Methodological guide for the segmentation of cycle routes

(Software used: GIS: QGIS 3.12 / spreadsheet: Microsoft Excel)

Methodology used for the EuroVelo 1 route in the AtlanticOnBike project

Document written by Corentin LEMAITRE - C-Mobilité on behalf of Vélo & Territoires and Eco-Compteur

Table des matières

Introduction

1.1 Context

The segmentation method presented was developed by C-Mobility as part of the evaluation of the economic impact along EuroVelo 1. The length of the itinerary does not allow for a segmentation into sections based on knowledge of the territory. For this reason, this method has been developed to carry out a breakdown based on the analysis of the data.

The method allows analysis along a linear geographical element and can be adapted to a different objective by changing the type of data considered. Most of the data used are available worldwide, but some data are only available on the European territory (Land use data and EuroVelo data).

1.2 Overall explanation of the method

The aim of segmentation is to divide the cycle route into continuous sections with a comparable context (e.g. a section in an urbanised area then a section in a rural area, etc.). To analyse the context of the itinerary, different characteristics can be considered depending on the objective (e.g. population density, type of development, territorial boundaries).

The method is composed of 3 steps:

1. Division of the route into unit segments, defining the level of granularity of the study (more than 10,000 units segments of 1 km within the framework of EuroVelo 1)
2. Adding context data along the route on each of the unit segments
3. Grouping of unit segments with a comparable context into unified sections

1.3. Why segmentation?

Cyclists travelling a cycle route incur expenses during their journey or outing. These expenditures are economic spin-offs for the territory they pass through. Knowing the economic spin-offs makes it possible to know the return on investment of actions to develop cycle tourism.

Within the framework of the evaluation of economic spin-offs, several "items" are used:

1. Automatic counts at several points along the itinerary
2. Surveys of route users at certain counting points on certain days for a sample of users
3. A knowledge of the complete itinerary to extrapolate one-off results over the entire route

To extrapolate one-off results to the entire route, the method uses a route represented by homogeneous sections. Based on these sections and the number of people using them, economic benefits can be estimated for the route. The purpose of this document is to present the method used to create these sections.

1.4 Glossary of terms

The notions of segments and sections will be used extensively throughout the rest of this document. The definitions of each of these terms are as follows:

- Segment: part of the route or network resulting from the segmentation and not exceeding one kilometre in length.

- Section: aggregation of homogeneous kilometre segments, i.e. with the same characteristics (population density, density of tourist beds, type of development, etc.).

1.5. Mention and re-use of the method

Any use or mention of this method must specify the following source: "Segmentation of cycle routes or networks - C-Mobility / Bicycle & Territories / Eco-Counter - 2020 - Funded by the European Union as part of the AtlanticOnBike project".

2. Method of operation

2.1. Preliminary note

Software requirements

The method presented is accessible by a spreadsheet and GIS software user. The treatments used on QGIS require version 3.12 or higher, the spreadsheet used for the study was Microsoft Excel but Libre Office Calc can also be used.

Hardware requirements

A storage space of 50 GB is required to carry out all the treatments. Some of the treatments may take several hours to complete. In order to obtain results in a reasonable time, it is recommended to use a high-performance computer (data on SSD, RAM greater than 8 GB, processor > 2017).

Choice of a reference coordinate system

To speed up processing, it is preferable that all data layers use the same reference coordinate system (SCR). **Before you start, choose the SCR** that will be used for all the rest of the work. The entire route must be within the validity range of the SCR and the unit of the SCR must be metric. The reproject layer tool allows you to change the projection system of a data layer.

GIS data management

To speed up processing time and simplify data management, **it is strongly recommended to use the Geopackage** file format to store all the layers produced. For the implementation of the method very large data volumes are used. The download time of these data must be anticipated. The data source section, below, allows you to identify the data used and upload it.

2.2. Data source

1. **European Cycling Federation.** Atlantic Coast Route. Eurovelo. [Online] https://en.eurovelo.com/ev1. The EuroVelo file integrating the road development status is to be requested from the ECF.

2. **Partners of the AtlanticOnBike project**. Traced within the different countries.

3. **Vélo & Territoires**. ON3V data in OpenData. [On line] https://www.velo-territoires.org/wp-content/uploads/2019/04/SHP-1.zip.

4. **OpenStreetMap contributors.** Waymarked Trails. [Online] https://cycling.waymarkedtrails.org/#route?id=2763798&map=4!57.9189!7.9873.

5. **Eurostat**. NUTS. [Online] https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts.

6. **Corine Land Cover.** [Online] 2018. https://land.copernicus.eu/pan-european/corine-land-cover/clc2018?tab=download.

7. **European Commission, Joint Research Centre (JRC)**. [Online] https://ghsl.jrc.ec.europa.eu/ghs_pop2019.php.

8. **Geofabrik**. OpenStreetMap Data Extracts. [Online] https://download.geofabrik.de/.

2.3. Division of the route into individual segments

Difficulty	The combination of data sources of different levels of accuracy, with different completion dates and using different reference frames for the definition of the route. The definition of the actual route and the correction of inaccurate geographic files can require verifications which can take a long time.
Medium	Software QGIS 3.12
Input	The different known routes of the itinerary
Output	Continuous route broken down into ordered unit segments

2.3.1. A continuous route

This step is optional, if you have a geographical file describing the route of the itinerary in a single "continuous line" entity. If this is the case, continue the method in paragraph "2.3.2. First level breakdown".

QGIS processing

Gather geographical files describing the cycle route or network from all available sources.

Open QGIS. Add the different routes as layers within your QGIS project. Create a new " Geopackage " layer named " chosen route " of the Multiline type which will contain the different parts of the route.

A visual comparison of the data sources must be carried out to determine which data source to use for each part of the route. The aim is to obtain the most reliable and complete route possible. Copy portions of the valid route into your chosen route layer. When adding elements to your layer, check that there is a common vertex between the different entities in the layer. If this is not the case, edit the layer to connect the entities by activating the snap on the summits of the active layer.

After bringing all parts of the route together use the "Shortest path (point to point)" algorithm with route chosen as the vector layer representing the network and selecting your start and end point of the route in North-South order (convention used by the European Cyclists' Federation (ECF)).

If the algorithm gives an error, check that the route is continuous. If not, fill in the gaps and then repeat the operation (using Ctrl + Alt + H allows you to restart the last algorithm used with the parameters retained).

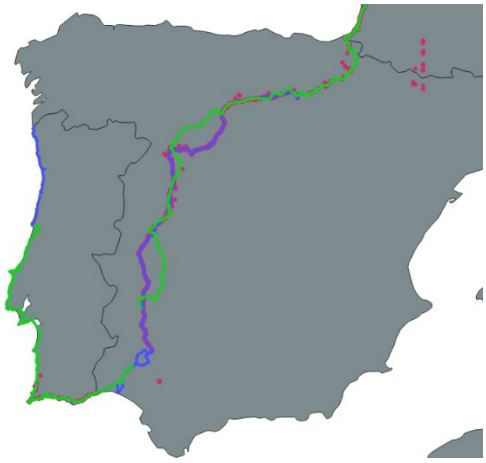
Convert the result of the treatment into a permanent layer under the name "continuous route".

Example for EuroVelo 1

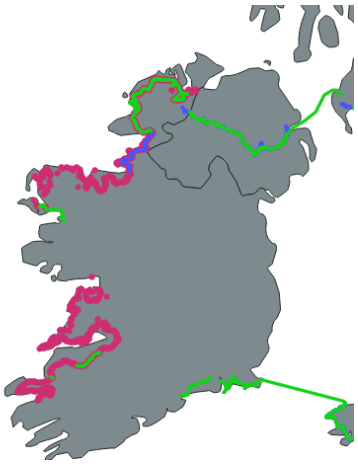
In the framework of the EuroVelo 1 analysis, several data are used: the route from the official EuroVelo website (Cf. source (1) - paragraph 2.2 of this document), the routes from the various organisations managing the route (Cf. source (2)), the route in France from ON3V (Cf. source (3)), the route route digitised by OpenStreetMap contributors that can be viewed and downloaded

from the waymarked trails site (Cf. source (4)) and the GPS survey points when a field survey has been carried out.

Despite the different sources used, the trail is not complete and there are big differences (see Map 1 and Map 2).



Map 1: Comparison of EV1 routes in Spain



Map 2: Comparison of EV1 routes in Ireland

2.3.2. First level breakdown

After having generated a continuous route the next step is to cut this route according to its natural boundaries. This cutting out is prior to any other cutting out. This can be an administrative boundary, a ferry crossing, a step-by-step breakdown. For the EuroVelo 1 study, the itinerary is divided up at each administrative boundary and at each ferry crossing.

QGIS processing

Add the administrative cutting layer within your QGIS project. Be careful, this layer must have a level of precision compatible with your layout (see map 3).

Use the "*intersection*" tool to cut the route with the administrative cutting layer.

If the intersection tool does not give a suitable result, use the tool to separate the entities by hand from the advanced digitisation toolbar by running the route in the North-South direction.

Once you have completed the split, add attributes to your parts of the route:

- Name: name of the route section
- Country: Country concerned
- Id_1: Indicate the order of the entities resulting from this first breakdown.

Use the "*order by expression*" tool to make the layer follow the order of Id_1 and *Convert the result of the treatment into a permanent layer* under the name "route by country".

Example for EuroVelo 1

The administrative breakdown used is that derived from NUTS 3 data to an accuracy of 1 m (s*ee Source (5) - paragraph 2.2 of this document*). This is not sufficient for the route layout and many "false breakdowns" are generated along the administrative border and at sea level. For this reason, a manual breakdown by country has been made. The NUTS 3 breakdown is only used at the stage of reconciliation with the context data.

EuroVelo 1 is divided into 9 parts.

Map 3: Itinerary on the edge of the administrative boundary cut out by mistake.

2.3.3. Breakdown into unit segments

QGIS treatment

Use the "*line division by maximum length*" tool with the maximum distance defined according to the size of the unit segments corresponding to the granularity of the study. Use the field calculator tool to define the "*id_segment*" field from the value $id. Convert the processing result into a permanent layer under the name "*unit segments*".

Example for EuroVelo 1

The maximum distance for the EuroVelo 1 study is 1000 m, which produced 10,762 segments of 1000 m or less.

2.3.4. Progress report

Layers produced:

- Continuous route
- Itinerary by country
- Unit segments

2.4. Reconciliation of segments with context data

Difficulty	Finding the processing to bring the context data closer to the route with the appropriate level of accuracy and method
Medium	Software QGIS 3.12
Input	Itinerary in unit segments + Geolocalised data to define the context
Output	Itinerary in unit segments with detailed context data

2.4.1. Possible types of joints

The route segment layers can be connected to the territory data in different ways. The choice of method depends on the purpose of the study and the representation of the link between the route and what the data represents. Here are the joining solutions used:

Joining attributes by "location": for each entity in the segment layer, the intersection with entities in another layer is checked. The information from the intersected entity is added to the segment entity. Example: define the administrative or common region for each of the segments. The advantage of this type of join is the speed of processing which makes it very useful to join a point layer with a surface layer; but one of the disadvantages is the need to have two layers with the same level of precision. For a joined line with surfaces, it is possible that the same line crosses several polygons.

Joining of attributes by "nearest": for each segment, we look for the N closest elements in another layer. A search distance limit and a number of elements to be searched for can be defined. This type of processing makes it possible, for example, to define for each segment the distance to

the nearest station or the nearest coast. The main advantage is that it is possible to compare data that do not have the same reference frame, the disadvantage is the possibility of multiple joins.

Joining of attributes by "summary location": This type of join allows to join to each entity in a layer A, the grouping of the elements present in a layer B. A grouping operation (sum, max, min, nb...) and one or more fields are selected. It is possible, for example, to count the number of hotels or the sum of hotel capacity for each region. The advantage is the grouping of information and the execution of an operation.

Example of join operations to collect data on unit segments:

Join by location	By closest	By summary location
Allows for each segment to collect information from a surface layer	*Allows you to collect the attributes and distance of the point element closest to the segment*	*Allows for each segment to calculate the number of cross tracks or other operations.*
example: For each segment, add the INSEE municipal data	*example: for each segment add the data of the nearest station*	*example: for each municipality add the sum of the capacity of the hotels present*

The data reconciliations carried out make it possible to take into account all types of context data. If you wish to use other context data, you can draw inspiration from the methods presented in the rest of the chapter. For example, this allows you to use INSEE data such as the permanent equipment database (BPE), the capacity of municipalities in tourist accommodation, the establishments in the SIRENE file, data from the national observatory for cycle routes and greenways (ON3V), etc.

2.4.2. Administrative regions

The transition from one country to another has been taken into account in the chapter "2.3.2 First level breakdown". The objective of this step is to take into account the administrative regions within the country in order to determine for each segment the region to which it belongs (without redrawing the unit segments).

Eurostat brings together and harmonises the administrative structure of European countries. Three scales are defined: NUTS 3 (equivalent to departments in France), NUTS 2 (equivalent to regions in France) and NUTS 1 (equivalent to a grouping of regions in France). The delineation of administrative boundaries is less precise than the route (see Map 3).

QGIS processing

Download the administrative boundary data layer with the best possible accuracy (*See source (5) - paragraph 2.2 of this document*). Integrate NUTS_RG_01M_2016_3857.shp which contains the administrative regions in polygon form and NUTS_RG_01M_2016_3857_LEVL_3.shp which contains only the NUTS 3 level one. Rename the first layer (NUTS_RG_01M_2016_3857.shp) to "NUTS_ALL_LEVEL" and then use the "*field calculator*" to generate the following fields in the second layer (NUTS_RG_01M_2016_3857_LEVL_3) :

- 3_NUTS_ID = "NUTS_ID".
- 3_NUTS_NAME = "NUTS_NAME".
- 2_NUTS_ID = left ("NUTS_ID" , 4)

- 2_NUTS_NAME = attributes (get_feature('NUTS_ALL_LEVEL' , 'NUTS_ID' , "2_NUTS_ID")) ['NUTS_NAME']]
- 1_NUTS_ID = left ("NUTS_ID" , 3)
- 1_NUTS_NAME = attributes (get_feature('NUTS_ALL_LEVEL' , 'NUTS_ID' , "1_NUTS_ID")) ['NUTS_NAME']]

After generating these fields, export the layer as "ADMIN_NUTS".

Use the "*join by location*" tool with unit segments as input layer and *admin_nuts* as the layer to be joined. Tick the *intersect* predicate, choose the previously created fields, and select the type of join *Take the attributes of the entity with the largest overlap*. Check the consistency of the results for selected segments and then convert to a permanent layer named *unit segments + admin*.

2.4.3 EuroVelo: crossing of other routes and progress status (line data)

The EuroVelo cycle routes are followed by the European Cyclists' Federation (ECF). The various routes have been mapped more or less precisely and their progress is also monitored at European level. By comparing this data with our itinerary, it is possible to define the parts that are common to several routes (for example: EuroVelo 1 and EuroVelo 6 share the same itinerary in Pays de la Loire) and the parts that have not yet been completed or have not yet been marked out (which has a direct impact on the number of people using the itinerary).

The routes are represented in the form of a line with the level of progress and the name of the route as an attribute. The route reference frame is not the same as the route of our itinerary, therefore the join by location between these routes and ours is impossible.

QGIS treatment

Open the file of the European Cyclists' Federation presenting the data of all EuroVelo (Cf. source (1) paragraph 2.2). Use the "*Group*" tool with the grouping fields corresponding to the route identifier and the progress status. Then use the "Buffer" tool to generate a 500 m buffer around each route section (the distance must be modified according to the difference in accuracy of the routes). Name the *network route buffer* result.

In order to retrieve the progress status from the EuroVelo files, use the "*select using an expression*" tool to select all the entities corresponding to your route (EuroVelo 1 in our case) in the route network buffer layer. Use the "*join by location*" tool with unit segments in the entry layer and the route network buffer in the layer to be joined. Tick the *selected entities only*, tick the "*cross*" predicate, select the field corresponding to the state of progress and select the type of join "*Take the attributes of the entity with the greatest overlap*". The result is your segment layer with the progress report from the EuroVelo data. Name this layer *result_1*.

In order to retrieve the sections shared with other EuroVelo routes, use the "*select using an expression*" tool to select all the entities that **do not correspond to your route in the route network** *buffer layer*. Use the "*join by location*" tool with *result_1* in the input layer, your *route network buffer layer* in the layer to be joined, tick the *selected entity(ies) only*, tick the predicate "inside", choose the field corresponding to the route identifier and select the type of join "one by one".

The result is your segment layer with two new fields, the status named status_ECF_2019 and the identifier of the routes having a common route named EuroVelo_en_commun. Check the consistency of the results for selected segments, then convert to a permanent layer named *unit segments + EuroVelo*.

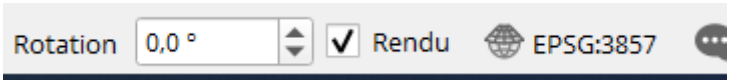
Note that if you wish to adapt the method with another GIS data source (e.g. for France, the ON3V database), it is at this stage that it should be taken into account.

2.4.4 Corine land-cover (surface data)

In Europe, a land use database is produced periodically (*see source (6) - paragraph 2.2 of this document*). It describes land use fairly precisely and makes it possible to know the type of space crossed by the segments. This data has been used for three types of information, land use at segment level, the distance between the segment and the sea, the distance between the segment and a river. The main feature of EuroVelo 1 is to follow the Atlantic coastline, which is a fairly strong feature in the context of the segments.

QGIS processing

Deactivate the rendering and then integrate the CorineLandCover layer into the QGIS project, this avoids too long a loading time. Set the minimum visibility scale to 1: 300000 then reactivate the rendering of the layers.



Use the '*extract by expression*' tool with expression `left ("Code_18", 2) = '52'` to generate a layer of maritime spaces and `left ("Code_18", 3) = '511'` to generate a layer of inland water spaces. Use the *'buffer'* tool over a distance of '· 20 m' to adjust the level of accuracy of the two layers in relation to the route. Use the '*join attributes by nearest'* tool with unit segments in the source layer, the maritime space layer in the layer to be joined, an arbitrary field for maritime space and start processing. Delete the attached field and the `feature_x`, `feature_y`, `nearest_x`, `nearest_y` fields and rename the `distance` field to `sea_distance`.

Use the tool "*join attributes by nearest*" with the previous result in the source layer, the *continental water* layer in the layer to join, an arbitrary field and start the processing. Delete the attached field and the fields `feature_x`, `feature_y`, `nearest_x`, `nearest_y` and rename the distance field to `river_distance`.

You should get the *unit segments* layer with additional fields named `sea_distance`, `river_distance`. Check the consistency of the results for selected segments and then convert to a permanent layer named *unit segments + CLC*.

2.4.5. Population data (raster data)

The data source used to define the population along the route is the global human settlement layer *(see source (7) - paragraph 2.2 of this document)*. The format is a raster and the resolution used is 250 m. The purpose of this step is to add for each segment the information on the population in the vicinity of the segment.

Please note: the method below works with population data in raster format. Another database for population data can be used. To use point or area data you can use the methods presented in the rest of the chapter.

QGIS processing

Generate a 5 km *buffer* around each unit segment. Use the "*zonal statistics*" tool with the population in the raster layer, the result of the buffer in the vector layer and the sum as the statistic to be calculated.

On the *unit segment* layer, make an attribute join with the previously generated layer to create the *unit segment + pop* layer. You must obtain the *unit segment* layer with an additional field indicating the total population at 5 km at the edge of the segment to be named population_5_km. Check the consistency of the result by identifying a segment and the population at 5 km around it.

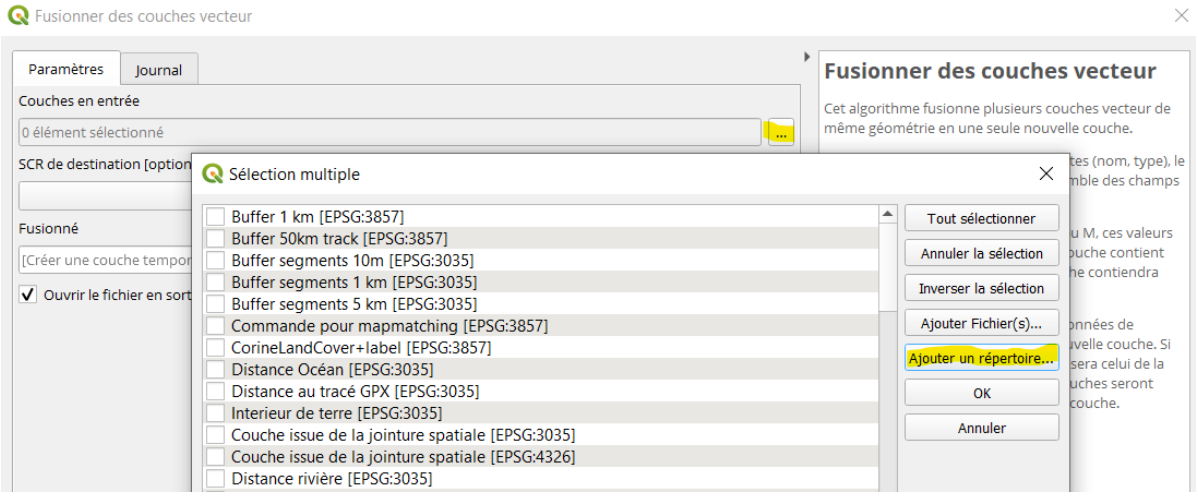
2.4.6. Railway stations and tourist accommodation (point data)

We use data from OpenStreetMap to complete our context analysis. OpenStreetMap is the free world map, it has been produced by a set of contributors, it is modifiable by all and also freely reusable. The level of quality of the data (accuracy, completeness, date updated) is not homogeneous and depends on locations and the involvement of local contributors.

We use this data source for two purposes: to know the location of tourist accommodation (hotels, campsites, etc.) and the typology of the road network along the route. All the elements mapped on OpenStreetMap are on the same basis, a significant filtering and refining work is necessary. Numerous online services make it easy to download OpenStreetMap data by area, region, or theme. We use the service offered by Géofabrik *(see paragraph 2.2. Data sources)* for its ease of use.

QGIS treatment

Download the data from Geofabrik in Shape format for all regions crossed *(Cf. source (8) paragraph 2.2)*. For each region, there is a compressed folder containing 15 shapefiles that separates the entities by type. Unzip them into one folder per type. Without loading the layers in QGIS, use the "*merge vector layers*" tool to generate a single layer grouping together the gis_osm_pois_free_1.shp layers of all the downloaded regions and reproject it in your coordinate system. Do the same treatment for gis_osm_pois_free_a_1.shp which contains the POIs as polygons, gis_osm_transport_a_free_1.shp which contains the transport locations as polygons and gis_osm_transport_free_1.shp which contains the transport locations as points[1].



Some hotels on OSM are represented as a polygon (a minority) which corresponds to their building. To complete the process, it is necessary to transform these polygons into dots in order to integrate them into the same layer as the rest of the hotels represented by dots. The polygon layer resulting from the merging of gis_osm_pois_a_free_1.shp into a point layer with the *"centroids*" tool. Copy the entities in the result layer to integrate them in the layer resulting from

[1] Names of the downloadable files from Geofabrik can change from time to time.

the merging of gis_osm_pois_free_1.shp. Use the "*extract by expression*" tool on POIs merged with the "fclass" IN formula ('hotel', 'camp_site', 'hostel', 'bed_and_breakfast'). Check the consistency of the results for selected points of interest and then convert to a permanent layer called *accommodation*.

Use the "*buffer*" tool on the unit segment layer with a distance of 5 km. Use the '*attach attributes by location (summary)*' tool with the previously created buffer in the source layer, in the layer to merge the accommodation layer and the operation counts on an arbitrary field. On the *unit segments* layer, make an attribute join with the layer previously generated from the id_segment field in order to add the attribute corresponding to the number of tourist accommodations at 5 km named hotel-camp_5km.

Transform the polygon layer resulting from the merging of gis_osm_transport_a_free_1.shp into a point layer with the "*centroids*" tool. Copy the entities in the result layer to integrate them in the layer resulting from the merging of gis_osm_transport_free_1.shp. Use the '*extract by expression*' tool with the 'fclass' IN ('railway_station') formula. Check the consistency of the results for selected stations and then convert to a permanent layer named *stations*.

Use the "*join attributes by nearest*" tool with *unit segments* in source layer and stations in layer to be joined, indicate 20 km as maximum distance. From the layer containing tourist accommodation information by segment (see above) make an attribute join from the field id_segment to add the attribute corresponding to the distance to the nearest station and name it distance_transport.

You should get the *unit segment* layer with additional fields named hotel-camp_5km and distance_transport. Check the consistency of the results for selected segments and then convert to a permanent layer named *unit segments + tourism*.

2.4.7. Roads (line data)

QGIS treatment

As in the previous chapter, use the data downloaded via Geofabrik and generate a single layer grouping together the gis_osm_roads_free_1.shp layers of the downloaded regions and reproject it in your coordinate system. Add this layer to your project and name it *Roads*.

The treatment of roads is more complex than that of point data (previous chapter), the method must be adapted according to the level of precision and the source of the route. Use the "*buffer*" tool on the unit segment layer with a distance of 10 m. Use the "*intersection*" tool with the result of merging the layers gis_osm_roads_free_1.shp as the source and the previous result as the overlay layer, choose *fclass* and *maxspeed* in the input fields to be kept and *id_segment* in the fields to be kept. Use "*extract by expression*" on the result with the formula $length > 25 (metres) to remove too small road sections.

Use the "*field calculator*" to create the following fields containing the length (in metres) per type:

- cycleway = CASE WHEN "fclass" = 'cycleway' THEN $length ELSE 0 END
- path = CASE WHEN "fclass" in ('bridleway', 'footway', 'path') THEN $length ELSE 0 END
- track = CASE WHEN "fclass" in ('track', 'track_grade1', 'track_grade2', 'track_grade3', 'track_grade4', 'track_grade5') THEN $length ELSE 0 END
- street = WHEN ("fclass" in ('living_street', 'pedestrian', 'residential', 'service') OR ("maxspeed" < 60) THEN 'street' THEN $length ELSE 0 END
- road = WHEN "fclass" in ('secondary', 'secondary_link', 'tertiary', 'tertiary_link', 'unknown') and "maxspeed" < 100 THEN $length ELSE 0 END

- other = CASE WHEN "cycleway" = 0 AND "path" = 0 AND "street" = 0 AND "road" = 0 THEN $length END

Use "*Attach attributes per location*" *(summary)* with the 10 m buffer in the source layer, the previous result in the layer to be attached, the previously created fields to be summarised and the sum as a summary to be calculated. The result gives the road linear by type for each unit segment (sum_cycleway, sum_path, sum_track, sum_street, sum_road, sum_other). Check the consistency of the results for selected segments and then convert to a permanent layer called unit segments + roadway.

2.4.8. Merging the results

QGIS treatment

Load the layers *unit segments + admin, unit segments + EuroVelo, unit segments + CLC, unit segments + pop, unit segments + tourism, unit segments + roads*.

Use the layer *unit segments* and in the layer, properties make joins to collect the fields listed in the table below from the loaded layers. Export the layer as an Excel file named unit segments + context.xlsx.

This file should contain the following fields:

Data source	Field names	Chapter	Type information
NUTS Eurostats	3_NUTS_ID, 3_NUTS_NAME, 2_NUTS_ID, 2_NUTS_NAME, 1_NUTS_ID, 1_NUTS_NAME	2.4.2	For each segment For each segment Identifier and names for 3 territorial scales.
EuroVelo	EuroVelo_en_commun, status_ECF_2019	2.4.3	Progress category and route identifier
Corine land·cover	sea_distance, river_distance	2.4.4	Distance in km
Global Human Settlement	population_5_km	2.4.5	Surrounding population
OpenStreetMap	hotel·camp_5km, distance_transport	2.4.6	Number of accommodations Distance to a train station
OpenStreetMap	sum_road, sum_street, sum_path, sum_track, sum_cycleway, sum_other	2.4.7	Total linear by type of road for each segment

2.5. Grouping of unit segments into unified sections

Difficulty	Finding the appropriate formulas to define the denominations of each section according to the type of variable studied and the distribution.
Medium	Spreadsheet and QGIS
Input	Itinerary in unit segments with detailed context data Excel file "unit segments + context.xlsx".
Output	Itinerary in coherent sections resulting from the grouping of unitary segments Excel file "unit segments + section id.xlsx".

Each unit segment is grouped together with the preceding or following unit segments according to the continuity of the context. The first criterion for grouping unit segments is their geographical continuity (two non-contiguous segments are not grouped together). This point is assessed before the coherence of the values of the context. We use for this purpose incremental calculations which are allowed by the use of the spreadsheet. We analyse the segments one after the other and follow the evolution of the context. If there is a strong change in the context, a break is defined, and a new section begins.

Two types of breaks have been defined:

- "**Hard cuts**" are changes in the context that lead to a mandatory cut. These modifications correspond to a change of administrative region, the crossing of another EuroVelo, a change in the status of EuroVelo (planned route, validated route, continuous route, marked route, certified route) or the crossing of an ocean (passage in a ferry).
- The "**soft cut**" which results in a cut-off if the segment exceeds a certain size. These cuts are determined by a contrast in population density, a strong change in the accommodation offer, a change in the type of infrastructure, passing close to a station, crossing a river or all these changes combined.

2.5.1. Definition of "hard cut

The values analysed are categorical values (as opposed to quantitative values). Changing the value between the current segment and the next segment results in a cut-off at the end of the current segment.

Spreadsheet processing

Open the file unit segments + context .xlsx with a spreadsheet program. Create an Admin_cut column which will take the value 1 when the segment is an end of section. Complete the column with the formula: SI (current 3_NUTS_ID = next 3_NUTS_ID ;0 ;1)

Create a EuroVelo_cut column which will take the value 1 when the segment is an end of section. Complete the column with the formula: SI (current EuroVelo_in_community = next EuroVelo_in_community ;0 ;1)

Create a column EuroVelo_status_cut which will take the value 1 when the segment is an end of section. Complete the column with the formula: SI (current status_ECF_2019 = next status_ECF_2019 ;0 ;1)

Create a column sea_cut which will take the value 1 when the segment is maritime. Complete the column with the formula SI (ET (sea_distance = 0; sum (sum_cycleway; sum_path; sum_road; sum_street; sum_track; sum_other) < 500) ;1 ;0). In this formula the value 500 corresponds to a distance in metres.

Create an empty manual_cut column but which is intended to collect manually defined cuts by adding a 1. These manual cuts can be useful to segment sections that are too long, to take into account a change of environment or to make them consistent with a pre-existing route breakdown (e.g. breakdown resulting from a traffic study).

Create a hard_cut column which will take the value 1 when the section is cut from the previous variables. Complete the column with the formula SI(OR (admin_cut = 1; EuroVelo_cut = 1; EuroVelo_status_cut = 1; current sea_cut <> next sea_cut; manual_cut = 1) ;1 ;0)

2.5.2. Definition of "soft cut" infra, river and station proximity

As mentioned above, soft cuts are linked to quantitative data, such as distance or change in density. The definition of "soft cuts" linked to the proximity of a station therefore falls within this field. For those related to the change in infrastructure typology, it is necessary to transform this categorical information into quantitative data by assigning a score to each segment. This score is calculated according to the length of each type of infrastructure of which it is composed.

The "soft cuts" are then determined by analysing the level of contrast of the quantitative values observed in relation to a given threshold. The generic formula for calculating the contrast is (value studied - value around) / average value.

Spreadsheet processing

In the file *unit segments + context.xlsx*, create four columns to handle the infrastructure variation:

sum_infra contains the sum of the road linear detected at 10 m, complete the column with the formula: SUM (sum_cycleway; sum_path; sum_track; sum_street; sum_road; sum_other)

note_mid_infra contains the rating of the road along the km, complete the column with the formula: (sum_cycleway*5; sum_path*4; sum_track*3; sum_street*2; sum_road*1 /sum_infra +1)

contrast_infra which follows the variation of the infra notation between segments. In order to avoid taking into account point variations, we evaluate both the variation of the current segment in relation to the next one and the variation of the 5 previous segments in relation to the next 5. Complete the column with the formula: ABS(current infra_mean_rating - next infra_mean_rating + AVERAGE (previous 5 previous infra_mean_rating) - AVERAGE (next 5 next infra_mean_rating))/2

discontinuity_infra contains 1 if there is a significant variation identified that could lead to a break. Complete the column with the formula: SI (ET (fra_contrast > 1.4; fra_contrast = MAX (fra_contrast 3previous :3following)); 1; 0)

Add a new river column that contains 1 if there is a river crossing to cause a cutoff. Complete the column with the following formula SI(ET(river_distance = 0; current river_distance < MIN (previous 4 river_distance); current river_distance -1 < MIN (next 4 river_distance));1;0).

Add a new column proximity_station to contain information about the proximity of a station being less than 2 km away. Complete the column with the following formula SI(ET(transport_distance < 2000; transport_distance < MIN(transport_distance 4previous and 4 following));1;0). In this formula the value 2000 corresponds to a distance in metres.

2.5.3. Definition of "soft cut" "population" and "hotel

The number of dwellings and the population per segment have a very high level of variation between towns, villages and rural areas. Our aim is to smooth out this variation and adapt the contrast calculations to take account of variations in the number of accommodation or population in a rural or urban context. A variation of 10 hotels for a segment with 15 hotels is larger than if the segment has 150 hotels. After several trials, we chose to calculate the contrast using the square root of the square of differences. This choice was inspired by the mathematical approach of transforming data to make them closer to a normal distribution.

Spreadsheet processing

Still in the file unit segments + context.xlsx, create 4 new columns for the calculation of the variation of the population and the variation of accommodation.

contrast_hotel contains the variation in the number of accommodations (hotels and campsites). If hotel·camp_5km is in column AH then the formula for row 6 is as follows: RACINE((((AH10-AH2)+(AH7-AH6)*3)/2*((AH10-AH2)+(AH7-AH6)*3)/2) / (1+MAX(AH2;AH5;AH6;AH7;AH10))). Then apply the formula to the whole column.

discontinuity_hotel contains 1 if there is a significant variation identified that could lead to a break. If hotel·camp_5km is in column AH then the formula for row 6 is as follows: SI(ET(MIN(AH2:AH10)>2; OR(ET(hotel_contrast > 2; hotel_contrast > MAX(hotel_contrast 4previous and 4following))));1;0)

contraste_pop contains the variation of the population along the route. If population_5km is in column AA then the formula for row 6 is as follows: ABS(((AA11-AA2) + (AA8-AA5)) / 2) / (1+MAX(AA2;AA5;AA8;AA11)). Then apply the formula to the whole column.

discontinuity_pop contains 1 if there is a significant variation identified that could lead to a cut-off. If population_5km is in column AA then the formula for row 6 is as follows: IF(ET(ET(MIN(AA2:AA10)>200;OR(ET(pop_contrast>0.6; pop_contrast > MAX(pop_contrast 4previous and 4following))));1;0).

2.5.4. Grouping of "soft cuts

After having defined the elements of discontinuities that can cause a cut, a change of section, we group them together and then we check that no hard cut occurs on a segment identified as a discontinuity in order to assign the soft cuts. Thus, if no "hard cut" is present in the 5 previous segments or in the next 5 segments, then the discontinuities ("soft cut") are analysed. To define the soft cuts, we study the discontinuities for the current segment, the 2 preceding segments and the 2 following segments, to identify the cut segment with the largest discontinuities.

Spreadsheet processing

Create a sum_soft column to group the different discontinuity indicators. Discontinuities being defined as the sum of the discontinuity_infra, discontinuity_pop, discontinuity_hotel and river. Complete the column with the following formula SUM (current discontinuities) + SUM (discontinuities 2previous; discontinuities 2following)/2 + current_station_proximity / 3.

Create a soft_cut column that changes to 1 if there is a break related to the discontinuities. Complete the column with the following formula IF(ET(ET(hard_cut 5previous and

5following) = 0; sum_soft >= 0.5; sum_soft >= MAX(sum_soft 5previous and 5following)); 1; 0)

2.5.5. Definition of the section identifier

After defining the "hard cut" and the "soft cut", they are brought together to define the unified section identifiers.

Spreadsheet processing

Create an all_cut column. Complete the column with the following formula IF (OR(hard_cut = 1; soft_cut = 1); 1; 0)

Create a section id_section column that will contain the section identifier and allow merging segments with the same section identifier. The presence of a break will cause the section id for the next unit segment to be changed. Complete the column with the following formula previous id_section + previous all_cut.

Check that the result is consistent, if necessary, adjust the cut rules, if sections seem too long add manual cuts. Then save the result as unit_segments + section_id.xlsx and unit_segments + section_id.csv.

2.5.6. Finalization of the treatment

A final treatment is necessary to produce the merged segments in geographical form, it must be carried out in QGIS.

Processing on QGIS

Open QGIS and add the *unit segment* layer and the CSV file obtained in the previous step. In the properties of the *unit segments* layer, make a join with the CSV from the id_segment field to add the fields id_section, all_cut, and sea_cut. Use the "*extract by expression*" tool with the formula sea_cut = 0 in order to discard the maritime sections.

Use the result of the previous processing to generate the segments merged into sections using the "*collect geometries*" tool and using the unique id_section identifier field. Use the "*line merge*" tool and then convert to a permanent layer named *route section* in the desired format and projection. Check that the result is in accordance with your expectations, in particular by checking the number of sections created compared to the number of non-marine sections in your file segments_units + section id.xlsx.

2.6. Clustering - Grouping of sections

Difficulty	Finding the right variables to study in order to respond to the similarity you want to address
Medium	R
Input	Itinerary in coherent section: unit_segments + id section.csv
Output	Route sections classified into different groups based on their similarities: sections_cluster.csv

From the sections of routes with homogenous characteristics, we will define "families" of sections that are similar (example: route sections with a high population density and little development). To define the families, we need to choose variables on which to study this proximity. In the case of EuroVelo 1, the values studied are: the population, the number of tourist beds in the vicinity of the route section and the share of exclusive right-of-way infrastructure of the section.

2.6.1. Preparation of the data

To study the similarity of the previously defined sections, the choice was made to focus on 3 indicators: the number of tourist beds available, the population and the share of infrastructure with exclusive right-of-way.

The number of tourist beds

For the whole segmentation process the data source used for accommodation is OSM. The disadvantage of this data source is that it does not give a view of the supply in terms of the number of establishments[2]. To get a more accurate view of the supply, it is necessary to convert this data into the number of beds. To do this, we recommend using Eurostat data at NUTS2 level to calculate average ratios of beds per type of accommodation and per geographical area, which will enable the OSM data to be transformed. The Eurostat data are available for downloading from the following address:

https://ec.europa.eu/eurostat/fr/web/tourism/data/database

The dataset includes the capacity in number of beds and in number of establishments per NUTS2 code for hotels, camping sites and short-stay establishments (e.g. youth hostels). From these data, a table must be constructed to calculate the average number of beds per type of accommodation and for each geographical area. The ratios obtained will be applied to the data from OSM. The result will feed the Bedplaces field.

The population

To carry out the similarity analysis of the sections, it is necessary to identify for each section the population at a distance of 5km on either side of the route. The data summarised by section is recorded in the Pop field.

The share of the route in exclusive right-of-way

Based on the km_segregate data used to define the sections, the weight of km_segragate that makes up each section is calculated. The result obtained will feed the Ratio_km_segragte field.

---

[2] In France, the national statistical service Insee offers a detailed file at a city level, but it is not available for all european countries. That is why OSM data (in number of establishment) have been chosen to conduct this transnational analysis.

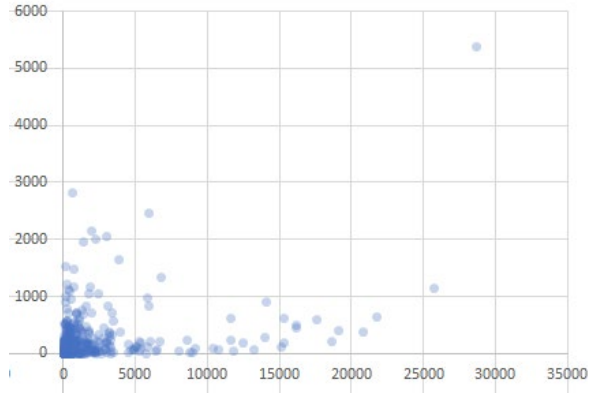
The result is a table that contains the following fields:

- Pop: the population at 5 km at the edge of each section.
- Bedplaces: the number of beds in the accommodations around the route. This value is calculated from OpenStreetMap and the average number of beds per establishment in each region from Eurostat data.
- Ratio_km_segregate: corresponds to the percentage of exclusive right-of-way detected along the route.

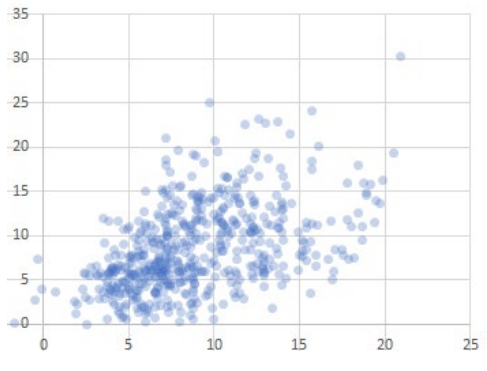
Save the table as sections_context.csv

2.6.2. Principle of value distribution

To understand the clustering approach, two variables can be represented on a scatter plot. The purpose of clustering is to group together points that are close to each other on this diagram. If we look at data such as population and number of tourist beds (see Figure 1 below), we can see that there is an irregular distribution on the diagram because of the distribution of this data. A transformation is necessary to adapt the distribution of the data series. In the second diagram (see Figure 2), we see the data after transformation of each of the variables via a Box-Cox transformation[3].



Graph 1: Distribution of sections according to the number of tourist beds and population



Graph 2: distribution of sections after data transformation

This approach makes it possible to obtain coherent groups which are not defined by thresholds in absolute values but use algorithmic processing to group them homogeneously.

[3] *Why using a box-cox transformation? Further explanation: https://www.minitab.com/fr-fr/Published-Articles/En-quoi-la-transformation-Box-Cox-peut-vous-%C3%AAtre-utile/*

2.6.3. Simplified clustering via R

A plugin named Mclust allows clustering in a simplified way on R. This plugin will automatically choose the type of data transformation to be applied, then produce clustering with the different known methods and choose the one that presents the best result. Beforehand, it is necessary to specify the variables to be studied and the number of clusters to be produced.

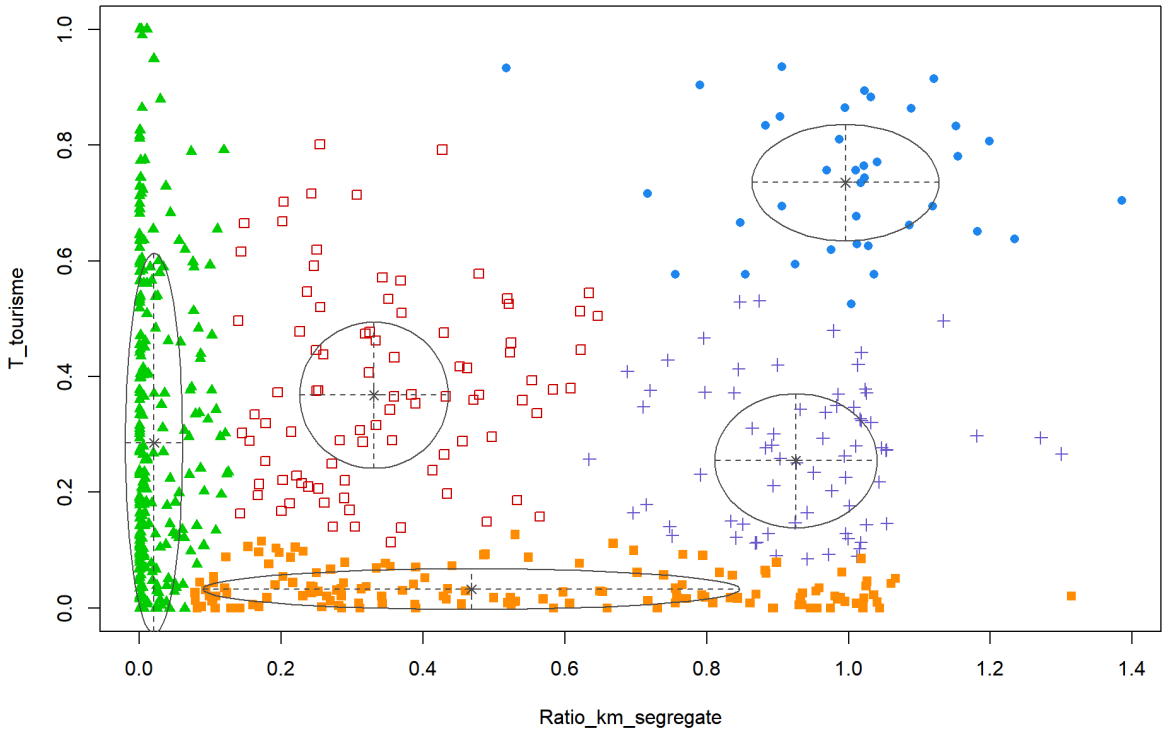
Processing on R Studio

Open Rstudio, create a new project and add in the folder the file section_ev1.xlsx which contains the following columns: "Pop", "Bedplaces", "Ratio_km_segregate" and "Country".

Using the interface or via the console, install the following packages and their dependencies: tibble, mclust and dplyr

Then execute the following code with a possible adaptation of the variables to be modelled and the number of groups to be produced.

```
# import of libraries
library(tibble)
library(mclust)
library(dplyr)

# data import and generation of new indicator
ev1 <-as_tibble(read.csv("sections_context.csv"))
X <- ev1[,c("Pop", "Bedplaces", "Ratio_km_segregate")] # Only 3 columns taken into account

# Calculation of column T_tourism, for the rate of tourism among pop+lit
X <- add_column(X, T_tourism = X$Bedplaces/ (X$Bedplaces + X$Pop))
X$T_tourism [is.nan(X$T_tourism)] <- 0

# Choice of variables to be studied from Pop, Bedplaces, Ratio_km_segregate, T_tourism
data <- dplyr::select(X, Ratio_km_segregate, T_tourism)

# Number of groups to be created (3:5 lets the algo choose between 3 and 5 groups).
n_groups <- 3:5

# Template creation
mod <- Mclust(data, G = n_groups)
plot(mod, what = "classification") # Representation of clustering

# Creation of result for the current result
result <- tibble(cluster_id = mod$classification)
result <- cbind(ev1,result)
write.csv(result, "sections_cluster.csv")
```



The result of the processing is a scatterplot diagram with a grouping represented by the colour and a result table saved in the project folder under the name sections_cluster.csv. This table contains the basic data and a new column named cluster_id which gives the identifier of the group or cluster section by section.

By grouping together sections with similar backgrounds, one-off surveys can be extrapolated to estimate patronage and economic impact at the scale of the entire route.

3. Criteria for success

3.1 Data quality

The quality of the data is one of the first criteria for the success of the approach. The accuracy of the route is one of the most important obstacles, particularly for the analysis of the road network. In fact, by shifting the route by 50 m, the type of development can be completely different. To overcome these difficulties, it is advisable to use the route from OpenStreetMap (possibly contribute to it if an update is necessary) or to have a GIS base that already describes all the infrastructure used by the route (such as the ON3V base in France).

3.2. Verification of treatments

The method presented in this guide has been used only once for the analysis of the EuroVelo 1 route. Map processing can sometimes lead to improbable results due to human error or a difference in data accuracy. The verification of the processing for certain sections should not be

neglected, otherwise the result and the analyses produced will be unusable. Sometimes some manual corrections are necessary, as was the case in the EuroVelo 1 analysis.

3.3. Adaptation of treatments

The treatments presented in this guide have been developed to adapt to the territorial contexts present along EuroVelo 1. All these treatments can be modified to simplify, analyse the context in a different way or use other data sources. Before any modification, it is essential to understand the reasons for each of the treatments and to compare the effect of the methodological adaptations.

4. Conclusion

The manual segmentation method usually used is based on the knowledge or interpretation of the territorial context by a person in charge of the economic analysis. The method presented in this guide, based on territorialized data, has been developed to make an economic analysis on a 10,000 km route crossing different countries. This method can be used for European routes or for smaller ones.

The advantage of a method based on data processing is to avoid the "human error" factor and to use only quantifiable information. On the other hand, one of the limitations of this method is the availability of certain data to accurately describe the context. There are no data available to describe the attractiveness of the landscape, the level of traffic or the level of respect for bicycle users. With a little thought and knowledge of the available data, indicators can compensate for this lack. For example, the tourist character of an area can be assessed through the density of tourist beds.