"Analysis of massive imports of open data in OpenStreetMap database: a study case for France"

A work of *Arnaud Le Guilcher, Ana-Maria Olteanu-Raimond, Mamadou Bailo Balde*

Presented by *Alexys Ren, student at ENSG*

May 2023

LaSTIG Université Gustave Eiffel ENSG Géomatique École nationale des sciences géographiques IGN

Table of contents

Context

- an important evolution in the production of Geographical Information: **volunteered geographic information VGI (Goodchild, 2017**)
- **The OpenStreetMap project** is one of the most prominent based on VGI
- Some studies with the goal to **compare both VGI & traditional spatial data produced by official agencies with authoritative spatial data**
- VGI's potential to produce **fresh and accurate data**, thus **reducing the time between 2 releases**
- In France, Spain or in the US, **an important proportion of the OSM features comes from authoritative data**
- 2 questions:
    - **How massive imports are evolving once integrated in the OSM databases?**
    - **How the enrichment, error corrections and updating of massive imports by the OSM community can benefit the original datasets which were integrated in OSM?**

Methodology
Classification of sources for OSM data (1/3)

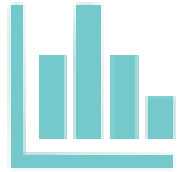
- **The "source" key**
- **Homogenization and classification process** (for the geometry of features):
  - Massive import: features coming from external spatial open data existing at a national scale, or a local scale and can be produced by national and local governments, third-party communities, NGO, citizen science communities
  - Photos analysis: features edited by OSM contributors by using geolocalised photos coming from different sources such as Yahoo, Mapillary
  - Vectorization: features edited by OSM contributors by using maps, aerial or satellite imagery
  - Satellite navigation receiver: features being collected by using equipment such as GNSS devices, smartphones with GPS included
  - No source: features having the "source" key not filled in

Methodology
Identification and analysis of massive imports (2/3)

- **Only OSM features belonging to "massive import" are considered further on**
- **Feature(t0):** feature at the time of its integration in the OSM database
- **Feature(t1), …, Feature(ti), …, Feature(tm) with i = 1, …, m.** m = total number of editions
- **The "changeset" =** changes between Feature(ti) and Feature (ti+1)
- **Typology of modifications:**
  - Geometry modification
  - Tag enrichment
  - Tag suppression
  - Tag modification
  - Geometry suppression
- **Intensity instead of magnitude**
- **No quality assessment!**

Methodology

Computation of the evolution pattern for massive imports (3/3)



- **The sequencing method**

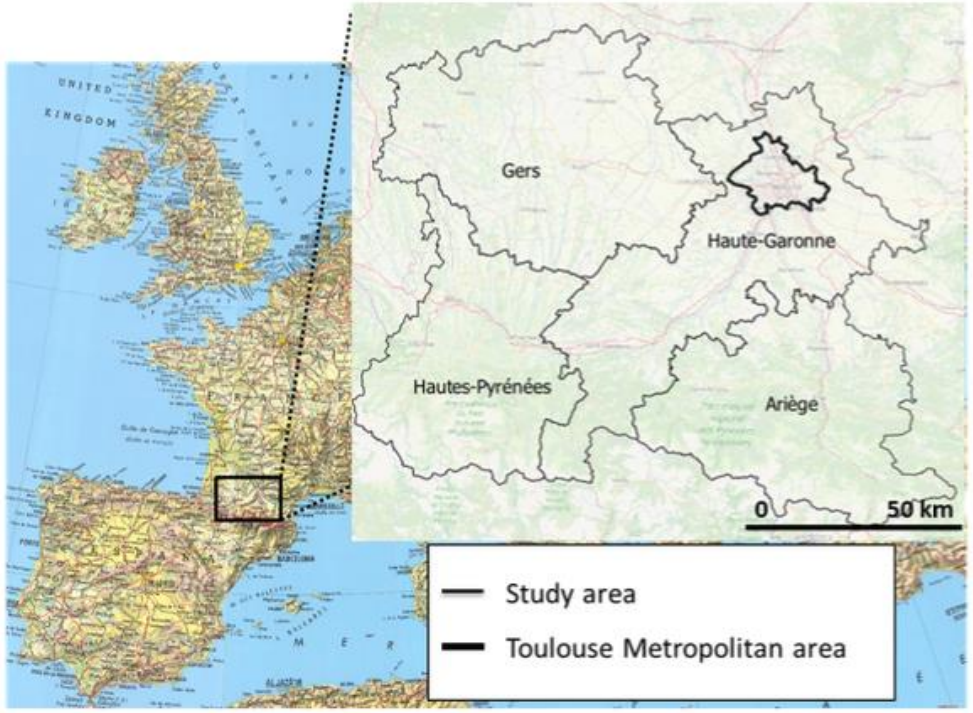
  1) Encoding the different categories of our typology of modifications with numbers from 1 to 4: **no geometry suppression!**

  2) A sequence is defined for each imported feature: creating a list of the codes of the different types of modifications underwent by the feature in increasing order, **for each changeset.** The complete sequence for the feature is then the concatenation of all the lists.

  3) Optimal Matching Analysis (OMA) (Abbott and Hrycak, 1990): counting the minimum number of modifications to be made between 2 sequences in order to obtain identical sequences. Those distances are stored in a triangular matrix, to which **Wards minimum variance (Dlouhy and Biemann, 2015)** is applied.

  => Features are assigned **clusters**
  => Features in the **same cluster** are supposed to follow the **same evolution pattern**!

Study site and datasets



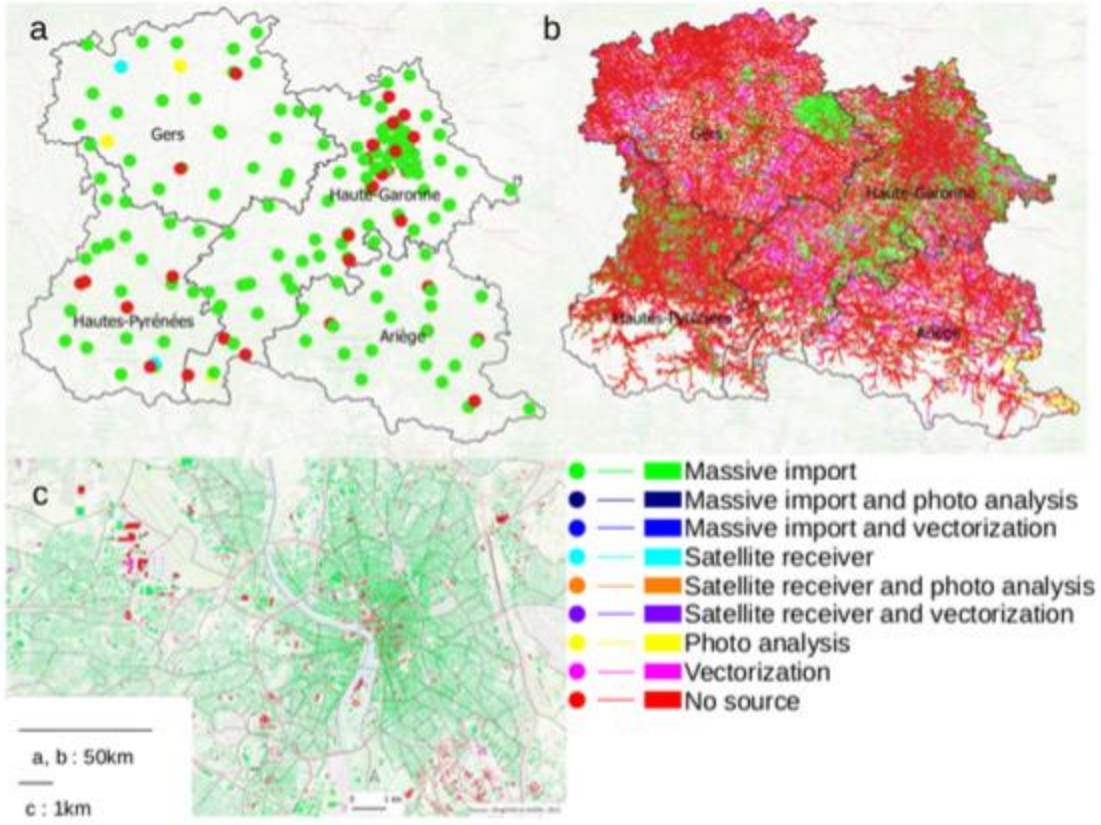
Sources: Géoportail, BDTOPO

- **4 French departments:** Ariège, Gers, Hautes-Pyrénées, and Haute Garonne (all in Occitanie region)
- **Area = 24,936 km2**
- **Various landscapes**
- **The Toulouse metropolitan**
- **Active local community**

3 types of vector data

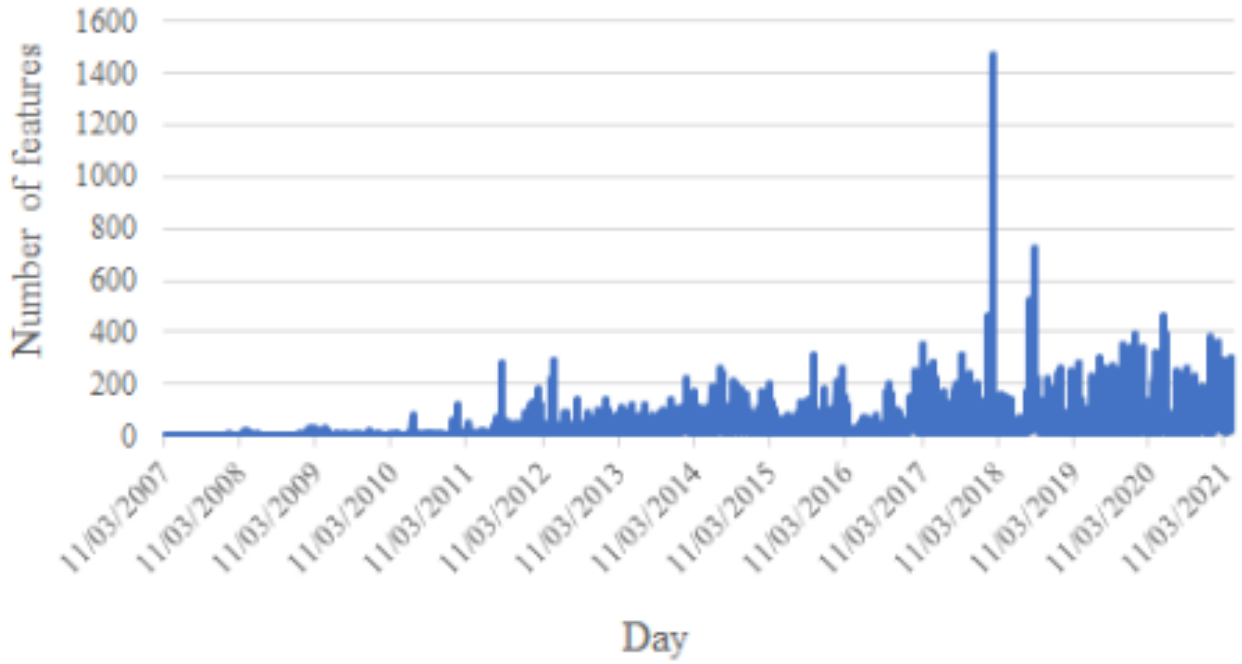
- Police Station ("POI"): points (186 features)
- Road network ("roads"): ways and relations (248,990 features)
- Building ("buildings"): nodes, lines and relations (400,063 features)

Results and discussion
Sources in the datasets (1/3)



Spatial distribution of the different types of sources for the "POI", "roads" and "buildings" datasets

Results and discussion
Sources in the datasets (1/3)



Distribution of the dates of creation of features with no source in "roads" dataset

Results and discussion

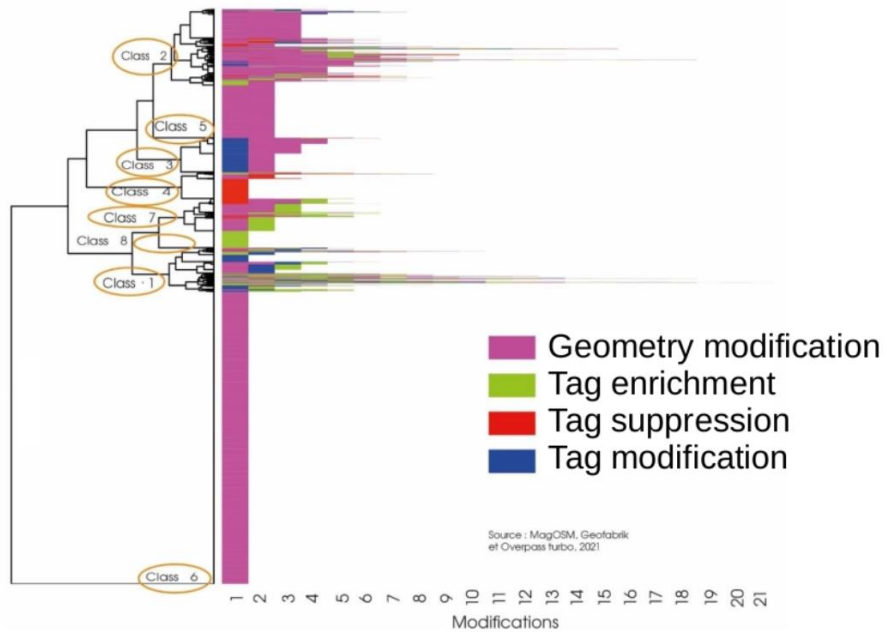
Unit modifications after massive imports (2/3)

	”POI”	”roads”	”buildings”
Geometry modification	5	39,869	11,189
Tag enrichment	160	23,751	2,370
Tag suppression	34	15,057	1,313
Tag modification	2	1,309	698
Total	201	79,986	15,570

Number of modifications for each type in "POI", "roads" and "buildings" datasets

Results and discussion
Editing sequences after a massive import (3/3)

- **Sequence definition:** PostgreSQL & R
- **Clustering analysis**: "TraMineR" library in R



Carpet of sequences for the "buildings" dataset, with a typology of 4 modifications

How we can apply this method to NLS's data