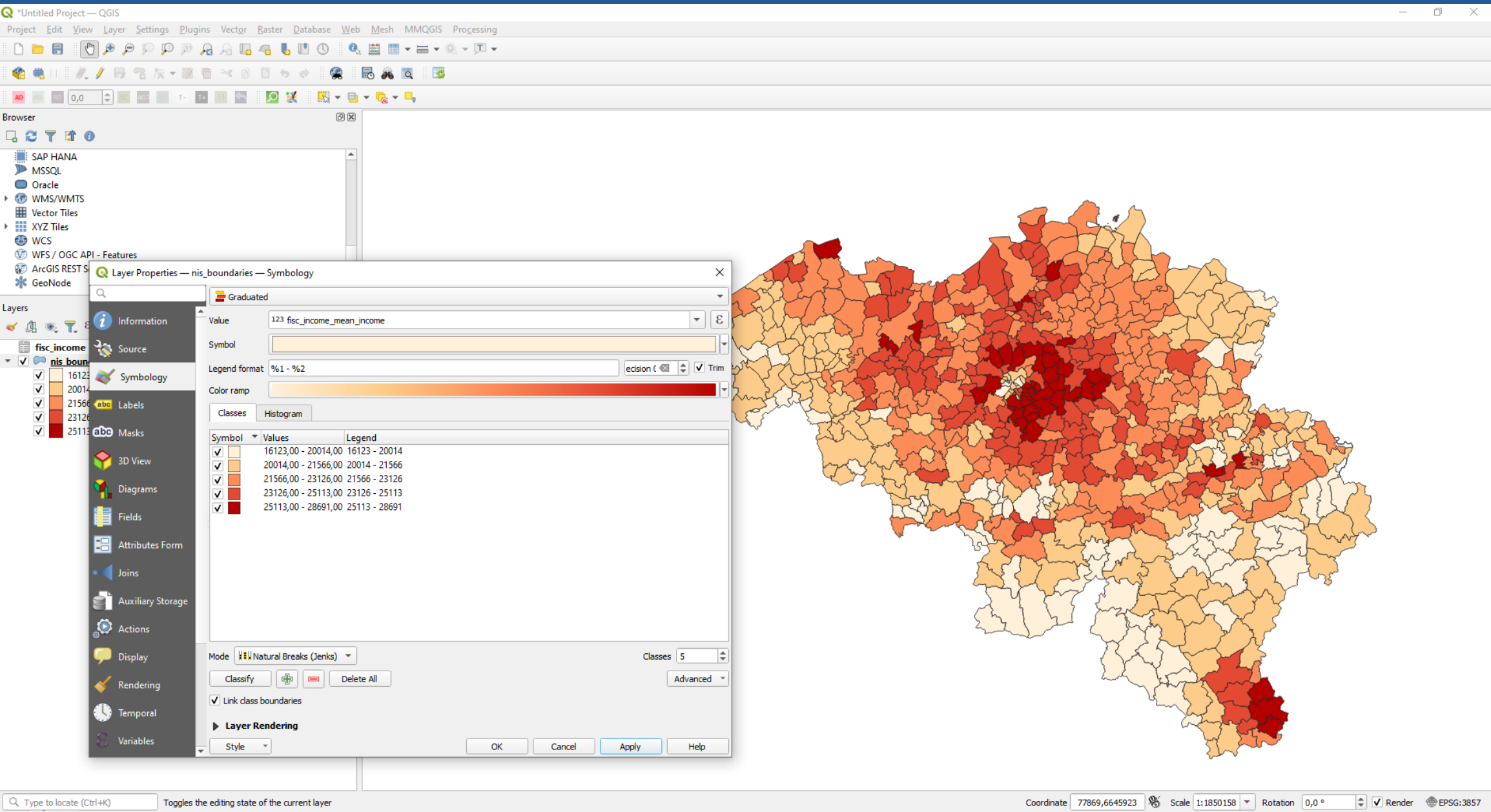


- On découpe les données en K (ici: 5) classes les plus « uniformes possibles »
- Minimise la variance au sein de chaque classe, la maximise entre les classes
- Version à 1 dimension de « K-means »

From	To	Width	Count
16123	20041	3918	82
20041	21566	1525	163
21566	23155	1589	168
23155	25113	1958	122
25113	28691	3578	46

Smals ICT for society

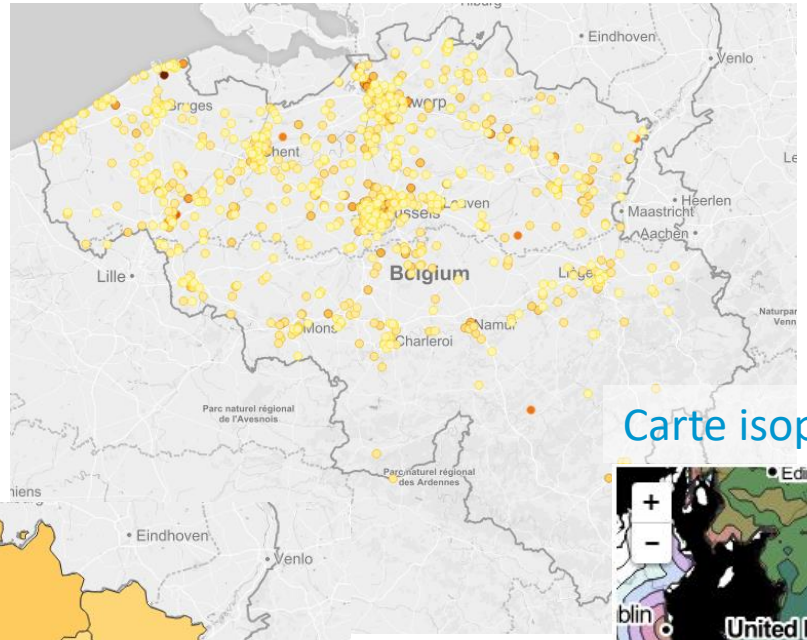
QGIS



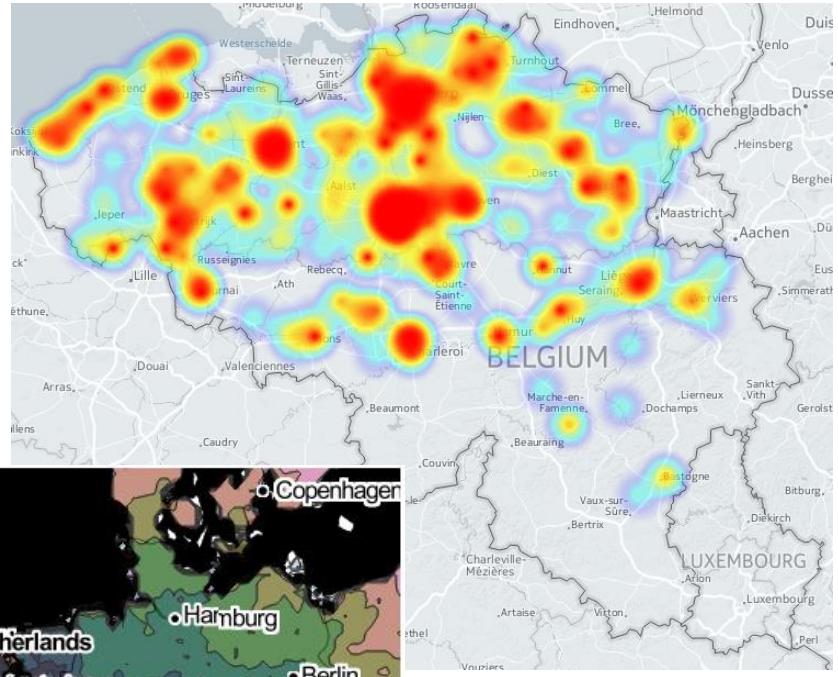
Smals
ICT for society

Autres visualisations

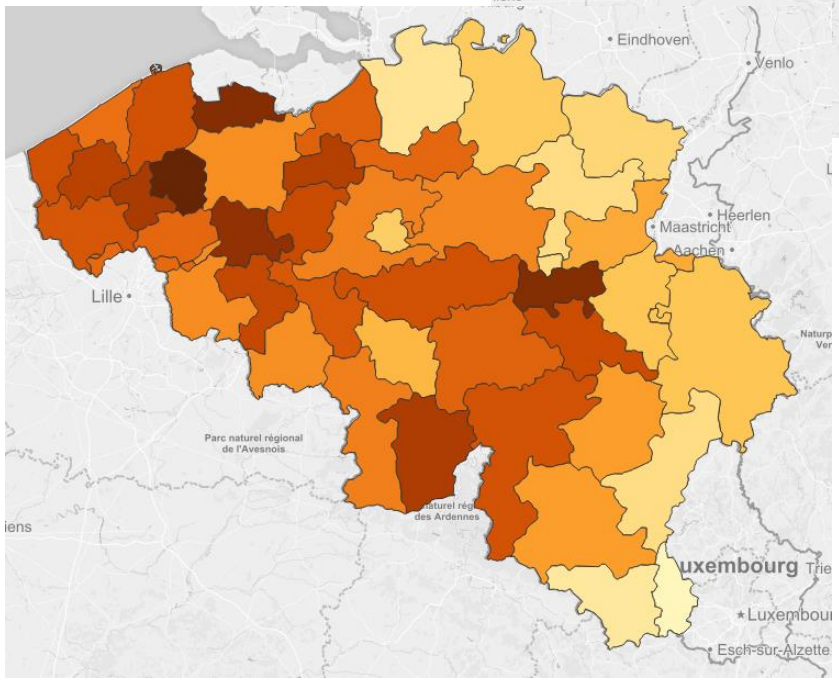
Carte par point



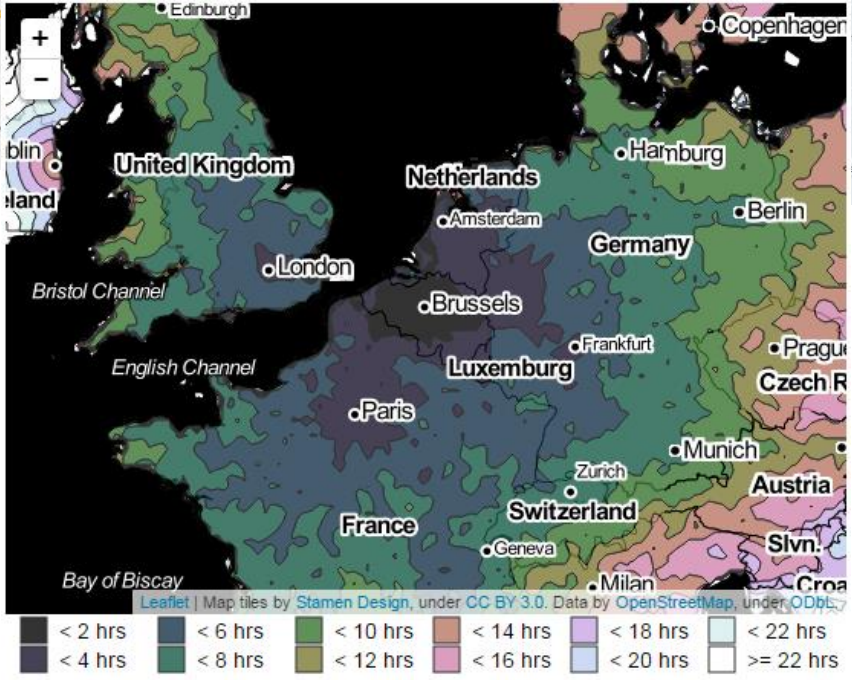
Heatmap (carte de chaleur)



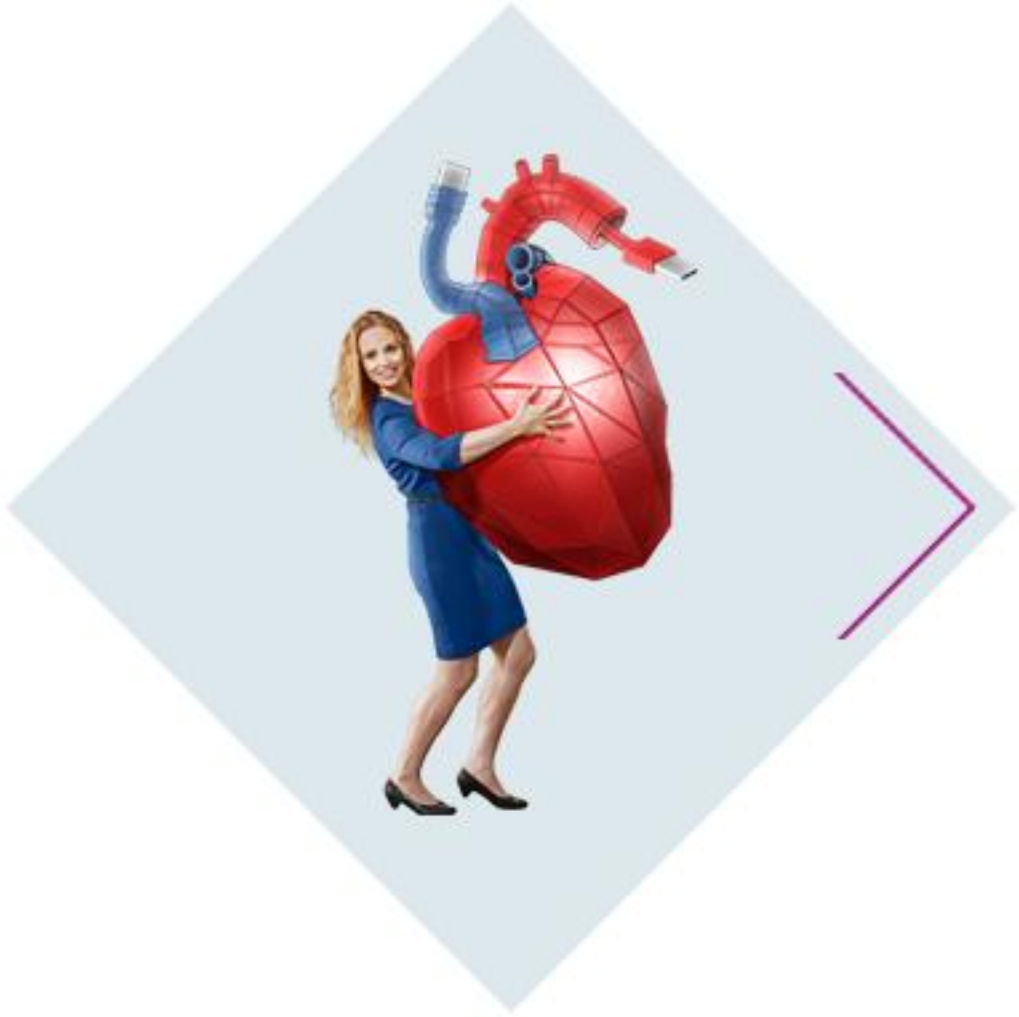
Carte choroplèthe



Carte isoplèthe



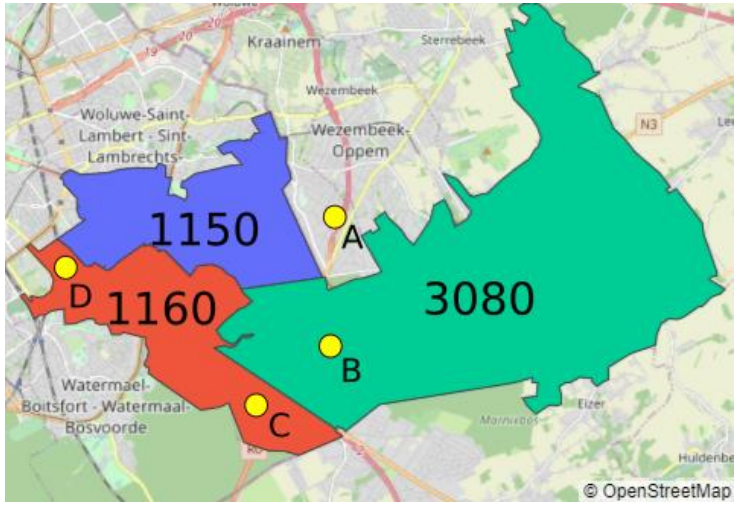
Smals
ICT for society



Jointure spatiale et bases de données géographiques

Jointure spatiale

- Jointure “classique”



observations

ID	ZIPCODE
A	
B	3080
C	1160
D	1160
…	…

zipcodes

ZIPCODE	PROVINCE
1150	Bruxelles-Capitale
1160	Bruxelles-Capitale
3080	Brabant Flamand
…	…

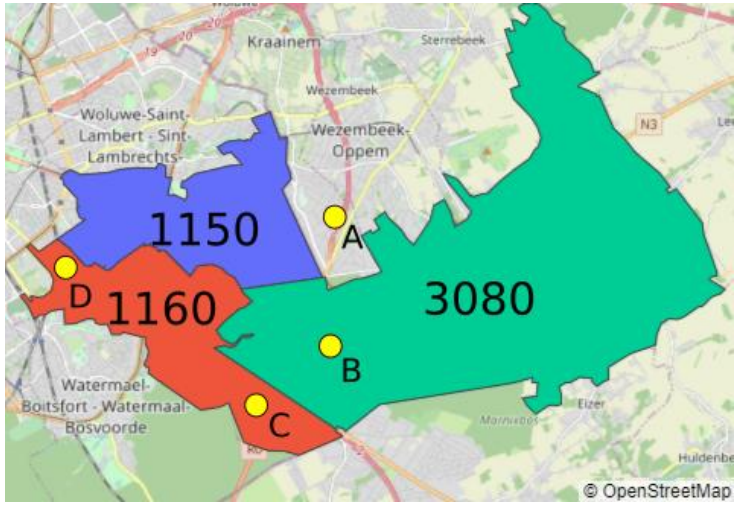
```
SELECT o.ID, o.zipcode, z.province
FROM observations o
LEFT JOIN zipcodes z
ON o.zipcode = z.zipcode
```

ID	ZIPCODE	PROVINCE
A		
B	3080	Brabant Flamand
C	1160	Bruxelles-Capitale
D	1160	Bruxelles-Capitale
…		

Smals ICT for society

Jointure spatiale

- Jointure “spatiale”



observations

ID	GEOM
A	point(x_a, y_a)
B	point(x_b, y_b)
C	point(x_c, y_c)
D	point(x_d, y_d)
…	…

provinces

PROVINCE	GEOM
Bruxelles-Capitale	polygon(pt_1, pt_2,…)
Brabant Flamand	polygon(…)
Braband Wallon	polygon(…)
…	…

```
SELECT o.ID, p.province
FROM observations o
LEFT JOIN provinces p
ON ST_Contains (p.geom, o.geom)
```

ID	PROVINCE
A	Brabant Flamand
B	Brabant Flamand
C	Bruxelles-Capitale
D	Bruxelles-Capitale
…	

Query “PostGIS” (extension de PostGreSQL)

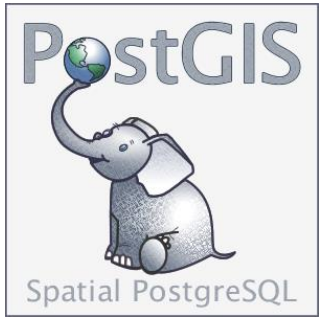
Smals ICT for society

PostGIS

- PostGIS (https://postgis.net/) is an extension of PostGreSQL

- Add new (spatial) data types :
  - POINT (+MULTIPOINT)
  - LINESTRING (+MULTILINESTRING)
  - POLYGON (+MULTIPOLYGON)
  - GEOMETRYCOLLECTION

- Add set of new (spatial) operations "ST_" (spatial type):
  - Comparison: ST_Contains, ST_Within, ST_Intersects
  - Aggregation: ST_Union
  - Combinations: ST_Intersection, ST_Difference
  - Measurement: ST_Distance, ST_Area
  - Projection handling: ST_Transform



- POINT = x, y (, z) + CRS
  - EPSG:4326 / WGS 84 (polar)
    - 4.3383, 50.8357 (longitude, latitude)
  - EPSG:3857 / WGS 84 pseudo-Mercator (metric)
    - 482 944.1, 6 592 292.4 (easting, northing)
  - EPSG:31370 / Belgian Lambert 72:
    - 147 859.2, 169 482.7 (easting, northing)

- (Minimal) Installation:
  - Install PostGreSQL
  - Install PostGIS binary (or build sources)
  - `CREATE EXTENSION postgis;`

Smals ICT for society

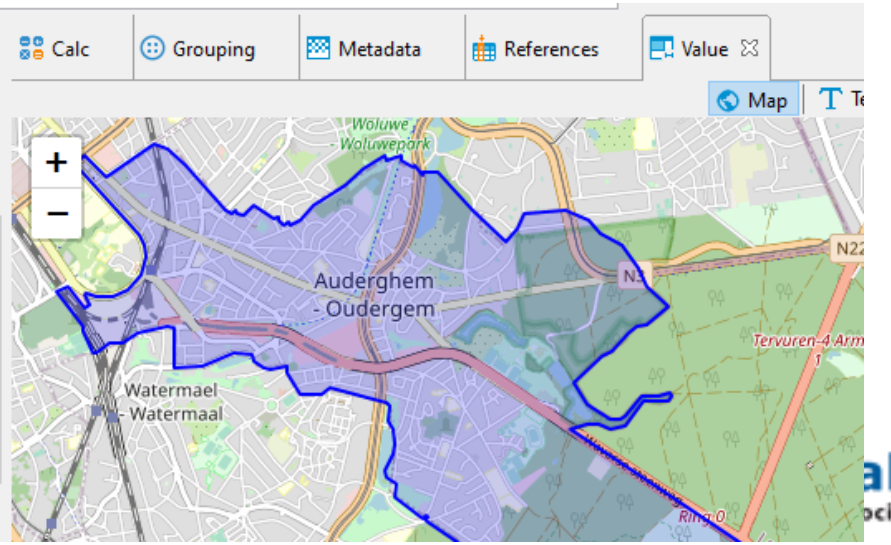
Data sources

- We will load 4 OpenData sources:

Subject	Source	Shapes	Format	URL
Zipcode list (mapping zipcode-main commune)	BPost	Non geo	Excel	https://www.bpost2.be/(...)/zipcodes_alpha_fr_new.xls
Zipcode boundaries	BPost	Polygons	Shapefile	https://bgu.bpost.be/(...)_x-shapefile_3812.zip
Zipcode centers	Agence du Numérique	Points	GeoJSON	https://www.odwb.be/(...)/code-postaux-belge/exports/geojson
NIS code boundaries (statistical code level)	Statbel	Polygons	Shapefile	https://statbel.fgov.be/(...)/sh_statbel_statistical_sectors_31370_20200101.shp.zip

- In DBeaver:

	zipcode	is_special	geometry
8	1083	[ ]	POLYGON Z((4.3141200710327965 50.879913314876895 0, 4.314364646248509 50.879910343655084 0, 4.3147113301714 41 50.879
9	1090	[ ]	POLYGON Z((4.327598207423981 50.893364785417354 0, 4.32626134037569 50.891408698066954 0, 4.32632830591178 50.8913892
0	1099	[v]	POLYGON Z((4.413543363024862 50.911582341339376 0, 4.413093431163375 50.91108304503646 0, 4.413061327321158 50.911042
1	1105	[v]	POLYGON Z((4.3531983480800145 50.85014122453475 0, 4.353247117590716 50.84997084005416 0, 4.353175537163507 50.849816
2	1110	[v]	POLYGON Z((4.42819723854411 50.882026337836 0, 4.429700045875309 50.8782804223708 0, 4.429484134859504 50.8777816956 0
3	1120	[ ]	MULTIPOLYGON Z(((4.406288376086285 50.913103136943214 0, 4.405861863346267 50.91250310244647 0, 4.405579633475274 50
4	1130	[ ]	POLYGON Z((4.421194590355868 50.90319073244135 0, 4.421300360278826 50.90318924713572 0, 4.42175262727196 50.90322401
5	1140	[ ]	POLYGON Z((4.402965863242295 50.885036114397074 0, 4.403236808800037 50.88495647629276 0, 4.403266394272282 50.884996
6	1150	[ ]	POLYGON Z((4.46453289128135 50.84543458084347 0, 4.464620997901788 50.845280038669415 0, 4.464660977110739 50.8452099
7	1160	[ ]	POLYGON Z((4.407530006463037 50.825526663960815 0, 4.408787206768317 50.82468003618179 0, 4.409639560497666 50.824134
8	1170	[ ]	POLYGON Z((4.406281879386394 50.81487173035492 0, 4.4063036984364095 50.81465642072515 0, 4.406273655148961 50.814529
9	1180	[ ]	POLYGON Z((4.367312449506682 50.816632364494566 0, 4.367301468664922 50.81657067518872 0, 4.367292880103075 50.816522
0	1190	[ ]	POLYGON Z((4.332631838354738 50.82868882934866 0, 4.33278910519502 50.82863412946709 0, 4.332855508690386 50.828611 03
1	1200	[ ]	POLYGON Z((4.436304387773484 50.8588848028416 0, 4.437371685036211 50.85850521008912 0, 4.437392187931652 50.858528 03
2	1210	[ ]	POLYGON Z((4.367165693806663 50.85791404502723 0, 4.367886918603596 50.8577594939246 0, 4.368046460119377 50.857725 30



Simple distance

```
SELECT
        a.locality,
        b.locality,
        ST_Distance(a.geometry, b.geometry),
        ST_DistanceSphere(a.geometry, b.geometry)
FROM
        gistest.zipcodes_centers a,
        gistest.zipcodes_centers b
WHERE a.zipcode='1160' AND b.zipcode='5190'
```

Geometric Euclidian distance (ok only for metric CRS)

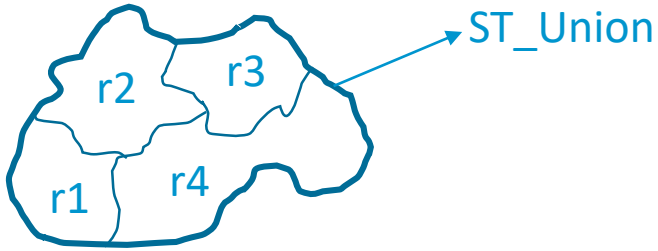
Geographical distance (in meters)

$$\sqrt{(4.433 - 4.995)^2 + (50.815 - 50.828)^2}$$

	locality	locality	st_distance	st_distancesphere
1	Auderghem	Moustier-Sur-Sambre	0.4378300821	43,089.00047741
2	Auderghem	Saint-Martin	0.3807369552	38,121.84729085
3	Auderghem	Balâtre	0.3778559061	38,134.34560066
4	Auderghem	Onoz	0.4006119939	39,630.80735948
5	Auderghem	Jemeppe-Sur-Sambre	0.4164304942	42,209.58949063
6	Auderghem	Ham-Sur-Sambre	0.4421927051	44,618.0699368
7	Auderghem	Mornimont	0.4508334992	44,379.13904389
8	Auderghem	Spy	0.4295900467	41,768.06793197

Aggregation

```
SELECT
        c_provi, t_provi_fr, t_regio_fr,
        ST_Union(geometry) as geometry
FROM gistest.statistical_sectors
GROUP BY c_provi, t_provi_fr, t_regio_fr
```



statistical_sectors 1

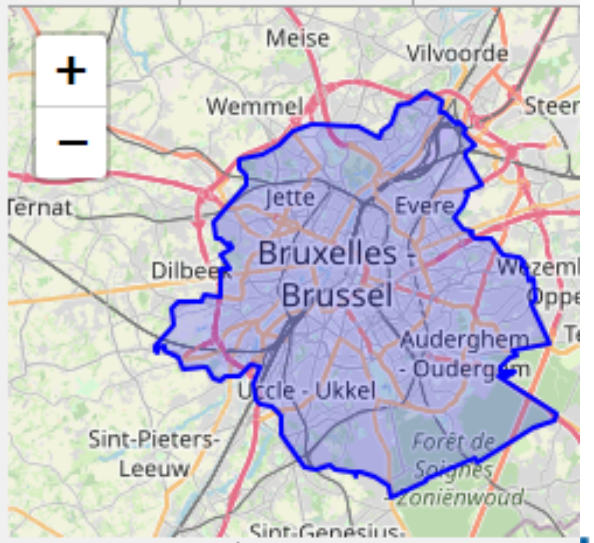
SELECT c_provi, t_provi_fr, t_regio_fr, ST_Union(geometry) as geome

Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text | Spatial | Record

	c_provi	t_provi_fr	t_regio_fr	geometry
1	40000	Province de Flandre orientale	Région flamande	POLYGON Z((3.8913699566800735 50.743034458685
2	80000	Province du Luxembourg	Région wallonne	POLYGON Z((5.619343985306024 49.5246340405451
3	10000	Province d'Anvers	Région flamande	MULTIPOLYGON Z(((4.356497945561555 51.0167519
4	70000	Province du Limbourg	Région flamande	MULTIPOLYGON Z(((5.152636643029442 50.6965789
5	60000	Province de Liège	Région wallonne	POLYGON Z((6.141075404446796 50.1658812945573
6	50000	Province du Hainaut	Région wallonne	MULTIPOLYGON Z(((4.25039772880804 49.96223026
7	20002	Province du Brabant wallon	Région wallonne	POLYGON Z((4.554173509041858 50.5361666493454
8	30000	Province de Flandre occidentale	Région flamande	POLYGON Z((3.191547205152201 50.7559780002177
9	90000	Province de Namur	Région wallonne	POLYGON Z((4.870313712271117 49.7924532777839
10	20001	Province du Brabant flamand	Région flamande	POLYGON Z((4.261234445222165 50.7000576337673
11	-	[NULL]	Région de Bruxelles-Capitale	POLYGON Z((4.4040908520136535 50.769799396396

Calc | Grouping | Metadata



ICT for society

Aggregation – commune boundaries

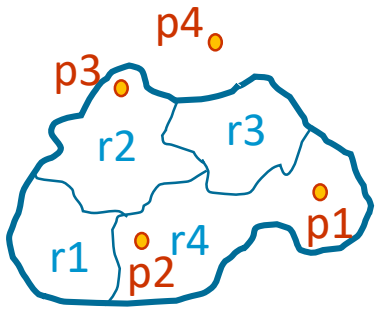
```
-- Create a table with NIS boundaries
CREATE TABLE gistest.nis_boundaries AS -- or: CREATE VIEW
    SELECT cnis5_2020, -- Code NIS5
           t_mun_fr,   -- Commune name
           ST_Union(geometry) AS geometry
    FROM gistest.statistical_sectors
    GROUP BY cnis5_2020, t_mun_fr


-- Create a table with postal boundaries
CREATE TABLE gistest.zip_commune_boundaries AS  -- or: CREATE VIEW
    SELECT main_commune,
           ST_Union(geometry) AS geometry
    FROM gistest.zipcodes zp
      JOIN gistest.zipcodes_boundaries zp_bnd
      ON zp.zipcode = zp_bnd.zipcode
    WHERE NOT zp.is_special
    GROUP BY main_commune
```

Spatial Join

table_polygons

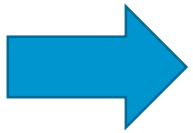
ID	geom
r1	POLYGON(…)
r2	POLYGON(…)
r3	POLYGON(…)
r4	POLYGON(…)



table_points

ID	geom
p1	POINT(…)
p2	POINT(…)
p3	POINT(…)
p4	POINT(…)

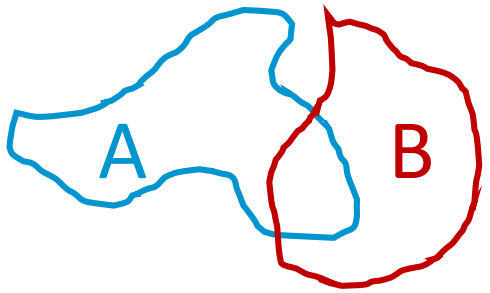
```
SELECT plg.ID, pnt.ID
FROM table_polygons plg
FULL JOIN table_points pnt
  ON ST_Contains(plg.geom, pnt.geom)
```



plg.ID	pnt.ID
r1	NULL
r2	p3
r3	NULL
r4	p1
r4	p2
NULL	p4

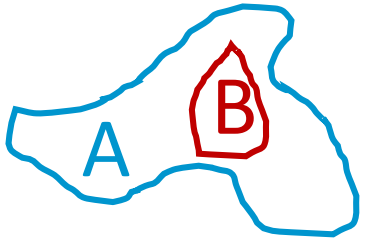
Smals ICT for society

Spatial operations



ST_Intersects(A, B)

ST_Touches(A, B)

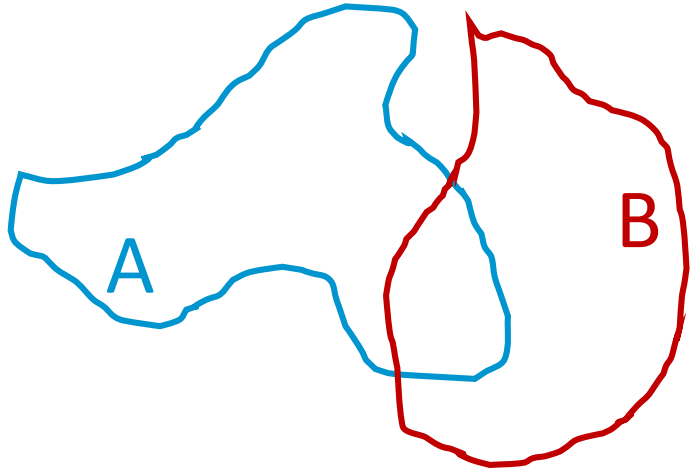


ST_Contains(A, B)
= ST_Within(B, A)

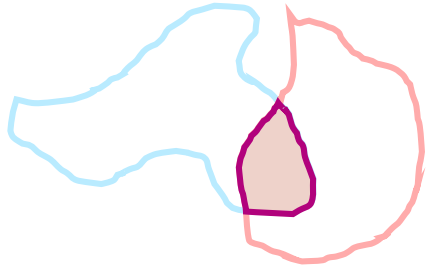
ST_Equals
ST_Overlaps
ST_Covers
ST_Disjoint
…

Smals
ICT for society

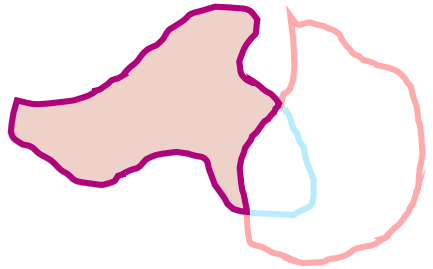
Spatial operations



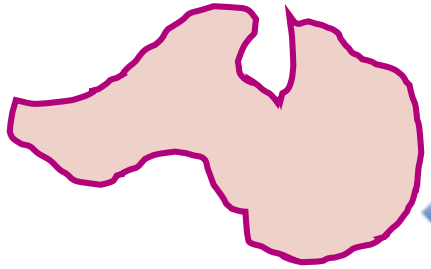
ST_Intersection(A, B):



ST_Difference (A, B):



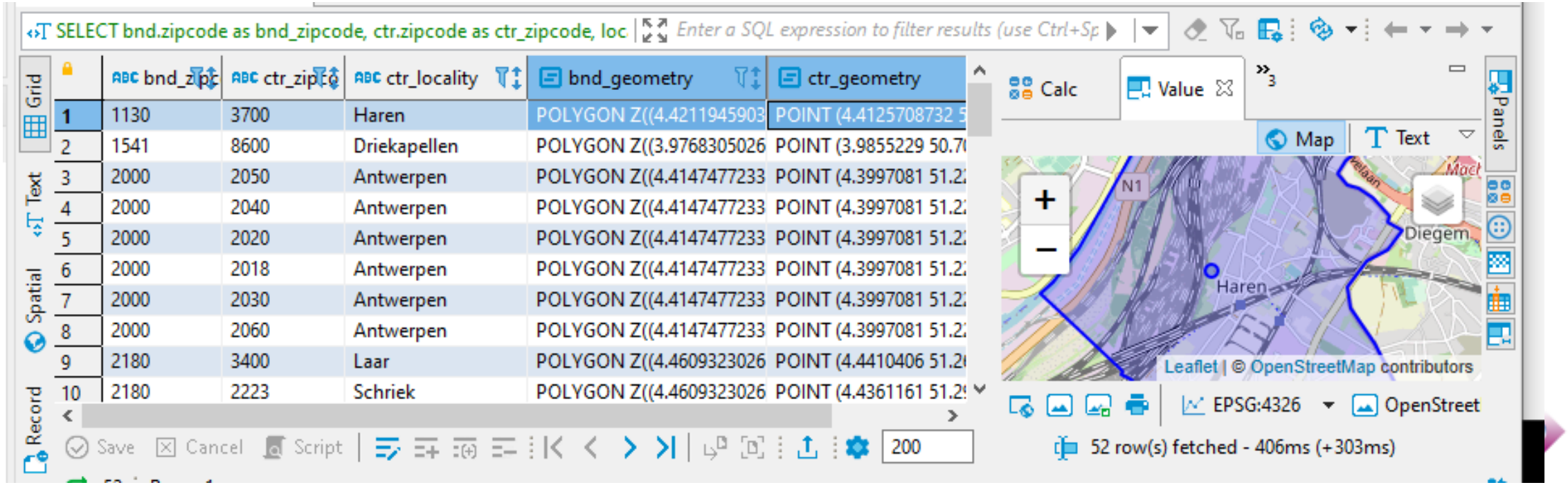
ST_Union (A, B):



Smals
ICT for society

Spatial join: Mismatch zip centers – zip boundaries

```
SELECT
       bnd.zipcode  as bnd_zipcode, ctr.zipcode  as ctr_zipcode,
       locality     as ctr_locality,
       bnd.geometry as bnd_geometry, ctr.geometry as ctr_geometry
FROM gistest.zipcodes_boundaries bnd
       JOIN gistest.zipcodes_centers ctr
       ON ST_Contains(bnd.geometry, ctr.geometry)
WHERE bnd.zipcode != ctr.zipcode
```



Smals
ICT for society

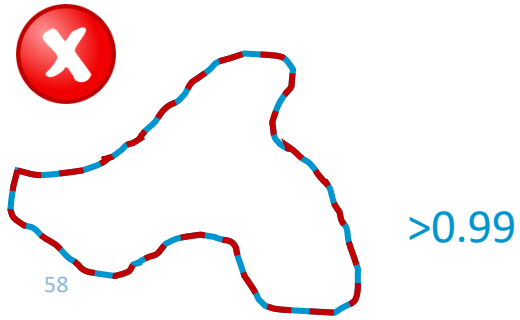
Spatial join – Internal mismatch @ BPost

```
SELECT zp1.zipcode, zp2.zipcode,
       zp1.geometry, zp2.geometry,
       ST_intersection(zp1.geometry, zp2.geometry),
       st_area(ST_intersection(zp1.geometry, zp2.geometry))/ST_Area(zp1.geometry)
              AS common_ratio
FROM gistest.zipcodes_boundaries zp1
JOIN gistest.zipcodes_boundaries zp2
ON ST_intersects(zp1.geometry, zp2.geometry)
       AND zp1.zipcode < zp2.zipcode
       AND ST_Area(ST_Intersection(zp1.geometry, zp2.geometry))/ST_Area(zp1.geometry) >0.01
ORDER BY common_ratio
```

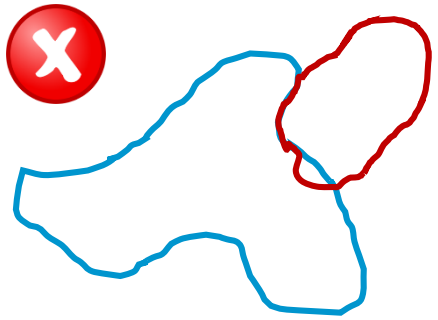

Smals
ICT for society

Spatial join: Mismatch NIS – Zip boundaries

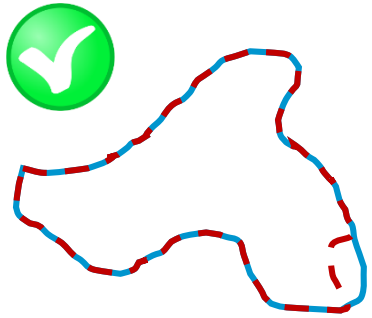
```
SELECT cnis5_2020, t_mun_fr, main_commune,
    nis_bnd.geometry as nis_geom,
    zip_bnd.geometry as zip_geom,
    ST_Difference(nis_bnd.geometry, zip_bnd.geometry) as only_nis_geom,
    ST_Difference(zip_bnd.geometry, nis_bnd.geometry) as only_zip_geom,
    ST_Intersection(nis_bnd.geometry, zip_bnd.geometry) as common_geom,
    ST_Area(ST_Intersection(nis_bnd.geometry, zip_bnd.geometry)) /
      ST_Area(nis_bnd.geometry) as overlap_ratio
FROM gistest.nis_boundaries nis_bnd
JOIN gistest.zip_commune_boundaries_filled zip_bnd
ON   ST_Intersects(nis_bnd.geometry, zip_bnd.geometry)
WHERE ST_Area(ST_Intersection(nis_bnd.geometry, zip_bnd.geometry)) /
       ST_Area(nis_bnd.geometry) between 0.5 and 0.99
```



>0.99



<0.5



Smals
ICT for society

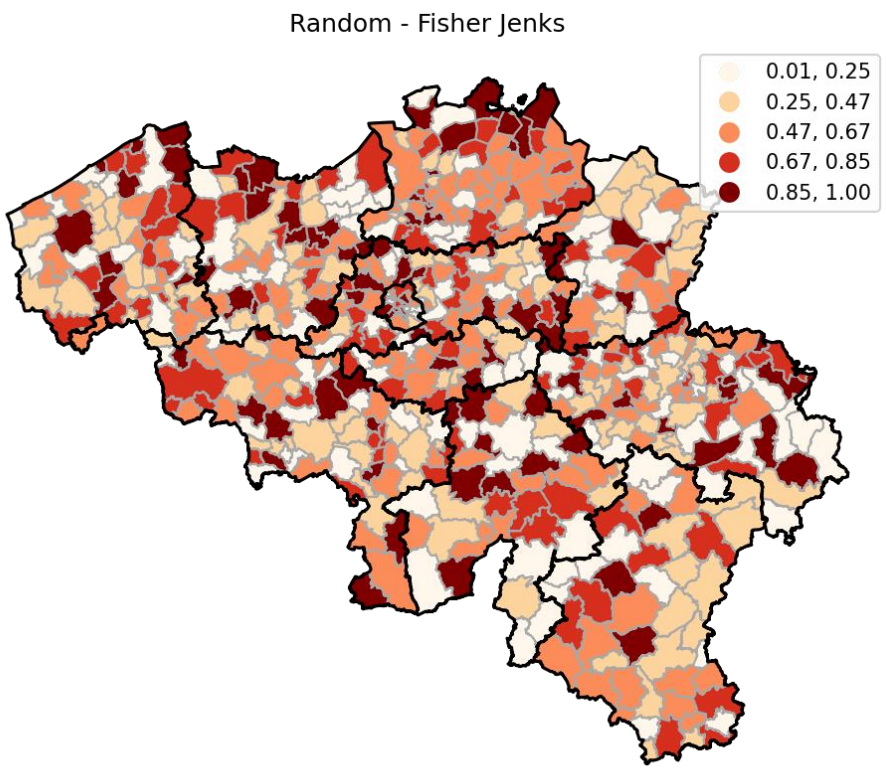
Applications

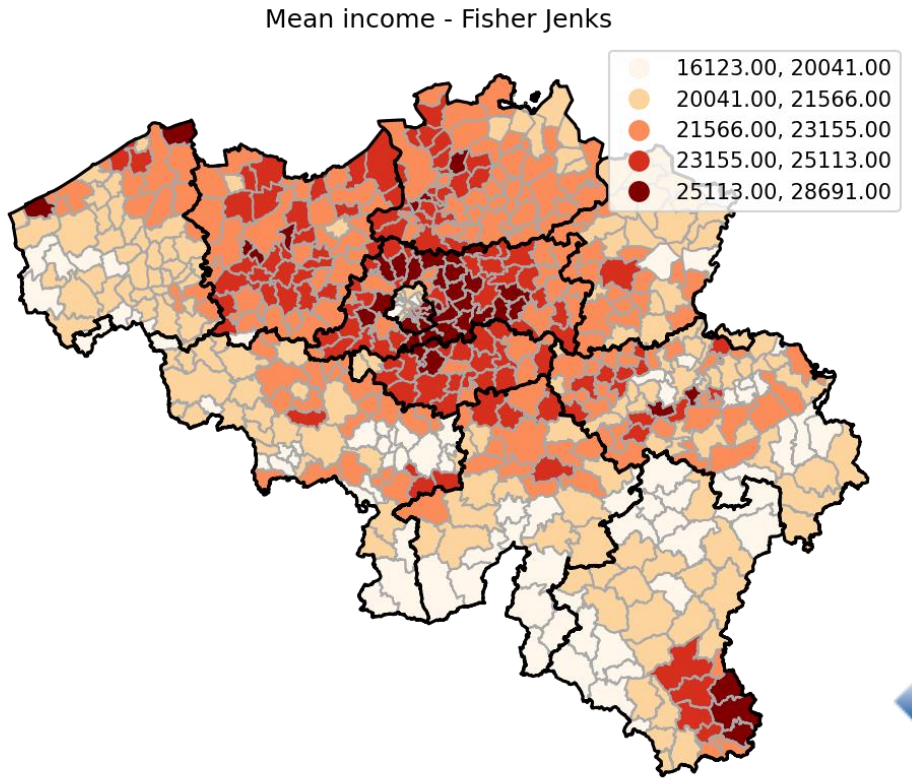
- Détection d'anomalies:
  - Centre de code postal pas situé dans le bon polygone (52 cas pour l'Agence du Numérique)
  - Superposition de polygones :
    - Intrasource: territoires attribués à deux codes postaux (Bpost: ~15 CP spéciaux pas correctement "extraits")
    - Intersource: territoires partagés entre une commune "postale" et une commune "statistique" (~25 incohérences)
- Attribuer un code postal, une commune, province, arrondissement… à chaque point d'une liste
- Combinaison avec une courbe isochrone: communes accessibles en X minutes à partir d'un point (cf plus loin)

Autocorrélation spatiale

Autocorrélation spatiale

- La valeur d'une observation d'une variable A est-elle liée aux observations voisines (géographiquement) ? Un indicateur a-t-il une composante géographique ?
- A-t-on plus de chance d'observer un salaire élevé près d'une commune avec un salaire élevé ?
- Le salaire moyen a-t-il un plus grande dépendance géographique que le niveau d'éducation ?
- Visuellement assez clair, mais comment le formaliser/quantifier ?

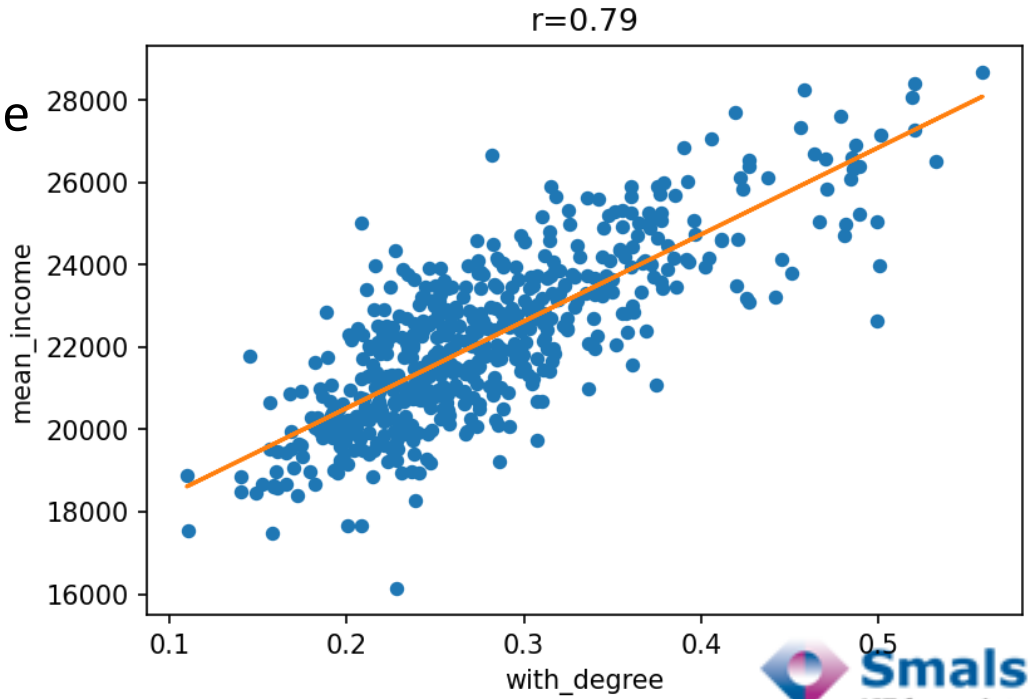




Smals
ICT for society

Corrélation

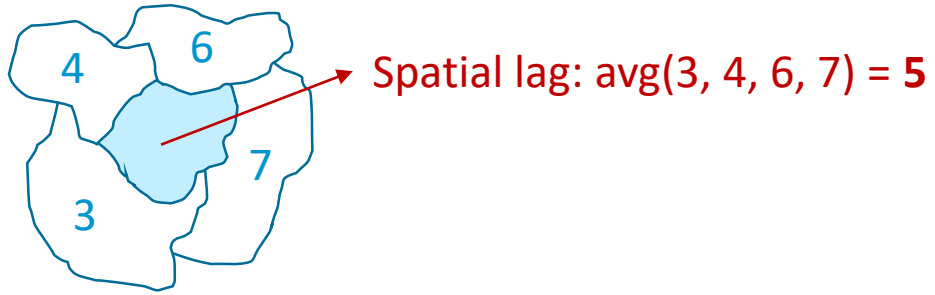
- Corrélation (classique) : caractérise le niveau de liaison entre deux « variables » (ou plus)
- Exemple: revenu moyen vs proportion avec un diplôme de l'ens. sup. (« degree »), par commune
- Corrélation n'implique pas causalité !
  - Bon revenu → accès à un enseignement de qualité ?
  - Bon diplôme → accès à des jobs bien rémunérés ?
  - Niveau/classe social(e) → accès à un enseignement de qualité ET à des jobs bien rémunérés ?
  - Dépendance cyclique ?
- Corrélation entre A et B si connaitre une valeur de A donne une indication de la valeur de B
- Causalité de A vers B si modifier A aura un impact sur B
- Coefficient de corrélation (Pearson's r) : 1 si corrélation parfaite, 0 si aucune corrélation, -1 si corrélation inverse

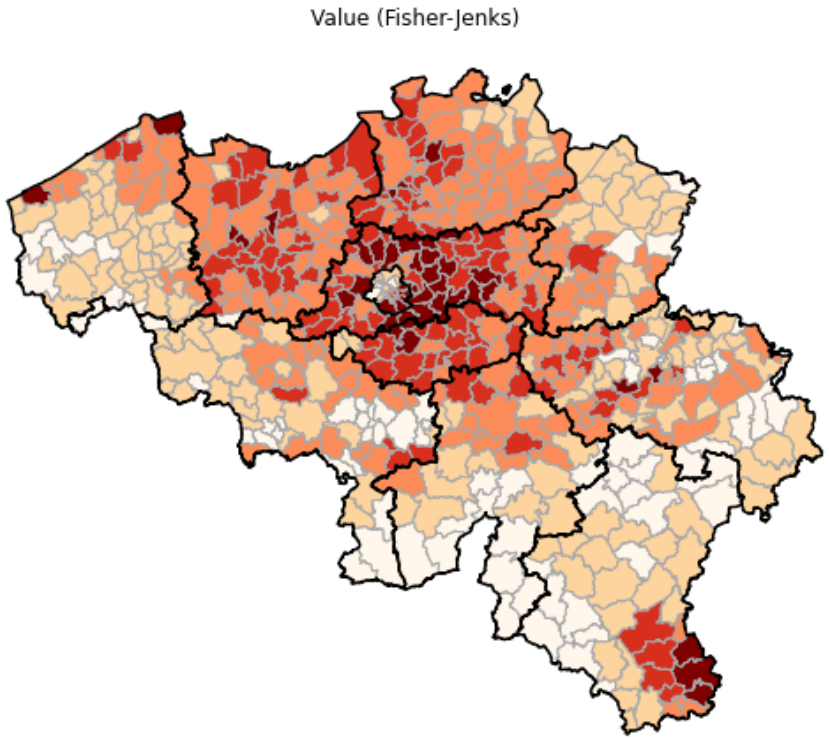


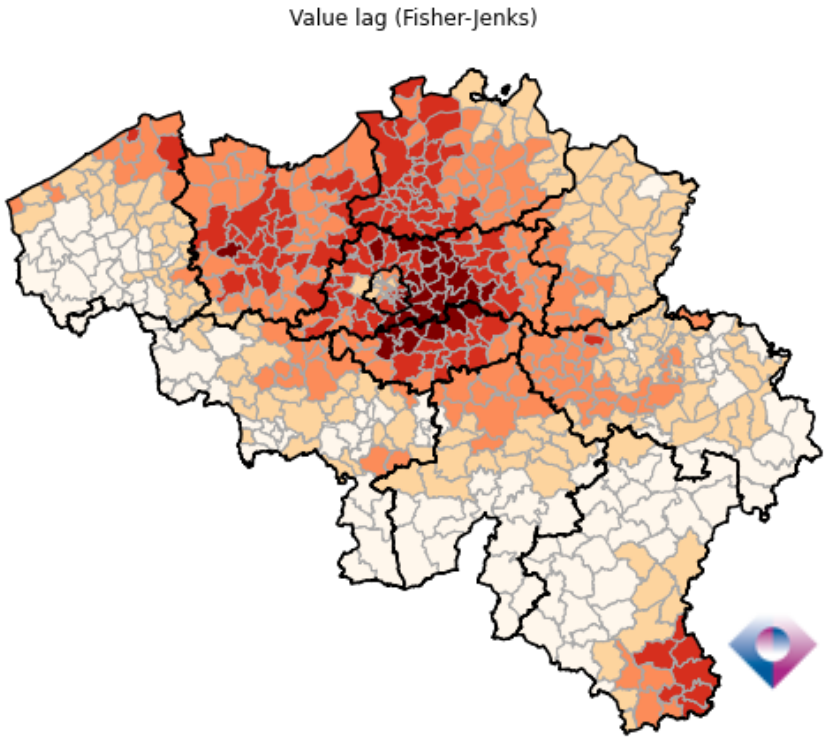
Sources: https://www.census2011.be/, https://statbel.fgov.be/

Autocorrélation locale : spatial lag

- Spatial lag (« décalage spatial ») : valeur moyenne des zones voisines de chaque zone
- Sans tenir compte de « soi-même »
- Permet de « lisser les valeurs »
- Nécessite un calcul de voisinage : quelle sont les voisins de chaque zone ? (Queen neighborhood)



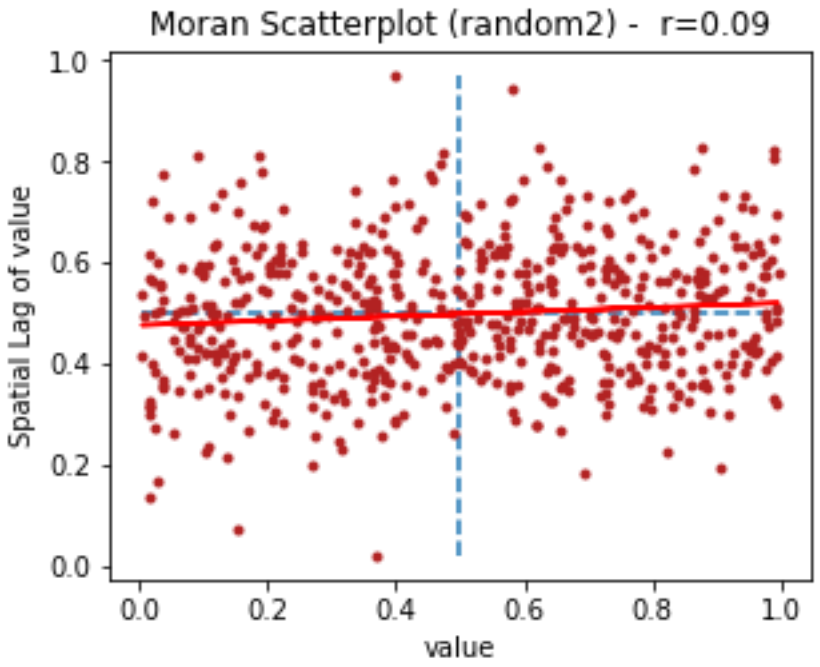


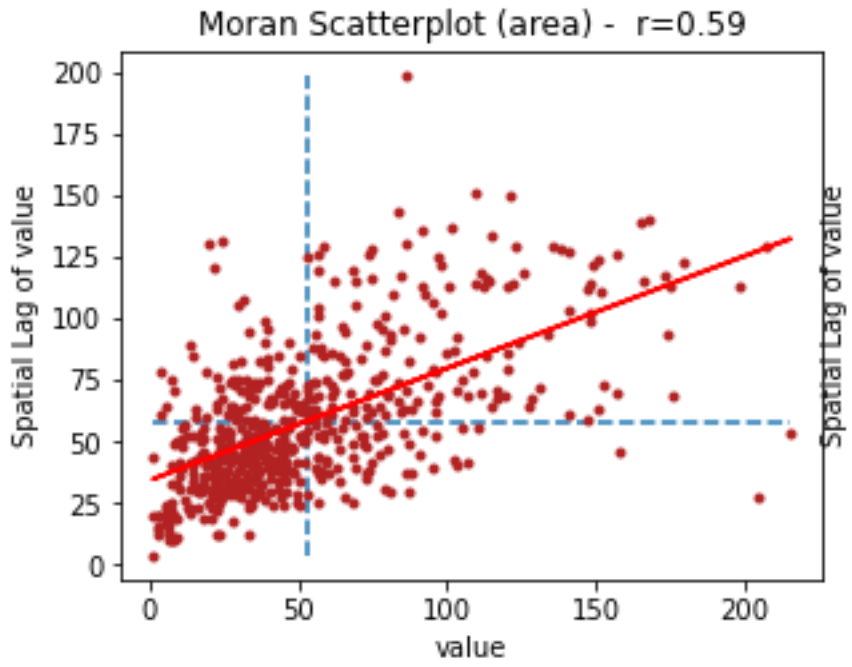


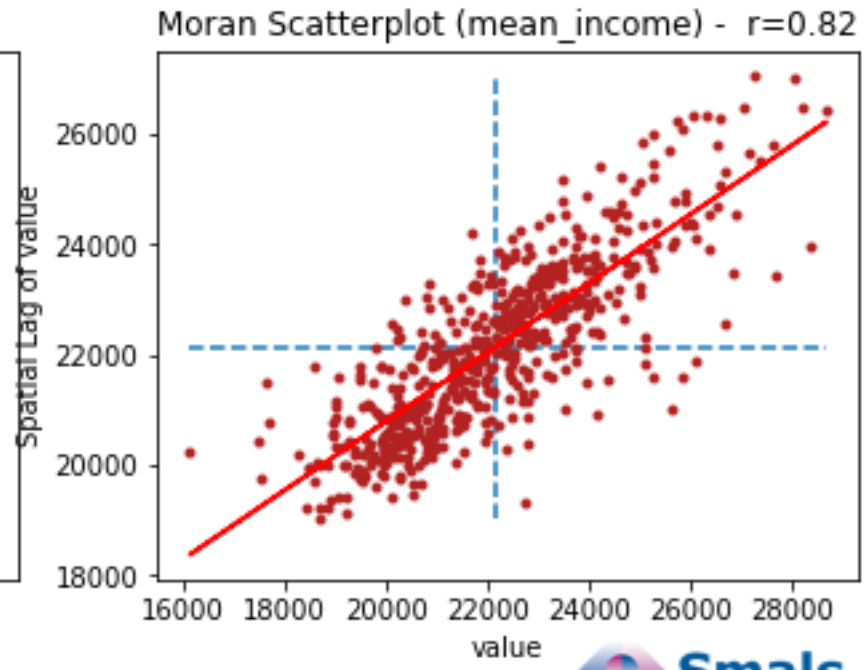
Smals
ICT for society

Autocorrélation locale : Moran scatterplot

- Indicateur de dépendance spatiale : corrélation entre la valeur et son « spatial lag »
- Aucune pour « random », très nette pour « mean income »→ indique une forte dépendance spatiale (valeur d'une zone corrélée à la valeur de ses voisines)



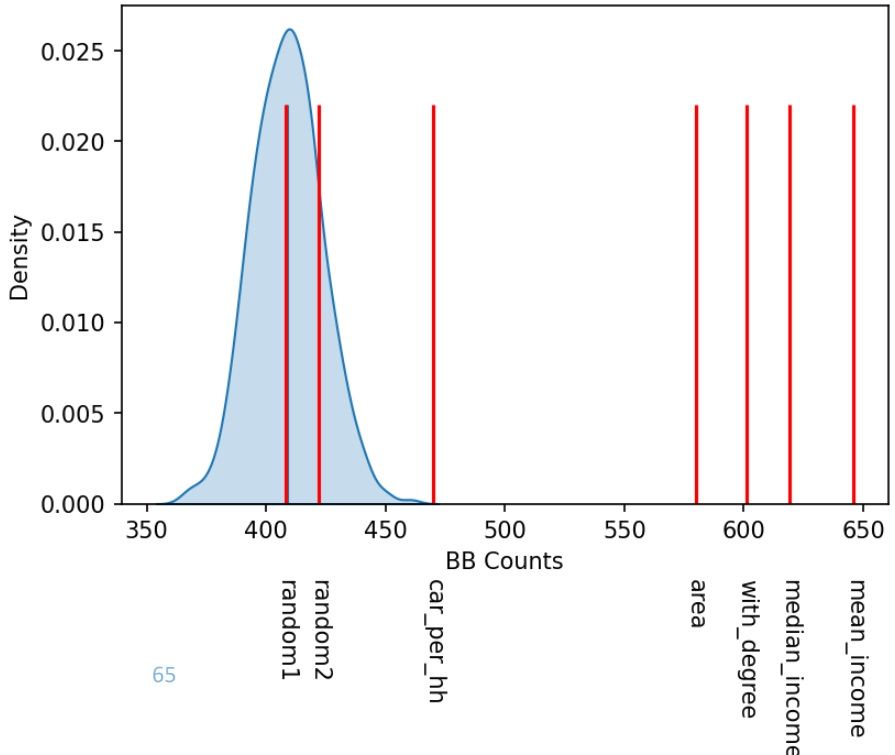


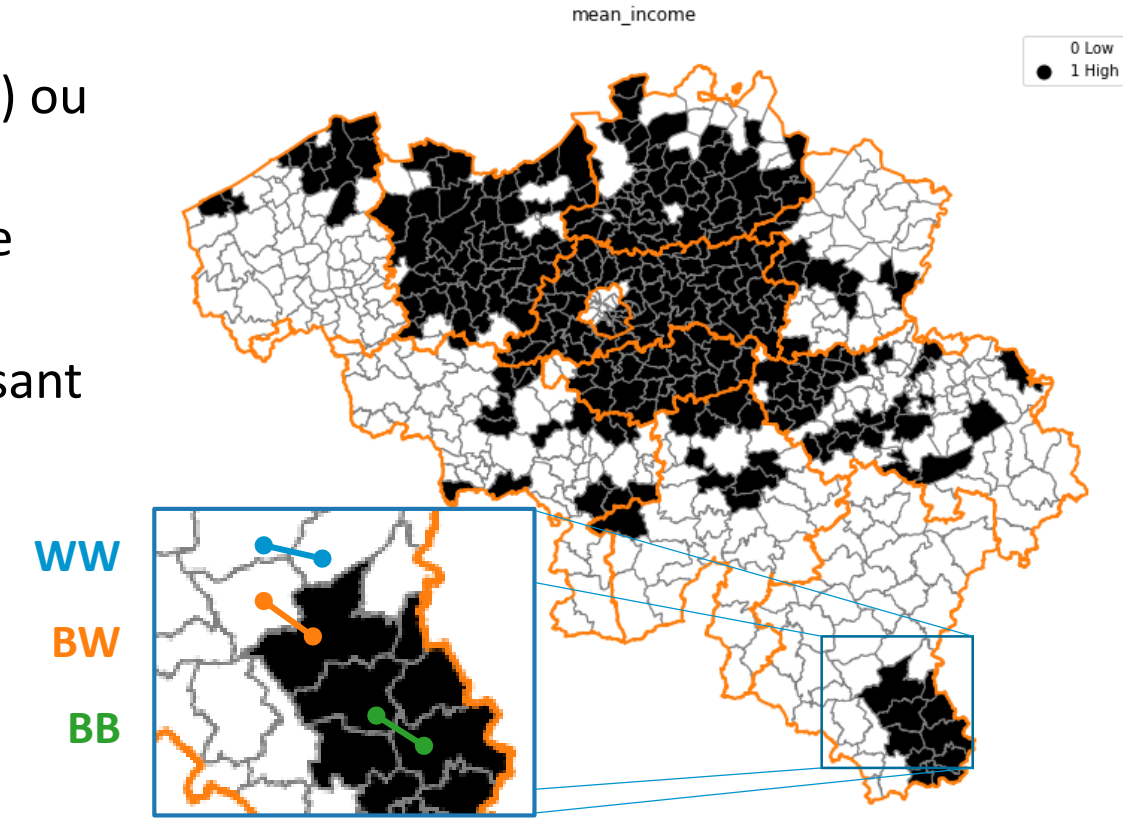


Smals
ICT for society

Autocorrélation globale : version binaire

- On convertit chaque valeur en 0 (≤ médiane, « White ») ou 1 (> médiane, « Black»)
- On regarde le nombre de « jonction Black-Black » (zone « Black » touchant une autre zone « Black »)
- On compare avec le nombre qu'on aurait eu en choisissant la couleur au hasard (pour la même géographie)





- Cloche: distribution d'un grand nombre de tirages aléatoires
- Plus on s'en éloigne, plus faible est la probabilité que la métrique soit aléatoire … donc plus grande est qu'elle soit « spatialement dépendante »
- Il existe une généralisation « continue » (Moran's I)

Smals ICT for society

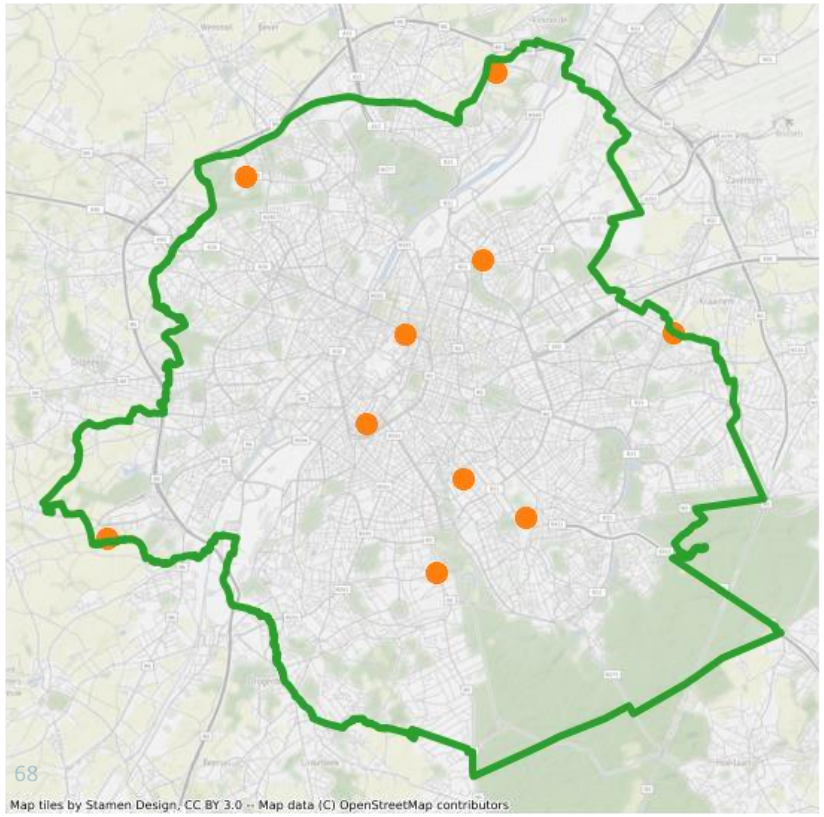
Autocorrélation spatiale : conclusion

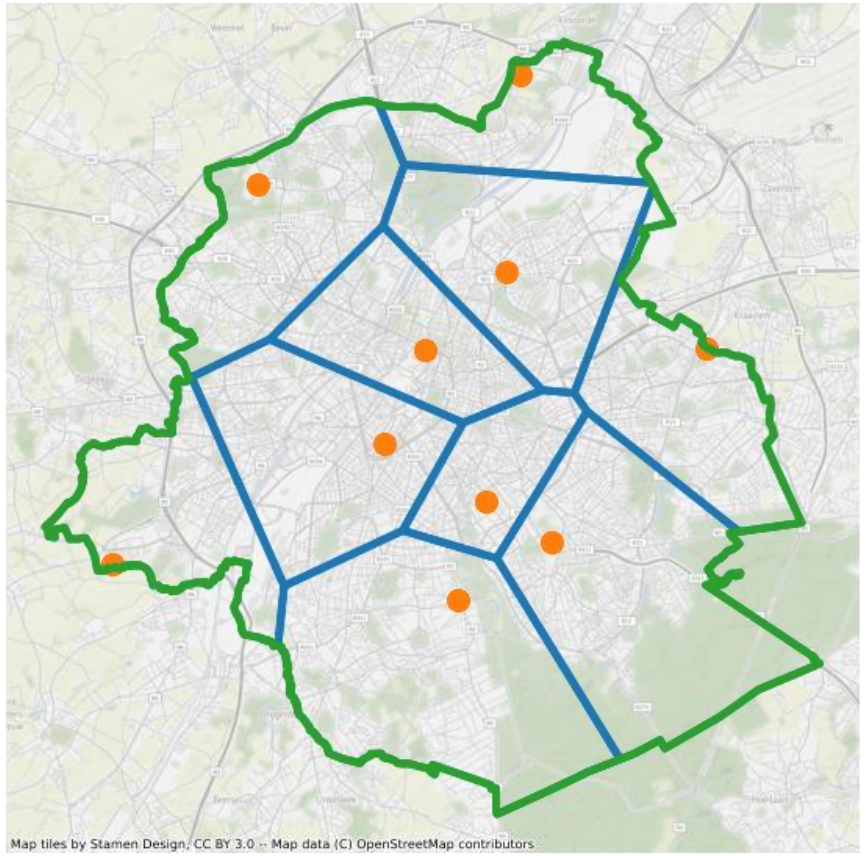
- L'autocorrélation permet quantifier/objectiver la dimension géographique d'une métrique
- Permet aussi d'identifier des anomalies (« hot spots », « donnuts »)
- Ne permet pas de déduire de causalité, mais est une première étape
- Métriques attribuées à des secteurs. Il existe aussi des méthodes pour des points. Exemple : corrélation entre le chiffre d'affaire/nombre de travailleurs/… d'un restaurant et celui des restaurant voisins ?

Diagramme de Voronoi

Diagramme de Voronoï

- Diagramme de Voronoï: À partir d'un ensemble de points (germe) dans un territoire (polygone), découpe le territoire en fonction du point le plus proche (zone d'influence)
- Détermine, pour chaque endroit du territoire, qui sera « le plus rapide à intervenir » (à vol d'oiseau), ou quelle est le « l'antenne la plus proche »
- Exemple : Polygone = Brucap, points = départs SMUR

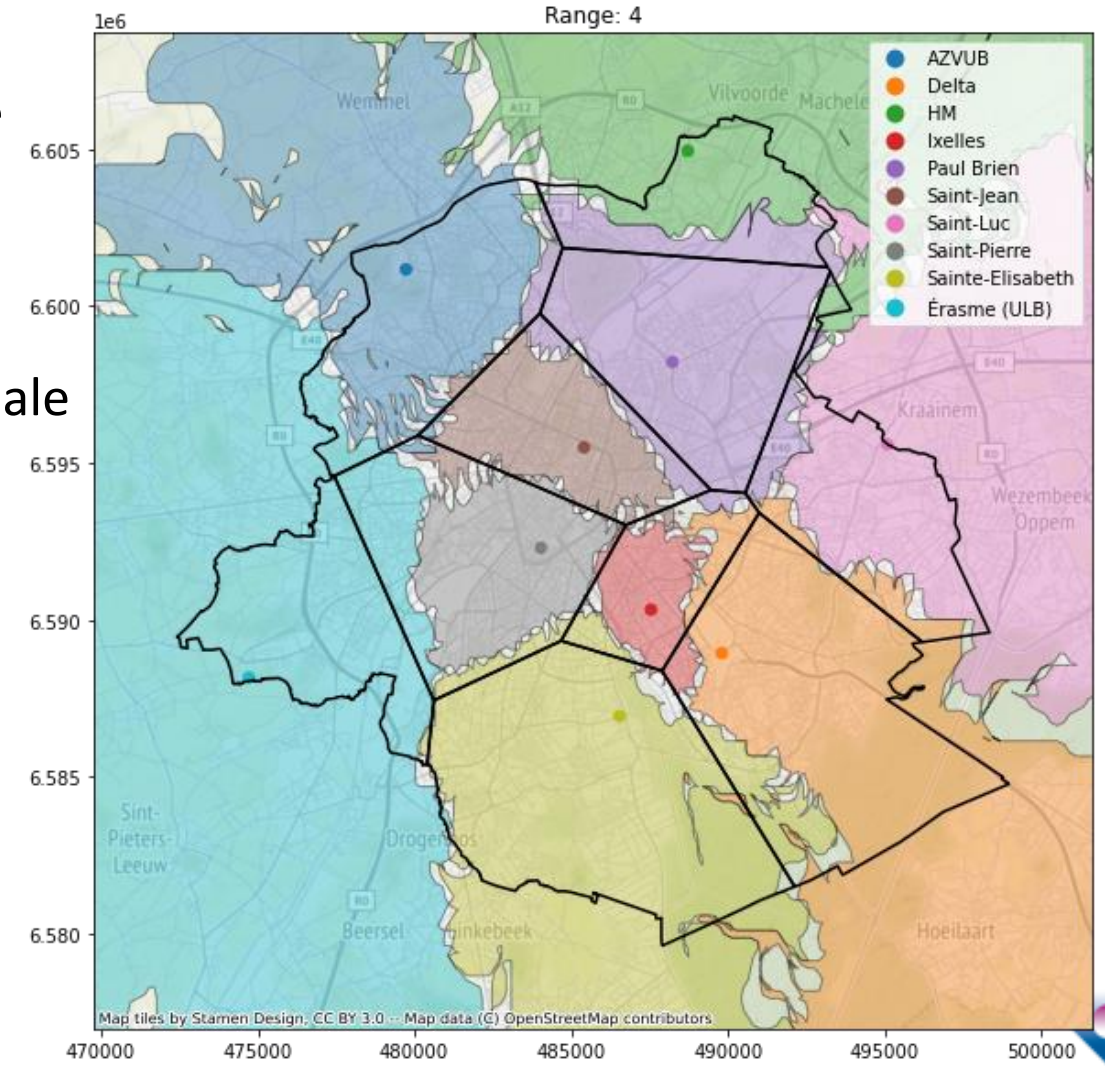




Smals ICT for society

Avec isolines

- S’il est nécessaire de tenir compte des temps de parcours, il est possible de combiner une série d’isolines
- Exemple, en extrayant les isolines de chaque départ, pour chaque pas de 30 secondes :
- Peu de différence par rapport à Voronoi (pour cet exemple)
- Peut être comparé à une attribution territoriale existante: à quel point est-elle efficace?



Multipoint reachability

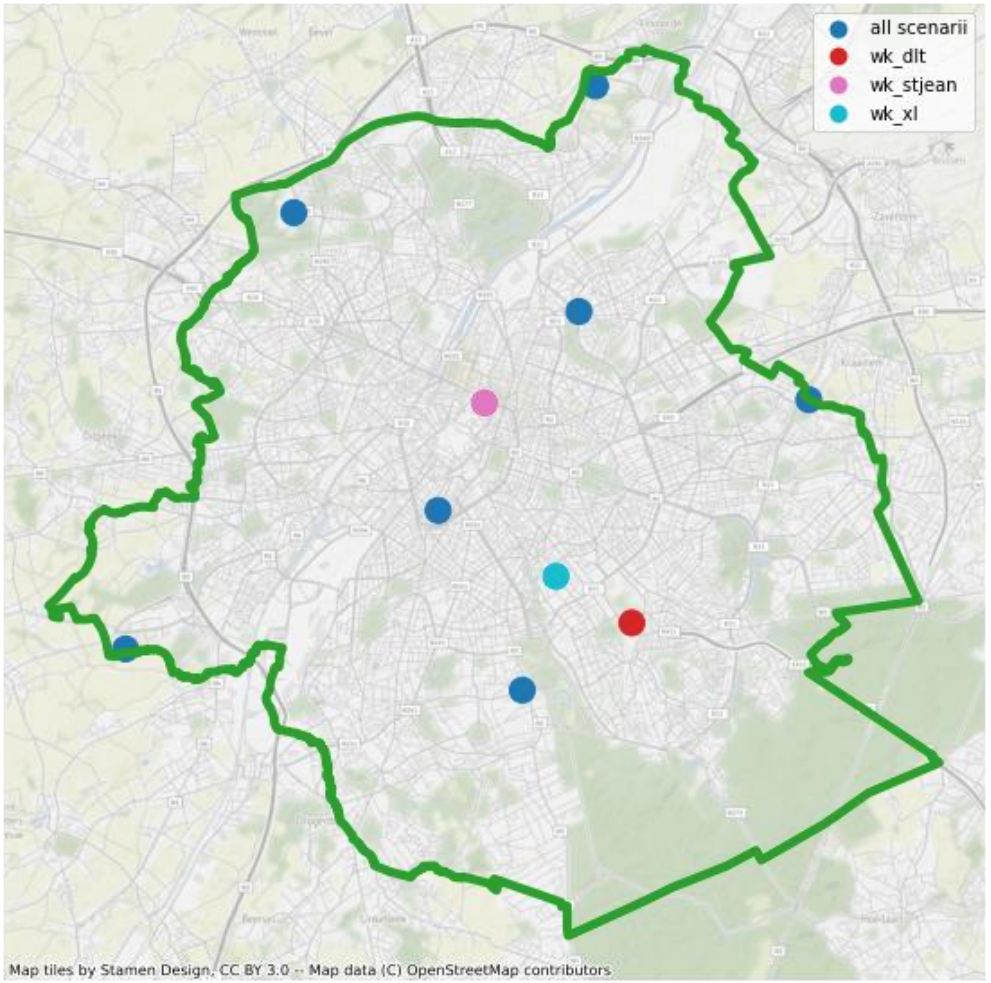
Multipoint reachability

- Question centrale : quel territoire ou quelle population est atteignable à partir d'un ensemble de points (hôpitaux, antenne locale d'une administration, domicile des agents de terrain…)
- En supposant un déplacement en voiture (possibles extensions : à pieds, en transport en commun…)
- On combine des « courbes isochrones »
- On va comparer plusieurs scénarios (ouverture/fermeture ou non d'un ou plusieurs départs)
- Usecase : SMUR (MUG) à Bruxelles : 10 départs, dont 3 en alternance un semaine/3
- « Disclaimer »: déplacement à allure « normale » et analyse simplifiée !
- Plus de détails :
  - https://www.smalsresearch.be/peut-on-toujours-atteindre-une-maternite-en-30-minutes/
  - https://www.smalsresearch.be/peut-on-toujours-atteindre-une-maternite-en-30-minutes-partie-2/

Départs

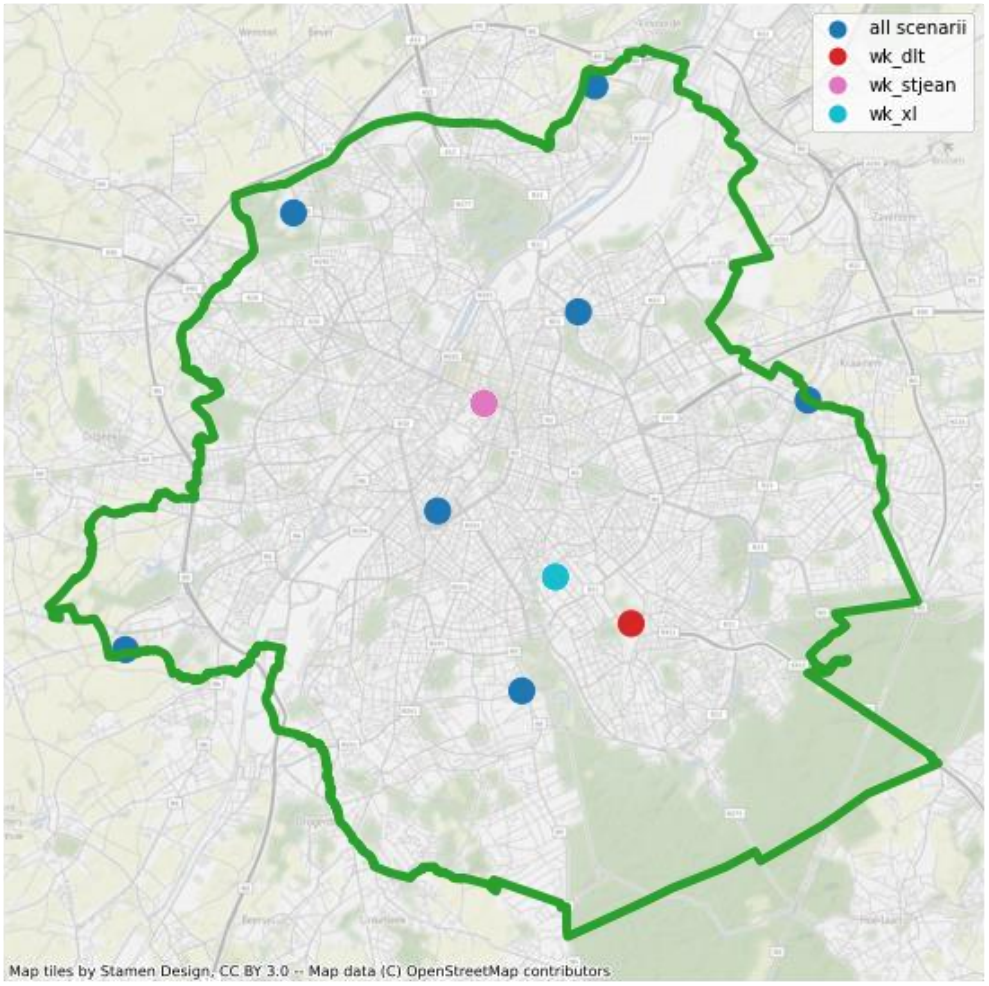
		wk_dlt	wk_stjean	wk_xl
Saint-Luc	Av. Hippocrate	x	x	x
Saint-Pierre	Rue haute	x	x	x
…	…			
Delta	Bvd du triomphe	x		
Saint-Jean	Rue du marais		x	
Ixelles	Rue Paquot			x

Géocodage

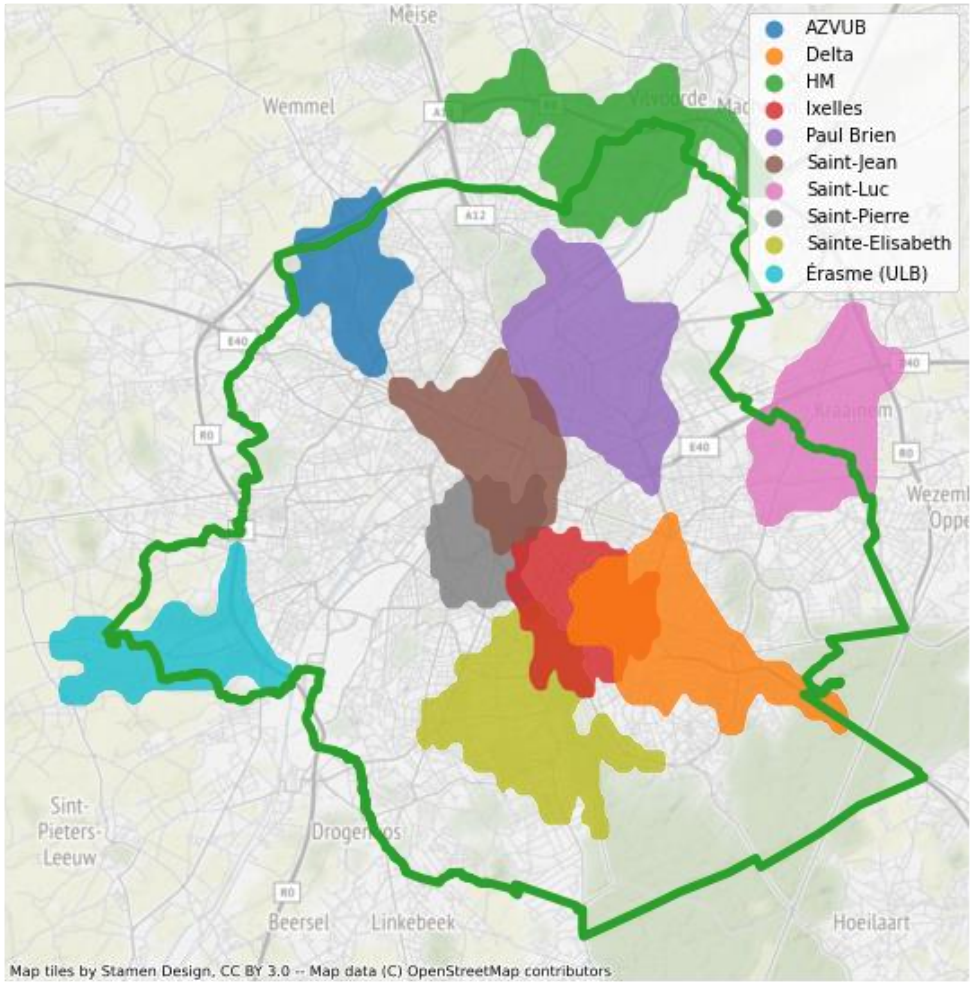


Smals
ICT for society

Courbes isochrones

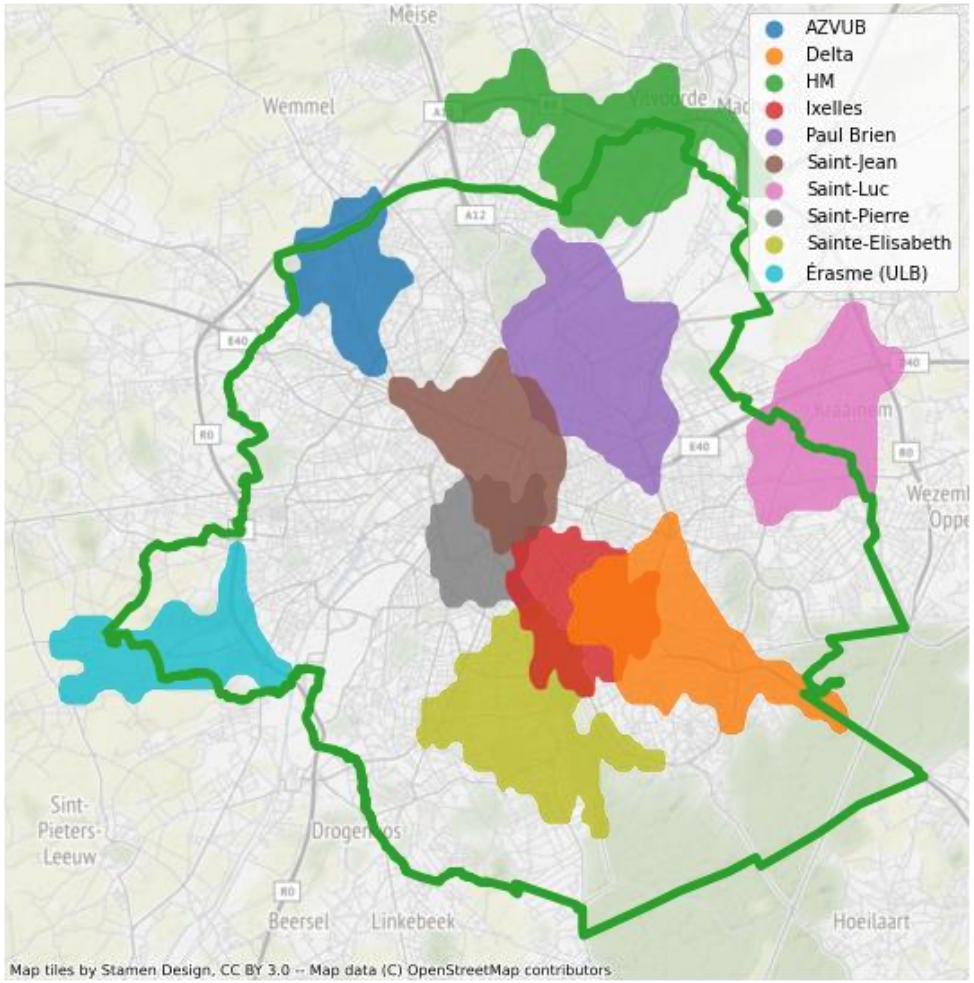


Isoline (5 min)

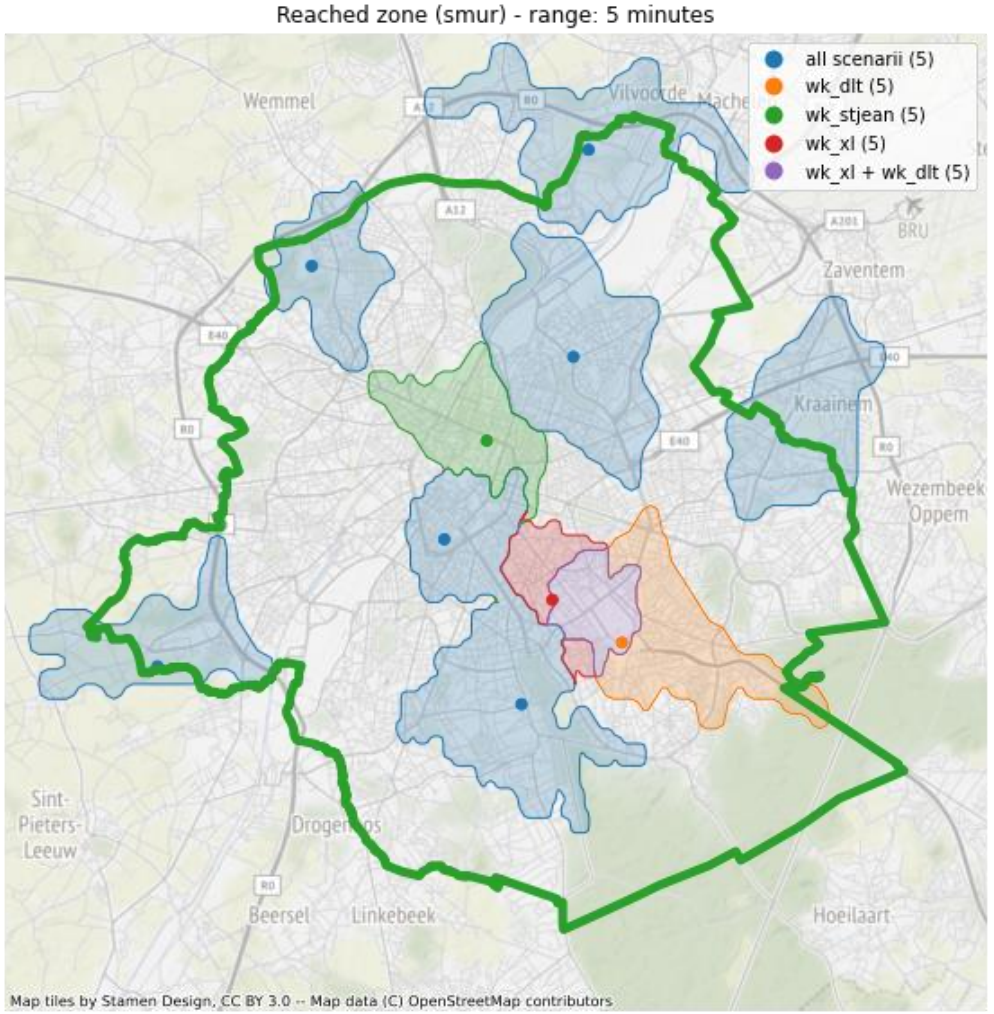


Smals
ICT for society

Pour un délai donné …

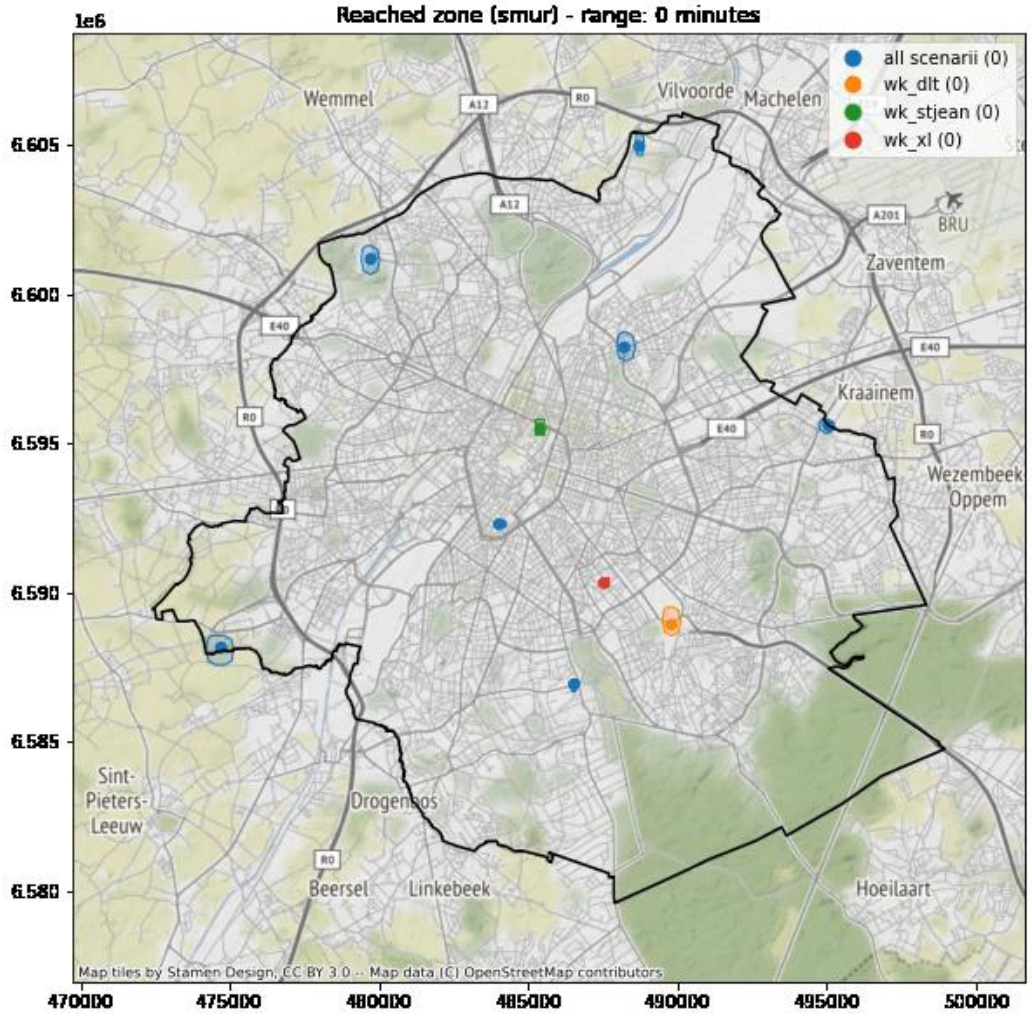


Spatial ops



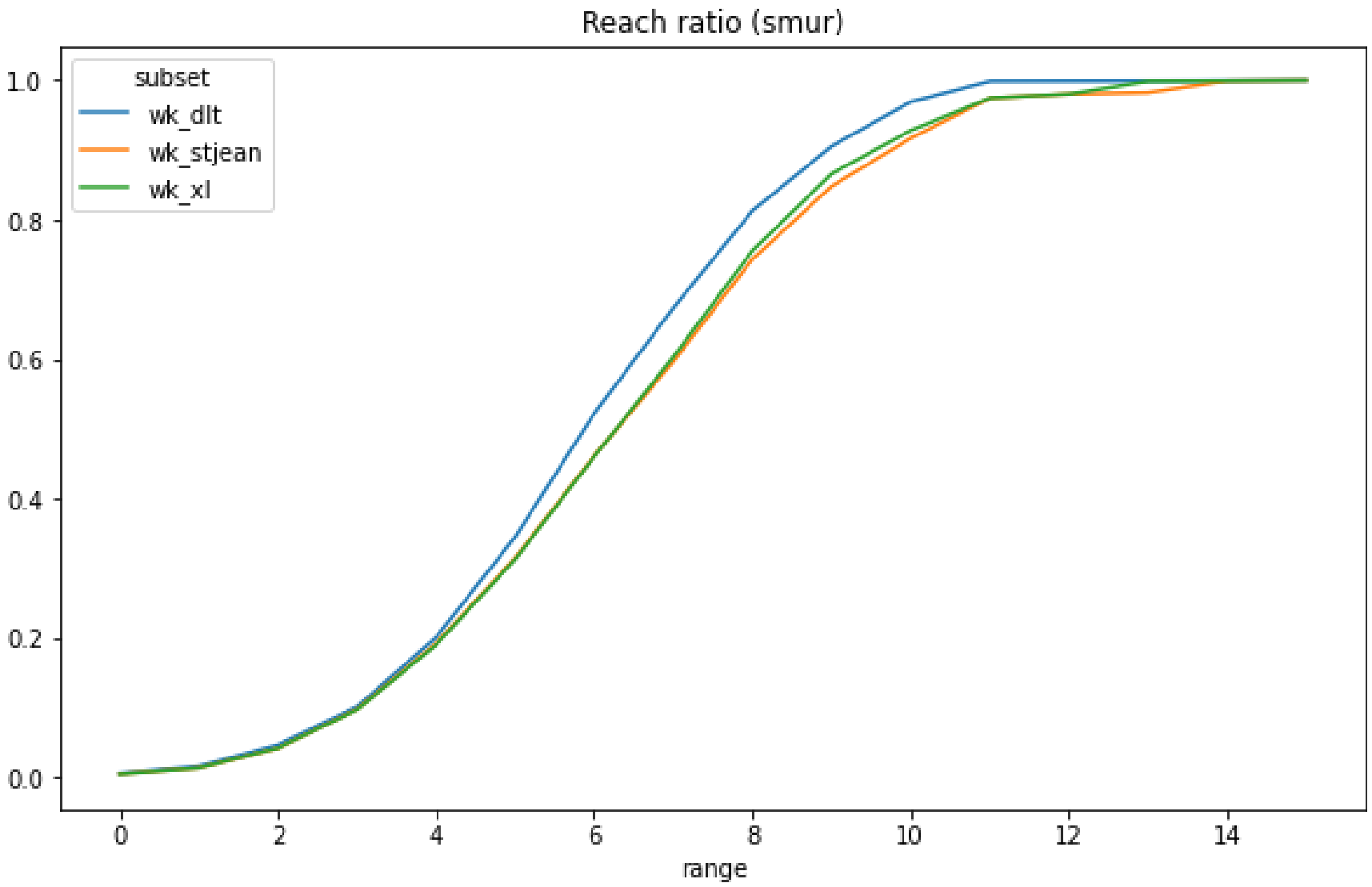
Smals ICT for society

Temps variable



Smals
ICT for society

Vue globale : reached ratio

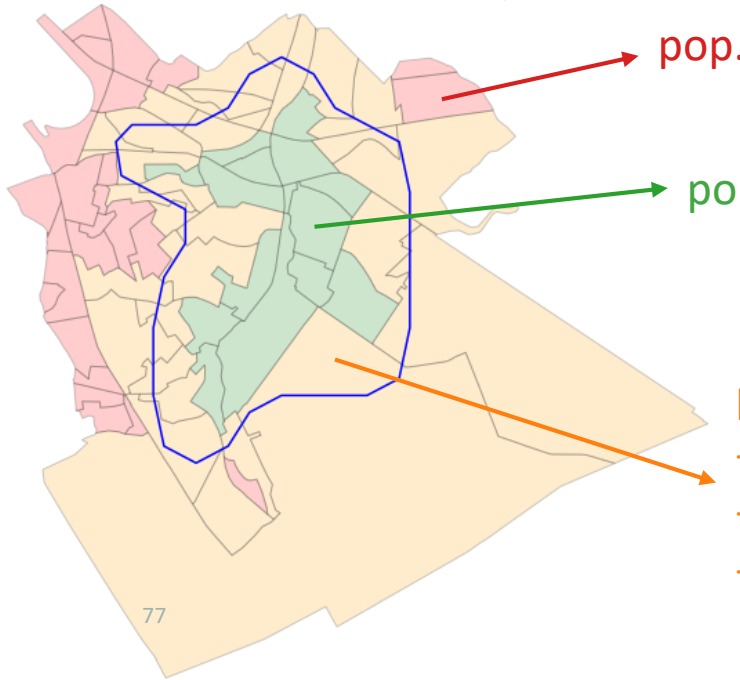


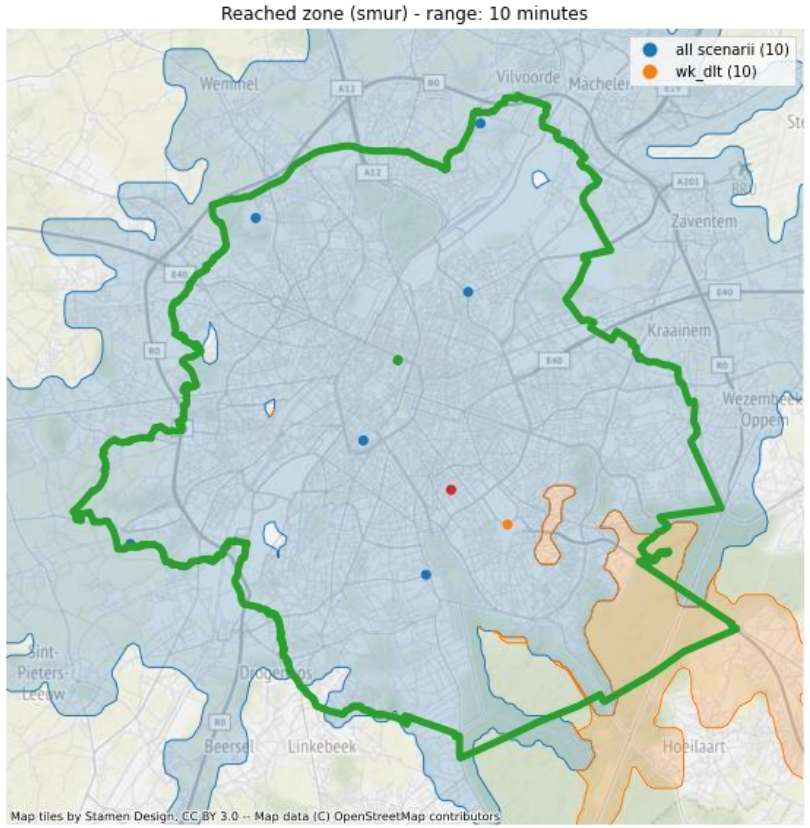
Version interactive : https://smalsresearch.be/rawfiles/reach_ratio_smur.html

Smals ICT for society

Couverture populationnelle

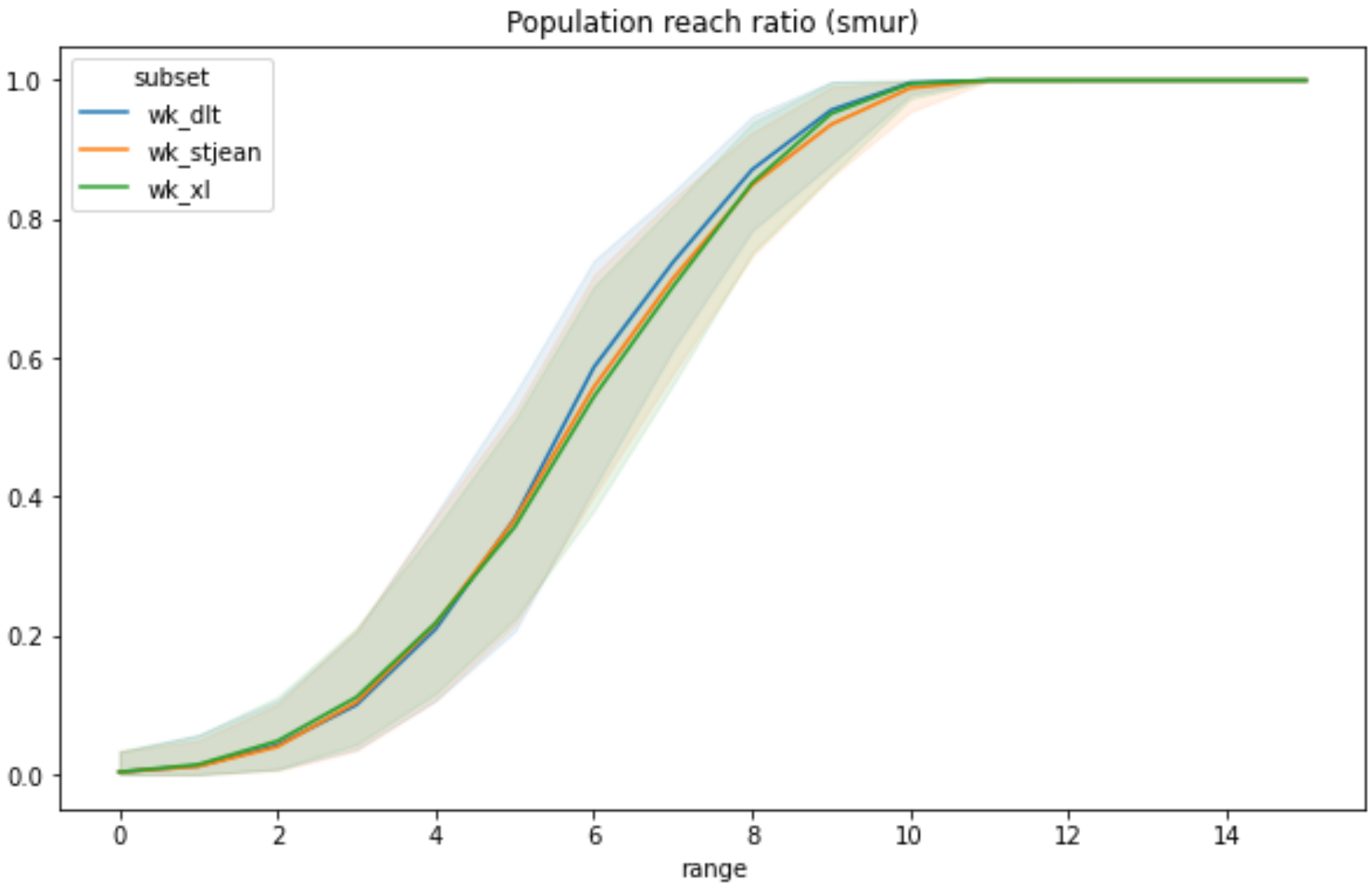
- Couverture territorial <> couverture populationnelle !
- Plus-value de la semaine “Delta”: couverture de …
  la forêt de Soignes !
- Connaissant la population par secteur statistique (Statbel), on peut approximer la population couverte par jointure spatiale
- “Approximer” car on ne connait pas la répartition au sein des secteurs statistiques





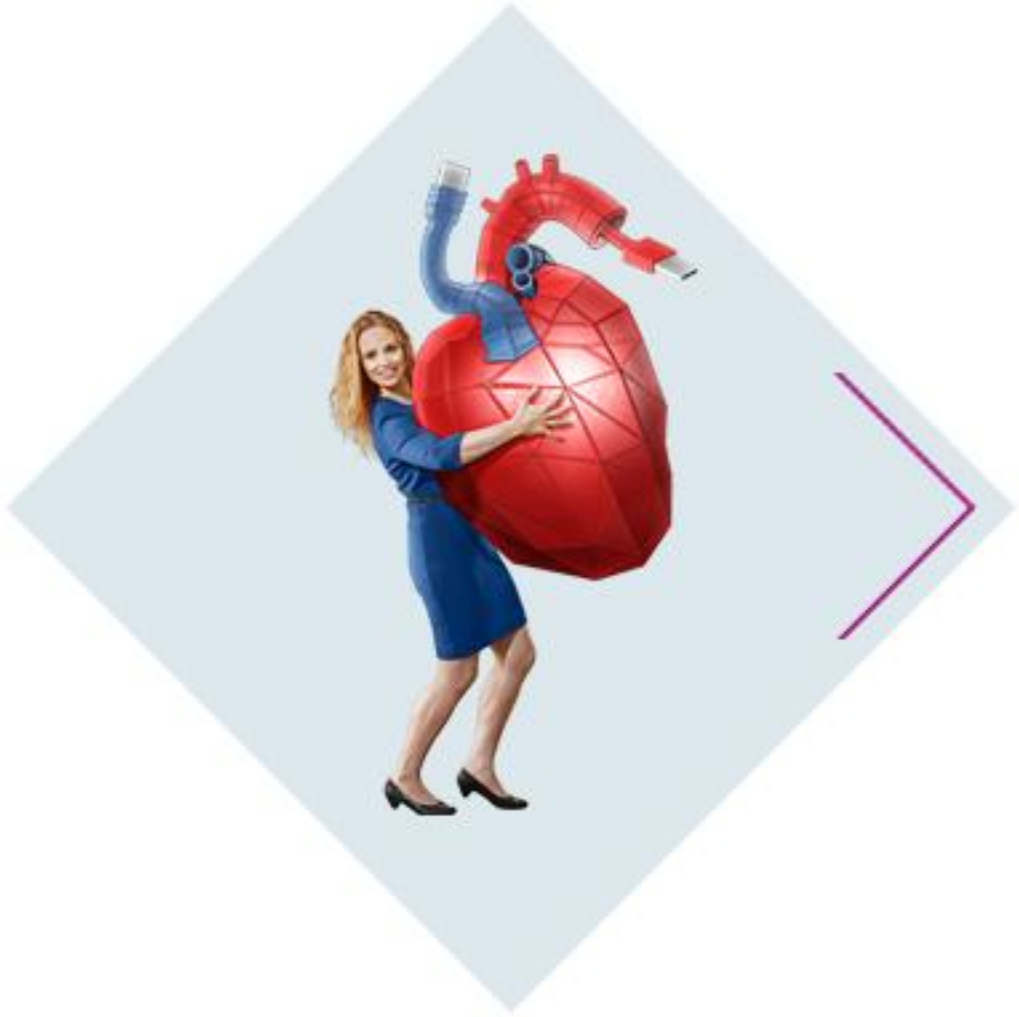
Smals ICT for society

Population reach ratio



Version interactive : https://smalsresearch.be/rawfiles/pop_reach_mean_ratio_smur.html

Smals
ICT for society



Use case “CAW”

Use case CAW

- **C**heck-in **A**t **W**ork: déclaration de présence sur un chantier (**P**lace **O**f **W**ork)
- Peut-être fait :
  - En direct : scan d'un QR code
  - À l'avance/distance
- Objectif : identifier les entreprises qui systématisent le principe de "check in everybody everywhere"
- Problèmes similaires:
  - Identifier un médecin qui envoie des certificats à des personnes "trop distantes" (INAMI ?)
  - Un indépendant qui déclare des chantiers trop distants (Fisc ?)

CAW : données en entrée

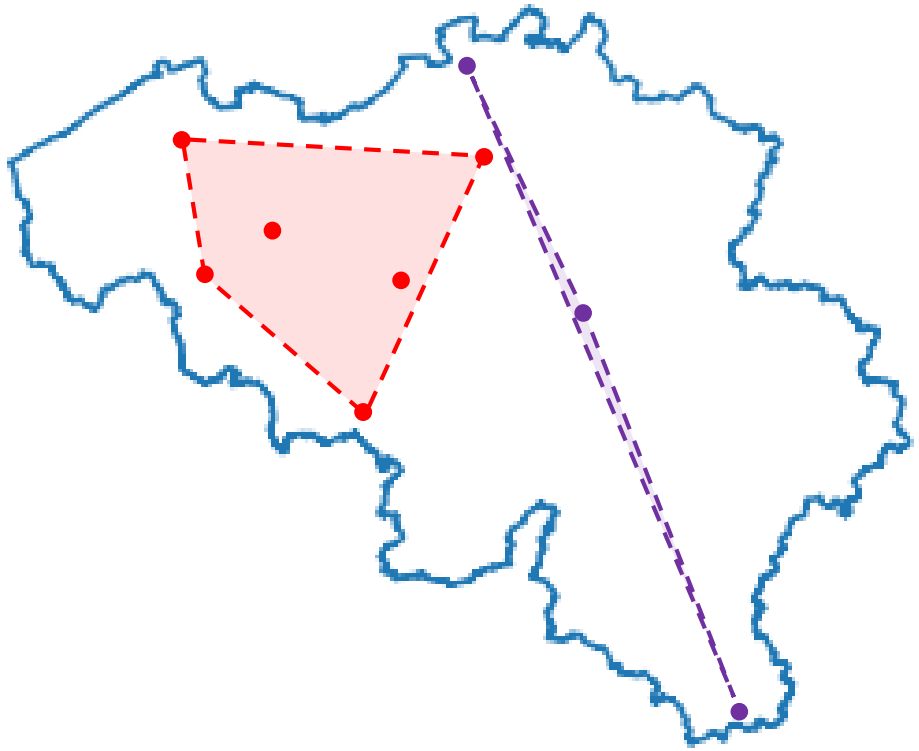
- Données en entrée (pour une date en particulier):

WORKER_ID	POW_ID	STATUS	ZIPCODE	GEOM
Wrk_1	Pow_1	OK	1060	POINT (4.34466 50.82946)
Wrk_1	Pow_2	OK	1160	POINT (4.43800 50.81019)
Wrk_2	Pow_3	NOK	5190	POINT (4.67619 50.47171)
Wrk_3	Pow_4	CANCELLED	6220	POINT (4.52807 50.47096)

- Pour éviter du géocodage, on se contente du code postal
- Jointure avec les données BPost

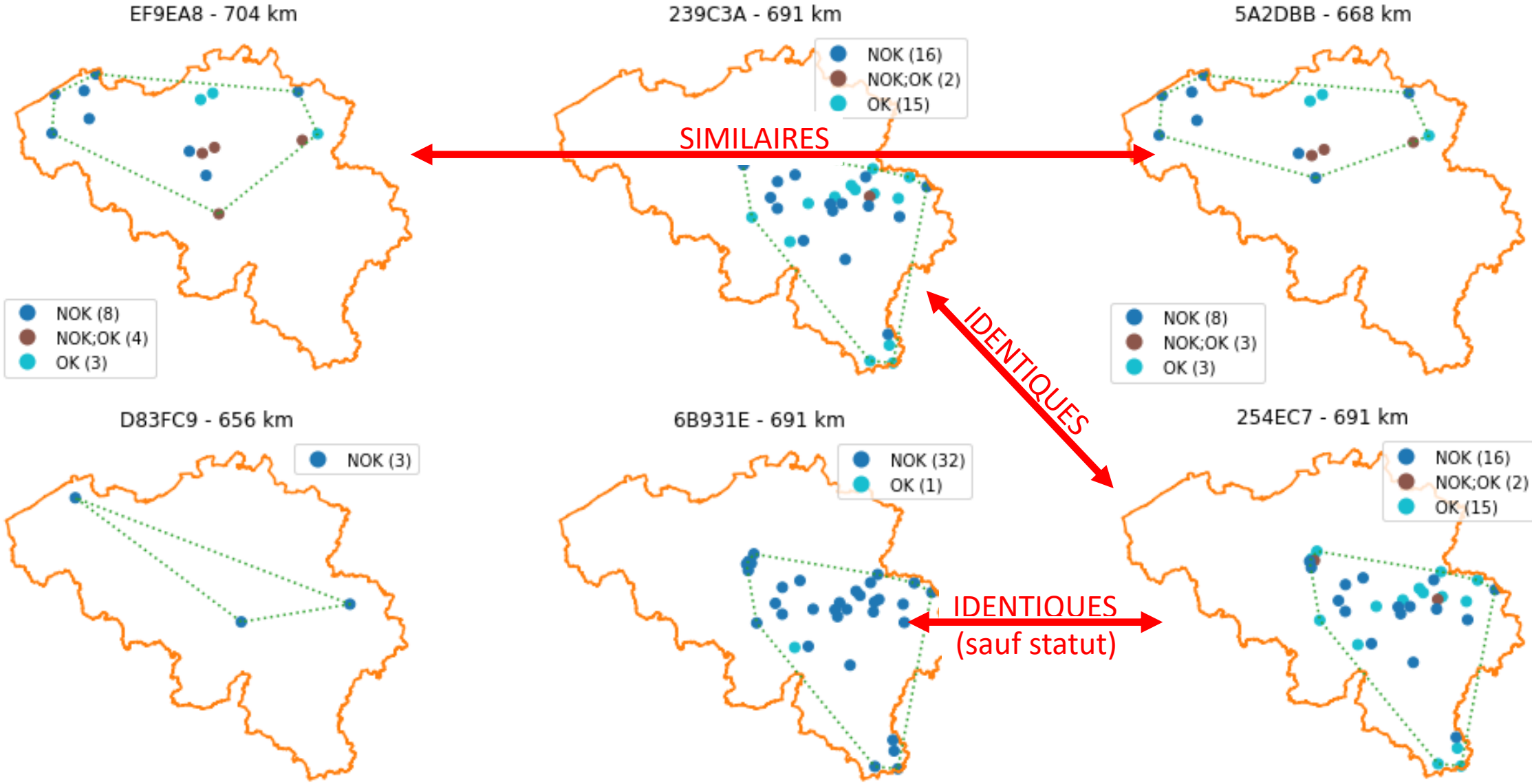
POW: Place Of Work

Enveloppe convexe



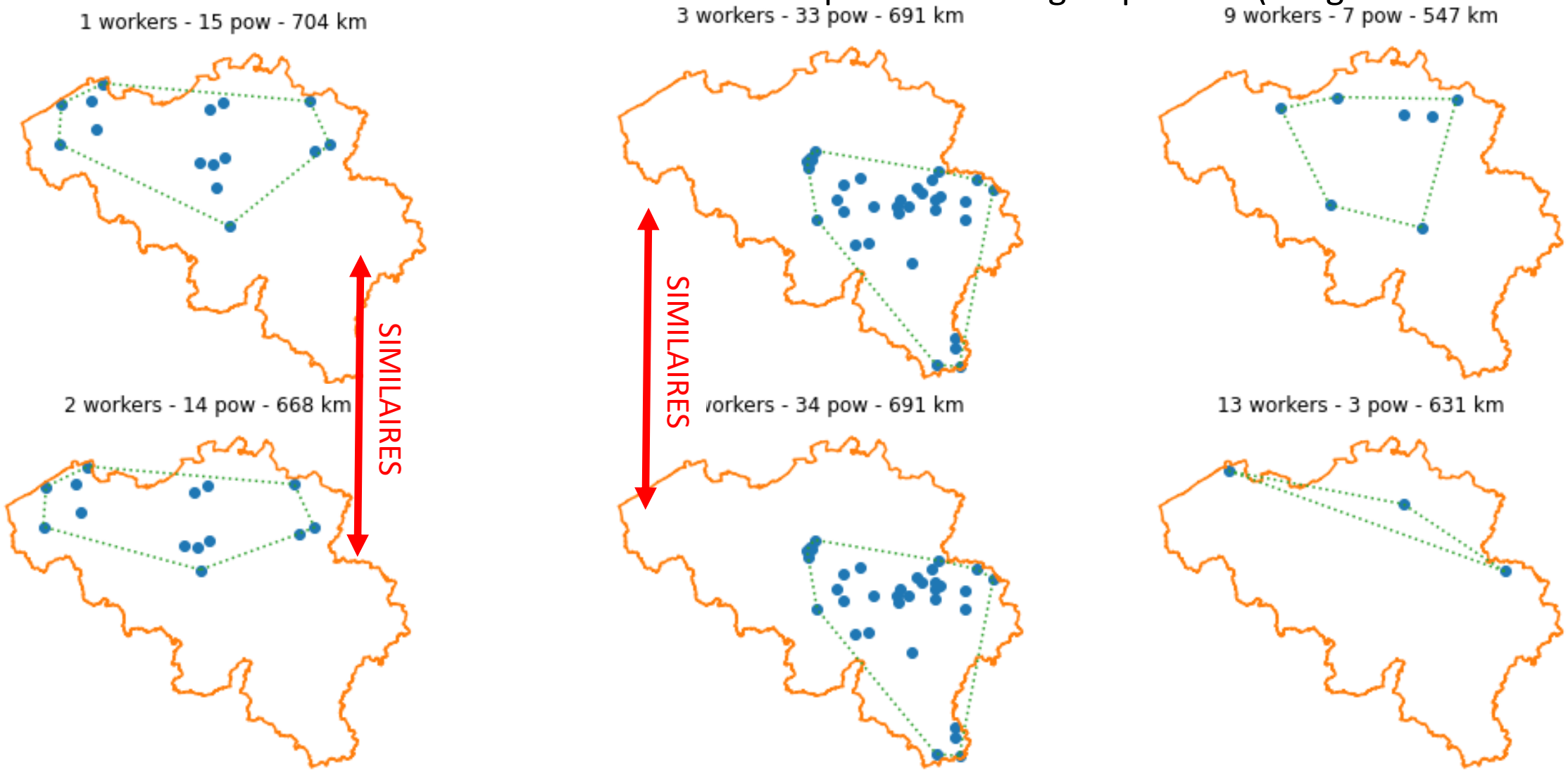
- Pour chaque travailleur, on rassemble tous les points
- On construit une “enveloppe convexe” (convex hull)
- Périmètre de l’enveloppe : borne inférieure de la distance à parcourir par le travailleur (retour au départ)
- Inférieure, car :
  - Ignore les points dans l’enveloppe
  - À vol d’oiseau
  - Supposant que le travailleur habite près d’un des points
- On peut aussi considérer l’aire de l’enveloppe, mais nulle si les points sont alignés (ou seulement 2 points)

Exemples


EF9EA8 - 704 km
NOK (8)
NOK;OK (4)
OK (3)
239C3A - 691 km
NOK (16)
NOK;OK (2)
OK (15)
SIMILAIRES
5A2DBB - 668 km
NOK (8)
NOK;OK (3)
OK (3)
IDENTIQUES
D83FC9 - 656 km
NOK (3)
6B931E - 691 km
NOK (32)
OK (1)
IDENTIQUES
(sauf statut)
254EC7 - 691 km
NOK (16)
NOK;OK (2)
OK (15)

Patterns identiques

Constat: de nombreux travailleurs ont le “même pattern” → regroupement (en ignorant le statut)



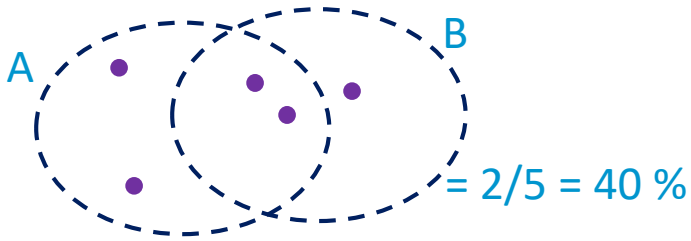
Patterns similaires

- On peut évaluer la proximité entre deux ensembles A et B par la “similarité de Jaccard” :

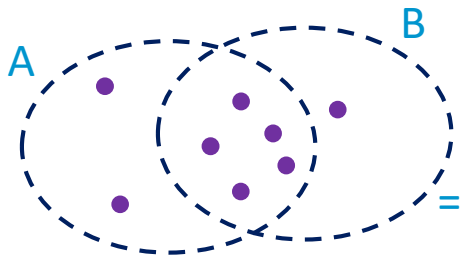
$$\frac{|A \cap B|}{|A \cup B|}$$

Taille de l’intersection

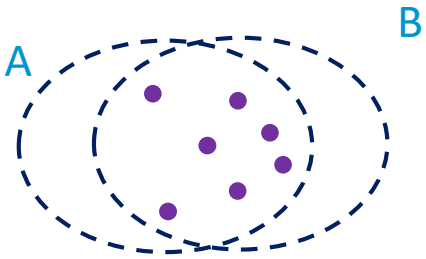
Taille de l’union



= 2/5 = 40 %



= 5/8 = 62,5 %



= 7/7 = 100 %

- “Point”= centre du code postal du POW

Smals ICT for society

Patterns similaires

Jaccard similarity: 97.1%

3 workers - 33 pow - 691 km

1 workers - 34 pow - 691 km

Points communs
Différences

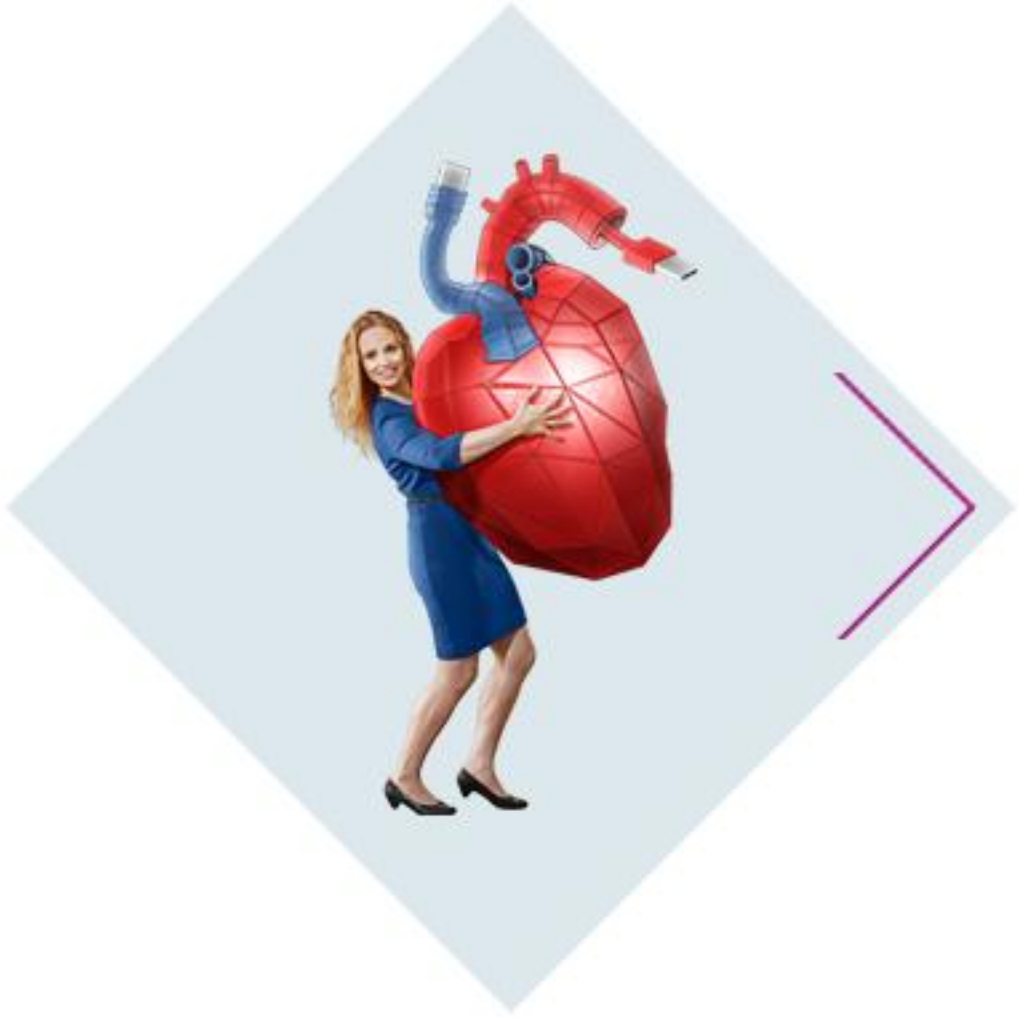
Jaccard similarity: 93.3%

2 workers - 14 pow - 668 km

1 workers - 15 pow - 704 km

Use case CAW: conclusions

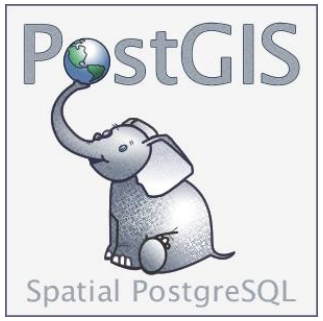
- Ceci est un simple POC “naïf”
- On se contente de la localisation du code postal
- On ignore le statut
- On ignore le “contenu de l’enveloppe"
- On ne regarde qu’un jour → l’aspect temporel est certainement crucial, mais “out of scope” de cette présentation
- Sans analytique géographique, ces cas auraient été difficiles à évaluer
- Extension possible : identifier les travailleurs qui ont (probablement) le même « patron »
- Distance totale à parcourir: problème du voyageur de commerce (Travelling salesman problem) !



Conclusions

Outils Open-source/gratuits

- Programmation :
  - Analytique : Python + Pandas + Geopandas/shapely
  - Visualisation (javascript) : leaflet.js, mapbox, google maps…
  - API : OSM, Here, Google maps…
- Database : Postgresql + PostGIS
- Desktop : QGIS



Smals
ICT for society

Conclusions

- Les données géographiques sont partout dans les données “administratives”
- Ne concerne pas que les géographes/géologues/épidémiologistes/militaires !
- Plus-value indéniable pour l’analyse :
  - Statistiques
  - Détections d’anomalies/fraudes
  - Analyse d’impact (ouverture/fermeture d’antenne)
  - Optimisation (tournées/affectation d’agents…)
- Des outils gratuits/Open-sources sont souvent suffisants
- À la portée de tout data-scientist/technicien curieux

Vandy BERTEN
vandy.berten@smals.be

Smals, ICT for society

02 787 57 11

Fonsnylaan 20 / Avenue Fonsny 20

1060 Brussel / 1060 Bruxelles