Assess the OpenStreetMap data quality starting from contribution history

SAGEO2017 - Workshop Crowdsourcing and voluntary geographical information

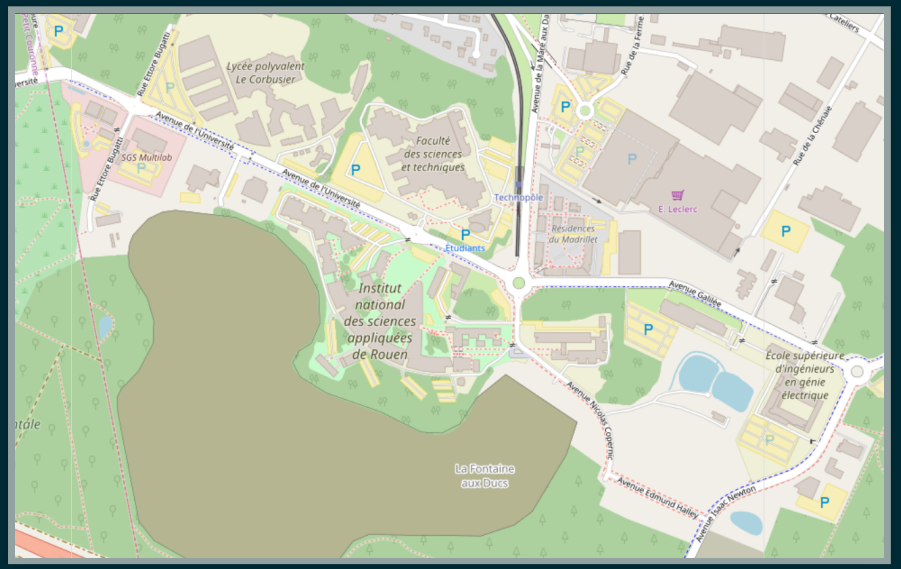
Raphaël Delhome(*), Damien Garaud, Hugo Mercier

Introduction

OpenStreetMap (OSM)

View of Europe

Here we are!



OpenStreetMap (OSM)

OSM: Map the world collaboratively

OSM data quality

Research question: *can we assess the OSM data quality?*

- Sparsity of alternative data sources (what about reference data?)
- Everyone can contributes, but does everyone knows how to contribute?
- Evaluate the OSM objects with their metadata

Outline

(Part 1) Research framework: get and prepare the data

(Part 2) Unsupervised learning applied on OSM metadata

(Part 3) Insights about OSM data quality

Research framework

Where to get OSM data

- Direct exports on OpenStreetMap API (small areas)
- `osm.pbf` file dowloading from:
  - GeoFabrik (`osh.pbf` history files),
  - Mapzen (large cities all around the world),
  - download.openstreetmap.fr
- `osmium` extracting tool (a small area from a larger one)

Then, processing with various tools (imposm, pyosmium, osm2pgsql...)

Data gathering

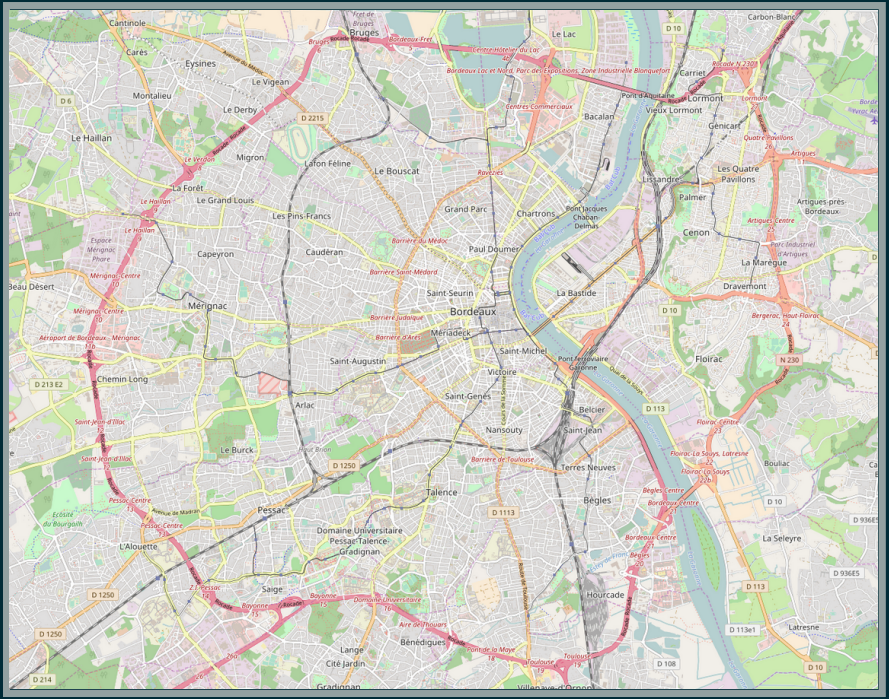
`Pyosmium`: a tool (among others) to parse OSM data

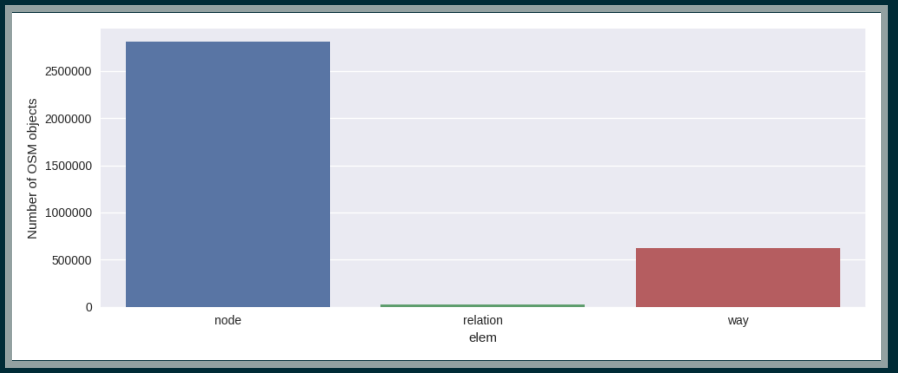
```
import osmium as osm

class TimelineHandler(osm.SimpleHandler):
    def __init__(self):
        osm.SimpleHandler.__init__(self)
        self.timeline = []
    def add_element(self, elem, elem_type):
        self.timeline.append([elem_type,
                              elem.id,
                              elem.version,
                              elem.whatever])
    def node(self, n):
        self.add_element(n, "node")
```

Recover the data in two lines:

```
timeline_handler = TimelineHandler()
timeline_handler.apply_file("osm_file.osh.pbf")
```

Data presentation





```
-<osm version="0.6" generator="CGImap 0.6.0 (6182 thorn-01.openstreetmap.org)"
copyright="OpenStreetMap and contributors" attribution="http://www.openstreetmap.org/copyright"
license="http://opendatacommons.org/licenses/odbl/1-0/">
   <node id="101439002" visible="true" version="6" changeset="25575532"
   timestamp="2014-09-21T09:49:32Z" user="Geofreund1" uid="179581" lat="44.5153863" lon="-
   0.8404732"/>
</osm>
```

Metadata extraction

From OSM data history, we get several versions per element, modified by different users.

-> How to exploit this information?

- Focus on change sets?
- Focus on contributors?
- Focus on elements themselves?

Metadata extraction

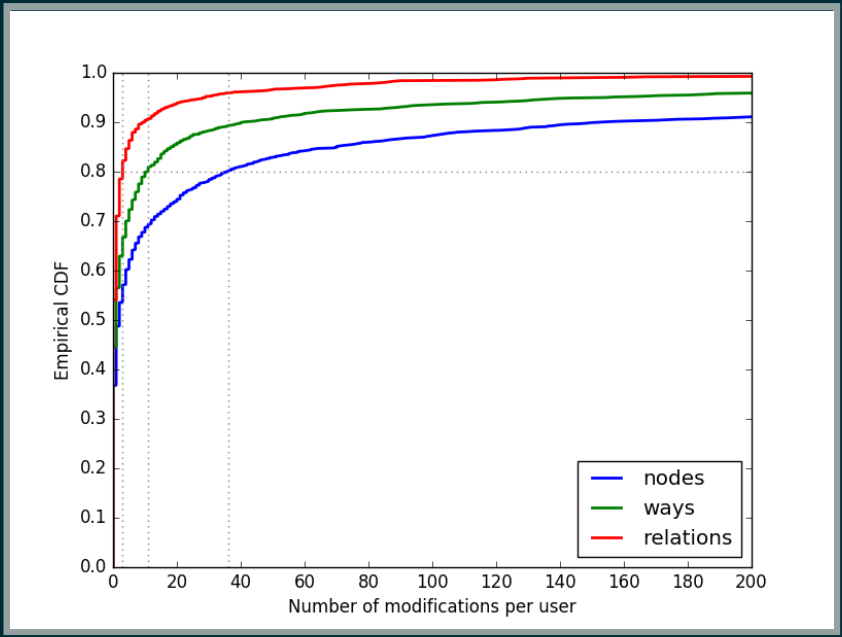
By focusing on OSM users (*ex:* Sylv1, id: 44978):

- temporal features
  - on OSM for 2136 days, 1 day of local activity
- change-set-related features
  - 19 change sets, among which 1 around Bordeaux
- modification-related features
  - 4 modifications, (3 node creations, 1 way improvement), corrected since
- editor-related features
  - 14 change sets done with Potlach, 3 with iD, 2 with JOSM

PCA and clustering on metadata

Metadata normalization

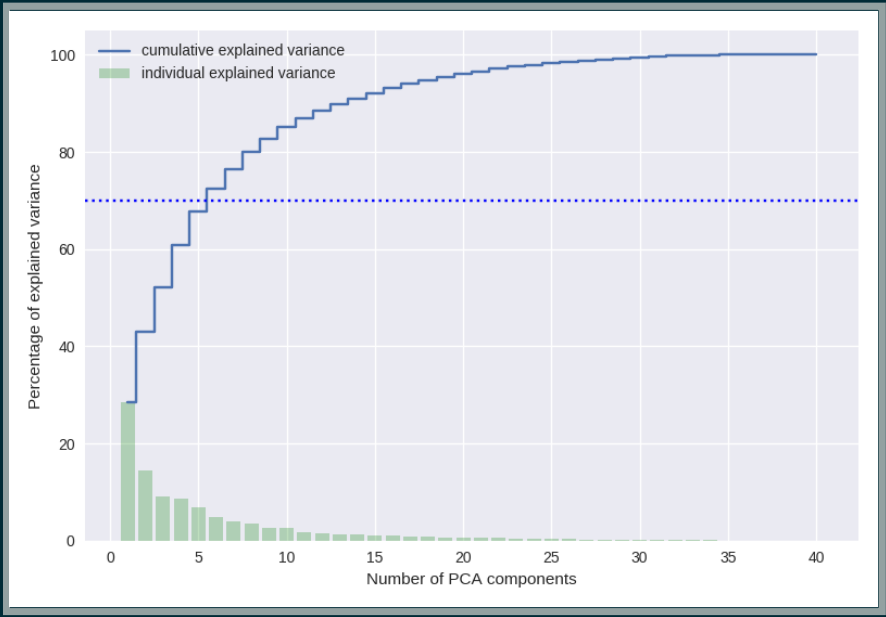
Some highly-skewed features...



- Modify feature definitions (%, empirical CDF)
- Min-Max scaler instead of standard scaler

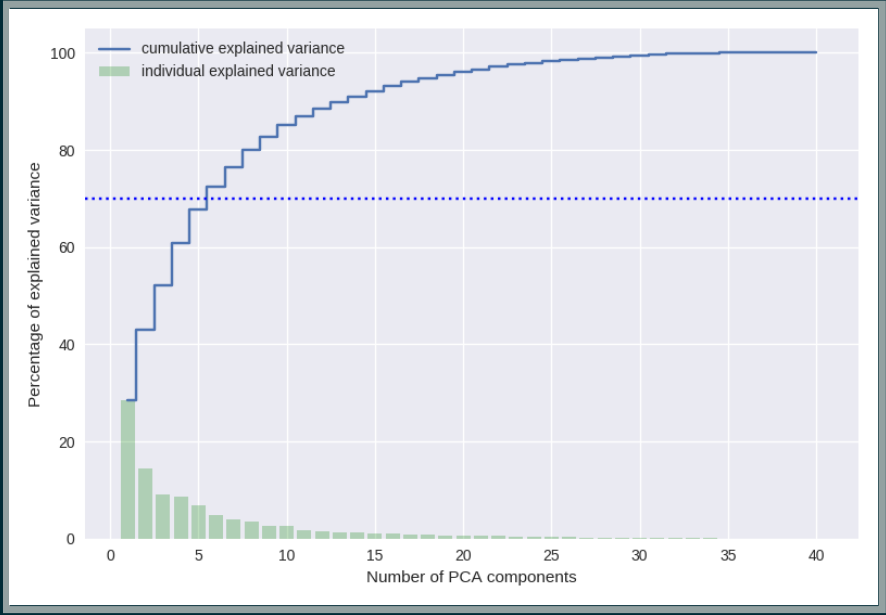
Dimensionality reduction (Principle Component Analysis)

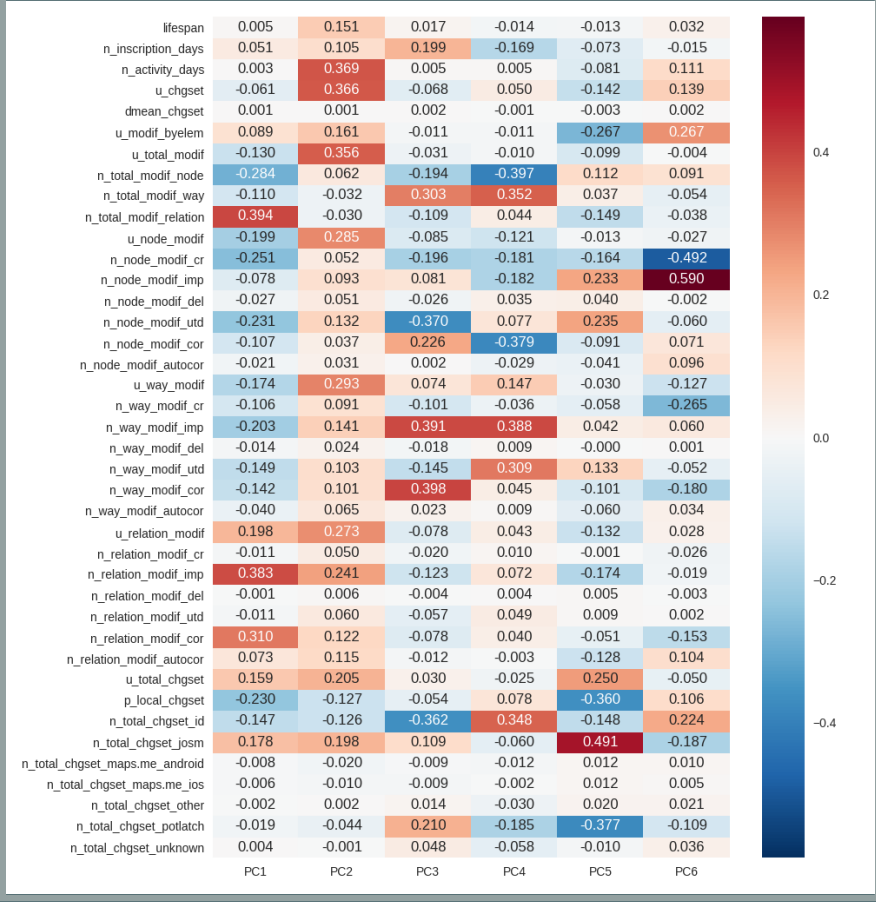
From 40 features to just a few ones: extract the most explainful components



Dimensionality reduction (Principle Component Analysis)

From 40 features to just a few ones: extract the most explainful components





Principle Component Analysis interpretation

C1 (28.5%) Experience on OSM on relation improvement, a few local contributions

C2 (14.5%) Experience on OSM, local and global activity, various contributions

C3 (9.1%) Old out-of-date contribution, way-focused

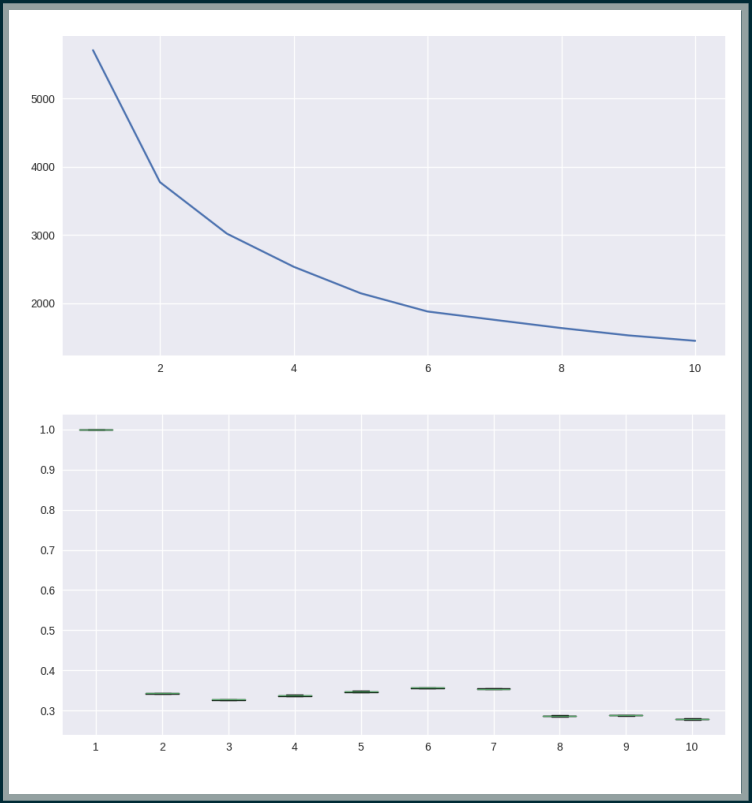
C4 (8.7%) Recent up-to-date contribution, mainly way improvement

C5 (6.9%) Experience on OSM, just a few local node-focused contributions

C6 (4.8%) Local specialization, node improvement (no creation), several contributions on the same OSM objects

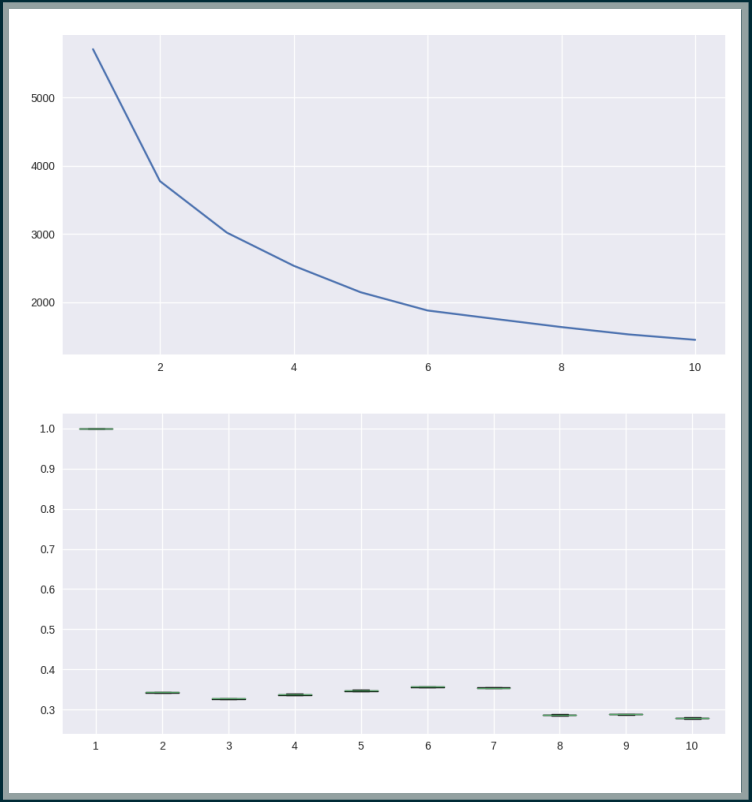
Users classification through KMeans

How many user types?



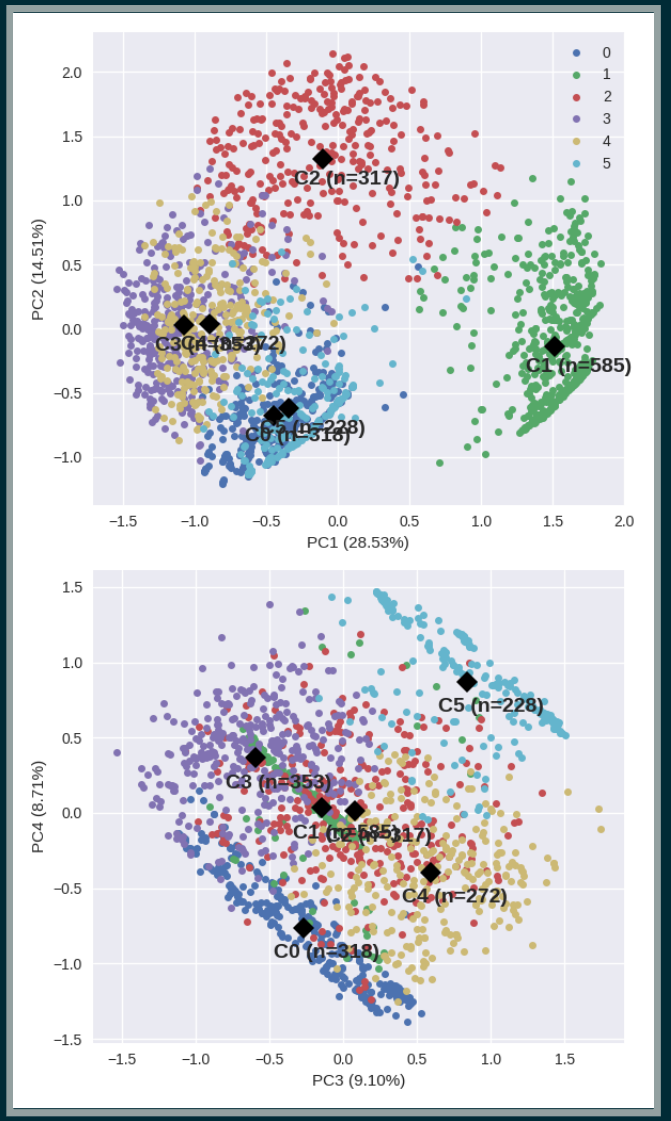
Users classification through KMeans

How many user types?



...6 clusters arise

User classification with KMeans



User classification with KMeans

G0 (n=318) Unexperienced node contributor

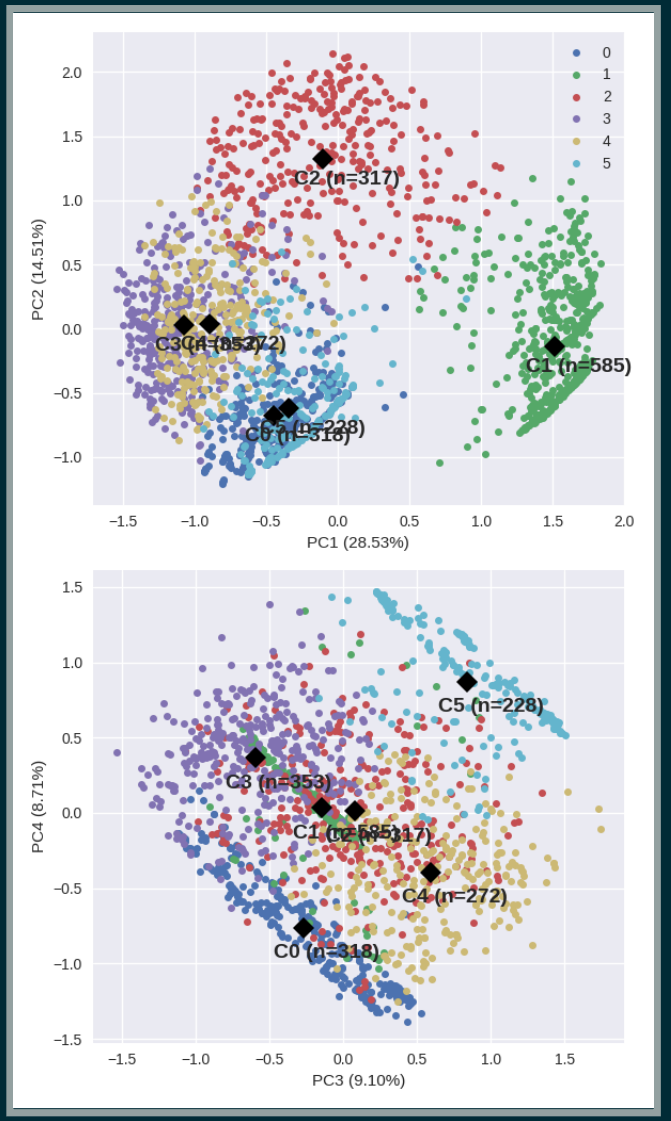
G1 (n=585) OSM experienced users, relation-focused on a unusual area

G2 (n=317) OSM experts, versatile users

G3 (n=353) New local way-focused contributors

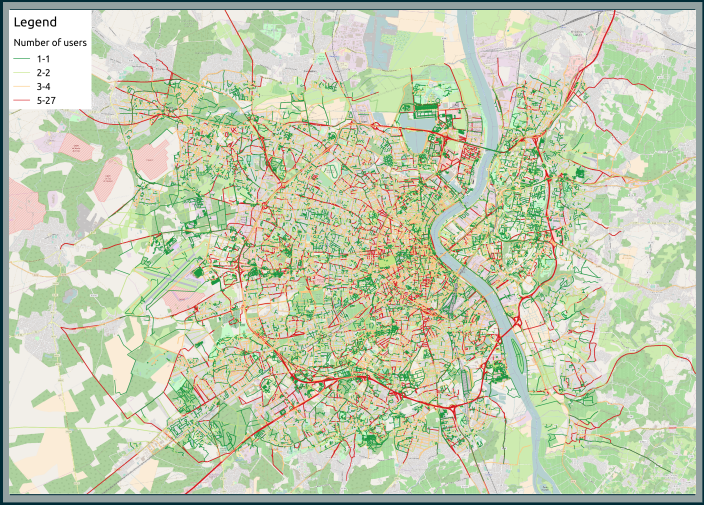
G4 (n=272) Old contributors, one-shotters

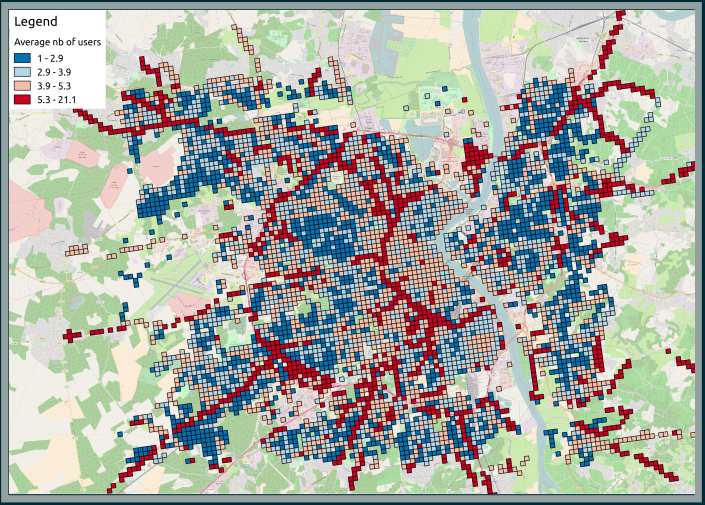
G5 (n=228) Way specialists, but little contributors

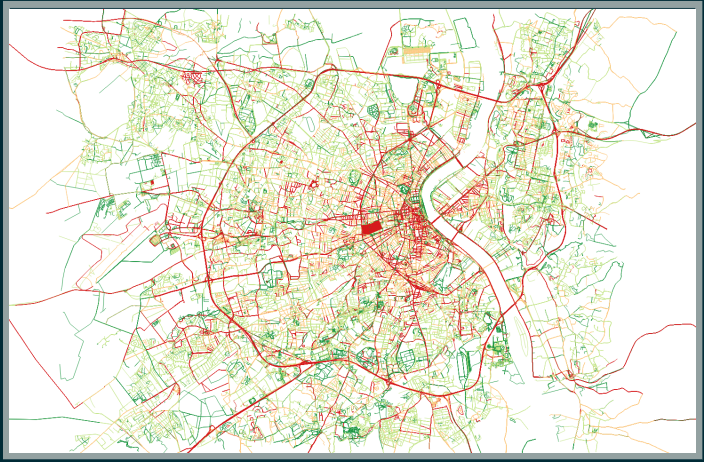


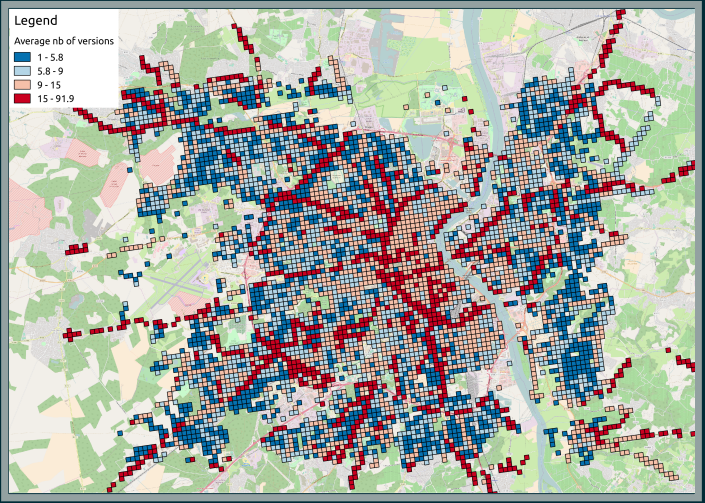
Address the OSM data quality issue

Mapping with metadata features







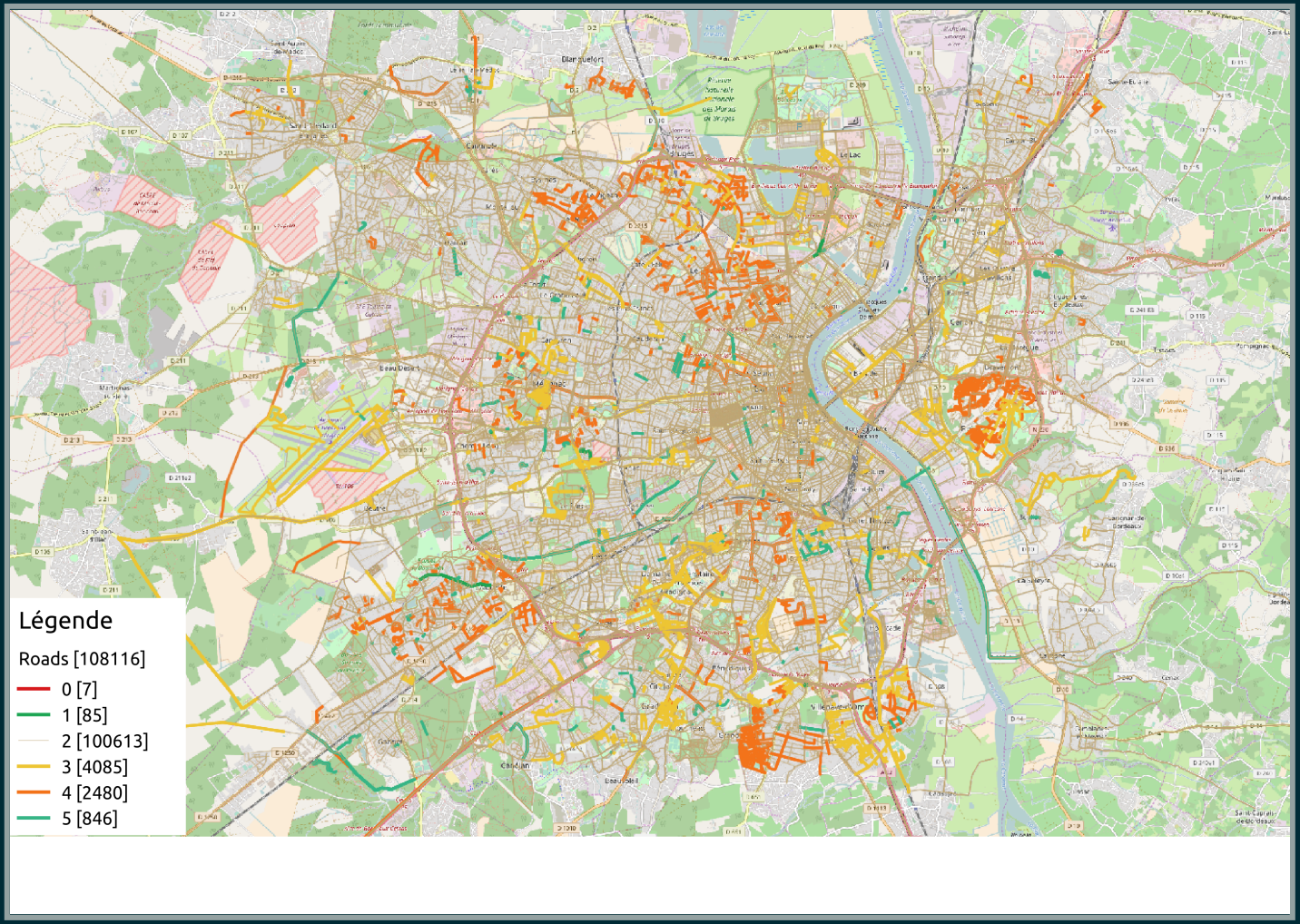


Integration of user classification

Example: OSM roads labelling -> experienced contributors produce good-quality elements

Users	Size	Last-modified objects
G0	318 (15.3%)	6690 (0.2%)
G1	585 (28.2%)	164 (0.0%)
G2	317 (15.3%)	2670391 (96.7%)
G3	353 (17.0%)	23287 (0.8%)
G4	272 (13.1%)	9993 (0.4%)
G5	228 (11.0%)	50474 (1.8%)

Map production



Conclusion

Conclusion about data quality

- Proposition of a methodology in order to characterize OSM metadata, after an unsupervised learning procedure

Conclusion about data quality

- Proposition of a methodology in order to characterize OSM metadata, after an unsupervised learning procedure
- How to characterize OSM data in terms of quality: the example of roads in Bordeaux (France) area

Opening: how to go further?

- Metadata completeness
  - Consider every users, and every contributions for each users (not only local ones)

Opening: how to go further?

- Metadata completeness
  - Consider every users, and every contributions for each users (not only local ones)
- Ground truth comparison
  - Compare the data with a ground truth -> IGN data, Bing orthophotos...?

Thanks for your attention!

Questions?

raphael.delhome@oslandia.com

See more on http://oslandia.com/en/blog/

Fun Facts

- more than 2000 contributors to design a medium-city like Bordeaux (over ~850k active OSM contributors)
- 50% of contributors around Bordeaux have produced less than 2 change sets, but the most "productive" contributor has produced 4428 change sets!
- Around Bordeaux, 73% of nodes, 44% of ways and 10% of relations have only one version!
- JOSM, iD and Potlach are the most used editors (75% of total number of change sets)