ANALYSIS OF MASSIVE IMPORTS OF OPEN DATA IN OPENSTREETMAP DATABASE: A STUDY CASE FOR FRANCE

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

LASTIG, Univ Gustave Eiffel, ENSG, IGN, F-94160 Saint-Mande, France
arnaud.le-guilcher@ign.fr

Commission IV, WG IV/4

KEY WORDS: OpenStreetMap, massive imports, evolution analysis, evolution patterns.

ABSTRACT:

Importing spatial open data in OpenStreetMap (OSM) project, is a practice that has existed from the beginning of the project. The rapid development and multiplication of collaborative mapping tools and open data have led to the growth of the phenomenon of importing massive data into OSM. The goal of this paper is to study the evolution of the massive imports over time. We propose an approach in three steps: classification of the sources used to edit features in the OSM platform including those massively imported, classification of modifications, and identification of evolution patterns. The approach is mixing global analysis (i.e. sources and modifications are classified) and feature based analysis (i.e. imported features are analyzed with respect to their evolution over time). The approach is applied on three datasets coming from OSM considered for their heterogeneity in terms of complexity, imports, and spatial and temporal characteristics. The results show that there is a sustained activity of edition on imported features, with a ratio between geometry editions and semantic editions depending on the type of the features, with roads being the features concentrating the most activity.

1. INTRODUCTION

An important evolution in the production of Geographical Information (GI) has been the rise of GI produced by citizens. Goodchild used the expression volunteered geographic information (VGI) to describe this tendency (Goodchild, 2007). Since then, VGI has been the subject of numerous research studies, relative to both the characteristics of the data and the dynamics of the communities of contributors. In particular, the interactions between VGI and traditional spatial data produced by official agencies have led to studies with the goal to compare it to authoritative spatial data to assess their quality, where much review papers are published (Senaratne et al., 2017; Yan et al., 2020) and update or enrich these authoritative spatial data (Liu et al., 2015; Ivanovic et al., 2020; Olteanu-Raimond et al., 2020).

Among the many forms VGI can take and the many projects based on VGI as mentioned by See et al. (2016), the OpenStreetMap (OSM) project is one of the most prominents. The OSM project was created with the goal to create a world-wide GI database, based on openness and flexibility to encourage wide participation in the project. Over its nearly two decades of existence, it has succeeded in creating an important community of contributors that help editing a world database that is continuously updated and improved, on the one hand and a research community studying different topics such as motivation of contributors, quality and the use of OSM data (Jokar Arsanjani et al., 2015), on the other hand. At a time where updating and correcting authoritative data is challenging for official agencies and there is a need for fresh data to sustain policy makers decisions, the evolution of the OSM project and other VGI initiatives is scrutinized for their potential to produce fresh and accurate data in order to be able to reduce the time between two releases.

At the same time, many countries have taken legal steps to make authoritative GI open and accessible to everyone. As a consequence of these legal steps, the GI contained in many authoritative datasets has been integrated in the OSM database, with OSM contributors adding them either by automated means or manual inputs. In some countries such as France, Spain, US for certain themes, such as buildings [1], [2] or roads (e.g. Tiger), an important proportion of OSM features comes from authoritative data.

Then, a crucial question is whether this evolution in the origin of data generates a change in the dynamics of the contributors' community. Do these massive imports of external data have the potential to affect contributors' implication? Does this addition of mostly accurate and complete databases discourage individual contribution, or does it encourage it? 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? The first two questions need social analysis such as in Bégin et al. (2018), the last two questions need statistical and computer science approaches.

The goal of this paper is to bring new insights on these last two questions by analyzing how features coming from massive external databases evolve in OSM and how OSM contributors remain involved by correcting these features or adding information about them. To do that, we mixed generic and feature based approaches in order to analyze the sources of the imported features and determine evolution patterns over time.

This paper is organized as follows. After the description of the OSM project in Section 2 and related work in Section 3, Section 4 describes our approach and the methods we used to identify the relevant features and analyze their evolution. Section 5 describes the area of study and the datasets we chose

* Corresponding author

[1] https://wiki.openstreetmap.org/wiki/Spanish_Cadastre/Buildings_import

[2] https://wiki.openstreetmap.org/wiki/WikiProject_Cadastre_Fran%C3%A7ais/Import_dans_OSM

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-4-2022-99-2022 | © Author(s) 2022. CC BY 4.0 License.

for our experiment and Section 6 gives the obtained results and their interpretation. Section 7 summarizes our findings and describes paths for further studies.

2. CHARACTERISTICS OF THE OSM PROJECT

OSM data are published according to the Open Database Licence (ODbL) share-alike licence, which authorizes re-use and modification of the database under the condition that the created databases must be shared with the same licence. With this licence, OSM contributors may re-use in their contributions external data published under the ODbL licence, or more permissive open licences. For example, public data published by the French administration mainly use the ODbL licence and the open license such as ODC-BY, CC-By 2.0) which also allows commercial re-use, so these public data may be re-used and modified in the OSM database.

OSM features are characterized by their geometry and a list of tags giving additional information about the feature.

The geometry is described depending on the type of the feature. There are three types of features: 'nodes' (i.e. points defined by their coordinates), 'ways' (i.e. polylines defined as sequences of 'nodes') and 'relations' (i.e. features grouping one or more 'node', 'way' or 'relation'). The polylines can be open polylines, or closed polylines if the last and first nodes are same, and simple polygons and surfaces are represented by closed polylines. Relations are generally used to represent complex or composite objects.

A feature may have any number of tags (some feature types have mandatory tags), and each tag has two elements, a 'key' and a 'value', which are both in free text format; the key generally describes the category of the tag, or the type of feature, while the value gives more specific information related to the key. There is no fixed list of possible keys, but the OSM wiki gives directions on when and how to use each key.

OSM contributors modify the OSM database by adding, suppressing or modifying features, and features modifications can consist in changing the geometry, adding, suppressing tags, or modifying the value associated to a tag key. These modifications are grouped in 'changesets'; the last contains all the modifications made by a given contributor in one session, between the beginning of the editing activity and the end of the editing session when the modifications are applied to the database.

Contributors do not necessarily indicate the source of their editions ; sources can be external data or sources private to the contributor. External data may be external vector databases with an open licence, or photographs or orthophotographs belonging to large photograph databases, or public GSP traces, while private sources are generally photographs or GPS traces.

The OSM community provides guideline as to how to use these sources to edit the OSM database.

3. RELATED WORK

From the beginning of OSM, research has been made to understand the OSM community, the contributors and their dynamics. Studies have looked over contributors' activity to identify contributor characteristics (Neis and Zipf, 2012) and define typical behaviours and interactions (Truong et al., 2019), and have analyzed how socio-economic factors influence contributions to OSM (Neis et al., 2013; Jokar Arsanjani and Bakillah, 2015).

On the more specific subject of massive imports, they are a relatively old phenomenon in OSM; among the first massive imports we can note in 2007 the integration of the TIGER road network in the US [3]. Research has been made to analyze the influence of imported data on OSM database and the results showed weak contributions on the imported data or linked to them (Zielstra et al., 2013). Similar results were found in Grchenig et al. (2014), and even more, the authors mentioned that the imported data had a negative influence on the OSM community. On the contrary, a recent study showed that contributor activity increased after a massive import (Witt et al., 2021). The question of how OSM ecosystem changes and will change, as the community of contributor expands to include industries or professional data producers, and massive external data co-exist with individual contributions in the databases, has been deeply studied in Hayat (2019). There, Hayat identifies two trajectories for the future of OSM. In the first scenario, the OSM project mixes external data and individual contributions and manages to preserve high methodological standards; in the second scenario, the multiple technologies that allow easy imports of external data favor the transformation of OSM into a wide-ranging geographical data aggregator.

To identify the massive imports as well as for some studies on OSM data and contributors, there is the need to identify the sources of features. In the OSM eco-system, communicating the source of a feature can be done in multiple ways, the most straightforward tool being the *source* key. For example, researchers have used the *source* key, in conjunction with other OSM tools, such as comments or the OSM wiki, to identify features imported from external databases, to link and compare features in common with another database (Juhász and Hochmair, 2016), or to infer the Level of Detail of imported features (Touya and Reimer, 2015). However, the last article also points out that the *source* key is seldom filled.

We saw that recent research allows to study the influence of massive imports on OSM community behavior. To the best of the authors' knowledge, few researches focused on the evolution of massive imports over time in term (Witt et al., 2021). Our paper contributes to this gap by proposing an approach to identify types of modifications and evolution patterns. Besides the relevance mentioned by Witt et al. (2021) of measuring the influence of massive imports on both contributors and data, we can add that such work can benefit the institutions which published the data integrated in OSM and enable the enrichment and update of the original data thanks to the modifications made by the OSM community.

4. METHODOLOGY

The proposed approach is composed of three steps: classification of the sources used for OSM data, identification and analysis of massive imports and finally, computation of the evolution pattern for massive imports. Each step is detailed here after.

4.1 Classification of sources for OSM data

The goal of the first step is to identify and classify the sources of information added to the OSM ecosystem. To reach this purpose, we used the key *source* which is used by the OSM community to describe the origins of the data modified by the contributors. There are different ways to fill in the value of the key such as adding it to the *changeset* where a change is made, adding it manually to a single feature or to a single attribute of

[3] https://wiki.openstreetmap.org/wiki/TIGER_fixup

a feature. Multiple values are also possible when the information is coming from different sources (e.g. GPS, orthophotos). Once the values of the key *source* are collected from OSM databases, we are manually homogenize and classify into generic and aggregated categories. The homogenization process is necessary due to the heterogeneous spelling for the name of the same sources of information. For example, spatial data imported from the french General Direction of Public Finances have different values for the *source* key : CADASTRE, cadastre, Direction Gnrale des Finances Publiques. Following the metadata regarding the source values from the OSM wiki [4] as well as the work proposed by Mooney et al. (2016) we defined a typology of sources (for the geometry of features) composed by five categories:

1. Massive import: this category describes features coming from external spatial open data existing at a national scale (e.g. General Direction of Public Finances, France electricity power line) or a local scale (e.g. Paris City Hall data, Toulouse City Hall data, Metro data, etc.) and can be produced by national and local governments, third-party communities, NGO, citizen science communities;

2. Photos analysis: this category contains features edited by OSM contributors by using geolocalised photos coming from different sources such as Yahoo [5], Mapillary [6];

3. Vectorization: this category considers features edited by OSM contributors by using maps, aerial or satellite imagery;

4. Satellite navigation receiver: features being collected by using equipment such as GNSS devices, smartphones with GPS include;

5. No source: features having the key *source* not filled in.

According to the value of the *source* key, we assigned a class category to each feature of our study area. Note that for multiple values in the key *source*, different categories are assigned.

4.2 Analysis of massive imports

The goal of this second step is to define a typology of types of modifications for massive imports. For this purpose, only the OSM features belonging to the category *massive import* defined in the previous section are considered further on. For these specific features the OSM history of edition is taken into account.

Let us consider a feature at the time t_0, which we note feature(t_0), corresponding to the moment of its integration in the OSM database. This feature can be modified at any time by an OSM contributor and produce a set of states feature(t_1), ..., feature(t_i),..., feature(t_m), with $i = 1, \dots, m$, where m represents the total number of editions. The natural elementary unit to study contributions to the OSM database is the changeset, a unit regrouping all the modifications done by a single contributor between the time they open their editing session and the time they close it and "push" their modifications in the OSM database. The change from feature(t_i) to feature(t_{i+1}) corresponds to a single changeset in which the feature under study is modified. In one changeset, a given OSM feature can undergo different types of modifications.

We defined a typology of modifications composed by five categories:

1. Geometry modification: this category corresponds to the modification of the geometry of the feature which can be a displacement or a change of at least one node, way, or relation;

2. Tag enrichment: it corresponds to the addition of one or several new tags giving supplementary information about the feature, such as opening hours, phone number, the source of the feature (for these examples, the tag keys would be 'opening_hours ', 'phone', and 'source' respectively);

3. Tag suppression: it refers to the removal of one or more tags between two versions;

4. Tag modification: it corresponds to a change regarding the spelling of a tag value. For example the tag "name"="national police station", the replacement of the lower cases with the upper cases (e.g. name=National Police Station), constitutes a tag modification.

5. Geometry suppression; its refers to a deleted feature;

This typology does not take into account the magnitude of the modifications. In particular, a big geometry change and a small geometry correction are treated in the same way, as are a spelling correction on a tag value and a complete modification of this value. As a result, this typology does not measure how much a feature changes, but rather the intensity of the contributor activity on this feature. Intensity is a trigger to measure the motivation of contributors to modify imported datasets, whereas the magnitude gives information about the quality of the imported dataset. The quality assessment is out of the purpose of this paper.
At the end of this step, each feature(t_i) is characterized by one or several categories of modification among those listed before, representing the modifications the feature underwent between its state feature(t_{i-1}) and its state feature(t_i).

4.3 Evolution pattern for massive imports

The final step of our proposed approach is the definition of evolution patterns. To analyse the evolution of the massive imports, the sequencing method is chosen. This method is inspired by that of 'coding theory and string editing' (Lesnard, 2006). It is a qualitative method for 'empirically applying a theoretical discourse on change' and was successfully used in different applications such as human mobility analysis (Viry and GAUTHIER, 2019) or career patterns (Dlouhy and Biemann, 2015). The method allows temporality to be integrated into the analysis of events. In our case, it consists in defining for each imported feature its sequence of modifications. To do this, we first proceed by encoding the different categories of our typology modifications with numbers from 1 to 4. Thus, the geometry modification category is assigned value 1, the tag enrichment category value 2, while the tag deletion and modification categories are assigned values 3 and 4 respectively. As we worked only on currently existing features, there is no geometry suppression in their sequences of modifications.

Second, a sequence is defined for each imported feature. For each changeset changing feature(t_{i-1}) into feature(t_i), the list c_i lists the codes of the different types of modifications underwent by the feature, in increasing order. For example, if the

[4] https://wiki.openstreetmap.org/wiki/Key:source
[5] https://images.search.yahoo.com/
[6] https://www.mapillary.com

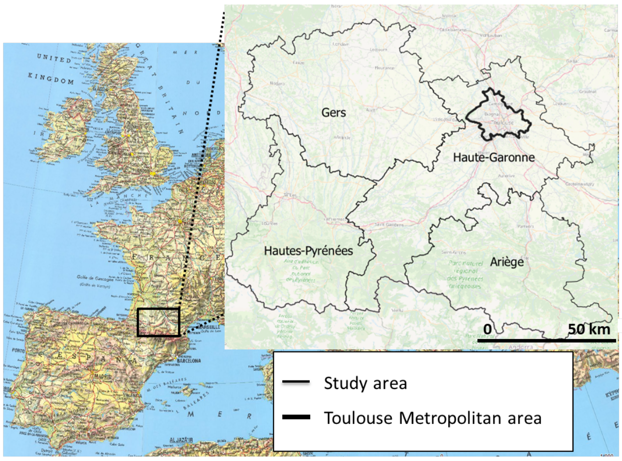
changeset modified the geometry of the feature, added a tag and modified the value of another tag, but made no tag suppression, the resulting sequence is $c_i = 1, 2, 4$. The complete sequence for a feature is the concatenation $c_1, \ldots, c_m$. We note that he numerical codes for each type of modification is significant and has an impact on the resulting sequences. Therefore, while the choice of the numerical codes is in some way arbitrary, we chose to separate geometry modifications (with the code 1) from semantic modifications (with the codes 2, 3 and 4). At the end of this step, a sequence is defined for each imported feature.

Third, a hierarchical classification of features is made based on their sequences. This step involves computing distances between the sequences. The method we used to compare sequences is the Optimal Matching Analysis (OMA) (Abbott and Hrycak, 1990) being considered the most suitable method for identifying regularities related to transitions between sequencing states (Aisenbrey and Fasang, 2010; Robette, 2011). It provides a similarity metric between sequences by counting the minimum number of modifications to be made between two sequences in order to obtain identical sequences. After computing metric distances for each pair of sequences stored in a triangular matrix, a hierarchical clustering algorithm is applied to the matrix using Wards minimum variance, following Dlouhy and Biemann (2015) recommendations. At the end of this step, the imported features are assigned clusters, where features in one cluster are assumed to follow the same evolution pattern.

5. STUDY SITE AND DATASETS

The study site is located in the Occitanie region [7] from south of France and corresponds to four french departments: Ariège, Gers, Hautes-Pyrénées and Haute Garonne (see Figure 1). The extent of the area is $24,936km^2$ and covers various landscapes such as urban, peri-urban and rural areas as well as mountainous areas. It is also characterized by an actively expanding urban area, the Toulouse Metropolitan. Moreover, from the analysis of the OSM forum and wiki, it looks that the local community is very active ; the osm wiki pages of the Gers [8] and Hautes-Pyrénées [9] departments and of Toulouse [10] explicitly mention the integration of public open data as a goal. We thus consider that these elements are appropriate to implement our approach on the described study site.

Regarding the type of data and based on our knowledge of the types of dumps massively imported in OSM we have chosen three themes corresponding to three types of vector data and downloaded the corresponding data in May 2021. The first theme is Police Station and it is part of the Points of Interest theme (POI) represented by points (186 features). According to OSM metadata description a police station "is a place that serves as a primary point of contact for the public, i.e. public-facing police facilities" and is recognized by "amenity=police". The tags used to describe a police station are : the address of this police station, opening hours, website of the police station, phone number, the e-mail address, the operator, information about the accessibility for wheelchairs. More information about the police station theme can be found on the OSM wiki [11] This dataset is noted "POI" thereafter.



Sources: Géoportail, BDTOPO

Figure 1. Study site description.

The second theme is road network data represented by ways and relations (248,990 features). These theme represents the transport infrastructure such as roads, routes, ways, etc. In OSM all these types of real entities are tagged with the tag key *highway*. Many semantic and thematic pieces of information are tagged such as the classification of the roads, names, speed limits, width, etc. More information about the road networks can be found on the OSM wiki [12]. This dataset is noted "roads" thereafter.

Finally, the third dataset is building data represented by nodes, lines and relations (400,063 features). This theme contains permanent constructions as well as mobile home, houseboats, etc. As for road networks, many semantic and thematic pieces of information are tagged such as the usage, the number of floors, entrance, house number etc. More information about the buildings in OSM can be found on the OSM wiki [13]. For the building dataset, only features located in Toulouse Metropolitan area (bold black limits illustrated in Figure 1) are used. This dataset will be noted "buildings".

These three datasets were chosen as they are examples of three important types of features in the OSM database : points of interests, which are represented as nodes, open polylines and closed polylines which are both represented as ways (actually, there can exist a very small number of nodes or relations among the features with the tag keys *highway* or *building* but they are rare). Among open polylines and closed polylines respectively, roads and buildings are arguably the types of features that are the most emblematic of the widespread development of the OSM project. We also wanted to have a smaller dataset to expose our concepts on fewer features, and police stations were chosen as *project of the month* in the French OSM community in November 2018. The goal of this initiative is to encourage the integration of a given category of features from an external database into the OSM database, while taking properly into account the features already present in OSM, and the coherence of the new features with the neighbouring features.

[7] https://en.wikipedia.org/wiki/Occitania_(administrative_region)

[8] https://wiki.openstreetmap.org/wiki/Gers

[9] https://wiki.openstreetmap.org/wiki/Hautes-Pyrénées

[10] https://wiki.openstreetmap.org/wiki/FR:Toulouse

[11] https://wiki.openstreetmap.org/wiki/Tag:amenity%3Dpolice

[12] https://wiki.openstreetmap.org/wiki/Highways

[13] https://wiki.openstreetmap.org/wiki/Buildings

6. RESULTS AND DISCUSSION

6.1 Analysis of sources in the datasets

The analysis of the homogenized source names shows the relevant number of sources of information imported in OSM for the three themes. The ”POI” and ”roads” datasets contain twenty-two sources of information (National Mapping agency, National Gendermerie, Services-Public.fr, European Union, fire brigade, General Direction of Public Finances) and the ”buildings” dataset twenty sources of information (e.g. Toulouse Metropolitan, Health Ministry, Education Ministry).

Our three datasets have very different proportions of source classes. In the ”POI” dataset, a majority of features (77%) are in the class ”Massive import”. An important proportion of objects (20.5%) has no source given. In the ”roads” dataset, a majority of the features (75.5%) has no source, and the remaining objects are mostly split between the class ”Massive import” (16%) and the class ”Vectorization” (7.5%). Let us mention that for the features without source information, a temporal analysis showed that some of them correspond to different changesets containing high numbers of features with no source information pushed in the same day, while others correspond to changesets made on days when few features without source information were created, as seen in Figure 3. For features with no source information that were created on days with a high number of such creations, it is possible that such a high number of features come from massive imports that were not properly sourced. For features created on days with less edits without source, it is not unreasonable to make the hypothesis that a large number of these features come from individual contributions (using maps, photographs or GNSS traces to trace roads).

For the ”buildings” dataset, a very large majority of the objects (97.5%) are in the class ”Massive import”, while 2% of the features have no source. Figure 2 shows how these different types of sources are spatially distributed on our three datasets.

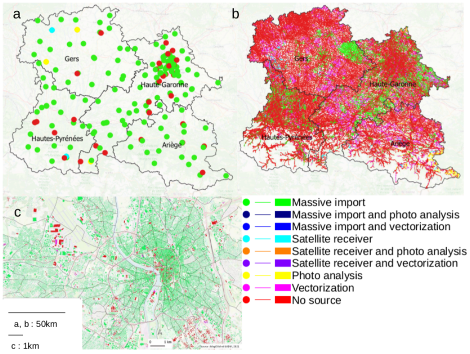

a
b
c
Gers
Haute-Garonne
Hautes-Pyrénées
Ariège
● 🟩 Massive import
● 🟦 Massive import and photo analysis
● 🟦 Massive import and vectorization
● 🟦 Satellite receiver
● 🟧 Satellite receiver and photo analysis
● 🟪 Satellite receiver and vectorization
● 🟨 Photo analysis
● 🟪 Vectorization
● 🟥 No source
a, b : 50km
c : 1km

Figure 2. Spatial repartition of the different types of sources for the ”POI”, ”roads”, and ”buildings” datasets

6.2 Analysis of unit modifications after massive imports

For all three datasets, we analyse the modifications that features coming from massive imports underwent, according to the typology defined in subsection 4.2. The results are given in the table below:

For the ”POI” dataset, only 2% of the objects have had no modification after the import. Among features that have been modified, the average number of modifications per feature is 4.0 and the most edited feature has been modified 8 times. Among these modifications, geometry modifications have been rare (2.5%), and most of the activity concentrated on semantic modifications : tag enrichment (79.6%), tag modification (16.9%), and tag suppression (1.0%). The rarity of geometry modifications seems reasonable knowing the crisp definition of POI location for buildings.

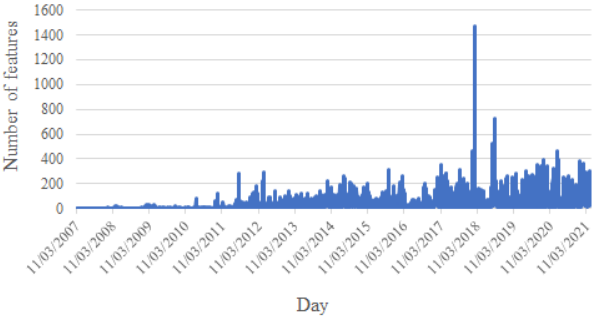


Figure 3. Repartition of the dates of creation of features with no source in our ”roads” dataset.

	”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

Table 1. Number of modifications for each type in the ”POI”, ”roads” and ”buildings” datasets

For the ”roads” dataset, 32% of the features have not been modified after the import. Among modified features, the average number of modifications per feature is 3,2. The proportion of geometry modifications in the ”roads” dataset is higher than for the ”POI” dataset being equals to (49.8%). The remaining modifications are divided between tag enrichments (29.7%), tag modifications (16.9%) and a few tag suppressions (1.6%).

For the ”buildings” dataset, 49% of the features have been modified. Among modified features, the average number of modifications per feature is 1.8. This low number of modifications per feature is explained by the fact that buildings undergo less semantic modifications that points of interest or roads. Indeed, 71.9% of the modifications are geometry modifications. Semantic modifications (tag enrichments (15.2% of the modifications), tag modifications (8.4%), tag suppressions (4.5%)) are fewer, with approximately 0.5 semantic modification per modified object.

6.3 Analysis of editing sequences after a massive import.

The analysis of editing sequences relies on the construction of clusters based on the editing history for each feature. For each feature, the computation of the editing sequence is done in PostgreSQL and then R. In PostgreSQL, our algorithm is applied to the table containing all the versions of the features (with one line per version) in the dataset, identifies the modifications underwent by the feature, and creates a new table, with one line per feature, containing, for each line, the ordered list of modifications for the feature, using the numerical code described in subsection 4.3. In R, this table is used to build a function seqdef() giving the editing history of each feature as a list of integers in $\{1, 2, 3, 4\}$.

The remainder of the clustering analysis is done in R, using the library TraMineR [14], implemented to compare sequences. The function seqdist() allow an ascending hierarchical classification of the features based on their editing histories. The method used to define the distance between two editing histories is a parameter of this function. We used the standard method OM (Optimal Matching), in accordance with our choices in subsection 4.3. As this accepts addition and suppression as elementary modifications, this definition is well-suited to compare sequences of different lengths, which is our case. The ascending hierarchical classification method produces a dendrogram, i.e. a tree where each feature is a leaf, and where leaves are regrouped in branches based on Ward distance. To determine effective clusters from this hierarchical classification, one must choose where the dendrogram must be cut ; to do this, we used a visual criterion to have both a reasonable number of classes and a low cutting height.

The clustering computation was applied for our three datasets. Here, we give a detailed analysis for the ”buildings” dataset. A first clustering for our initial typology with four different modifications are carried out. The clusters are represented on the carpet of sequences in Figure 4.

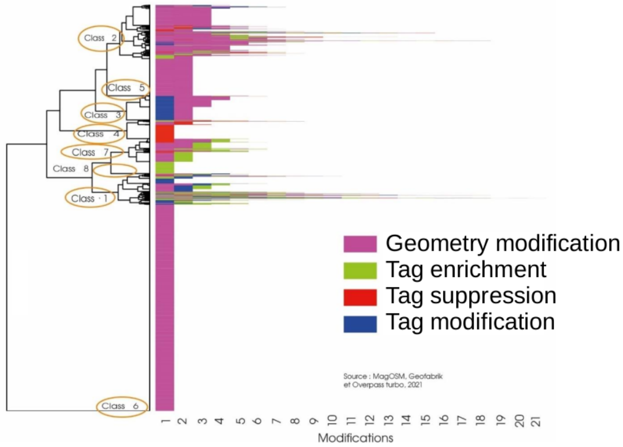


Figure 4. Carpet of sequences for the ”buildings dataset”, with a typology of four modifications.

The clustering method was used on 8,464 features. The biggest class is class 6, with 4,303 elements (50.8%). It contains features that have underwent exactly one geometry modification and no semantic modification after the import. The second biggest class is class 2 (1,119 features, 13.2%), which regroups heterogeneous features that have had a moderate or high number of modifications, a high proportion of these modifications bearing on their geometry. The other classes are relatively smaller, between 249 and 776 elements. Class 5 (776 features, 9.2%) contains features with 2 geometry modifications and no semantic modification. Class 1 (644 features, 7.6%) regroups features with varying numbers of modifications (between 1 and 18 modifications) and a high proportion of semantic modifications. Class 3 (505 features, 6.0%) regroups features for which the first modification was a modification of a tag value, followed by one or a few geometry modifications. Class 7 (478 features, 4.6%) regroups features with 2 and 6 modifications, with a tag enrichment late in the sequence. Class 4 (390 features, 4.6%) contains features that underwent a tag suppression, with a majority of them having had no other edition. Class 8 (249 features, 2.9%) contains features with a unique tag enrichment.

While this method gives an interpretable set of clusters, we found that this detailed classification choices tend to blur the distinction between geometric modifications and semantic modifications, by adopting a model that considers that a tag enrichment and a tag modification are as dissimilar as they are both to a geometry modification. To give more weight to the distinction between geometry and semantic, a second clustering with a simplified typology with only two types (geometry modification and semantic modification) is applied. The resulting clusters and carpet of sequences is given in Figure 5.

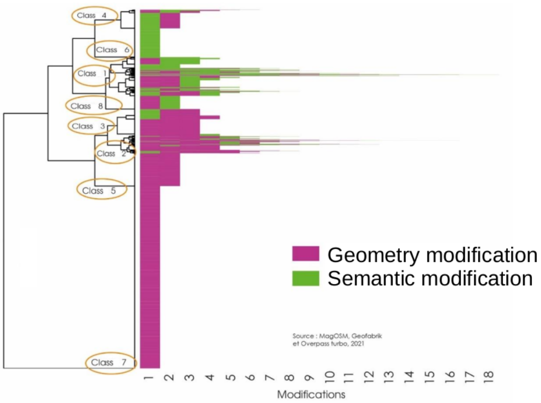


Figure 5. Carpet of sequences for the ”buildings” dataset, with a typology of two modifications.

This clustering is more straightforwardly interpretable. It roughly separates features with one modification (classes 6 and 7), two modifications (classes 5 and 8, and a large majority of class 4), and more modifications (class 3 contains features with 3 or 4 modifications, while class 1 and 2 are more heterogeneous). It also separates features with a large majority of geometry modifications (classes 2, 3, 5 and 7) and features for which semantic modifications are more present (classes 1, 4, 6 and 8).

Here after, we give a short summary of the results of the clustering for the ”POI” and ”roads” datasets.

For the ”POI” dataset, the total number of features that evolved after their integration is equals to 50. As geometry modifications are very rare, we consider the second typology (with only 2 different types) to highlight their presence in sequences. We obtained 7 clusters. The table below gives the number of features in each cluster (column # f), the percentage of features (column %), and prominent characteristics of the cluster, the typical number of modifications (column # m) and the types of modifications (column type). As the type ”semantic modification” makes for a large majority of modifications, we concentrate on the presence of geometry modifications in the sequences.

# f	%	# m	geometry modifications
2	4 %	1	no
2	4 %	2	no
12	24 %	3	no
13	26 %	4	no
12	24 %	5 to 6	in some sequences
4	8 %	6 to 7	in some sequences
5	10 %	4 to 6	in all sequences

Table 2. Sizes and characteristics of clusters for the ”POI” dataset

For the ”roads” dataset, we used the more detailed typology, as the ratio between semantic and geometry modifications is

[14] https://larmarange.github.io/analyse-R/analyse-de-sequences.html

more balanced and it can be interesting to keep the distinction between tag enrichment, modification and suppression. The clustering was applied on 25,289 features and nine clusters have been obtained. The percentage of features for each class is depicted in the table below.

class	# f	%
class 1	2,723	10.8 %
class 2	3,508	13.9 %
class 3	1,665	6.6 %
class 4	3,522	13.9 %
class 5	3,587	14.2 %
class 6	1,743	6.9 %
class 7	3,574	14.1 %
class 8	1,968	7.8 %
class 9	2,999	11.9 %

Table 3. Sizes of clusters for the ”roads” dataset

Class 1 regroups features with a moderate to high number of modifications, with many geometry changes. Class 2 is somewhat heterogeneous class, with a higher proportion than average proportion of tag modifications. For class 3, the features have a low to moderate number of modifications, tag suppression being more frequent than the average. In class 4, tag enrichments are more frequent. Features in class 5 have much to moderate amount of modifications, with a majority of geometry changes. Class 6 contain features having a high proportion of tag modifications and geometry changes. Class 7 mostly contains features with a low amount of changes and a prominent presence of tag enrichments. Finally classes 8, and 9 regroup respectively features having exactly one tag modification, and exactly one geometry change.

Counting the different types of modifications and computing clusters of features for each dataset show that contributors do not stop modifying imported features, but the intensity of contributor implication depends on the dataset considered. For the ”POI” dataset, almost all the imported feature have been edited after the import, with a mean of 4 modifications per feature, and a large majority of tag enrichment. This analysis shows that OSM contributors make efforts to add information on imported features in this dataset and that the initiatives of the OSM community to keep contributors involved in the improvement of the data. For the ”roads” dataset, two thirds of the imported features have underwent modifications, with a lot of diversity in the number of modifications, types of modifications, and trajectory of the feature after the import. This also shows that OSM contributors remain involved in the improvement of road features after the import, be it to improve their geometry or to enrich or correct its semantic information. The conclusion is more nuanced for the ”buildings” dataset, where one half of the features are not modified after the import, and the other generally have few modifications. Geometry modifications constitute a large majority. Features with a significant proportion of semantic modifications are a minority, and most of them have a small number of modifications (less than 2). For this dataset, contributor involvement may not be sufficient to correct and update imported buildings. The disparity between roads and buildings may be explained by the long-lasting interest of the OSM community for the road network, and the fact that the very high number of building features makes a systematic completion of semantic information difficult.

7. CONCLUSIONS

This study presents an approach to analyse the evolution of massive imports in OpenStreetMap. To reach this goal we first proposed a classification of existing sources in OSM based on the *source* tag, a classification of modifications applied on imported features over time and a sequencing and hierarchical classification based method to define evolution patterns over time.

The first results show that contributors remain involved in the evolution of massive imports, features are still edited after a massive import. The editing activity on imported police stations is mostly semantic modifications, while the activity on imported buildings mostly affects their geometry. On imported roads, the activity is balanced between semantic and geometric modifications, and road is the theme for which the average number of contributions after the import is the highest. Activity is less sustained on imported buildings. In terms of the scenarios described by Hayat (2019), these results tend to indicate that contributor involvement is sufficient to maintain a high methodological standard by mixing external imports and individual contributions for the police station and road themes, while being less conclusive for the building theme, were the quantity of features to edit and update is even more important.

Our typology of modifications only captures contributor activity on an imported feature, but it can not be used to evaluate how much the current version of a feature differs from its version at the time of the import. To do so, a measure for the magnitude of each change would be needed, and the typology would have to take it into account. More research is needed to determine if the quality of imported data is improved, as a consequence of the actions of the OSM community after the massive import. For an integration of OSM data in the original datasets different issues need to be managed such as license, data model, specification rules.

Concerning the limits of our study, we identified the following elements. First, the conclusions of our study regarding the evolution patterns are strictly limited to the study area and to the feature types we selected. At this point, it is difficult to say if the same conclusions can be raised on other study (different areas or themes). We intend to apply the proposed approach to other study areas on both rural and urban areas as well as on other themes such as land cover data, public transportation data. The authors also belive that the proposed approach can be easily applied to other countries.

For the evolution analysis, only features having the source tag filled in were used. Nevertheless, in our datasets, we noticed that the source tag is not always filled in. For the ”POI” dataset, and even more for the ”buildings” dataset, the ”massive import” class contains most of the features coming from massive imports ; the number of missing features is probably very small. Thus, we consider that the conclusions for the study area are stable for these two themes. This could possibly be different for the ”roads” dataset, where much features (77%) have no given source and relevant features could be missed by considering only features in the ”massive import” class. In the future, more research will be made to analyse features without source information in order to precisely identified if massive imports are part of those category.

Finally, another future work we identified is to extend this work to update and enrich the original data sets by using the last version of the imported features. This requires the definition of a data matching process between the original data set and the modified imported features and processes to transfer the complementary information from one source to another (i.e. OSM to external source).

References

Abbott, A., Hrycak, A., 1990. Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *American journal of sociology*, 96(1), 144–185.

Aisenbrey, S., Fasang, A. E., 2010. New Life for Old Ideas: The ”Second Wave” of Sequence Analysis Bringing the ”Course” Back Into the Life Course. *Sociological Methods & Research*, 38(3), 420-462. https://doi.org/10.1177/0049124109357532.

Bégin, D., Devillers, R., Roche, S., 2018. The life cycle of contributors in collaborative online communities -the case of OpenStreetMap. *International Journal of Geographical Information Science*, 32(8), 1611-1630. https://doi.org/10.1080/13658816.2018.1458312.

Dlouhy, K., Biemann, T., 2015. Optimal matching analysis in career research: A review and some best-practice recommendations. *Journal of Vocational Behavior*, 90, 163-173.

Goodchild, M., 2007. Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69, 211-221.

Grchenig, S., Brunauer, R., Rehrl, K., 2014. Digging into the history of VGI data-sets: results from a worldwide study on OpenStreetMap mapping activity. *Journal of Location Based Services*, 8(3), 198-210. https://doi.org/10.1080/17489725.2014.978403.

Hayat, F., 2019. Production des biens communs numériques et usages cartographiques. Theses, Université de Paris.

Ivanovic, S. S., Olteanu-Raimond, A.-M., Mustière, S., Devogele, T., 2020. Potential of crowdsourced traces for detecting updates in authoritative geographic data. P. Kyriakidis, D. Hadjimitsis, D. Skarlatos, A. Mansourian (eds), *Geospatial Technologies for Local and Regional Development*, Springer International Publishing, Cham, 205–221.

Jokar Arsanjani, J., Bakillah, M., 2015. Understanding the potential relationship between the socio-economic variables and contributions to OpenStreetMap. *International Journal of Digital Earth*, 8(11), 861–876.

Jokar Arsanjani, J., Zipf, A., Mooney, P., Helbich, M., 2015. An introduction to openstreetmap in geographic information science: Experiences, research, and applications. *OpenStreetMap in GIScience*, Springer, 1–15.

Juhász, L., Hochmair, H. H., 2016. Cross-linkage between Mapillary street level photos and OSM edits. *Geospatial Data in a Changing World*, Springer, 141–156.

Lesnard, L., 2006. Optimal Matching and Social Sciences. working paper or preprint.

Liu, C., Xiong, L., Hu, X., Shan, J., 2015. A Progressive Buffering Method for Road Map Update Using OpenStreetMap Data. *ISPRS International Journal of Geo-Information*, 4(3), 1246–1264. https://www.mdpi.com/2220-9964/4/3/1246.

Mooney, P., Minghini, M., Laakso, M., Antoniou, V., Olteanu-Raimond, A.-M., Skopeliti, A., 2016. Towards a Protocol for the Collection of VGI Vector Data. *ISPRS International Journal of Geo-Information*, 5(11). https://www.mdpi.com/2220-9964/5/11/217.

Neis, P., Zielstra, D., Zipf, A., 2013. Comparison of volunteered geographic information data contributions and community development for selected world regions. *Future internet*, 5(2), 282–300.

Neis, P., Zipf, A., 2012. Analyzing the contributor activity of a volunteered geographic information project - The case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1(2), 146–165.

Olteanu-Raimond, A.-M., See, L., Schultz, M., Foody, G., Riffler, M., Gasber, T., Jolivet, L., le Bris, A., Meneroux, Y., Liu, L., Poupe, M., Gombert, M., 2020. Use of Automated Change Detection and VGI Sources for Identifying and Validating Urban Land Use Change. *Remote Sensing*, 12(7). https://www.mdpi.com/2072-4292/12/7/1186.

Robette, N., 2011. *Explorer et décrire les parcours de vie: les typologies de trajectoires*. CEPED.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.-Y., Milinski, G., Niki, M., Painho, M., Pdr, A., Olteanu-Raimond, A.-M., Rutzinger, M., 2016. Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. *ISPRS International Journal of Geo-Information*, 5(5). https://www.mdpi.com/2220-9964/5/5/55.

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., Haklay, M., 2017. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31(1), 139–167.

Touya, G., Reimer, A., 2015. Inferring the scale of OpenStreetMap features. *OpenStreetMap in GIScience*, Springer, 81–99.

Truong, Q. T., De Runz, C., Touya, G., 2019. Analysis of collaboration networks in OpenStreetMap through weighted social multigraph mining. *International Journal of Geographical Information Science*, 33(8), 1651–1682.

Viry, G., GAUTHIER, J.-A., 2019. L'analyse de séquence pour étudier les comportements de mobilité spatiale dans le parcours de vie. *RTS - Recherche Transports Sécurité*, 2019, 18p. https://hal.archives-ouvertes.fr/hal-02146243.

Witt, R., Loos, L., Zipf, A., 2021. Analysing the Impact of Large Data Imports in OpenStreetMap. *ISPRS International Journal of Geo-Information*, 10(8). https://www.mdpi.com/2220-9964/10/8/528.

Yan, Y., Feng, C.-C., Huang, W., Fan, H., Wang, Y.-C., Zipf, A., 2020. Volunteered geographic information research in the first decade: a narrative review of selected journal articles in GIScience. *International Journal of Geographical Information Science*, 34(9), 1765-1791. https://doi.org/10.1080/13658816.2020.1730848.

Zielstra, D., Hochmair, H. H., Neis, P., 2013. Assessing the Effect of Data Imports on the Completeness of OpenStreetMap–A United States Case Study. *Transactions in GIS*, 17(3), 315–334.