Analysis of cycling network evolution in OpenStreetMap through a data quality prism

Raphaël Bres, Veronika Peralta, Arnaud Le Guilcher, Thomas Devogele, Ana-Maria Olteanu-Raimond, Cyril de Runz

AGILE: GIScience Series, 4, 3, 2023. https://doi.org/10.5194/agile-giss-4-3-2023
Proceedings of the 26th AGILE Conference on Geographic Information Science, 2023.
Editors: P. van Oosterom, H. Ploeger, A. Mansourian, S. Scheider, R. Lemmens, and B. van Loenen.
This contribution underwent peer review based on a full paper submission.

Analysis of cycling network evolution in OpenStreetMap through a data quality prism

Raphaël Bres [1,2], Veronika Peralta [1], Arnaud Le-Guilcher [2], Thomas Devogele [1], Ana-Maria Olteanu Raimond [2], and Cyril de Runz [1]

[1]Univ de Tours, LIFAT, BDTLN, Blois, France
[2]Univ Gustave Eiffel, ENSG, IGN, LASTIG, F-94160 Saint-Mandé, France

Correspondence: Raphaël Bres (raphael.bres@univ-tours.fr)

Abstract. Cycling practice has been constantly increasing for several years and the COVID crisis has just accelerated the process. Indeed, more and more municipalities have developed new cycle paths to facilitate cycling. Considering this increasing interest for cycling, it makes sense to study how this recent evolution is reflected in the underlying representation of the cycling network in the geographic databases. Main studies analysing the evolution of the road network focus on the motor vehicle network in the major cities of the world. These studies do not seem applicable to cycling network specially to some low population density areas or even to smaller cities. This paper analyses the changes in the cycling network through OSM data from a data freshness perspective. These changes can be either updates from changes in the real-world network or upgrades to the network. To these end, we propose a method using a Monte Carlo simulation (MCS) to analyse the frequency of changes in cycling routes in several areas with different population density, all in the Loire Valley region in France. We also define the cycling network, which is a very complex concept and we explain how it is represented in OSM data and suffers from different data quality issues. Results show that the number of changes across time are similar in areas having a similar population density, while being lower in low population density areas. These phenomena is higher in the cycling network compared to other networks.

Keywords. cycling, OpenStreetMap, data quality, mobility

Acknowledgements. We greatly thank the Centre Val-de-Loire region for financing this research.

1 Introduction

Cycling has gained in popularity these last ten years and has become a serious alternative to car mobility. In the context of the energy economy, CO_2 emissions reductions and the high price of gasoline, finding some new ways of transport becomes vital for a lot of people. In low population density areas, electric bikes represent a new asset in favour of bike mobility. Electric kick scooters also develop very quickly and they mostly use the cycling network since both travel at the same speed.

Despite the development of the cycling infrastructures, they are sub-represented in the main geographic databases compared to the pedestrian network or the motor vehicles network. For example, the French national mapping agency produces road data with major completeness lack for the cycle lanes (only 4 objects present in the Centre Val-de-Loire region in September of 2022).

To work on the cycling network data changes over time and on how people approach this network, we focus our work on the main Volunteered Geographical Information project: OpenStreetMap (OSM). Even though OSM offers possibilities for studying the cycling network, its data suffers from several data quality issues such as heterogeneous geographic/semantic precision. Despite such issues, with its large and dynamic community of contributors, OSM has the potential to become an up-to-date geographic database in terms of the cycling network. Therefore, the freshness of OSM data can be very interesting for an evolution analysis.

Issues

According to Peralta (2006), one aspect of data freshness, called currency or currentness, is how stale is data with respect to the real-world. In order to analyse it, we categorise OSM changes into two categories: updates and upgrades. An update consists in adding, deleting or modifying an object of the network according to the real-world changes.

An upgrade consists in adding, deleting, or modifying an object of the network because there is a data quality problem. If a real-world change in the cycling network is not reflected in the representation of the network within a given period, a data freshness problem emerges and the required update of the network becomes an upgrade. Our hypothesis is: *if data in an area has some updates but nearly no upgrades, then its overall quality (including freshness) is good.* .

This thinking leads to the two following questions:"How does the cycling network evolve over time in OSM?", and "How informative are the changes on data freshness and more generally on data quality?"

Therefore, our first idea is to study the network changes over time and, in particular, to identify changes according to the cycling usage of the network. The cycling network, for which this paper proposes a definition, forms a connected component with large numbers of vertices (e.g., crossroads) and edges (e.g., roads, cycle paths). The complete comparison, edge by edge, is difficult and could be less informative to real-life usage than the comparison between representative routes linking the same geographical points, year by year. According to this, we consider the use of a Monte-Carlo simulation to generate the representative cycling routes, like the approach done by Schmidl et al. (2021) to study the Vienna motor vehicle network in OSM.

Moreover, the dynamics of the changes vary according to the geographical zones. This could be due to the densities of populations. The higher the population density, the bigger the network development, and in particular for OSM data, the greater the possible number of contributors. Therefore, this paper studies different subareas with various population densities using the previous processes. The changes of the generated routes over the different subareas are classified according to their impact on the route length (insignificant, significant and major) and on their nature (update/upgrade). These two kinds of labels are informative on the subarea data freshness and thus on data quality.

Contributions

This paper proposes three contributions. The main contribution is a method to extract and compare representative routes for cycling usage year by year for several subareas. According to the route comparison, we propose, as the second contribution, to identify changes that are significant and major in terms of path length differences and whether they are updates or upgrades. Both criteria give information about data freshness in a given area. These contributions are applied to the cycling network which needs a clear definition and its OSM representation. Both the cycling network definition and representation in OSM compose the third contribution of this paper.

Section 2 introduces the related work on OSM road network data quality assessment. Section 3 defines the cycling network and the issues regarding its representation. Section 4 presents the method proposed to analyse the cycling network. Section 5 shows our results and section 6 summarises our findings and proposes some future works.

2 Related work

This section shows the existing works on temporal data quality in OSM and routing

2.1 Temporal OSM data quality

Many researchers interested in the quality of OSM data studied the road network and its quality, in part because the main key corresponding to roads (highway) is the third most used tag in the project according to OSM taginfo global page[1] just behind source and building. A lot of research work exist on this topic.

The main approach used to evaluate the OSM road network data is a comparison between OSM data and some authoritative data (Girres and Touya, 2010; Zhao et al., 2015) that are considered as a ground truth. Ferster et al. (2019) use some open data on smaller area that is really close to the ground truth but which is complicated to adapt it to a bigger area because of the time needed to create the data.

Data is not available for every country in the world so another type of analysis is needed for these areas, relying on intrinsic characteristics of the OSM data. Arsanjani et al. (2013) show how data quality for a road network can be assessed thanks to the contributor's other contributions.

These articles focus on the most used roads such as motorways or primary roads. The highway objects dedicated to other means of transport are not analysed in detail.

Some research papers focus on analysing OSM data quality overtime trough different snapshots, e.g. (Zhao et al., 2015). Another completely different approach to assess the evolution of the road network is described in Schmidl et al. (2021). The main idea is to generate couples with one starting point and one ending point using a Monte Carlo approach. A shortest path algorithm is then performed between each couple in each version of the network (one per year between 2014 and 2020 in the limit of Geofabrik when the experiment was done). These routes are then analysed across time. The number of distinct routes differing from their final version has been computed alongside the percentage of route length increase across time.

The evolution of OSM data regardless of the type of object is also an interesting input for our study. Novack et al. (2022) show that the number of OSM data contributions is increasing following a linear model in different major cities in the world.

A study with a route generation was conducted by Bres et al. (2022) to show some data quality problems by the generation of one route with six different tools. That ex-

[1] https://taginfo.openstreetmap.org/keys

perience showed that both OSM and Google data are not perfect.

But the one major problem for developing cycling resides in the routes that are not specific to it.

2.2 Routing

As an exhaustive study of each change on the network is time consumming and as MCSs generally produce good trend information, Schmidl Schmidl et al. (2021) use it to assess the OSM motorway network freshness and evolution. They generate several routes using OSRM from different OSM snapshot and then compare the obtained routes (path length) to see the frequency of OSM data updates for road networks. To implement this method on the cycling network, one can use the OSRM routing engine. The OSRM engine uses some OSM data (Luxen and Vetter, 2011) and it already has a specific version for cycling routes. Nevertheless, the OSRM cycling option allows cyclists to access ways that are forbidden to them in several countries such as France. Therefore, the cycling network definition used by OSRM needs to be adapted.

The computation of the ideal cyclist trajectory is a multi-criteria optimisation problem. This is discussed in Sauvanet (2011) and the web application "GéoVélo", used in our preliminary work, came out thanks to this work. GéoVélo is a routing engine used by some important French cities like Lyon or Tours to present the cycleways and generate a route based on some criteria. It uses OSM data and proposes different routes based on different criteria like safety or distance. Several platforms, such as GeoVelo, propose tools to compute cyclist trajectories based on OSM and complementary data. It could be interesting to use them for the trajectory generation according to the fact that adding safe infrastructure in the trajectory computation, even if longer, is an important aspect of the cycling network evolution. However, those platforms do not open their codes and they generally exploit data not coming from OSM. Therefore, their use is not adapted to study the OSM cycling network evolution.

For our approach reproducibility and robustness, we will use the OSRM tool with the cyclist option and adapt it to our cycling network definition that is introduced in the next section.

3 Cycling network construction

This section defines the cycling network and describes different data quality problems encountered in OSM cycling network.

3.1 Definition

In the literature, the cycling network is defined as the set of roads which have some cycling infrastructures (Vybornova, 2021). This definition was given in an article whose study area was the city of Copenhagen, Denmark. This city is known as one of the most "bikeable" city in the world but it would be impractical to adapt this definition everywhere. Based on this definition, it would be nearly impossible to find an route between two given points, because, very few routes have specific cycling infrastructures such as cycling lanes. Indeed, typical cyclists' daily routes include many roads without cycling infrastructure since the everyday cyclists often look for short, quick and flat routes. A more complete definition of the cycling network should not be limited to cycling infrastructures but include every road where a cyclist is allowed to go, even without a dedicated place for a cyclist. The definition given here is close to the one used by the website wandrer.earth[2] but this website proposes only an OSM representation of the cycling network.

There are two main types of cycling infrastructures, namely cycle paths and cycle lanes. The cycle paths are safe infrastructures, physically separated from roads. They are frequently used by occasional cyclists, especially when they have children with them. The cycle lanes are specific lanes on the motor vehicles network, dedicated to cycling. In an urban area, this is a common infrastructure to segregate bikes and some slow motor vehicles, such as scooters. Another quickly developing infrastructure is a motor vehicle road where only one way is available for cars but both ways for cyclists.

Let be $\mathcal{N}_{IC}$ the set of all cycling infrastructures, $\mathcal{N}_{CP}$ the set of all cycle paths, $\mathcal{N}_{CL}$ the set of all cycle lanes, $\mathcal{N}_{CTWS}$ the set of all cycle two-ways street, $\mathcal{N}_P$ the pedestrian network, $\mathcal{N}_M$ the motor vehicles network, $\mathcal{N}_{MW}$ the motorway network and $\mathcal{N}_{HW}$ the set of high-speed ways. We define the cycling network (the set of all roads allowed to cyclists), $\mathcal{N}_C$, as the union between $\mathcal{N}_{IC}$ (defined in Equation 1), $\mathcal{N}_P$, and $\mathcal{N}_M$ where $\mathcal{N}_{MW}$ and $\mathcal{N}_{HW}$ are withdrawn. Equation 2 gives the formal definition of $\mathcal{N}_C$.

$$\mathcal{N}_{IC} = \mathcal{N}_{CP} \cup \mathcal{N}_{CL} \cup \mathcal{N}_{CTWS} \tag{1}$$

As cyclists can take pedestrian ways if they walk alongside their bikes, $\mathcal{N}_C$ also includes sidewalks, even though they are not suited for this use. The steps are also included into the analysis but with a decreased speed since it is complicated to use steps with a bike. These are generally avoided by cyclists because walking alongside their bike is a major loss of time but it is sometimes inevitable.

$$\mathcal{N}_C = \mathcal{N}_{IC} \cup \mathcal{N}_P \cup (\mathcal{N}_M \setminus (\mathcal{N}_{MW} \cup \mathcal{N}_{HW})) \tag{2}$$

With this larger notion of a cycling network in mind, the next subsection describes how the cycling network is represented in OSM.

Cycling network	Formalisation	Composition
cycle paths	$\mathcal{N}_{CP}$	$highway = cycleway$
cycle lanes	$\mathcal{N}_{CL}$	$highway! = cycleway + cycleway = lane$
cycle two ways streets	$\mathcal{N}_{CTWS}$	$oneway = yes + oneway : bicycle = no$
cycling infrastructures	$\mathcal{N}_{IC}$	$\mathcal{N}_{CP} \cup \mathcal{N}_{CL} \cup \mathcal{N}_{CTWS}$
pedestrian network	$\mathcal{N}_{P}$	$highway IN footway, living\_street, pedestrian, path, track, steps$
motor vehicles network	$\mathcal{N}_{M}$	$highway NOT IN cycleway, footway, living\_street, pedestrian, path, track, steps$
motorways	$\mathcal{N}_{MW}$	$highway IN motorway, motorway\_link$
high-speed ways	$\mathcal{N}_{HW}$	$highway IN trunk, trunk\_link$

Table 1. Table of correspondences between cycling network and OSM tags

3.2 OSM representation and data quality

The rapid evolution of cycling infrastructures and the open model of OSM result in some heterogeneity in the representation of cycling infrastructure.

The cycle paths network ($\mathcal{N}_{CP}$) is represented in OSM by the tag $highway = cycleway$ and, if a road is close from the path, the latter should be segregated by a large space or an elevation to be called a cycle path. According to the world page of OSM taginfo, there are more than 1.5 million cycleways in the world. In some isolated areas, data are missing but this is not a problem specific to the road network or cycle paths in general. OSM taginfo also presents a plot with the number of objects tagged with $highway = cycleway$ overtime. The website page[3] shows a first increase at the beginning of the project that flattens a little between 2015 and 2019 but the COVID crisis really made a huge jump in the number of cycleways. Only on $\mathcal{N}_{CP}$, there are some data completeness problems with for example the key $surface$ that tells us if a road is paved, made of cobbles or sand, there only are 57% of completeness for one of the most important attributes describing a cycling infrastructure.The completeness is even lower for the keys $oneway$ (21%) and $width$ (7%). A major problem comes when some cycle paths are represented on a way belonging to the $\mathcal{N}_{MV}$. Nowadays, the good practice is to create another way for the cycle path but this rule is not always respected. The pedestrian sidewalks also have this problem. $\mathcal{N}_{CL}$ is not separated from the road, and information about them has to be added to the highway feature, that is generally made for motor vehicles. The main tag to find a cycle lane is $cycleway = lane$ but the key can also contain information about the position of the lane. That information is conveyed by the key being $cycleway : both, cycleway : left$ or $cycleway : right$, indicating on which side of the road the cycle lane is. There are more than 700 000 ways tagged with $cycleway = lane$ or a declination of it. This number seems low compared to the development of this kind of infrastructure in a lot of cities.

$\mathcal{N}_{CTWS}$ has two possible representations in OSM. The first one is to use the tag $oneway : bicycle = no$ and the second one is to use the tag $cycleway = opposite$. Each time, the tag $oneway = yes$ is needed to indicate that cars can go in only one direction. None of these two has been chosen as the standard rule yet. There are 1 820 ways tagged with the first representation and approximately 70 000 ways tagged with the second proposition including the more precise tags evoked when defining the cycle lanes.

Every component of $\mathcal{N}_C$ with the OSM tags composition are listed in Table 1.

With the cycling network presented, we describe the data quality problems we expect to find in the data. Firstly, the cycling network undergoes rapid evolution but draws less interest to the OSM community than the general road network, which can lead to lack of completeness, and a lack of geometric or semantic accuracy. This last aspect can be a false value on a tag for the semantic accuracy or a road not well represented or connected to the network for the geometric and topologic accuracy.

The notion of data freshness (Peralta, 2006) is also important; the COVID crisis has lead to the creation of a lot of new infrastructures, some of them being temporary.

4 Method and experiments

This section presents the method used for the experiments, the different parameters defined for the experiments and the area of study. The data and software accessibility are exposed too.

4.1 Method and experiments

Our method customs Schmidl et al. (2021) to study the OSM cycling network and compute new statistics for several areas of interest (AoIs). Figure 1 presents the entire pipeline of the proposed method for only one area of interest. That pipeline can be done in parallel for multiple areas of interest.

The first step consists in downloading an OSM data snapshot per period of interest from Geofabrik's internal download server. We consider n periods of interest. Each snapshot contains every elements showed in Table 1.

[2] https://wandrer.earth

[3] https://taginfo.openstreetmap.org/tags/highway=cycleway#chronology

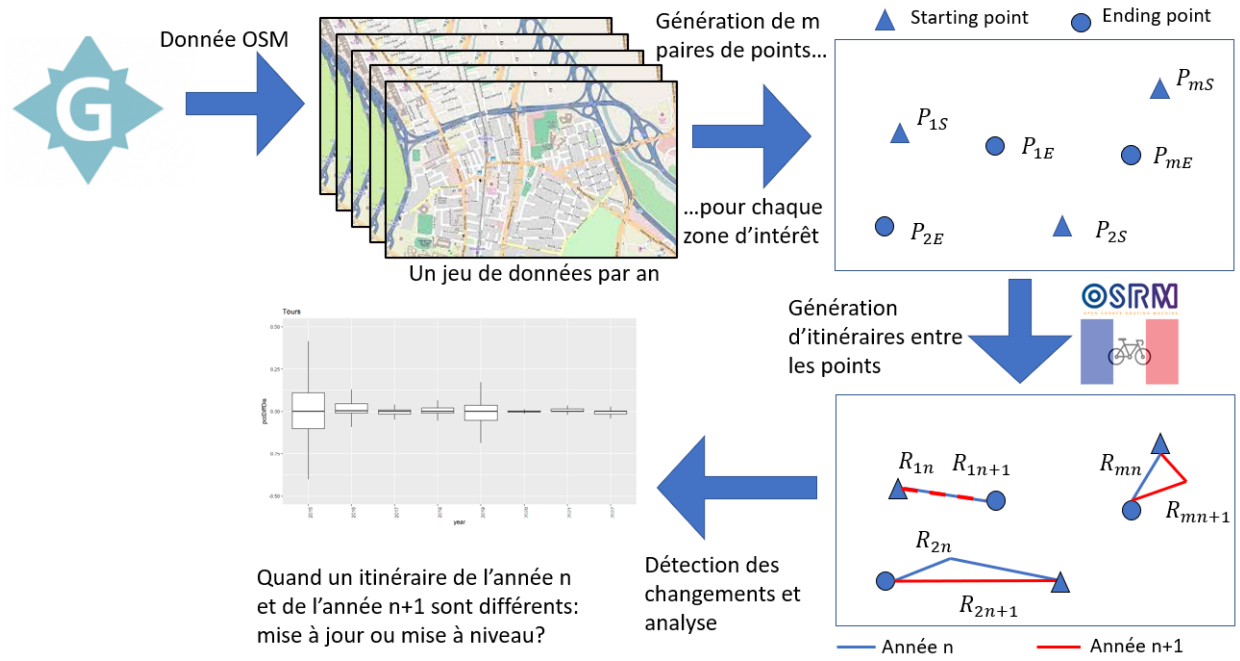

Figure 1. Pipeline of the proposed method for one area of interest

To be able to use OSRM with our definition of $\mathcal{N}_C$, we adapt the profile *bike.lua* by adding the $trunk$ and $trunk\_link$ to the list of forbidden values for the highway key.

According to this modification, and considering a snapshot in entry, OSRM is now in capacity to match point to the cyclist network and to generate cyclist routes in this snapshot.

In these snapshots, we will study k AoIs.

Then, the route generation step can begin. For each AoI, we compare $m \times n$ different routes, m is the number of the generated routes for a snapshot in an AoI. In (Schmidl et al., 2021), the starting and ending point coordinates were chosen thanks to a uniform law and their matching to the network may differ according to the snapshot. In this study, the area was a big city (Vienna) with a high density of roads. Thus the distances of the generated points to the network are not very useful. As we want to study different-density population areas, the distance between points, the distance between generated points to the network, and the unicity of the network matching points must be considered.

For that, we consider first the generation of m pair of points according to the following process. Firstly, the starting (X_{is}) and ending (X_{ie}), $i \in [1, m]$ points are randomly generated with a uniform law in the AoI and matched to the closest point on $\mathcal{N}_C$. A point is kept only if it is matched at the same place in $\mathcal{N}_C$ for every year (isTheSame). The distance $distPNc$ between each point to the network $\mathcal{N}_C$ should be lower than a maximum distance ($distPNcMax$). When the routes are computed for each snapshot, at least one of them should have a length $distPP$ taking a value between $distPPMin$ and $distPPMax$. These parameters are addressed in the next subsection.

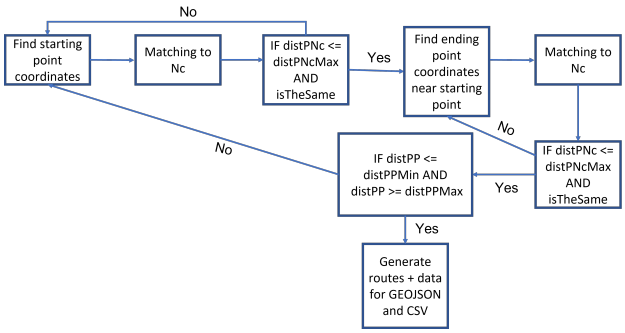

Figure 2. Algorithm to create one route

Figure 2 resumes the algorithm to create one route for one AoI.

This paper focuses on the length since the distance was computed with OSRM. The percentage of length change compared to the previous year is the main variable analysed in Section 5. In fact, we will study different type of changes: small changes that produce at least a difference of $slc\%$ and major changes that produce at least a difference of $mlc\%$ in route length. Then we study the number of those two kinds of changes for the different AoIs.

The next subsection is about some parameters of the route generation algorithm.

4.2 Parameters

This subsection presents the different parameters used for our experiments.

We consider 9 snapshots ($n = 9$). Each one represents a year from 2014 to 2022 (due to GeoFabrik limits) and is extracted the 1^st of January.

In the area of study, we consider 6 AoIs, chosen according to their population densities ($k = 6$). The choices of the AoIs are explained in Section 4.3.

In each AoI, to make the MCS representative, 1000 starting and ending points are generated for each OSM snapshot ($m = 1000$).

To achieve a reasonable balance between AoIs with a dense network and areas with a sparser network, distance to the network is set at 100 meters ($distNcMax = 100m$).

The chosen interval $[distPPMin, distPPMax]$ is $[300m, 5000m]$ accordingly with Litman (2010).

The percentage length difference compared to the last year is the heart of the analysis.

We set empirically the threshold in distance length difference for **significant changes** (slc) as 1%, and the one for the major changes (mlc) 20%. Some other thresholds have been tested but those values are informative for us. For instance, taking a threshold higher than 1% for significant changes could not allow identifying the adding of cycle paths close to the motor vehicle roads.

Our experiments were running on a virtual machine running on an Intel Core 7 with 32GB of RAM, without any multi-thread optimisation.

The next subsection explains the chosen area of study and the different AoIs taken for the analysis presented above.

4.3 Area of study

The study area is a part of Loire Valley in the Centre Val-de-Loire region which is located in the centre of France (see figure 3). This is an agricultural area with many rural areas and some medium-sized cities at the France scale like Tours and Orléans. This is a very touristic area having famous Renaissance castles such as Chambord[4] or Chenonceau[5]. Recently, the region developed its cycling network for tourism mainly based on "The Loire by bike" [6] which is a bike route following the Loire river. To represent every aspect of the region, our 6 AoIs are chosen as follows. Two AoIs include the most dynamic cities of the region which are Tours (294 220 inhabitants in the metropolis) and Orléans (288 229 inhabitants). Then, two medium-sized cities at the scale of the region are selected with Blois (105 286 inhabitants) and Bourges (102 679 inhabitants). The last two areas are bigger zones with

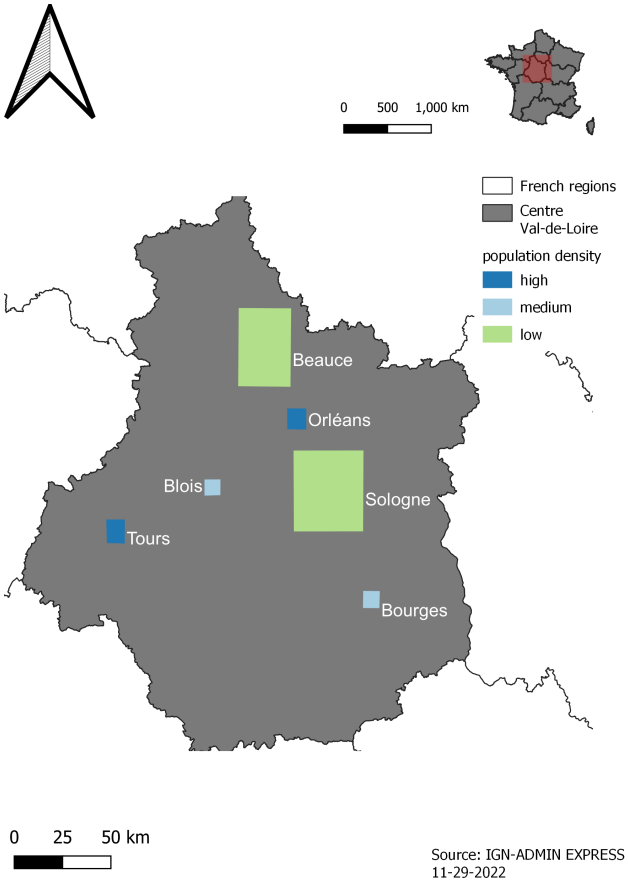

Figure 3. Localisation of the study area

low population densities. They are the agricultural Beauce plain and the Sologne forest. Beauce has a lot of roads that often cross between them. On the contrary, Sologne is a forest area where there are long straight roads without any crossing for several kilometres and some packed villages. With its famous castles, Sologne attracts many tourists that are potential OSM contributors.

4.4 Data and software availability

The experiments done in this paper only required OSM data that are free under the Open Data Commons Open Database License (ODbL).

OSM data was downloaded on Geofabrik. OSRM is free and open source under a 2-Berkeley Software Distribution License. Python 3.8 was used to generate the routes with the packages NumPy, CSV, requests and GeoJSON. R was then used to analyse the data on each route with the packages dplyr and ggplot2. From this script, both Tables 2 and 3 and Figure 4 were generated. Figure 5,6 and 7 were obtained thanks to QGIS version 3.20.

The results obtained and presented in this paper are reproducible. The deposit (https://github.com/raphael-bres/AGILE2023-reproduction) contains all the data, tools, explanations about these experiments, and results.

[4] https://www.chambord.org/en
[5] https://www.chenonceau.com/en
[6] https://www.loirebybike.co.uk/

Table 2. Percentage of route length change by at least 1%

Population density	Zone name	2014/2015	2015/2016	2016/2017	2017/2018	2018/2019	2019/2020	2020/2021	2021/2022	Average
High	Tours	40.4	27.1	17.7	19.7	20.8	11.0	12.2	15.5	20.55
	Orléans	29.1	9.5	32.2	18.3	34.6	34.0	34.5	30.5	27.84
Medium	Blois	18.2	14.1	5.5	6.1	24.6	10.9	27.2	37.1	17.96
	Bourges	21.1	30.8	11.9	9.1	22.9	11.9	21.0	28.0	19.59
Low	Beauce	4.9	3.7	12.2	6.0	2.4	1.6	0.8	4.8	4.55
	Sologne	4.9	5.8	2.2	2.7	4.4	5.9	1.2	1.6	3.59

Table 3. Percentage of route length change by at least 20%

Population density	Zone name	2014/2015	2015/2016	2016/2017	2017/2018	2018/2019	2019/2020	2020/2021	2021/2022	Average
High	Tours	1.6	0.8	0.2	0.6	0.3	0.0	0.4	0.1	0.5
	Orléans	1.4	0.3	1.3	0.4	0.7	0.7	0.3	0.4	0.69
Medium	Blois	1.1	1.4	0.1	0.0	2.2	0.7	5.5	2.7	1.71
	Bourges	0.5	1.0	0.5	0.6	0.7	0.7	1.0	1.5	0.81
Low	Beauce	1.5	1.3	5.3	2.3	0.5	0.9	0.0	1.2	1.63
	Sologne	1.5	2.9	0.7	0.8	1.1	1.9	0.1	0.6	1.2

5 Results and discussion

That section is divided into two parts. The first one shows the evolution of OSM cycling network and the second one categorises some changes seen in the first part.

5.1 Evolution of OSM cycling network

After applying the method presented in section 4.1, the first result computed is how much the routes' lengths have evolved. To this purpose, a plot of the percentage of routes that got a 1% or more length change is created. The increases and the decreases are not dissociated. The 1% change value is used to indicate a significant change to not capture small changes such as the transformation of a crossroad into a roundabout or a geometry improvement of a single lane.

Table 2 shows that the low population density areas have in general some smaller change rates compared to the four cities.

Tours has many changes as well as Orléans. Bourges and Blois have a quantity of changes similar to Tours with an increase of small changes over the years. In Tours, there are many changes at the beginning of the study but they are reduced in the end because the possible changes in the city centre are limited. The low population density areas have fewer changes but for Sologne, the hypothesis of the desired length (between 300 to 5000 meters) was not enough to capture some well suited routes for that analysis. The same table was computed with a 20% change to only capture the major changes. The results for this rate are shown in Table 3.

Table 3 shows that the proportion of routes with changes by at least 20% is much smaller and that these proportions are not much different between cities and rural areas. This may be due to the fact that cities have fewer possibilities for a major improvement of route lengths, whereas network densification in rural areas can make several routes much shorter. These changes are not often present so the numbers presented here are close to reality, unlike the 1% change plot where an average and standard deviation analysis could verify the values with several repetitions of this experiment.

The dispersion of the percentage of length change was analysed and is presented with the boxplots in figure 4.

With these boxplots, the dynamics of the changes are very different for Tours and Orléans. The means are close but the changes are not done during the same years. The plots for Blois, Bourges, Beauce and Sologne were omitted because there were too few changes to produce visually significant boxplots. The changes had some small variation for the medium-sized cities but for the low population density areas, the mean, the median, the first, and the third quarters were all equal to zero. This leads to the question of whether the parameters chosen were adapted for every study area. The main area where parameters could be modified to achieve more significant results is Sologne. Sologne is composed of packed villages linked with some straight roads that are often longer than five km. It is possible that allowing longer routes would have given better results for this area. The distance to the network is also a question here because the road network is not dense at all, so the distance to the network could have been higher for this area. Another hint about the problem with Sologne and the parameters is found in the execution time of the route generation script. According to our configuration, the 1000 routes in Sologne were generated in 18 minutes and 25 seconds while the five other areas (including Beauce, a low population density area) took between 6 and 10 minutes each. The quantitative analysis shows some differences between the different cities and the low population density areas. The low population density areas have a similar evolution which is not what we expected because there are a lot more tourists (and thus more possible contributors) visiting Sologne than Beauce. These OSM-impacting changes can be categorised into two groups with the changes caused by recent real-world changes and the data freshness problems. That step is important to assess

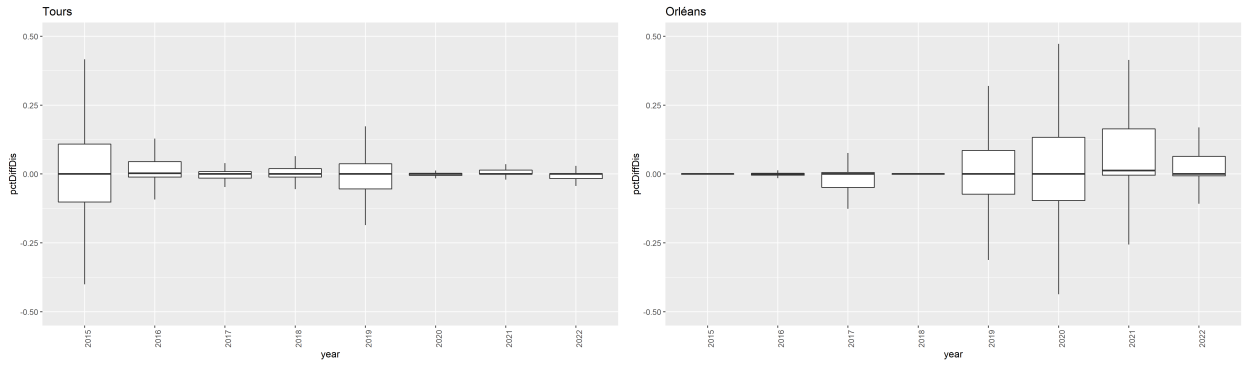

Figure 4. Boxplots of the percentage of route length change per year for Tours and Orléans

the overall data quality of an area. The next subsection offers a classification by hand of some changes.

5.2 Evolution of network typology

This subsection focuses on Blois AoI. The main idea is to present some data quality problems and see how they can be solved.

In this analysis, we consider that if a real-world change has been reflected in the OSM data in less than 3 years, it is an update, else it is an upgrade. The three-year period must allow normally the OSM community to reflect real-world changes. If it lasts more than 3 years, we consider that it reveals a quality issue.

The first quality problem we identified is about a road that is forbidden to cyclists. We could not find the exact date when the road has become forbidden but it was certainly before 2010. Many routes could potentially use this road, because Blois is divided in two parts by the Loire river and there are only three bridges to cross it. One of these bridges only allows cyclists and pedestrians from both sides without any segregation. In our experience, 47 routes in 2020 use this road. In 2021, this road has been upgraded as forbidden to cyclists, and for every route that used the road in 2020, the percentage of length difference was analysed. This percentage is at least 1% for 37 routes among 47 but only 12 routes are longer in 2021 than in 2020. Figure 5 shows in blue, one of the 12 routes use the forbidden road section in 2020. The green route is the one for 2021 avoiding the forbidden route section. The 2021 route was not perfect yet and suffers from the absence of one cycle lane. This problem required at least an upgrade of the network because the last roadworks on that road were finished in 2018. As, since the end of January 2021, a cycle path fully recovers the street, which is still not reflected in OSM, there is a clear need of a network update.

Another problem on the same route is found near the ending point. The problem is a way not well connected to other ways in the OSM network, which leads to a detour. The problem is shown in figure 6. The change that should be done here is a topological upgrade on the network.

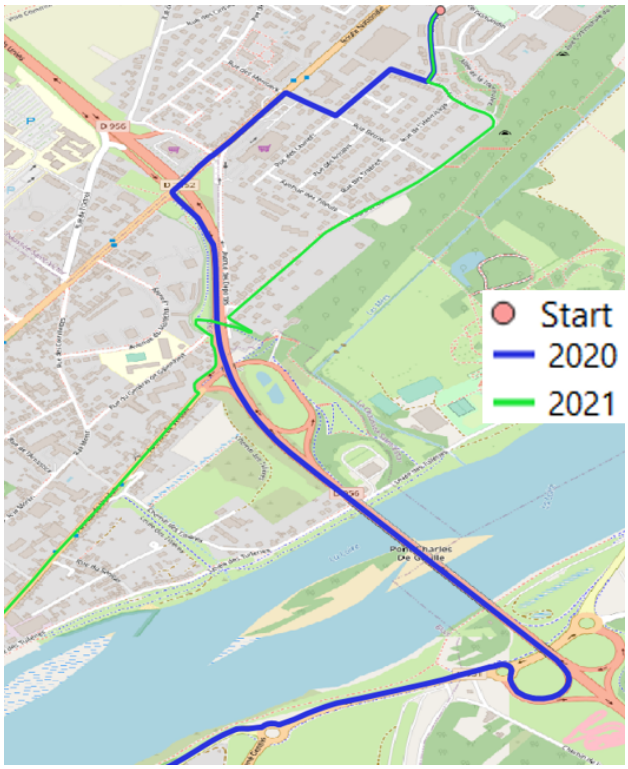

Figure 5. Example of a route using a forbidden road in 2020

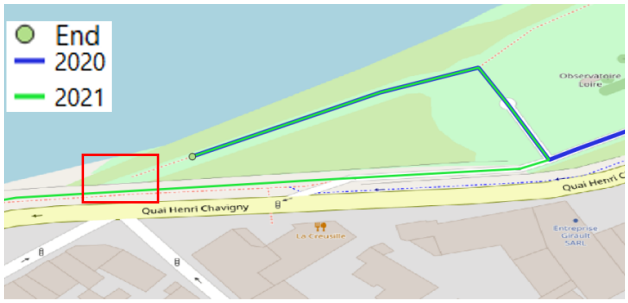

Figure 6. Example of a network connectivity problem

There is one more data quality problem illustrated in figure 7 with the geometric accuracy. The 2014 route (black) that has a straight line whereas the other routes have some

curves. This is a geometric precision problem with the 2014 route that did not represent reality. The road has been upgraded between the 2014 and the 2015 snapshot. At the beginning of the route, the 2022 route (brown) is the only one that can use the new cycle path parallel to the main road of the area. The update of the network has been done here even before the opening of the cycle path.

Figure 7. Example of a route using a road that has a modified geometry over time

Thanks to this analysis, there are many data quality problems in the data, and OSM data is not completely suitable for bike routing yet. These problems can be corrected through upgrades of the data. But since the real-world cycling network is always evolving, OSM cycling network requires some updates to consider the latest changes. A required update in the network becomes an upgrade if the real-world change was done at least three years ago.

6 Conclusion

This paper proposed a definition of the cycling network as the set of all roads allowed to cyclists. We then studied the evolution of this network in six areas in the Loire Valley with the adaptation of an existing method. The work we have done gives a method to analyse cycling data and a state of the network in our study area. The study is also original with an analysis of a new network in multiple population-density areas. In fact, most of the related work in OSM data quality shows major cities but we also investigated a few rural areas in the Loire valley.

This study shows that the low population density areas and the cities have a different evolution for the representation of their cycling network in OSM. The cities with the same population density show some differences at the moment when they got their route changes but they have a similar average over time. The main example is shown between Tours and Orléans where Tours got the most changes in the first years of the study but Orléans' network was more changed in the last years of the study. The typology brings two main types of changes, update and upgrade. It is now impossible to automatically differentiate an update and an upgrade but this question is important to know the quality of the data. It is thus an important research field to open. The study is applicable in every region of the world and some tests in other regions could bring some new variables. Our study areas are mainly flat, but everyday cyclists often choose a route with the least amount of climbing. Reproducing this study in a mountainous area could add the slope variable where a route with a gentle slope will be used instead of a direct route with a difficult hill.

A major future work in this field concerns bikeability, studied for instance in (Kellstedt et al., 2021). We will work on comparing generated routes according to their changes through the measures of their bike friendliness in place of the distance used in this paper.

References

Arsanjani, J.-J., Barron, C., Bakillah, M., and Helbich, M.: Assessing the Quality of OpenStreetMap Contributors together with their Contributions, in: AGILE' 2013, 2013.

Bres, R., Peralta, V., Le-Guilcher, A., Devogele, T., Olteanu-Raimond, A.-M., and de Runz, C.: Spécification et qualité du réseau cyclable, application à la recherche d'itinéraires, in: INFORSID, 2022.

Ferster, C., Fischer, J., Manaugh, J., Nelson, T., and Winters, M.: Using OpenStreetMap to inventory bicycle infrastructure: A comparison with open data from cities, International Journal of Sustainable Transportation, https://doi.org/10.1080/15568318.2018.1519746, 2019.

Girres, J.-F. and Touya, G.: Quality Assessment of the French OpenStreetMap Dataset, Transactions in GIS, 14, 435–459, https://doi.org/10.1111/j.1467-9671.2010.01203.x, 2010.

Kellstedt, D., Spengler, J., Foster, M., Lee, C., and Maddock, J.: A Scoping Review of Bikeability Assessment Methods, Journal of Community Health, 46, 211–224, https://doi.org/10.1007/s10900-020-00846-4, 2021.

Litman, T.: Analysis of Shorter Trips Using National Personal Travel Survey Data, 2010.

Luxen, D. and Vetter, C.: Real-time routing with OpenStreetMap data, ACM SIGSPATIAL 2011, pp. 513–516, https://doi.org/10.1145/2093973.2094062, 2011.

Novack, T., Vorbeck, L., and Zipf, A.: An investigation of the temporality of OpenStreetMap data contribution activities, Geo-spatial Information Science, pp. 1–17, https://doi.org/10.1080/10095020.2022.2124127, 2022.

Peralta, V.: Data Quality Evaluation in Data Integration Systems, Ph.D. thesis, Université de Versailles-Saint Quentin en Yvelines ; Université de la République d'Uruguay, Versailles, France; Montevideo, Uruguay, 2006.

Sauvanet, G.: Recherche de chemins multiobjectifs pour la conception et la réalisation d'une centrale de mobilité destinée aux cyclistes, Ph.D. thesis, Université François Rabelais - Tours, Tours, France, 2011.

Schmidl, M., Navratil, G., and Giannopoulos, I.: An Approach to Assess the Effect of Currentness of Spatial Data on Routing Quality, AGILE: GIScience Series, 2, 1–12, https://doi.org/10.1234/56789, 2021.

Vybornova, A.: Identifying and classifying gaps in the bicycle network of Copenhagen, Ph.D. thesis, University of Copenhagen Environmental Science, Copenhagen, Denmark, 2021.

Zhao, P., Jia, T., Qin, K., Shan, J., and Jiao, C.: Statistical analysis on the evolution of OpenStreetMap road networks in Beijing, Physica A: Statistical Mechanics and its Applications, 420, 59–72, https://doi.org/10.1016/j.physa.2014.10.076, 2015.