BikeDNA Report

Copenhagen: GeoDanmark

Report generated on: 2022-12-09 10:15:40

1a. Load and Process OSM data

This notebook:

- Loads the polygon defining the study area and then creates a grid overlay for the study area.
- Downloads street network data for the study area using OSMnx.
- Creates a network only with bicycle infrastructure (with queries defined in `config.yml`).
- Creates additional attributes in the data to be used in the analysis.

Sections

- Load data for study area and define analysis grid
- Load and process OSM data

Load data for study area and define analysis grid

This step:

- Loads settings for the analysis from the config file.
- Reads data for the study area.
- Creates a grid overlay of the study area, with grid cell size as defined in configuration file `config.yml`.

Read in data for study area

The study area is defined by the user-provided polygon. It will be used for the computation of *global* results, i.e. for the entire study area.

Create analysis grid

Load and process OSM data

This step:

- Downloads data from OpenStreetMap using OSMnx.
- Projects the data to the chosen CRS.
- Creates a subnetwork consisting only of bicycle infrastructure.
- Classifies all edges in the bicycle network based on whether they are protected or unprotected bicycle infrastructure, how they have been digitized, and whether they allow for bidirectional travel or not

- Simplifies the network *(to read more about the modified OSMnx simplification (Boeing, 2017) used here, we refer to this GitHub repository which contains both the simplification functions, explanation of the logic and a demonstration)*.
- Creates copies of all edge and node datasets indexed by their intersecting grid cell.

1b. Intrinsic Analysis of OSM Bicycle Network Data

This notebook analyses the quality of OSM data on bicycle infrastructure for a given area. The quality assessment is *intrinsic*, i.e. based only on the input data set, and making no use of information external to the data set. For an extrinsic quality assessment that compares the OSM data set to a user-provided reference data set, see the notebooks 3a and 3b.

The analysis assesses the *fitness for purpose* (Barron et al., 2014) of OSM data for a given area. Outcomes of the analysis can be relevant for bicycle planning and research - especially for projects that include a network analysis of bicycle infrastructure, in which case the topology of the geometries is of particular importance.

Since the assessment does not make use of an external reference dataset as the ground truth, no universal claims of data quality can be made. The idea is rather to enable the those working with bicycle networks based on OSM to assess whether the data is good enough for their particular use case. The analysis assists in finding potential data quality issues, but leaves the final interpretation of the results to the user.

The notebook makes use of quality metrics from a range of previous projects investigating OSM/VGI data quality, such as Antoniou & Skopeliti (2015), Fonte et al. (2017) and Fester et al. (2020).

Familiarity required
For a correct interpretation of some of the metrics for spatial data quality, some familiarity with the area is necessary.

Sections

Data completeness

Network Density

In this setting, network density refers to the length of edges or number of nodes per square kilometer (i.e., the definition of network density usually used when looking at street networks, which is distinct from the definition usually found in graph theory). Network density without comparing to a reference dataset does not in itself indicate spatial data quality. For anyone familiar with the study area, network density can however indicate whether parts of the area appear to be under- or over-mapped and is thus included here.

Method

The density here is not based on the geometric length of edges, but instead on the computed length of the infrastructure. For example, a 100-meter-long bidirectional path contributes with 200 meters of bicycle infrastructure. With `compute_network_density`, the number of elements (nodes; dangling nodes; and total infrastructure length) per unit area is calculated. The density is computed twice: first for the study area for both the entire network ('global density'), then for each of the grid cells ('local density'). Both global and local densities are computed for the entire network and for respectively protected and unprotected infrastructure.

Interpretation

Since the analysis conducted here is intrinsic, i.e., it makes no use of external information, it cannot be known whether a low-density value is due to incomplete mapping, or due to actual lack of infrastructure in the area. However, a comparison of the grid cell density values can provide some insights, for example:

- lower-than-average infrastructure density indicates a locally sparser network
- higher-than-average node density indicates a that there are relatively many intersections in the grid cell
- higher-than-average dangling node density indicates that there are relatively many dead ends in the grid cell

Global network density

Local network density

Densities of protected and unprotected infrastructure

OSM tag analysis

For many practical and research purposes, more information than just the presence/absence of bicycle infrastructure is of interest. Information about e.g., the width of the infrastructure, speed limits, streetlights etc. can be of high relevance, for example when evaluating the bike friendliness of an area or individual network segment. The presence of these tags (describing attributes of the bicycle infrastructure) is however highly unevenly distributed in OSM, which poses a barrier to evaluations of bikeability and traffic stress. Likewise, the lack of restrictions on how OSM features can be tagged often result in conflicting tags, which undermines the evaluation of cycling conditions.

The section includes analyses of a. missing tags (edges with tags that lack information), b. incompatible tags (edges with tags labelled with two or more contradictory tags), and c. tagging patterns (the spatial variation of what tags are being used to describe bicycle infrastructure).

Note that for the evaluation of tags, the non-simplified edges should be used to avoid issues with tags that have been aggregated in the simplification process.

Missing tags

The information that is required or desirable to obtain from the OSM tags will depend on the use case - for example, the tag `lit` for a project looking at light conditions on cycle paths. The workflow below allows to quickly analyze the percentage of network edges that have a value for the tag of interest.

Method

We analyze all tags of interest as defined in the `existing_tag_analysis` section of `config.yml`. For each of these tags, `analyze_existing_tags` is used to compute the total number and the percentage of edges that have a corresponding tag value.

Interpretation

On study area level, a higher percentage of existing tag values in principle indicates a higher quality of the data set; however, note that this is different from an estimation of whether the existing tag values are truthful. On grid cell level, lower-than-average percentages for existing tag values can indicate a more poorly mapped area. However, note that the percentages are less informative for grid cells with a low number of edges: for example, if a cell contains one single edge that has a tag value for `lit`, the percentage of existing tag values is 100% - but given that there is only 1 data point, this is less informative than, say, a value of 80% for a cell that contains 200 edges.

Global missing tags

Local missing tags

Incompatible tags

Given that the tags in OSM data lack coherency at times and there are no restrictions in the tagging process (cf. Barron et al., 2014), incompatible tags might be present in the data set. For example, an edge might be tagged with the following two contradicting key-value pairs: `bicycle_infrastructure = yes` and `bicycle = no`.

Method

In the `config.yml` file, a list of incompatible key-value pairs for tags in the `incompatible_tags_analysis` is defined. Since there is no limitation as to which tags a data set could potentially contain, the list is, by definition, non-exhaustive, and can be adjusted by the user. In the section below, `check_incompatible_tags` is run, which identifies all incompatibility instances for a given area, first for study area level and then for grid cell level.

Interpretation

Incompatible tags are an undesired feature of the data set and render the corresponding data points invalid; there is no straightforward way to resolve the arising issues automatically, making it necessary to either correct the tag manually or to exclude the data point from the data set. A higher-than-average number of incompatible tags in a grid cell suggests local mapping issues.

Global incompatible tags (total number)

Local incompatible tags (per grid cell)

Plotting incompatible tag geometries

Tagging patterns

Identifying bicycle infrastructure in OSM can be tricky due to the many different ways in which the presence of bicycle infrastructure can be indicated. The OSM Wiki is a great resource for recommendations for how OSM features should be tagged, but some inconsistencies and local variations do remain. The analysis of tagging patterns allows to visually explore some of the potential inconsistencies.

Regardless of how the bicycle infrastructure is defined, examining which tags contribute to which parts of the bicycle network allows to visually examine patterns in tagging methods. It also allows to estimate whether some elements of the query will lead to the inclusion of too many or too few features.

Likewise, 'double tagging' where several different tags have been used to indicate bicycle infrastructure can lead to misclassifications of the data. For this reason, identifying features that are included in more than one of the queries defining bicycle infrastructure can indicate issues with the tagging quality.

Method

The section below first plots individual subsets of the OSM data set for each of the queries listed in `bicycle_infrastructure_queries`, as defined in the `config.yml` file. The subset defined by a query is the set of edges for which this query is True. Since several queries can be True for the same edge, the subsets can overlap. In the second step below, all overlaps between 2 or more queries are plotted (i.e. all edges that have been assigned several, potentially competing, tags).

Interpretation

The plots for each tagging type allow for a quick visual overview of different tagging patterns present in the area. Based on local knowledge, the user may estimate whether the differences in tagging types are due to actual physical differences in the infrastructure or rather an artefact of the OSM data. Next, the user can access overlaps between different tags; depending on the specific tags, this may or may not be a data quality issue. For example, in case of `'cycleway:right'` and `'cycleway:left'`, having data for both tags is valid, but other combinations such as `'cycleway'='track'` and `'cycleway:left=lane'` gives an ambiguouos picture of what type of bicycle infrastructure is present.

Tagging types

Multiple tagging

Network topology

This section explores the geometric and topological features of the data.

These are, for example, network density, disconnected components, dangling (degree one) nodes; it also includes exploring whether there are nodes that are very close to each other but do not share an edge - a potential sign of edge undershoots - or if there are intersecting edges without a node at the intersection, which might indicate a digitizing error that will distort routing attempts on the network.

Due to the fragmented nature of most networks of bicycle infrastructure, many metrics, such as missing links or network gaps, simply reflect the true extent of the infrastructure (Natera Orozco et al., 2020). This is different for car networks, where e.g., disconnected components could more readily be interpreted as a data quality issue.

Therefore, the analysis only takes very small network gaps into account as potential data quality issues.

Subsections:

- Simplification outcome
- Dangling nodes
- Under/Overshoots
- Missing intersection nodes

Simplification outcome

When converting a set of geocoded linestrings (polygonal chains) to graph format, not all vertices (nodes) are of equal meaning. For geometry of the infrastructural element, all nodes are needed as an ordered list. For the topology of the network, however, only those nodes that are endpoints or intersection points with other edges are needed, while all other (so-called 'interstitial') nodes do not add any information. To compare the structure and true ratio between nodes and edges in a network, a simplified network representation which only includes nodes at endpoints and intersections, or where the value of important attributes changes, is required. Therefore, in notebook 1 the bicycle network was simplified by removing all interstitial nodes from the graph object (retaining, however, the complete node lists in the geometry attribute of each edge). An additional advantage of simplifying the network is the resulting substantial reduction of the number of nodes and edges, which makes computational routines much faster.

Comparing the degree distribution for the networks before and after simplification is a quick sanity check for the simplification routine. Typically, the big majority of nodes in the non-simplified network will be of degree two; in the simplified network, however, most nodes will have degrees other than two. Degree two nodes are retained in only two cases: if they represent a connection point between two different types of infrastructure; or if they are needed in order to avoid self-loops (edges whose start and end points are identical) or multiple edges between the same pair of nodes.

As part of the simplification routine, in cases where there are several edges between the same pair of nodes ('parallel edges' or 'multiedges'), only one of the edges is retained. Within the routine, the number edges removed in this way are counted.

Method

The degree distributions before and after simplification are plotted below.

Interpretation

Typically, the degree distribution will go from high (before simplification) to low (after simplification) counts of degree two nodes, while it will not change for all other degrees (1, or 3 and higher).

Further, the total number of nodes will see a strong decline. If the simplified graph still maintains a relatively high number of degree two nodes, or if the number of nodes with other degrees changes after the simplification, this might point to issues either with the graph conversion or with the simplification process.

Dangling nodes

Dangling nodes are nodes of degree one - in other words, nodes that have only one single edge attached to them. Most networks will naturally contain a number of dangling nodes. Dangling nodes can occur at actual dead-ends (representing a cul-de-sac) or at the endpoints of certain features (e.g., when a bicycle path ends in the middle of a street). However, dangling nodes can also occur as a data quality issue in case of over/undershoots (as described in detail in the next section). The number of dangling nodes in a network does to some extent also depend on the digitization method, as shown in the illustration below.

Therefore, the presence of dangling nodes is in itself not a sign of low data quality. However, a high number of dangling nodes in an area that is not known for suffering from many dead-ends can indicate digitization errors and problems with edge over/undershoots.

Dangling nodes occur where road features end (left), but when separate features are joined at the end (right), there will be no dangling nodes

Method

Below, a list of all dangling nodes is obtained with the help of `get_dangling_nodes`. Then, the network with all its nodes is plotted. The dangling nodes are shown in purple; all other nodes are shown in black.

Interpretation

We recommend a visual analysis in order to interpret the spatial distribution of dangling nodes, with particular attention to areas of high dangling node density. It is important to understand where dangling nodes come from: are they actual dead-ends or digitization errors (e.g., over/undershoots)? A higher number of digitization errors points to a lower quality of the data.

Under/overshoots

When two nodes in a simplified network are placed within a distance of a few meters, but do not share a common edge, it is often due to an edge over/undershoot or another digitizing error. An overshoot occurs when two features meet and one of them extends beyond the other. An undershoot occurs when two features are supposed to meet, but instead are just in close proximity to each other. See the image below for an illustration of an overshoot (left) and an undershoot (right). For a more detailed explanation of over/undershoots, see the GIS Lounge website.

Overshoots refer to situations where a line feature extends too far beyond at intersecting line, rather than ending at the intersection (left). Undershoots happen when two line features are not properly joined, for example at intersection (right)

Method

Overshoots: First, the `length_tolerance` (in meters) is defined in the cell below. Then, with `find_overshoots`, all network edges that have a dangling node attached to them and that have a maximum length of `length_tolerance` are identifed as overshoots, and the results are plotted.

Undershoots: First, the `length_tolerance` (in meters) is defined in the cell below. Then, with `find_undershoots`, all pairs of dangling nodes that have a maximum of `length_tolerance` distance between them, are identified as undershoots, and the results are plotted.

The workflow for over/undershoot detection below is inspired by Neis et al. 2012.

Interpretation

Note that over/undershoots are not necessarily always a data quality issue - they might be instead an accurate representation of the network conditions or of the digitization strategy (for example, a cycle path might end abruptly soon after a turn, which results in an overshoot; protected cycle paths are often digitized in OSM as interrupted at intersections, which results in intersection 'undershoots').

The interpretation of the impact of over/undershoots on data quality is context dependent. For certain applications, such as routing, overshoots do not present a particular challenge; they can, however, pose an issue for other applications such as network analysis, given that they skew the network structure. Undershoots, on the contrary, are a serious problem for routing applications, especially if only bicycle infrastructure is considered; they also pose a problem for network analysis, for example for any path-based metric, such as most centrality measures (e.g., betweenness centrality).

Missing intersection nodes

When two edges intersect without having a node at the intersection - and if neither edges are tagged as a bridge or a tunnel - there is a clear indication of a topology error.

Method

The worfklow below is inspired by Neis et al. 2012. First, with the help of `check_intersection`, each edge which is not tagged as either tunnel or bridge is checked for any *crossing* with another edge of the network. If this is the case, the edge is marked as having an intersection issue. The number of intersection issues found is printed and the results are plotted for visual analysis.

Interpretation

A higher number of intersection issues points to a lower data quality. However, it is recommended with a manual visual check of all intersection issues with a certain knowledge of the area, in order to determine the origin of intersection issues and confirm/correct/reject them.

Network components

Disconnected components

Disconnected components do not share any elements (nodes/edges). In other words, there is no network path that could lead from one disconnected component to the other. As mentioned above, most real-world networks of bicycle infrastructure do consist of many disconnected components (Natera Orozco et al., 2020). However, when two disconnected components are very close to each other, it might be a sign of a missing edge or another digitizing error.

Method

First, with the help of `return_components`, a list of all (disconnected) components of the network is obtained. The total number of components is printed and all components are plotted in different colors for visual analysis. Next, the component size distribution (with components ordered by the network length they contain) is plotted, followed by a plot of the largest connected component.

Interpretation

As with many of the previous analysis steps, knowledge of the area is crucial for a correct interpretation of component analysis. Given that the data represents the actual infrastructure accurately, bigger components indicate coherent network parts, while smaller components indicate scattered infrastructure (e.g., one single bicycle path along a street that does not connect to any other bicycle infrastructure). A high number of disconnected components in near vicinity of each other could indicate digitization errors or missing data.

Number of components

Number of components per grid cell

Distribution of network length per component

Largest connected component

Potential missing links between disconnected components

In the plot of potential missing links between components, all edges that are within the specified distance of an edge on another component are plotted. The gaps between disconnected edges are highlighted with a marker. The map thus highlights edges which, despite being in close proximity of each other, are disconnected and where it thus would not be possible to bike on cycling infrastructure between the edges.

Component connectivity

Visualizing differences between how many cells can be reached from each cell.

This is a crude measure for network connectivity but has the benefit of being computationally cheap and thus able to highlight stark differences in network connectivity in very little time.

Summary

2a. Load and Process Reference Data

This notebook:

- Loads the polygon defining the study area and then creates a grid overlay for the study area.
- Loads the reference data.
- Processes the reference data to create the network structure and atttributes needed in the analysis.

Sections

- Load data for study area and define analysis grid
- Load and process reference data

Load data for study area and define analysis grid

This step:

- Loads settings for the analysis from the configuration file.
- Reads data for the study area.
- Creates a grid overlay of the study area, with grid cell size as defined in configuration file `config.yml`.

Read in data for study area

The study area is defined by the user-provided polygon. It will be used for the computation of **global** results, i.e. for the entire study area.

Create analysis grid

Load and process reference data

This step:

- Creates a network from the reference data.
- Projects it to the chosen CRS.
- Clips the data to the polygon defining the study area.
- Measures the infrastructure length of the edges based on the geometry type and whether they allow for bidirectional travel or not.

- Simplifies the network.
- Creates copies of all edge and node datasets indexed by their intersecting grid cell.

2b. Intrinsic Analysis of Reference Bicycle Network Data

This notebook analyses the quality of a user-provided reference bicycle infrastructure data set for a given area. The quality assessment is *intrinsic*, i.e. based only on the input data set, and making no use of information external to the data set. For an extrinsic quality assessment that compares the reference data set to corresponding OSM data, see the notebooks 3a and 3b.

The analysis assesses the *fitness for purpose* (Barron et al., 2014) of the reference data for a given area. Outcomes of the analysis can be relevant for bicycle planning and research - especially for projects that include a network analysis of bicycle infrastructure, in which case the topology of the geometries is of particular importance.

Since the assessment does not make use of an external reference dataset as the ground truth, no universal claims of data quality can be made. The idea is rather to enable the those working with bicycle networks to assess whether their data is good enough for their particular use case. The analysis assists in finding potential data quality issues, but leaves the final interpretation of the results to the user.

The notebook makes use of quality metrics from a range of previous projects investigating OSM/VGI data quality, such as Antoniou & Skopeliti (2015), Fonte et al. (2017) and Fester et al. (2020).

> **Familiarity required**
> For a correct interpretation of some of the metrics for spatial data quality, some familiarity with the area is necessary.

Sections

Data completeness

Network Density

In this setting, network density refers to the length of edges or number of nodes per square kilometer (i.e., the definition of network density usually used when looking at street networks, which is distinct from the definition usually found in graph theory). Network density without comparing to a reference dataset does not in itself indicate spatial data quality. For anyone familiar with the study area, network density can however indicate whether parts of the area appear to be under- or over-mapped and is thus included here.

Method

The density here is not based on the geometric length of edges, but instead on the computed length of the infrastructure. For example, a 100-meter-long bidirectional path contributes with 200 meters of bicycle infrastructure. With `compute_network_density`, the number of elements (nodes; dangling nodes; and total infrastructure length) per unit area is calculated. The density is computed twice: first for the study area for both the entire network ('global density'), then for each of the grid cells ('local density'). Both global and local densities are computed for the entire network and for respectively protected and unprotected infrastructure.

Interpretation

Since the analysis conducted here is intrinsic, i.e., it makes no use of external information, it cannot be known whether a low-density value is due to incomplete mapping, or due to actual lack of infrastructure in the area. However, a comparison of the grid cell density values can provide some insights, for example:

- lower-than-average infrastructure density indicates a locally sparser network
- higher-than-average node density indicates a that there are relatively many intersections in the grid cell
- higher-than-average dangling node density indicates that there are relatively many dead ends in the grid cell

Global network density

Local network density

Densities of protected and unprotected infrastructure:

Network topology

This section explores the geometric and topological features of the data.

These are, for example, network density, disconnected components, dangling (degree one) nodes; it also includes exploring whether there are nodes in close proximity, that do not share an edge - a potential sign of edge undershoots - or if there are intersecting edges without a node at the intersection, which might indicate a digitizing error that will distort any attempts at routing on the network.

Due to the fragmented nature of most networks of bicycle infrastructure, many metrics, such as missing links or network gaps, simply reflect the true extent of the infrastructure (Natera Orozco et al., 2020). This is different for car networks, where e.g., disconnected components could more readily be interpreted as a data quality issue.

Therefore, the analysis only takes very small network gaps into account as potential data quality issues.

Subsections:

- Simplification outcome
- Dangling nodes
- Under/Overshoots

Simplification outcome

When converting a set of geocoded linestrings (polygonal chains) to graph format, not all vertices (nodes) are of equal meaning. For geometry of the infrastructural element, all nodes are needed as an ordered list. For the topology of the network, however, only those nodes that are endpoints or intersection points with other edges are needed, while all other (so-called 'interstitial') nodes do not add any information. To compare the structure and true ratio between nodes and edges in a network, a simplified network representation which only includes nodes at endpoints and intersections, or where the value of important attributes changes, is required. Therefore, in the notebook `01_load_data` the bicycle network was simplified by removing all interstitial nodes from the graph object (retaining, however, the complete node lists in the geometry attribute of each edge). An additional advantage of simplifying the network is the resulting substantial reduction of the number of nodes and edges, which makes computational routines much faster.

Comparing the node degree distribution for the networks before and after simplification is a quick sanity check for the simplification routine. Typically, the big majority of nodes in the non-simplified network will be of degree two; in the simplified network, however, most nodes will have degrees other than two. Degree two nodes are retained in only two cases: if they represent a connection point between two different types of infrastructure; or if they are needed in order to avoid self-loops (edges whose start and end points are identical) or multiple edges between the same pair of nodes.

As part of the simplification routine, in cases where there are several edges between the same pair of nodes ('parallel edges' or 'multiedges'), only one of the edges is retained. Within the routine, the number edges removed in this way are counted.

Method

The node degree distributions before and after simplification are plotted below.

Interpretation

Typically, the node degree distribution will go from high (before simplification) to low (after simplification) counts of degree two nodes, while it will not change for all other degrees (1, or 3 and higher). Further, the total number of nodes will see a strong decline. If the simplified graph still maintains a relatively high number of degree two nodes, or if the number of nodes with other degrees changes after the simplification, this might point to issues either with the graph conversion or with the simplification process.

Dangling nodes

Dangling nodes are nodes of degree one - in other words, nodes that have only one single edge attached to them. Most networks will naturally contain a number of dangling nodes. Dangling nodes can occur at actual dead-ends (representing a cul-de-sac) or at the endpoints of certain features (e.g., when a bicycle path ends in the middle of a street). However, dangling nodes can also occur as a data quality issue in case of over/undershoots (as described in detail in the next section). The number of dangling nodes in a network does to some extent also depend on the digitization method, as shown in the illustration below.

Therefore, the presence of dangling nodes is in itself not a sign of low data quality. However, a high number of dangling nodes in an area that is not known for suffering from many dead-ends can indicate digitization errors and problems with edge over/undershoots.

Dangling nodes occur where road features end (left), but when separate features are joined at the end (right), there will be no dangling nodes

Method

Below, a list of all dangling nodes is obtained with the help of `get_dangling_nodes`. Then, the network with all its nodes is plotted. The dangling nodes are shown in orange; all other nodes are shown in black.

Interpretation

We recommend a visual analysis in order to interpret the spatial distribution of dangling nodes, with particular attention to areas of high dangling node density. It is important to understand where dangling nodes come from: are they actual dead-ends or digitization errors (e.g., over/undershoots)? A higher number of digitization errors points to a lower quality of the data.

Dangling nodes

Under/overshoots

When two nodes in a simplified network are placed within a distance of a few meters, but do not share a common edge, it is often due to an edge over/undershoot or another digitizing error. An overshoot occurs when two features meet and one of them extends beyond the other. An undershoot occurs when two features are supposed to meet, but instead are just in close proximity to each other. See the image below for an illustration of an overshoot (left) and an undershoot (right). For a more detailed explanation of over/undershoots, see the GIS Lounge website.

Overshoots refer to situations where a line feature extends too far beyond at intersecting line, rather than ending at the intersection (left). Undershoots happen when two line features are not properly joined, for example at intersection (right)

Method

Overshoots: First, the `length_tolerance` (in meters) is defined in the cell below. Then, with `find_overshoots`, all network edges that have a dangling node attached to them and that have a maximum length of `length_tolerance` are identifed as overshoots, and the results are plotted.

Undershoots: First, the `length_tolerance` (in meters) is defined in the cell below. Then, with `find_undershoots`, all pairs of dangling nodes that have a maximum of `length_tolerance` distance between them, are identified as undershoots, and the results are plotted.

The workflow for over/undershoot detection below is inspired by Neis et al. 2012.

Interpretation

Note that over/undershoots are not necessarily always a data quality issue - they might be instead an accurate representation of the network conditions or of the digitization strategy (for example, a cycle path might end abruptly soon after a turn, which results in an overshoot; protected cycle paths are often digitized in OSM as interrupted at intersections, which results in intersection 'undershoots').

The interpretation of the impact of over/undershoots on data quality is context dependent. For certain applications, such as routing, overshoots do not present a particular challenge; they can, however, pose an issue for other applications such as network analysis, given that they skew the network structure. Undershoots, on the contrary, are a serious problem for routing applications, especially if only bicycle infrastructure is considered; they also pose a problem for network analysis, for example for any path-based metric, such as most centrality measures (e.g., betweenness centrality).

Under/overshoots

Network components

Disconnected components

Disconnected components do not share any elements (nodes/edges). In other words, there is no network path that could lead from one disconnected component to the other. As mentioned above, most real-world networks of bicycle infrastructure do consist of many disconnected components (Natera Orozco et al., 2020). However, when two disconnected components are very close to each other, it might be a sign of a missing edge or another digitizing error.

Method

First, with the help of `return_components`, a list of all (disconnected) components of the network is obtained. The total number of components is printed and all components are plotted in different colors for visual analysis. Next, the component size distribution (with components ordered by the network length they contain) is plotted, followed by a plot of the largest connected component.

Interpretation

As with many of the previous analysis steps, knowledge of the area is crucial for a correct interpretation of component analysis. Given that the data represents the actual infrastructure accurately, bigger components indicate coherent network parts, while smaller components indicate scattered infrastructure (e.g., one single bicycle path along a street that does not connect to any other bicycle infrastructure). A high number of disconnected components in near vicinity of each other could indicate digitization errors or missing data.

Number of components

Number of components per grid cell

Distribution of network length per component

Largest connected component

Potential missing links between disconnected components

In the plot of potential missing links between components, all edges that are within the specified distance of an edge on another component are plotted. The gaps between disconnected edges are highlighted with a marker. The map thus highlights edges which, despite being in close proximity of each other, are disconnected and where it thus would not be possible to bike on cycling infrastructure between the edges.

Component connectivity

Visualizing differences between how many cells can be reached from each cell.

This is a crude measure for network connectivity but has the benefit of being computationally cheap and thus able to highlight stark differences in network connectivity in very little time.

Summary

3a. Extrinsic Analysis: Comparison of OSM & Reference Data

This notebook compares the OSM data on bicycle infrastructure for a given area with a provided reference data set in a so-called extrinsic quality assessment. To run this part of the analysis, a reference data set must be available for comparison.

The analysis is based on comparing OSM data to the reference dataset and highlighting how and where they differ, both in terms of *how much* bicycle infrastructure is mapped in the two datasets, but also when it comes to *how* the infrastructure is mapped (e.g., looking at differing network topology and structure).

All differences are computed with OSM values as the baseline. For example, the difference in network density is computed by subtracting the reference density from the OSM density: *OSM - reference = difference*. Hence, positive difference values (over 0) indicate how much higher the OSM value is; negative difference values (below 0) indicate how much lower the OSM value is. Accordingly, if differences are given in percent, the OSM value is taken to be the total value (100%).

While the analysis is based on a comparison, the analysis makes no a priori assumptions about which dataset is better. The same goes for the identified differences: the workflow does not lead to an automatic conclusion as to which data set is of better quality, but instead requires the user to interpret the meaning of the differences found, e.g., whether differing features are results of errors of omission or commission, and which dataset is more correct.

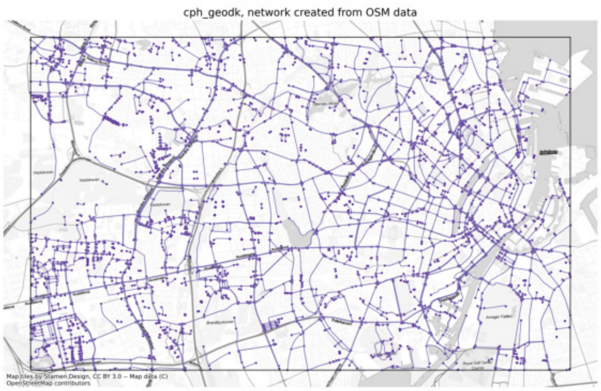
The goal is that the identified differences can be used by the user to both assess the quality of the OSM and the reference datasets, and to support the decision of which dataset should be used for further analysis.

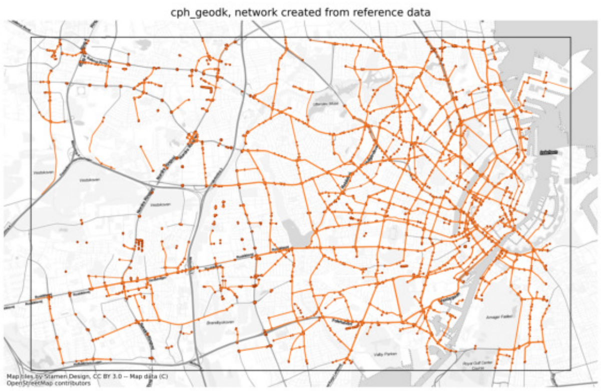
> **Familiarity required**
> For a correct interpretation of some of the metrics for spatial data quality, some familiarity with the area is necessary.

Sections

- Data completeness
  - Network length
  - Network density
- Network topology
  - Simplification outcome
  - Alpha, beta, and gamma indices
  - Dangling nodes

OSM & reference networks


cph_geodk, network created from OSM data
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors


cph_geodk, network created from reference data
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors

Data completeness

This section compares the OSM and reference datasets in terms of data completeness.

The goal is to identify whether one dataset has more bicycle infrastructure mapped than the other, and if so, whether those differences are concentrated in some areas.

The section starts with a comparison of the total length of the infrastructure in both data sets. Then, infrastructure, node and dangling node densities (i.e., the length/number of infrastructure/nodes per square kilometer) is compared first at a global (study area) and at local (grid cell) level. Finally, density differences for protected and unprotected bicycle infrastructure are compared separately.

Method

To account for differences in how bicycle infrastructure has been mapped, the computation of network length and density is based on the infrastructure length, not the geometric length of the network edges. For example, a 100 meter long **bi**directional path (geometric length: 100m) contributes with 200 meters of bicycle infrastructure (infrastructure length: 200m).

Interpretation

Density differences can point to incomplete data. For instance, if a grid cell has a significantly higher edge density in the OSM than in the reference data set, this indicates unmapped grid cell features in the reference data set (or potentially that a street mistakenly has been tagged as bicycle infrastructure.)

Network length

1b. Network Densities

Global network densities

Global average network densities (per km2):

Local network densities

Please note that the plots of densities for respectively the OSM and reference data are produced in the intrinsic notebooks. The plots will thus not necessarily have the same value range for the color bars.

Local differences in network densities

The densities in the OSM data set are taken as base line for comparison. Hence, positive values indicate that the OSM density of the infrastructure type is higher than the reference density; negative values indicate that the OSM density is lower than the reference density.

Densities of protected and unprotected bicycle infrastructure

Global network densities for protected/unprotected infrastructure

Local network densities for protected/unprotected infrastructure

The densities in the OSM data set are taken as base line for comparison. Hence, positive values indicate that the OSM density of the infrastructure type is higher than the reference density; negative values indicate that the OSM density is lower than the reference density.

Differences in infrastructure type density

Network topology

After having compared data completeness, i.e., *how much* infrastructure is mapped, in the section above, we now will focus on differences in network *topology*, which can give some information about *how* the infrastructure is mapped in both datasets. Here we also analyze the extent to which network edges are connected to one or more other edges, or if they on the other hand end in a dangling node. The extent to which edges are properly connected to adjacent edges are important for, for example, analyses of accessibility and routing.

When working with data on bicycle networks, a dataset without gaps between actually connected network elements is preferred - while of course reflecting the real conditions. Identifying the dangling nodes in a network are a quick and easy way to identify edges that end in a 'dead end'. Under and overshoots offer a more precise picture of respectively network gaps and overextended edges, that give a misleading count of dangling nodes.

Method

To identify potential gaps or missing links in the data, first the dangling nodes in both datasets are plotted. Then, the local percentage of dangling nodes out of all nodes in each dataset is plotted separately followed by a plot showing the local difference in the percent of nodes categorized as dangling.

Under and overshoots in both OSM and reference data are finally plotted together in an interactive plot for further inspection.

Interpretation

If an edge ends in a dangling node in one dataset but not the other, this indicates a problem with the data quality and that there either is a missing connection in the data, or that two edges have been connected erroneously. Similarly, different local rates in the share of dangling nodes indicates differences in how the bicycle networks have been mapped - although differences in data completeness of course should be considered in the interpretation.

Undershoots are clear indications of misleading gaps in network data - although they might also represent actual gaps in bicycle infrastructure. Comparing undershoots in one dataset with another dataset can help identify whether it is a question of data quality of the quality of the actual infrastructure. Vast differences in the presence of undershoots or gaps across intersections might be an indication in differing digitizing strategies, since some approaches will map a bike lane crossing a street as a connected stretch, while others will introduce a gap in the width of the crossing street. While both approaches can be considered correct, datasets created with the former method are more suited for network-based analysis.

Overshoots will often be less consequential for analysis, but a high number of overshoots will introduce false dangling nodes and distort measures for network structure based on e.g., node degree or the ratio between nodes and edges.

Simplification outcomes

Node degree distribution

Alpha, beta & gamma indices

In this subsection, we compute and contrast the three aggregated network metrics alpha, beta, and gamma. These metrics are often used to describe network structure, but as measures of data quality, they are only meaningful when compared to the values of a corresponding dataset. For this reason, alpha, beta, and gamma are only part of the extrinsic analysis and not included in the intrinsic notebooks.

While no conclusion can be made about data quality based on any of the three metrics by itself, a comparison of the metrics for the two data sets can indicate differences in network topology, and hence differences in how the infrastructure has been mapped.

Method

All three indices are computed with `eval_func.compute_alpha_beta_gamma`.

The **alpha** value is the ratio of actual to possible *cycles* in the network. A network cycle is defined as a closed loop - i.e. a path that ends on the same node that it started from. The value of alpha ranges from 0 to 1. An alpha value of 0 means that the network has no cycles at all (e.g. a tree-like structure); an alpha value of 1 means that the network is fully connected, which is very rarely the case.

The **beta** value is the ratio of existing edges to existing nodes in the network. The value of beta ranges from 0 to N-1, where N is the number of existing nodes. A beta value of 0 means that the network has no edges; a beta value of N-1 means that the network is fully connected (see also gamma value of 1). The higher the beta value, the more different paths (on average) can be chosen between any pair of nodes.

The **gamma** value is the ratio of existing to *possible* edges in the network. Any edge that connects two of the existing network nodes is defined as "possible". Hence, the value of gamma ranges from 0 to 1. A gamma value of 0 means that the network has no edges; a gamma value of 1 means that every node of the network is connected to every other node.

For all three indices, see Ducruet and Rodrigue, 2020. All three indices can be interpreted in respect to network connectivity: The higher the alpha value, the more cycles are present in the network; the higher the beta value, the higher the number of paths and thus the higher the complexity of the network; and the higher the gamma value, the fewer edges lie between any pair of nodes.

Interpretation

These metrics do not say much about the data quality itself, nor are they useful for a topological comparison of networks of similar size. However, some conclusions can be made through a comparison. For example, if the indices are very similar for the two networks, despite the networks e.g., having very different geometric lengths, this suggests that the data sets have been mapped in roughly the same way, but than one simply includes more features than the other. On the other hand, if the networks have roughly the same total geometric length, but the values from alpha, beta and gamma differ, this can be an indication that the structure and topology of the two datasets are fundamentally different.

Dangling nodes

Dangling nodes in OSM & reference networks

Local values for dangling nodes

Dangling nodes as percentage of all nodes:

Local differences in percent dangling nodes:

Under/overshoots

Over and undershoots in OSM and reference networks

Network components

This section takes a close look at the network component characteristics for each of the two data sets.

Disconnected components do not share any elements (nodes/edges). In other words, there is no network path that could lead from one disconnected component to the other. As mentioned above, most real-world networks of bicycle infrastructure do consist of many disconnected components (Natera Orozco et al., 2020). However, when two disconnected components are very close to each other, it might be a sign of a missing edge or another digitizing error.

Method

To compare the number and pattern of disconnected components in OSM and reference data, all component results from the intrinsic analyses are juxtaposed and two new plots showing respectively components gaps for OSM and reference data and the difference in component connectivity are produced.

Interpretation

The fragmented nature of many bicycle networks make it hard to assess whether disconnected components are a question of a lack of data quality of a lack of properly connected bicycue infrastructure. Comparing disconnected components in two datasets enables a more accurate assessment of whether a disconnected component is a data or a planning issue.

Subsections

- Disconnected components
- Component size distribution
- Largest connected component
- Potentially missing links
- Number of components per grid cell
- Component connectivity

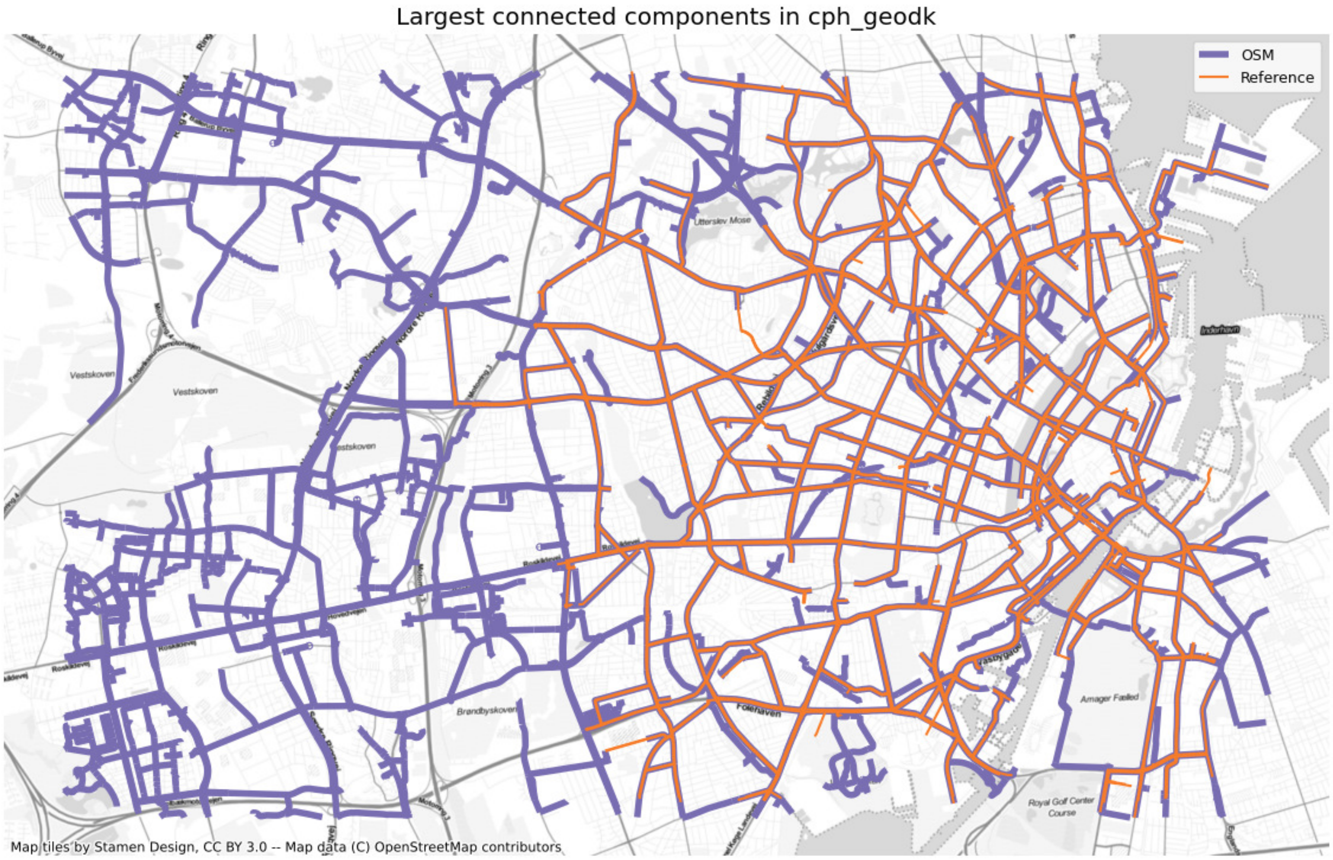
Disconnected components

Component size distribution

TODO: add explanation of plot (after plot has been redone)

Largest connected component

Overlay of largest connected component in OSM and reference networks



Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors

Potentially missing links

In the plot of potential missing links between components, all edges that are within the specified distance of an edge on another component are plotted. The gaps between disconnected edges are highlighted with a marker. The map thus highlights edges which, despite being in close proximity of each other, are disconnected and where it thus would not be possible to bike on cycling infrastructure between the edges.

Number of components per grid cell

The plots below show the number of components intersecting a grid cell. A high number of components in a grid cell is generally an indication of poor network connectivity - either due to fragmented infrastructure or because of problems with the data quality.

Component connectivity

Visualizing differences between how many cells can be reached from each cell. In the plot showing the relative difference in percent cells reached, positive values means that more cells can be reached from this particular cell using the OSM data, while negative values indicate a higher connectivity using the reference data set.

The metric is a crude measure for network connectivity but has the benefit of being computationally cheap and thus able to highlight stark differences in network connectivity in very little time.

Summary