BikeDNA Report

Copenhagen

Bicycle Infrastructure Data & Network Assessment

Report generated on 2022-12-20 16:45:14
with BikeDNA · github.com/anerv/BikeDNA

About BikeDNA

This report was automatically generated by BikeDNA (Bicycle Data & Network Assessment) by converting Jupyter notebooks into pdf. BikeDNA is a tool for assessing the quality of OpenStreetMap (OSM) and other bicycle infrastructure data sets in a reproducible way. It provides planners, researchers, data maintainers, cycling advocates, and others who work with bicycle networks a detailed, informed overview of data quality in a given area.

BikeDNA is maintained at https://github.com/anerv/BikeDNA by Ane Rahbek Vierø, Anastassia Vybornova, and Michael Szell. It is available under the AGPL 3.0 license.

A fair amount of research projects on OpenStreetMap and other forms of volunteered geographic information (VGI) have already been conducted, but few focus explicitly on bicycle infrastructure. Doing so is important because paths and tracks for cyclists and pedestrians often are mapped last and are more likely to have errors (Barron et al., 2014, Neis et al. 2012). Moreover, the spatial distribution of dips in data quality in crowdsourced data are often not random but correlate with population density and other characteristics of the mapped area (Forghani and Delavar, 2014), which requires a critical stance towards the data we use for our research and planning, despite the overall high quality of OSM.

Data quality covers a wide range of aspects. The conceptualization of data quality used here refers to *fitness-for-purpose* (Barron et al., 2014) - this means that data quality is interpreted as whether or not the data fulfils the user needs, rather than any universal definition of quality. To particularly support network-based research and planning, BikeDNA provides insights into the topological structure of the bicycle network apart from data coverage.

The purpose is not to give any final assessment of the data quality, but to highlight aspects that might be relevant for assessing whether the data for a given area is fit for use. While BikeDNA can make use of a reference dataset to compare with OSM, if one is available, BikeDNA cannot give any final assessment of the quality of a refernece data compared to OSM. However, OSM data on bicycle infrastructure is often at a comparable or higher quality than governmental datasets, and the interpretation of differences between the two requires adequate local knowledge.

1a. Initialize 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 create analysis grid

This step:

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

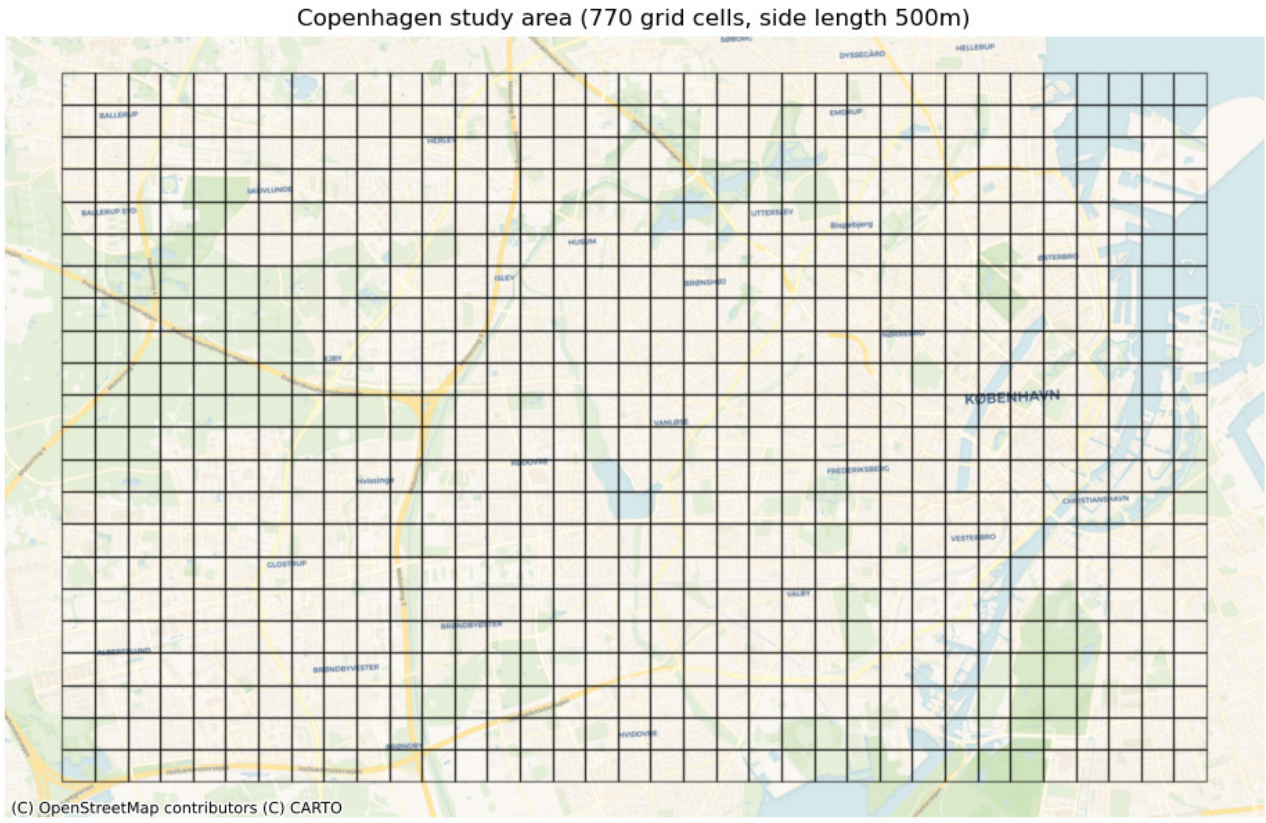
Load data for study area

The study area is defined by the user-provided polygon. It will be used for the computation of **global** results, i.e. quality metrics based on all data in the study area.

```
The size of the study area is 181.38 km2.
```

Create analysis grid

```
The grid contains 770 square cells with a side length of 500 m and an area of 0.25 km2.
This grid will be used for local (grid cell level) analysis:
```



Download and preprocess 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.
- Creates copies of all edge and node data sets indexed by their intersecting grid cell.

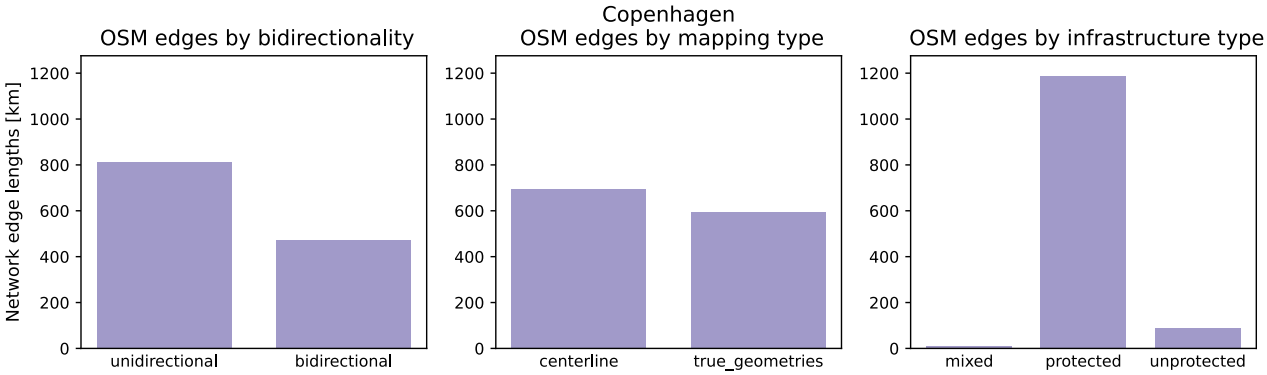
OSM data model

In OSM, street network data are stored using *nodes* (points) and *ways* (lines). In BikeDNA, OSM data are converted to a network structure consisting of *nodes* and *edges* (we use the terminology used in OSMnx). Edges represents the actual infrastructure, such as bike lanes and paths, while nodes represents the start and end points for the edges, as well as all intersections. For further details, read more about the OSM data model and the network data model.

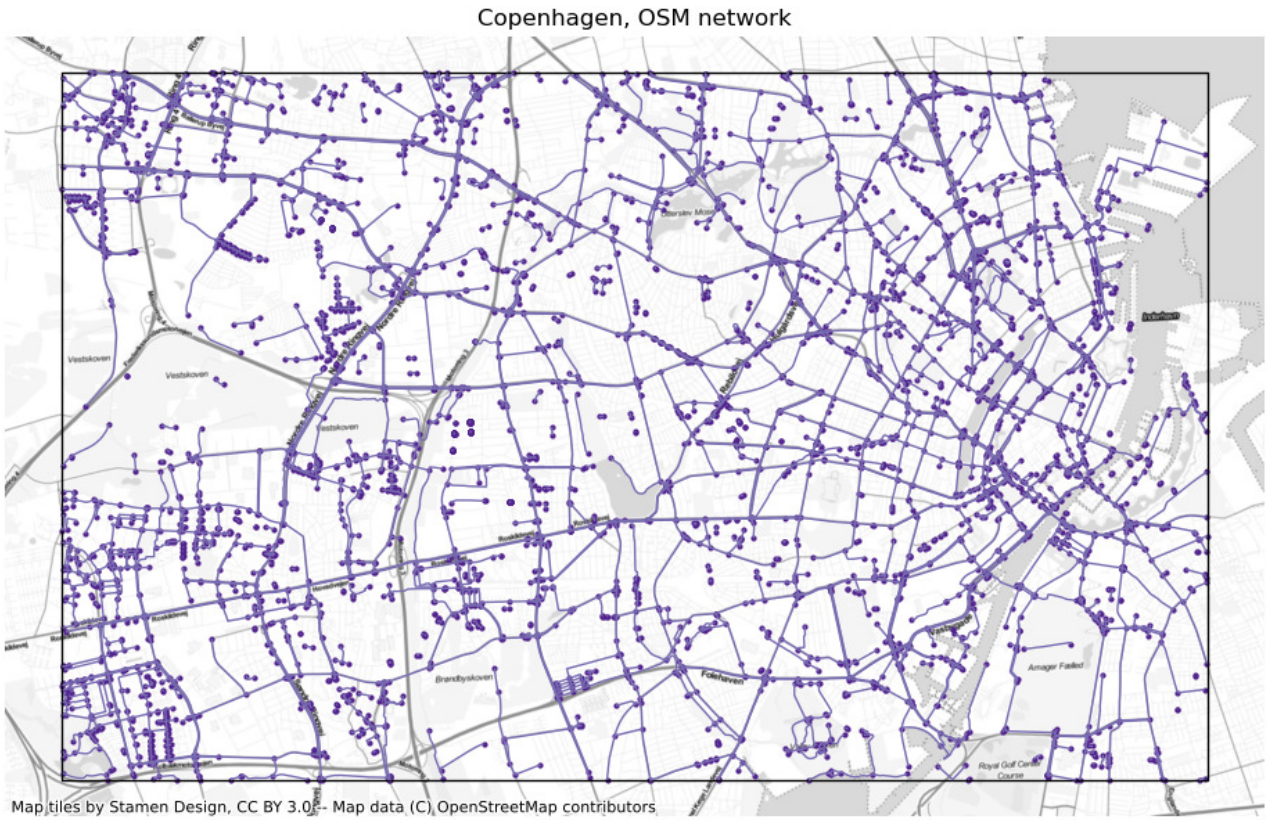
```
Edges where 'bicycle_bidirectional' is False: 33534 out of 50959 (65.81%)
Edges where 'bicycle_bidirectional' is True: 17425 out of 50959 (34.19%)
```

```
Edges where the geometry type is 'centerline': 25742 out of 50959 (50.52%)
```

```
Edges where the geometry type is 'true_geometries': 25217 out of 50959 (49.48%)


Edges where the protection level is 'protected': 46583 out of 50959 (91.41%)
Edges where the protection level is 'unprotected': 3713 out of 50959 (7.29%)
Edges where the protection level is 'mixed': 663 out of 50959 (1.3%)
```




```
The graph covers an area of 179.70 km2.
The length of the OSM network with bicycle infrastructure is 1063.18 km.
```



1b. Intrinsic Analysis of OSM Data

This notebook analyzes the quality of OSM bicycle infrastructure data for a given area. The quality assessment is *intrinsic*, i.e. based only on the one input data set without makeing use of external information. For an extrinsic quality assessment that compares the OSM data 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 data set as the ground truth, no universal claims of data quality can be made. The idea is rather to enable those working with OSM-based bicycle networks to assess whether the data are 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 Ferster et al. (2020), Hochmair et al. (2015), Barron et al. (2014), and Neis et al. (2012).

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 km2. This is the usual definition of network density in spatial (road) networks, which is distinct from the *structural* network density known more generally in network science. Without comparing to a reference data set, network density 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.

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 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:

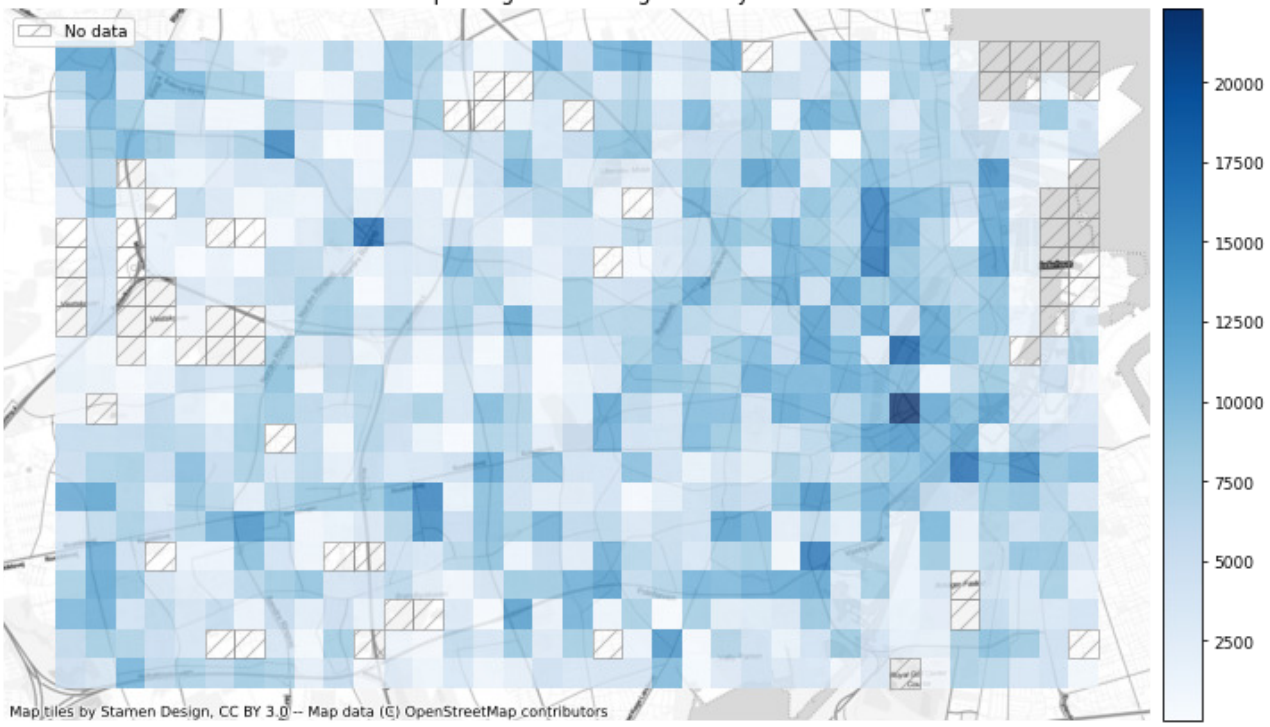
- lower-than-average infrastructure density indicates a locally sparser network
- higher-than-average node density indicates that there are relatively many intersections in a grid cell
- higher-than-average dangling node density indicates that there are relatively many dead ends in a grid cell

Global network density

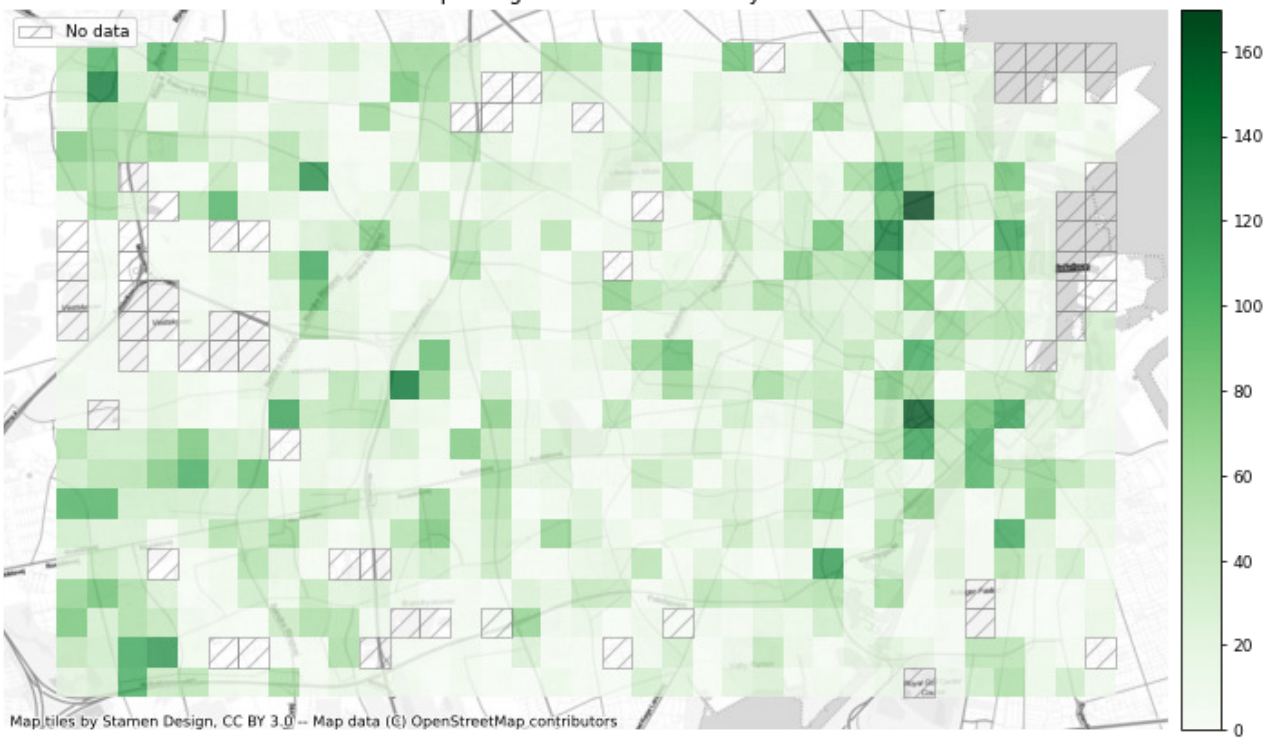
```
For the entire study area, there are:
- 5861.46 meters of bicycle infrastructure per km2.
- 27.15 nodes in the bicycle network per km2.
- 10.02 dangling nodes in the bicycle network per km2.
- 5302.84 meters of protected bicycle infrastructure per km2.
- 499.41 meters of unprotected bicycle infrastructure per km2.
- 59.21 meters of mixed protection bicycle infrastructure per km2.
```

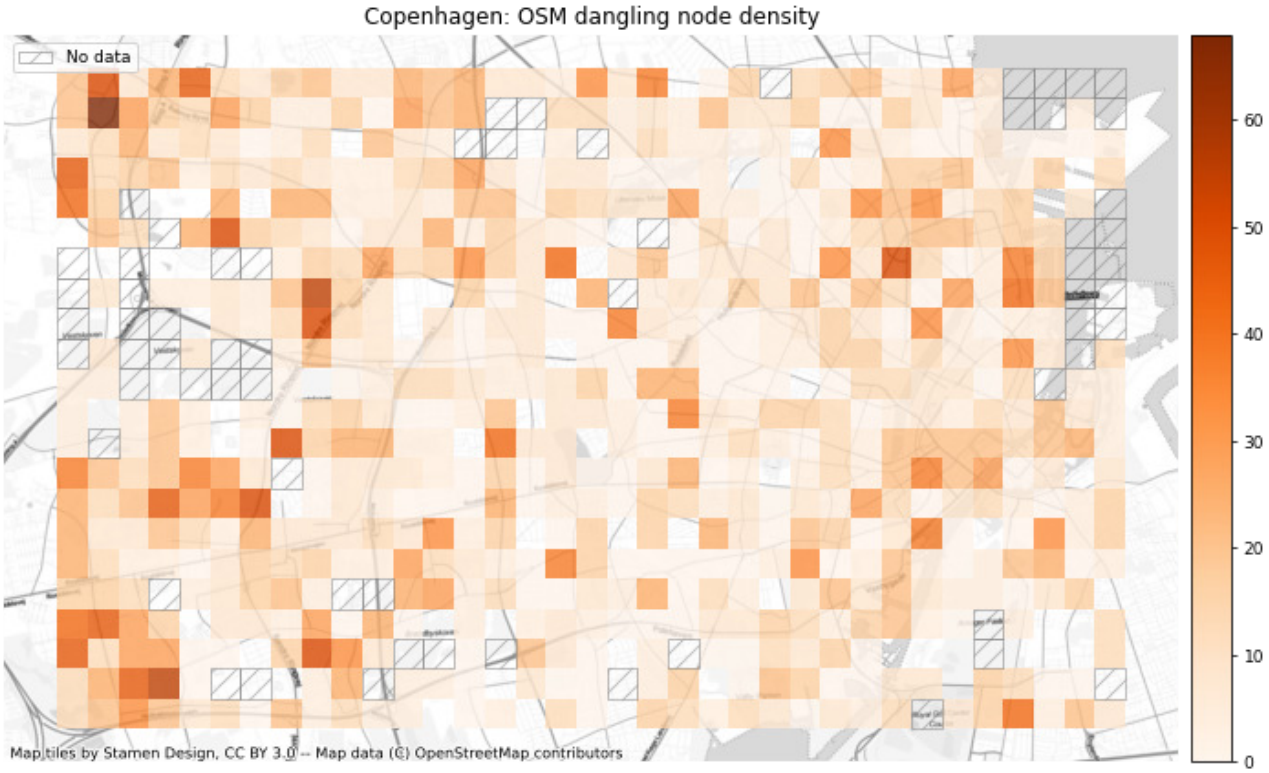
Local network density

Copenhagen: OSM edge density



Copenhagen: OSM node density



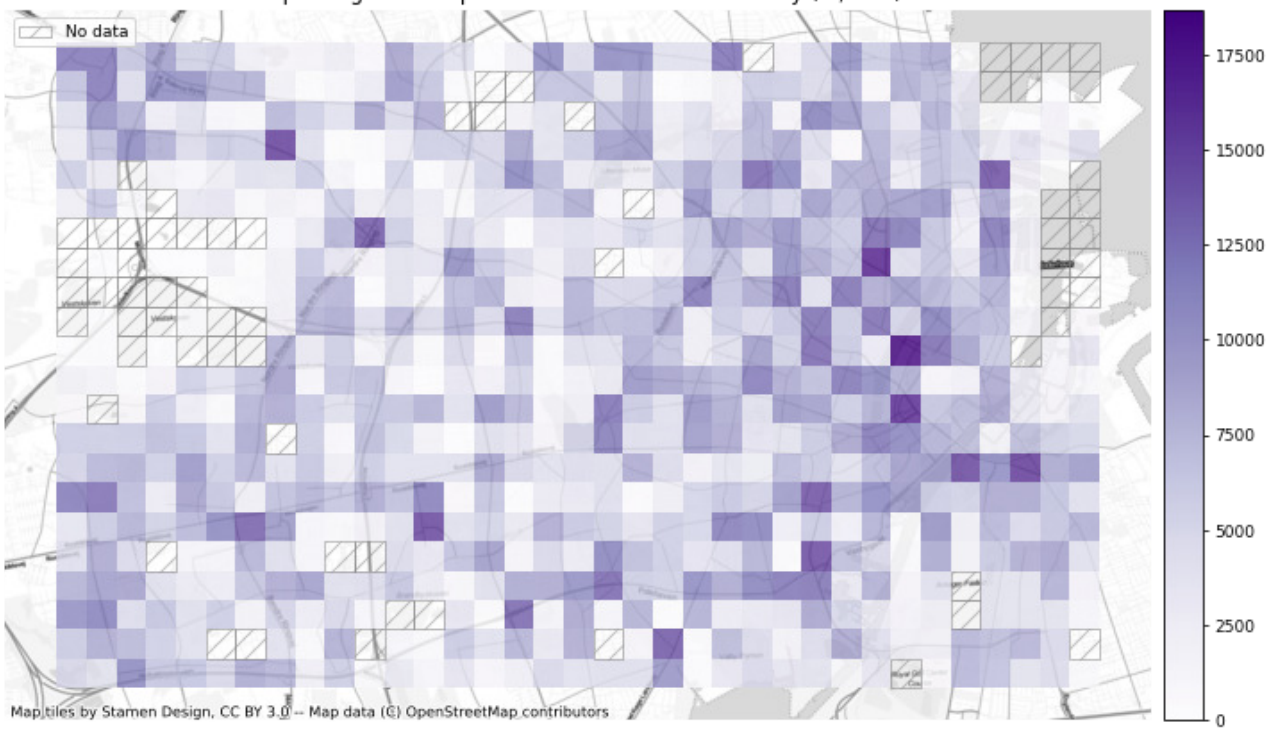


Densities of protected and unprotected infrastructure

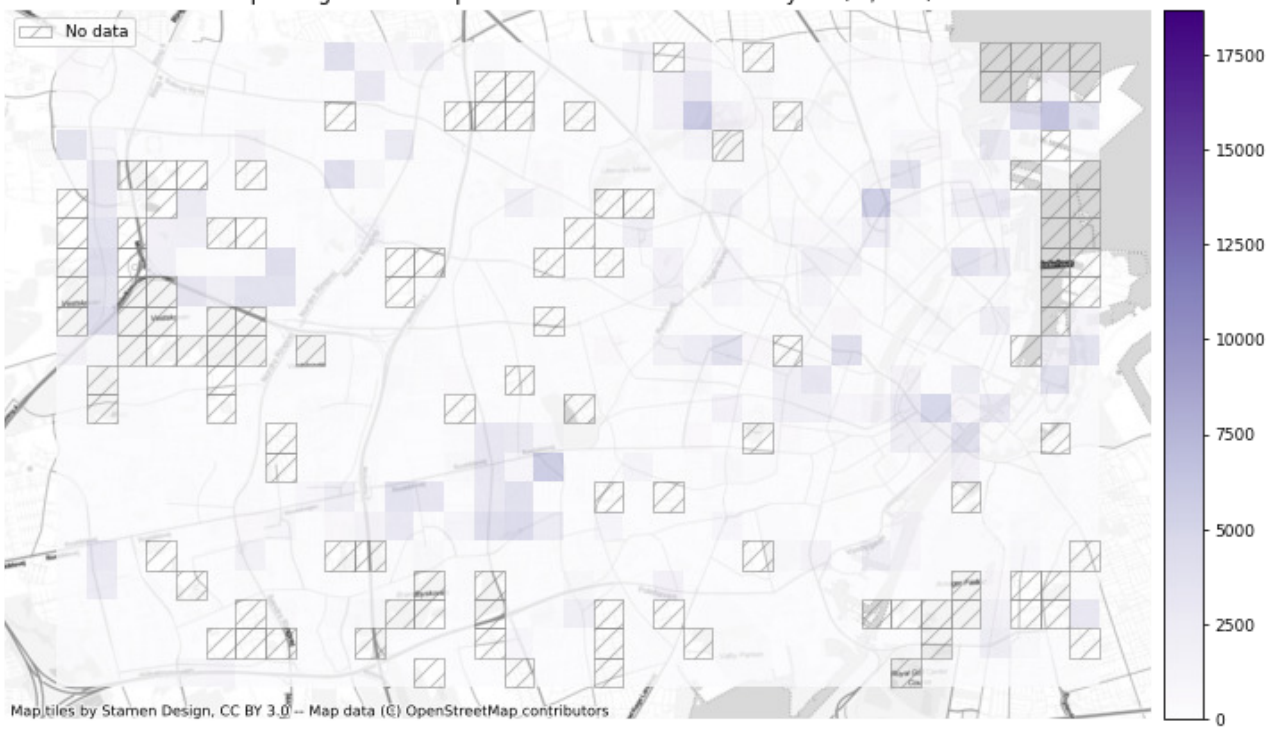
In BikeDNA, *protected infrastructure* refers to all bicycle infrastructure which is either separated from car traffic by for example an elevated curb, bollards, or other physical barriers, or for cycle tracks that are not adjacent to a street.

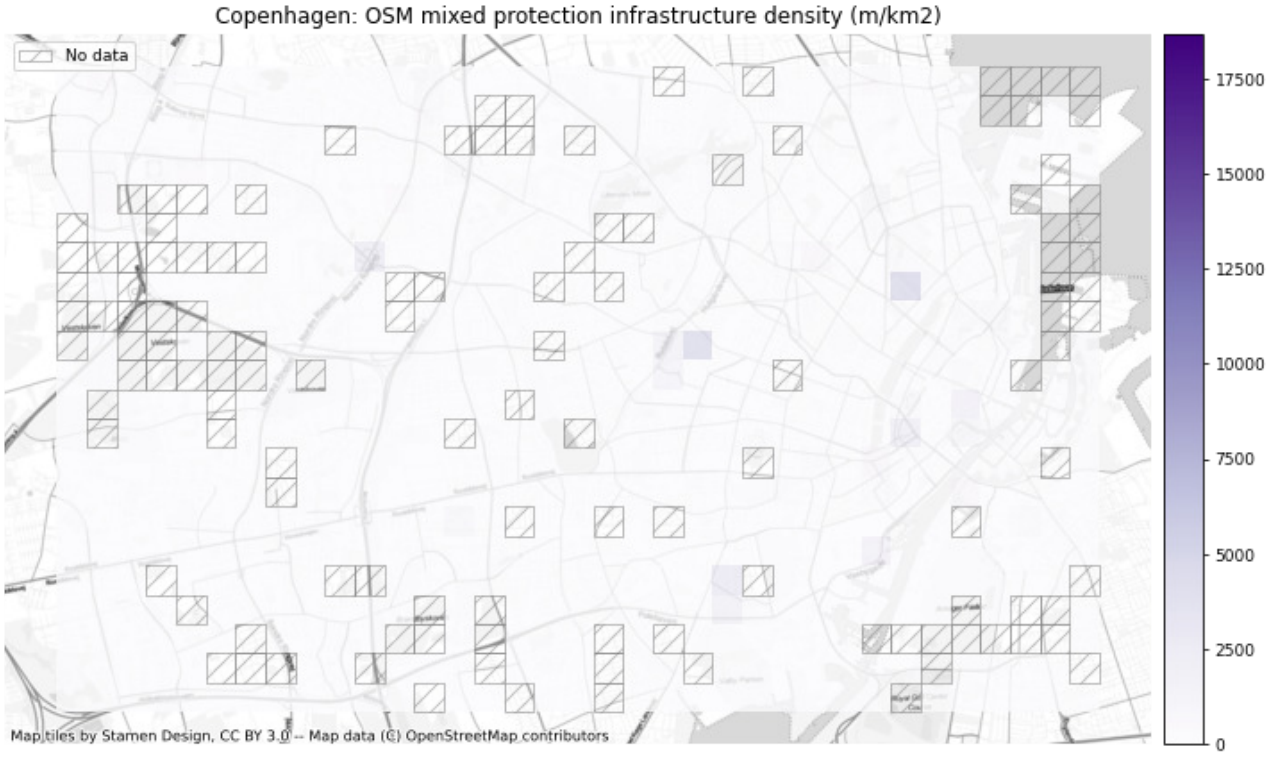
Unprotected infrastructure are all other types of lanes that are dedicated for bicyclists, but which only are separated by car traffic by e.g., a painted line on the street.

Copenhagen: OSM protected infrastructure density (m/km2)



Copenhagen: OSM unprotected infrastructure density for (m/km2)





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 an 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 sometimes result in conflicting tags which can undermine the evaluation of cycling conditions.

This section includes analyzes of missing tags (edges with tags that lack information), incompatible tags (edges with tags labelled with two or more contradictory tags), and tagging patterns (the spatial variation of which tags are being used to describe bicycle infrastructure).

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 depends on the use case - for example, the tag `lit` for a project that studies light conditions on cycle paths. The workflow below allows to quickly analyze the percentage of network edges that have a value available 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

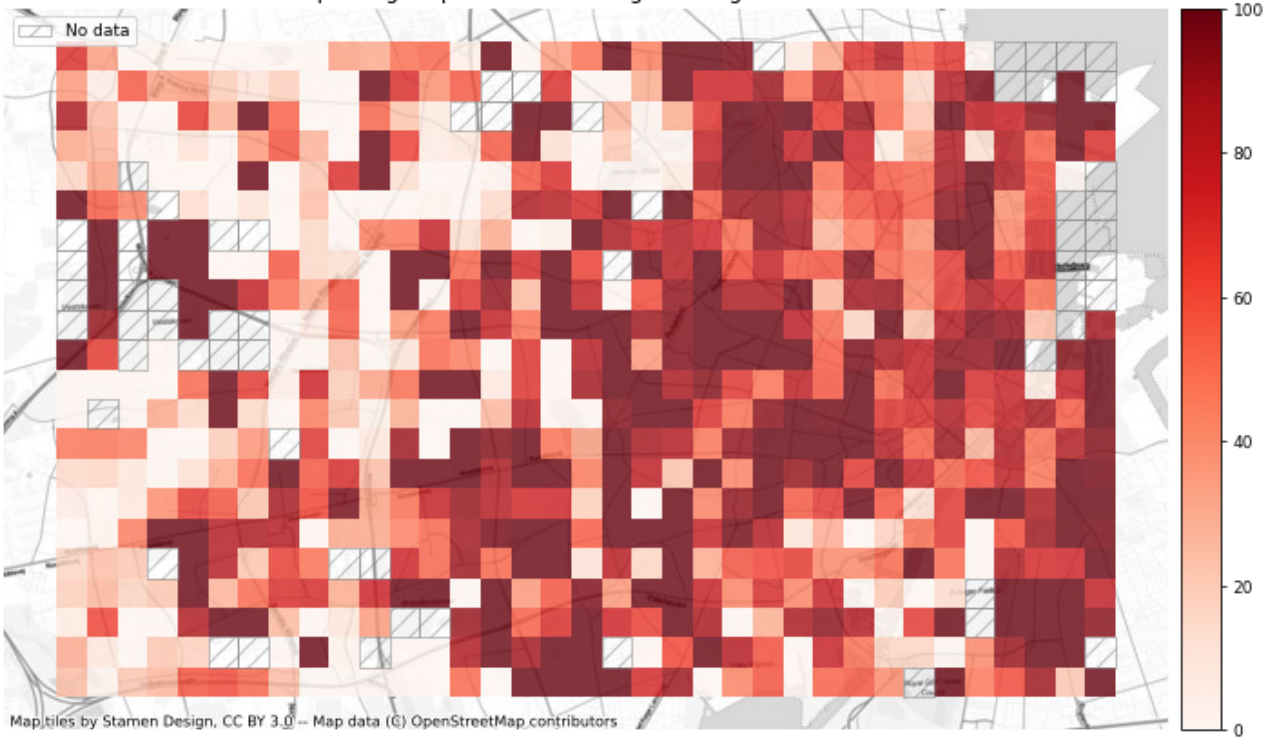
On the study area level, a higher percentage of existing tag values indicates in principle a higher quality of the data set. However, this is different from an estimation of whether the existing tag values are truthful. On the grid cell level, lower-than-average percentages for existing tag values can indicate a more poorly mapped area. However, 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

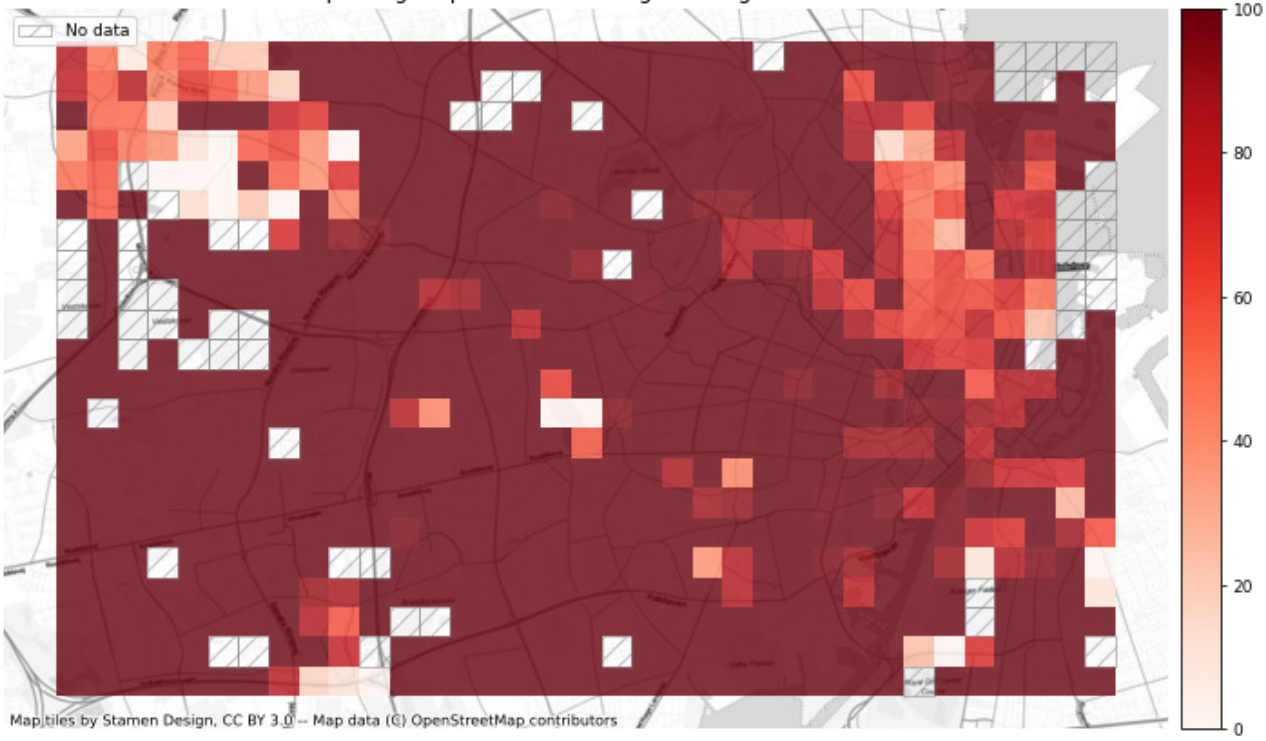
```
Analysing tags describing:
surface – width – speedlimit – lit –

surface: 23325 out of 50959 edges (45.77%) have information.
surface: 552 out of 1285 km (42.95%) have information.


width: 5015 out of 50959 edges (9.84%) have information.
width: 97 out of 1285 km (7.56%) have information.


speedlimit: 25348 out of 50959 edges (49.74%) have information.
speedlimit: 684 out of 1285 km (53.21%) have information.


lit: 39318 out of 50959 edges (77.16%) have information.
lit: 1008 out of 1285 km (78.45%) have information.
```

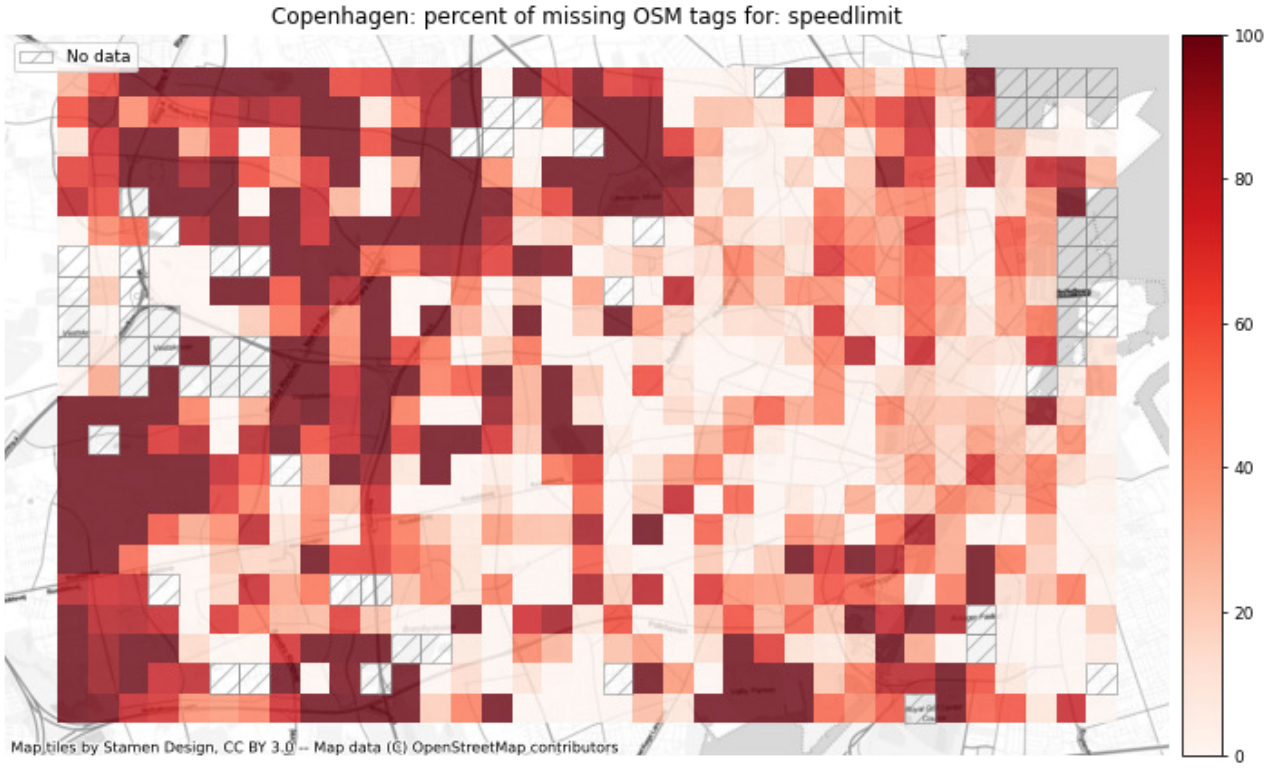
Local missing tags

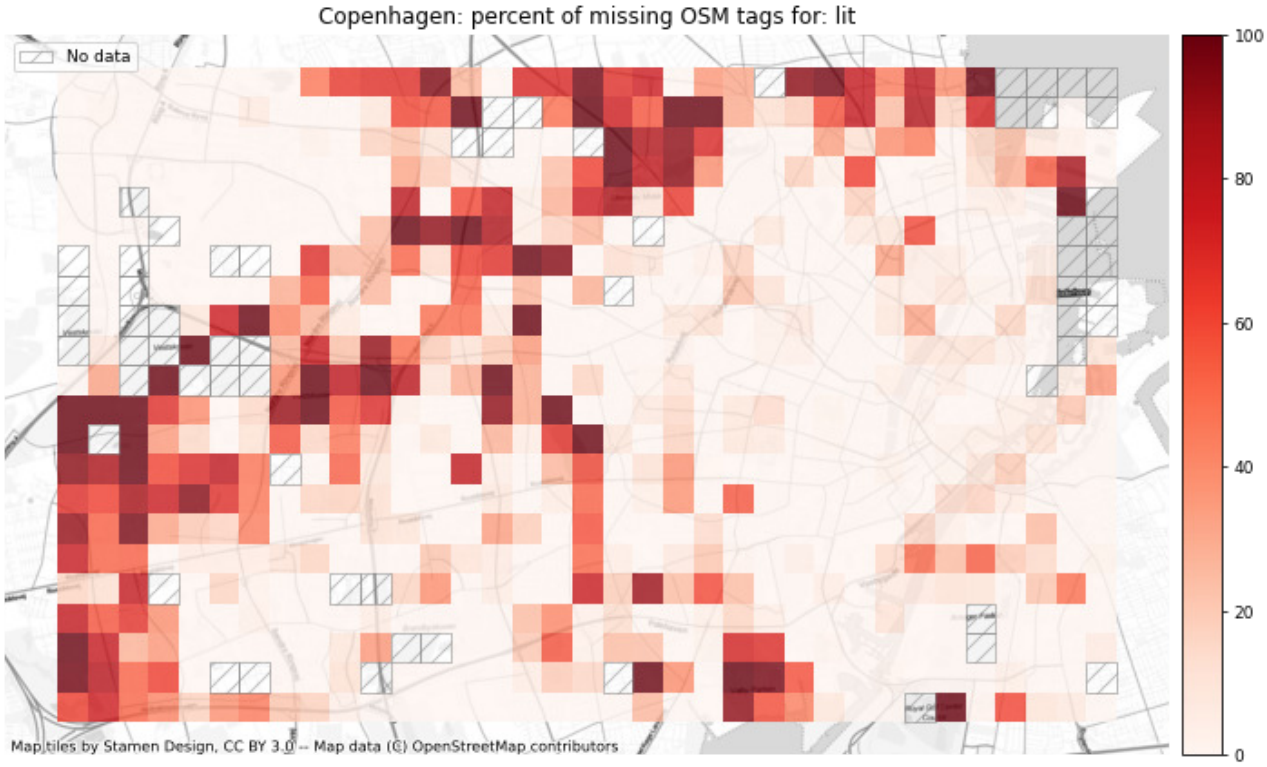
Copenhagen: percent of missing OSM tags for: surface



Copenhagen: percent of missing OSM tags for: width







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

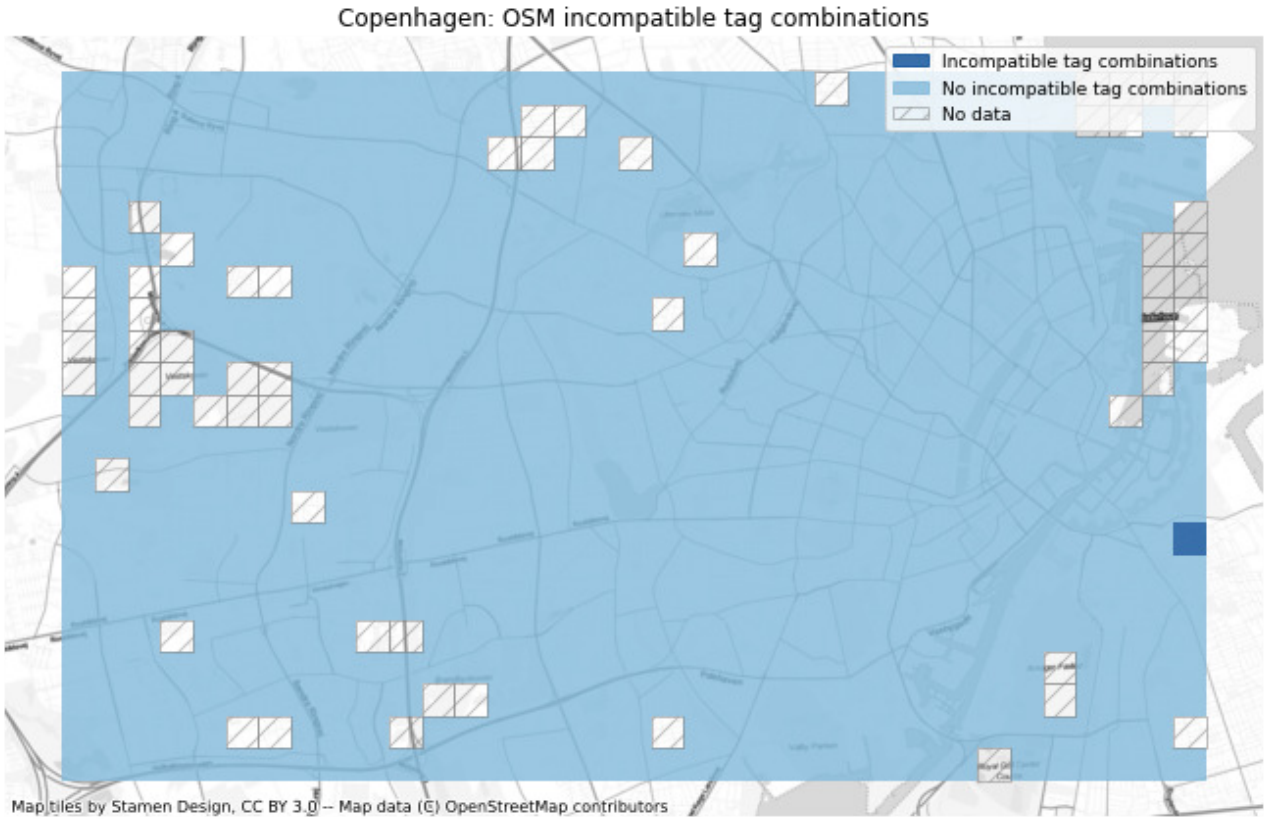
In the `config.yml` file, a list of incompatible key-value pairs for tags in the `incompatible_tags_analysis` is defined. Since there is no limitation to which tags a data set could potentially contain, the list is, by definition, non-exhaustive, and can be adjusted by the user. In the section below,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

```
In the entire data set, there are 2 incompatible tag combinations (of those defined in the confi
guration file).
```

Local incompatible tags (per grid cell)



Plotting incompatible tag geometries

```
Interactive map saved at results/OSM/cph_geodk/maps_interactive/tagsincompatible_osm.html
```

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 can 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

We first plot 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

```
Interactive map saved at results/OSM/cph_geodk/maps_interactive/taggingtypes_osm.html
```

Multiple tagging

Copenhagen: OSM bicycle infrastructure defined with tags: cycleway_left + cycleway_right

Copenhagen: OSM bicycle infrastructure defined with tags: cycleway + cycleway_both

Copenhagen: OSM bicycle infrastructure defined with tags: highway + cycleway

Copenhagen: OSM bicycle infrastructure defined with tags: cycleway_left + cycleway_both

Copenhagen: OSM bicycle infrastructure defined with tags: cycleway + cycleway_right

Copenhagen: OSM bicycle infrastructure defined with tags: cycleway_right + cycleway_both


```
Interactive map saved at results/OSM/cph_geodk/maps_interactive/taggingcombinations_osm.html
```

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 bicycle networks, many metrics, such as missing links or network gaps, can simply reflect the true extent of the infrastructure (Natera Orozco et al., 2020). This is different for road 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.

Simplification outcome

To compare the structure and true ratio between nodes and edges in the network, a simplified network representation which only includes nodes at endpoints and intersections was created in notebook `1a` by removing all interstitial nodes.

Comparing the degree distribution for the networks before and after simplification is a quick sanity check for the simplification routine. Typically, the vast 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.

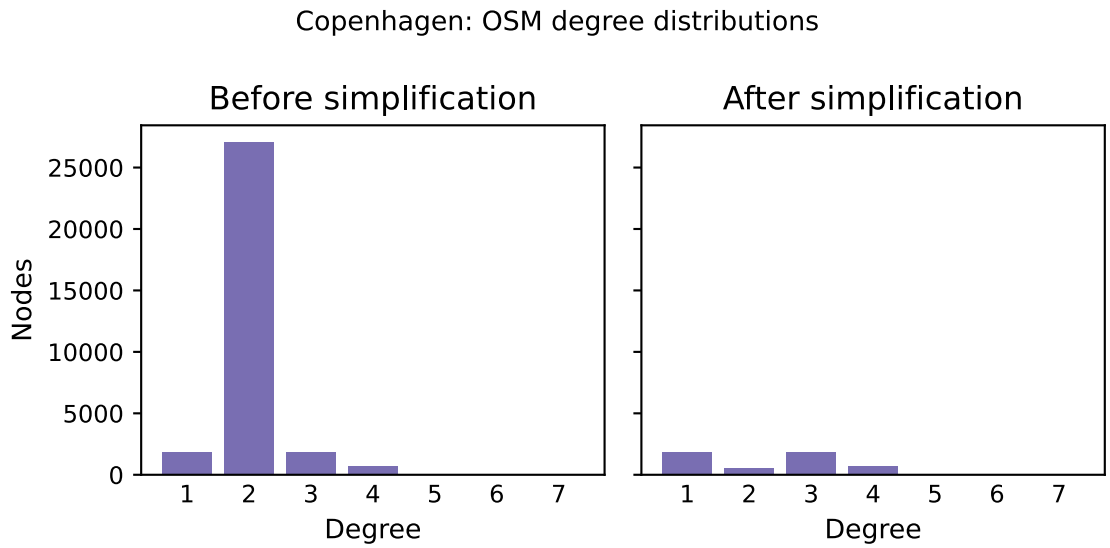
Method

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.

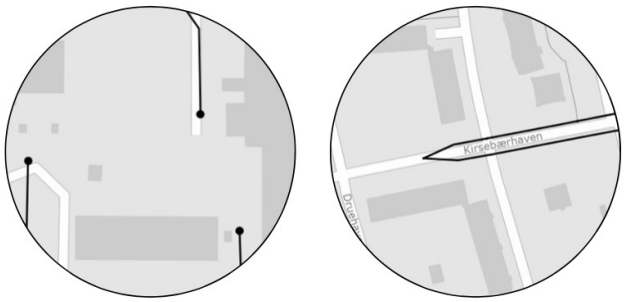
```
Simplifying the network decreased the number of edges by 88.9% and the number of nodes by 84.3%.
```



Dangling nodes

Dangling nodes are nodes of degree one, i.e. they 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 (see 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 containing many dead-ends can indicate digitization errors and problems with edge over/undershoots.



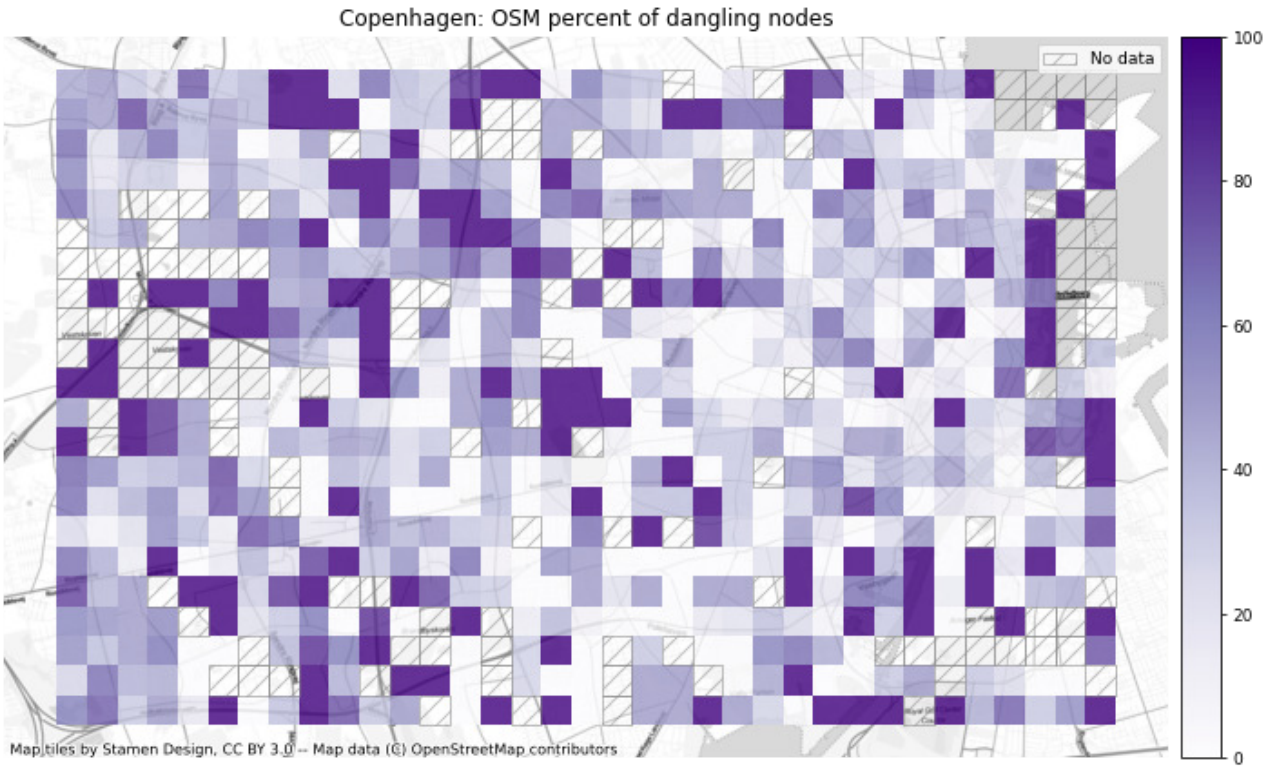
Left: Dangling nodes occur where road features end. Right: However, when separate features are joined at the end, there will be no dangling nodes.

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 color, all other nodes are shown in black.

Interpretation

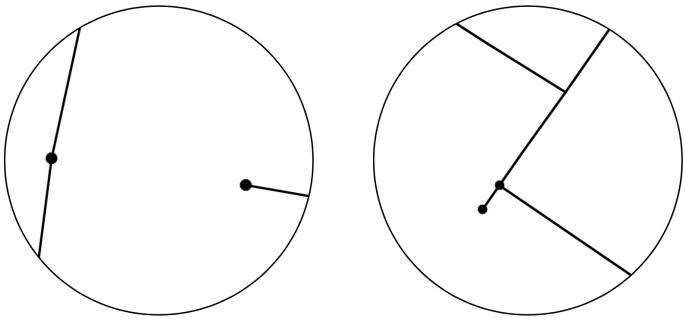
We recommend a visual analysis in order to interpret the spatial distribution of dangling nodes, with particular attention to areas of high dangling node density. It is important to understand where dangling nodes come from: are they actual dead-ends or digitization errors (e.g., over/undershoots)? A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




```
Interactive map saved at results/OSM/cph_geodk/maps_interactive/danglingmap_osm.html
```

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 undershoot occurs when two features are supposed to meet, but instead are just in close proximity to each other. An overshoot occurs when two features meet and one of them extends beyond the other. See the image below for an illustration of an undershoot (left) and an overshoot (right). For a more detailed explanation of over/undershoots, see the GIS Lounge website.



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

Method

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.

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.

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

Interpretation

Under/overshoots 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 like betweenness centrality.

```
9 potential overshoots were identified using a length tolerance of 3 m.
14 potential undershoots were identified using a length tolerance of 3 m.
```

```
Interactive map saved at results/OSM/cph_geodk/maps_interactive/underovershoots_3_3_osm.html
```

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

```
1 place(s) appear to be missing an intersection node or a bridge/tunnel tag.

Interactive map saved at results/OSM/cph_geodk/maps_interactive/intersection_issues_osm.html
```

Network 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.

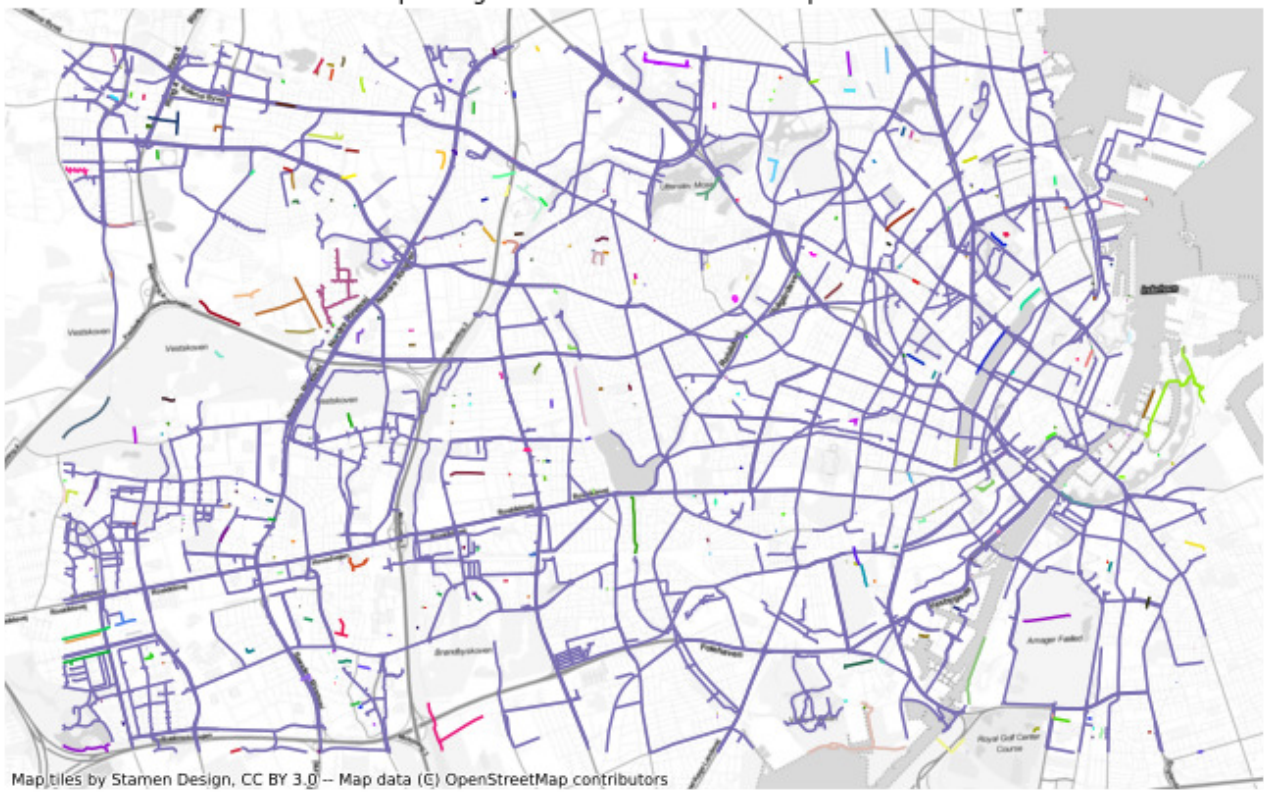
Interpretation

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.

Disconnected components

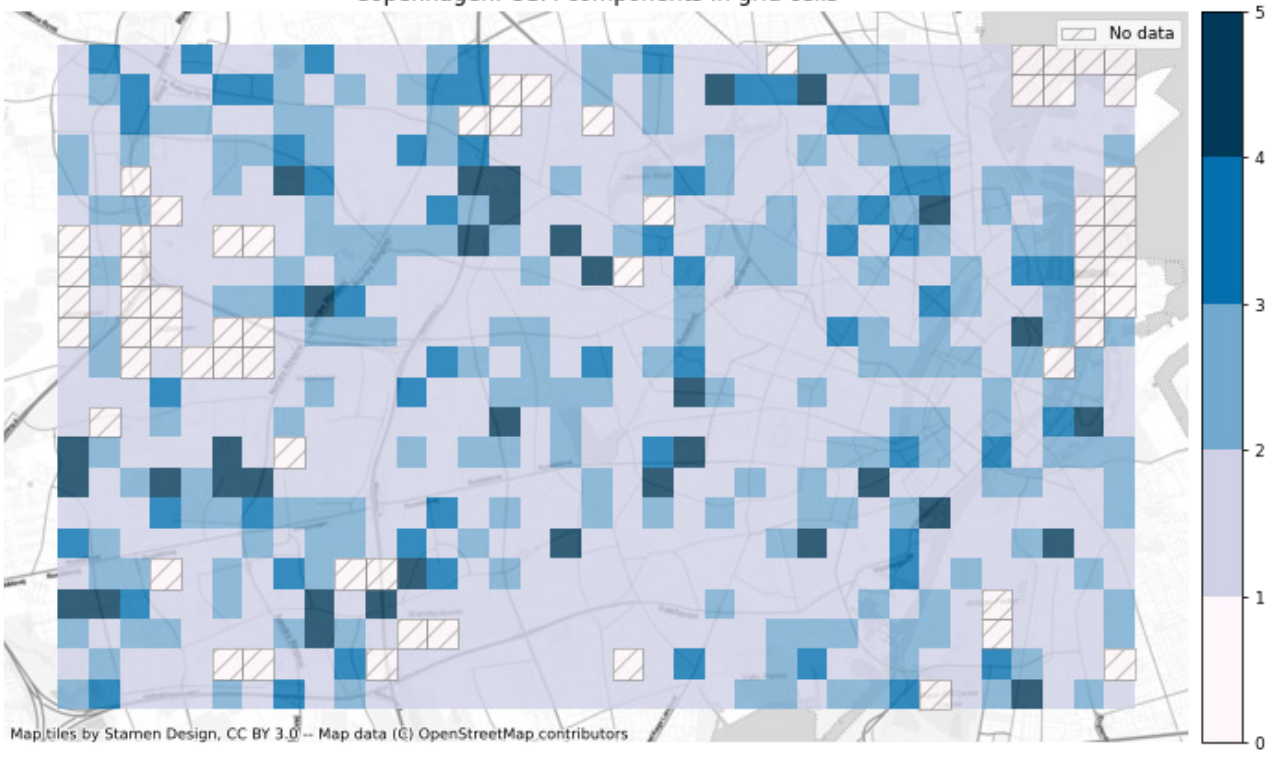
```
The network in the study area has 352 disconnected components.
```

Copenhagen: OSM disconnected components



Components per grid cell

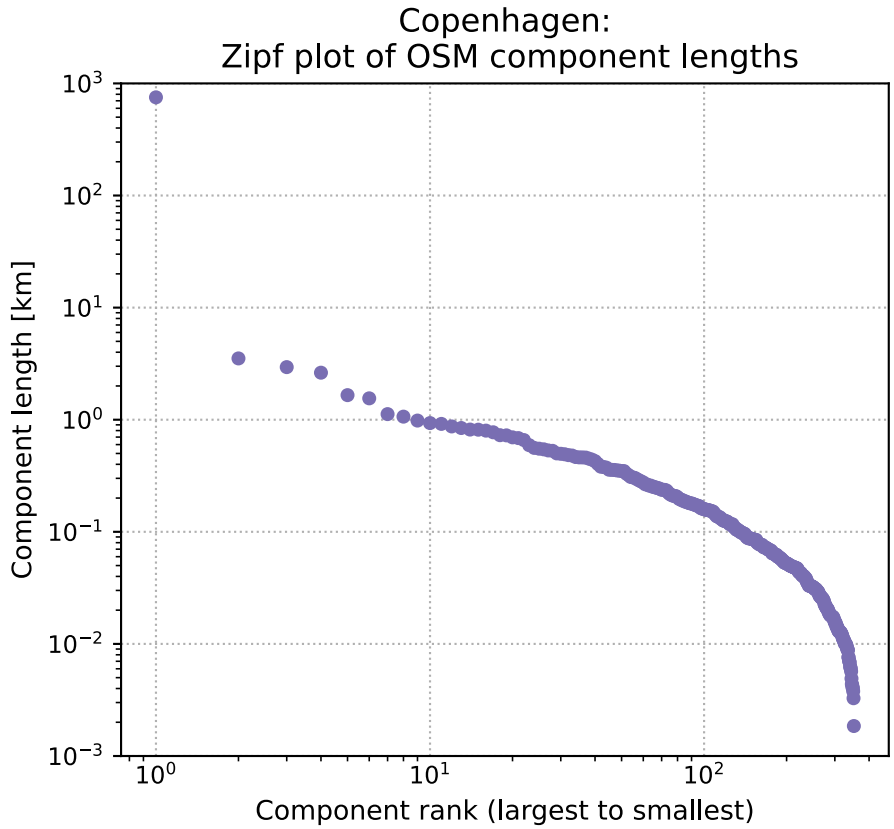
Copenhagen: OSM components in grid cells



Component size distribution

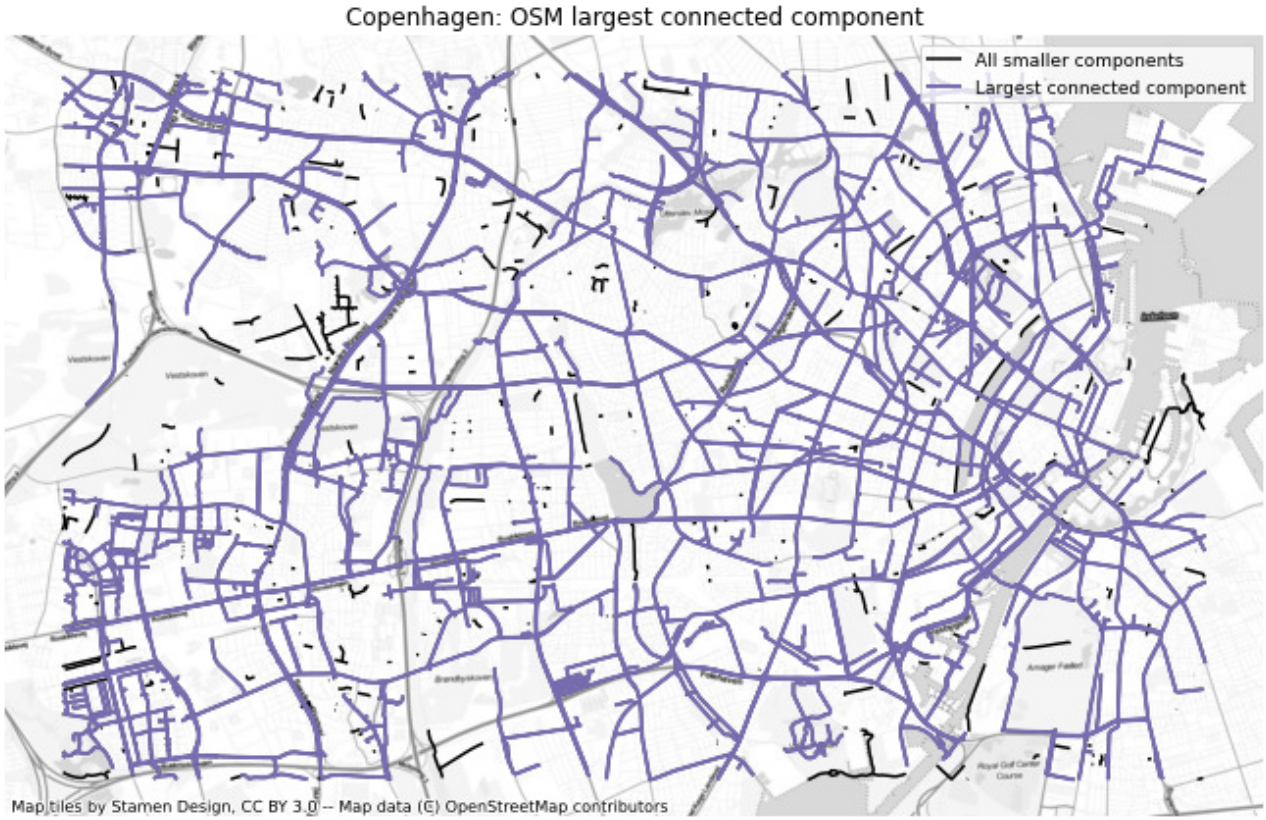
Many empirical distributions are skewed and often follow a power law, i.e. a straight line in a log-log plot, due to natural processes such as multiplicative network growth (Clauset et al., 2009). The network component size distribution (where size is length) can be visualized with a so-called Zipf plot, which plots the frequency of a component versus its rank (from largest to smallest). When a Zipf plot follows a straight line in log-log scale, it means that there is an much higher chance to find small disconnected components than expected by a distribution from an exponential family (like a normal distribution). This can mean that there has been no consolidation of the network, only piece-wise or random additions (Szell et al., 2022).

However, it can also happen that the largest connected component (the leftmost marker in the plot at rank $1 = 10^0$) is a clear outlier, while the rest of the plot follows a different shape. This can mean that a consolidation *has* taken place, and that either a central planner has deliberately targeted to connect the network, or that the data are of high enough quality to have overcome many gaps.



Largest connected component

```
The largest connected component contains 92.30% of the network length.
```



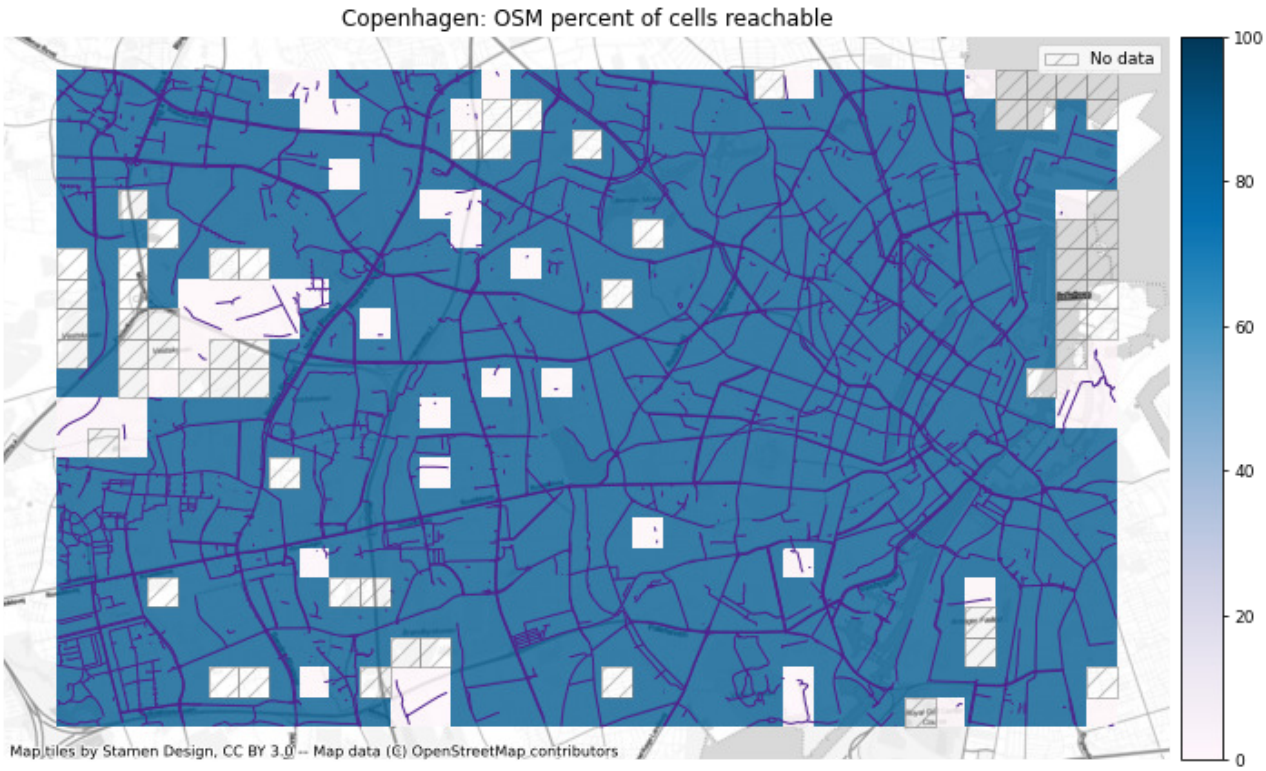
Missing links

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.

```
Analysis with component distance threshold of 10 meters:


Interactive map saved at results/OSM/cph_geodk/maps_interactive/component_gaps_10_osm.html
```

Component connectivity

Here we visualize 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 quickly highlight stark differences in network connectivity.



Summary

Intrinsic Quality Metrics - OSM data

Total infrastructure length (km)	1,063
Protected bicycle infrastructure density (m/km2)	5,303
Unprotected bicycle infrastructure density (m/km2)	499
Mixed protection bicycle infrastructure density (m/km2)	59
Bicycle infrastructure density (m/km2)	5,861
Nodes	4,925
Dangling nodes	1,818
Nodes per km2	27
Dangling nodes per km2	10
Incompatible tag combinations	2
Overshoots	9

Undershoots	14
Missing intersection nodes	1
Components	352
Length of largest component (km)	752
Largest component's share of network length	92%
Component gaps	91

2a. Initialize 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 attributes needed in the analysis.

Sections

Load data for study area and create analysis grid

This step:

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

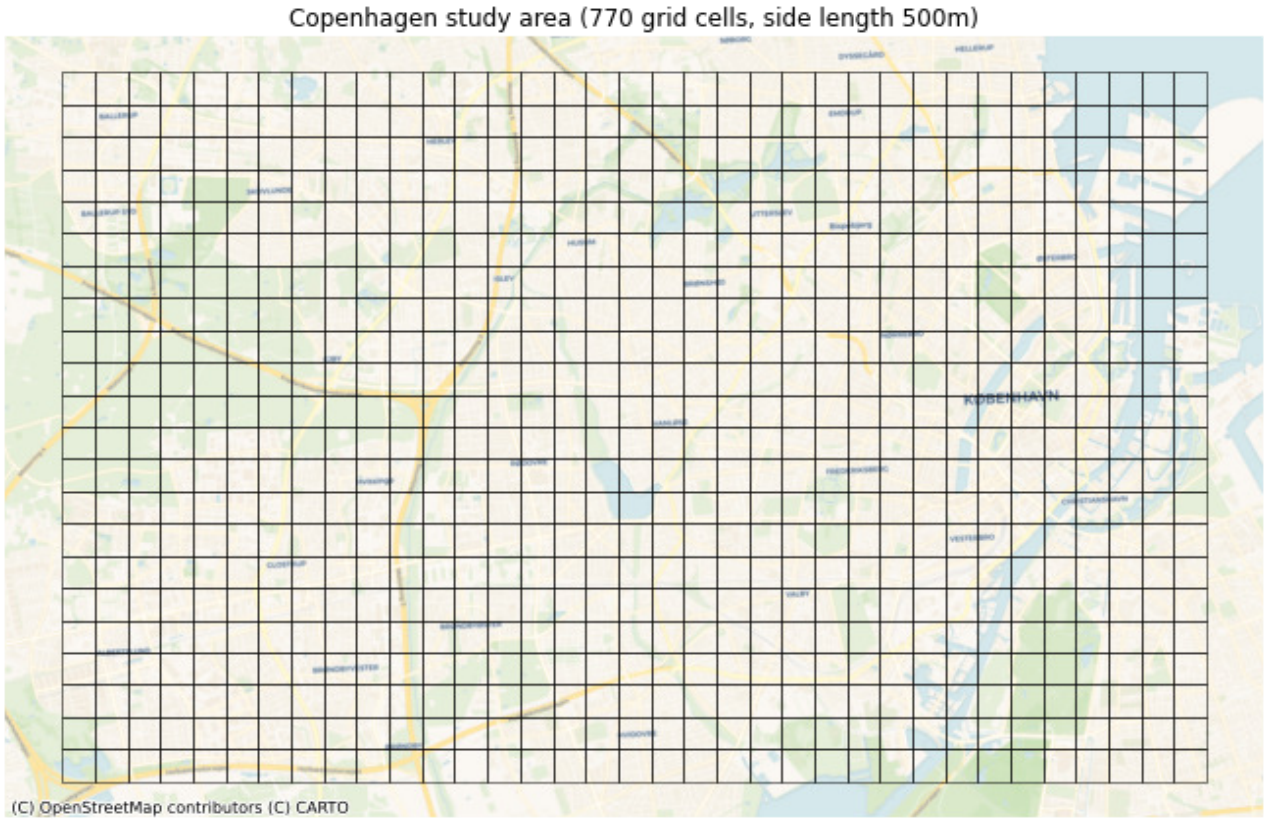
Load data for study area

The study area is defined by the user-provided polygon. It will be used for the computation of **global** results, i.e. quality metrics based on all data in the study area.

```
The size of the study area is 181.38 km2.
```

Create analysis grid

```
The grid contains 770 square cells with a side length of 500 m and an area of 0.25 km2.
This grid will be used for local (grid cell level) analysis:
```



Load and preprocess 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 data sets indexed by their intersecting grid cell.

Network data model

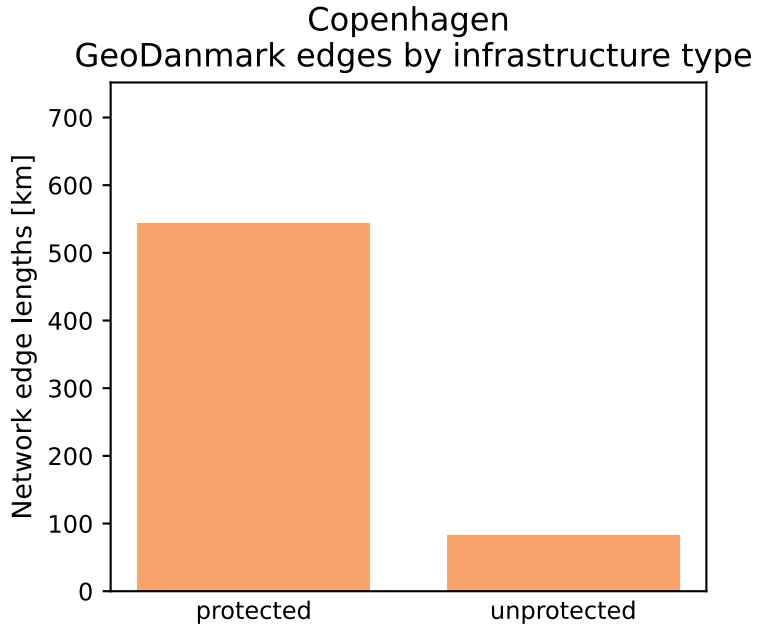
In BikeDNA, all input data are converted to a network structure consisting of *nodes* and *edges*. Edges represents the actual infrastructure, such as bike lanes and paths, while nodes represents the start and end points for the edges, as well as all intersections. For further details, read more about the network data model.

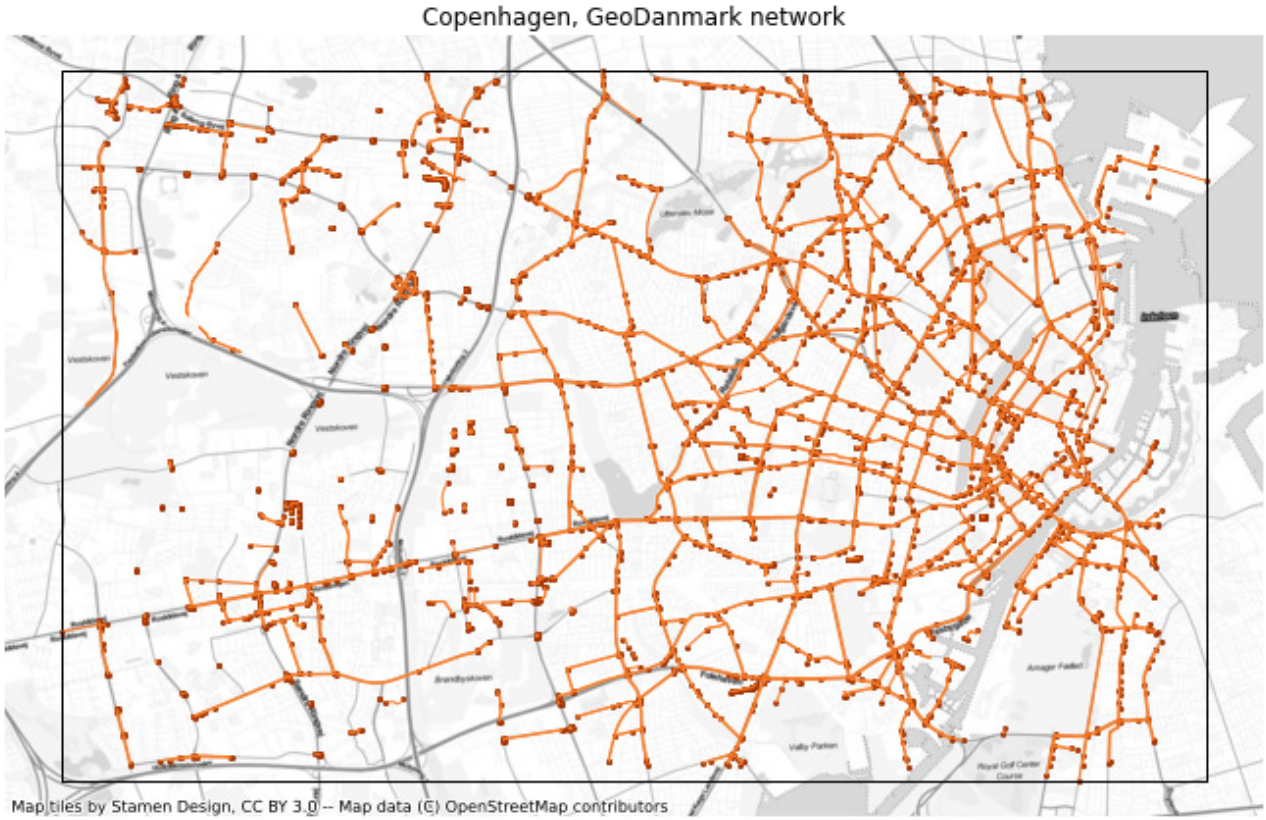
▶ **Network simplification**

```
The GeoDanmark data covers an area of 169.76 km2.
```

```
Edges where the protection level is 'protected': 46097 out of 53580 (86.03%)
Edges where the protection level is 'unprotected': 7483 out of 53580 (13.97%)
```

```
Using global settings for cycling direction.

Using global settings for geometry type.

The length of the GeoDanmark network is 626.48 km.
```




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

2b. Intrinsic Analysis of Reference 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 one 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 data set as the ground truth, no universal claims of data quality can be made. The idea is rather to enable those working with bicycle networks to assess whether their data are 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 Ferster et al. (2020), Hochmair et al. (2015), Barron et al. (2014), and Neis et al. (2012).

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

Data completeness

Network density

In this setting, network density refers to the length of edges or number of nodes per km2. This is the usual definition of network density in spatial (road) networks, which is distinct from the *structural* network density known more generally in network science. Without comparing to a reference data set, network density 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.

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 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:

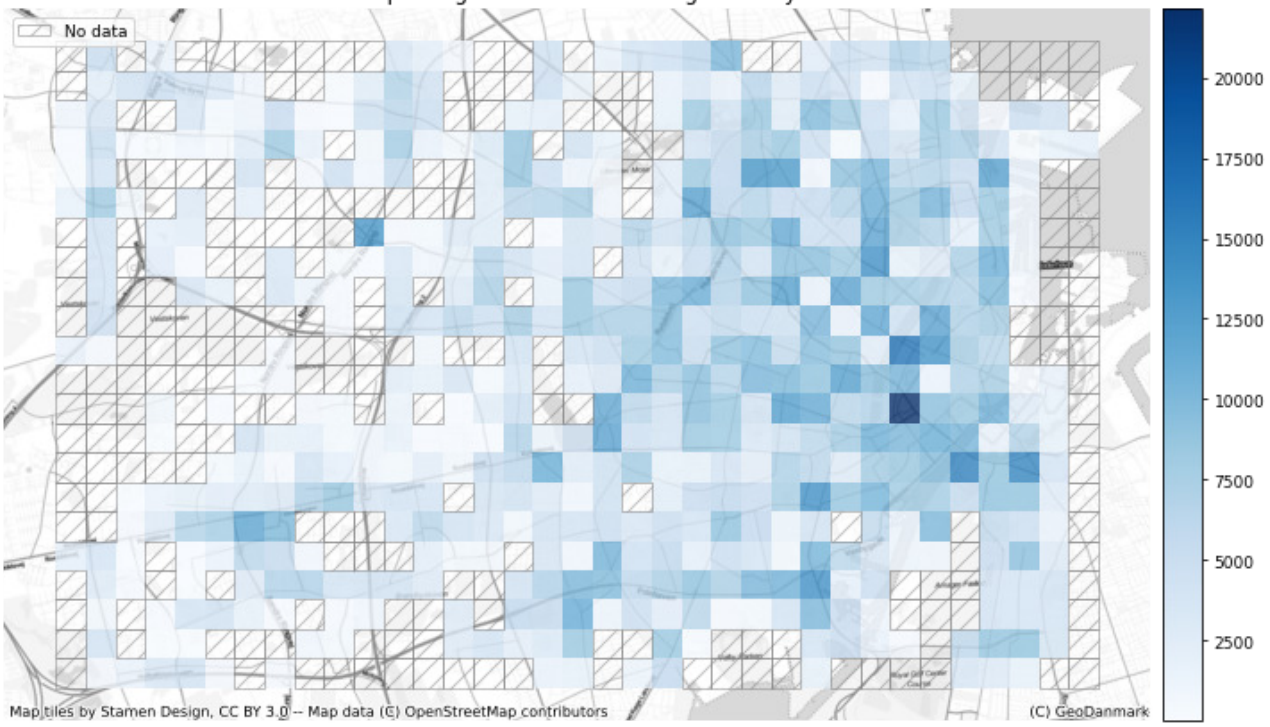
- lower-than-average infrastructure density indicates a locally sparser network
- higher-than-average node density indicates that there are relatively many intersections in a grid cell
- higher-than-average dangling node density indicates that there are relatively many dead ends in a grid cell

Global network density

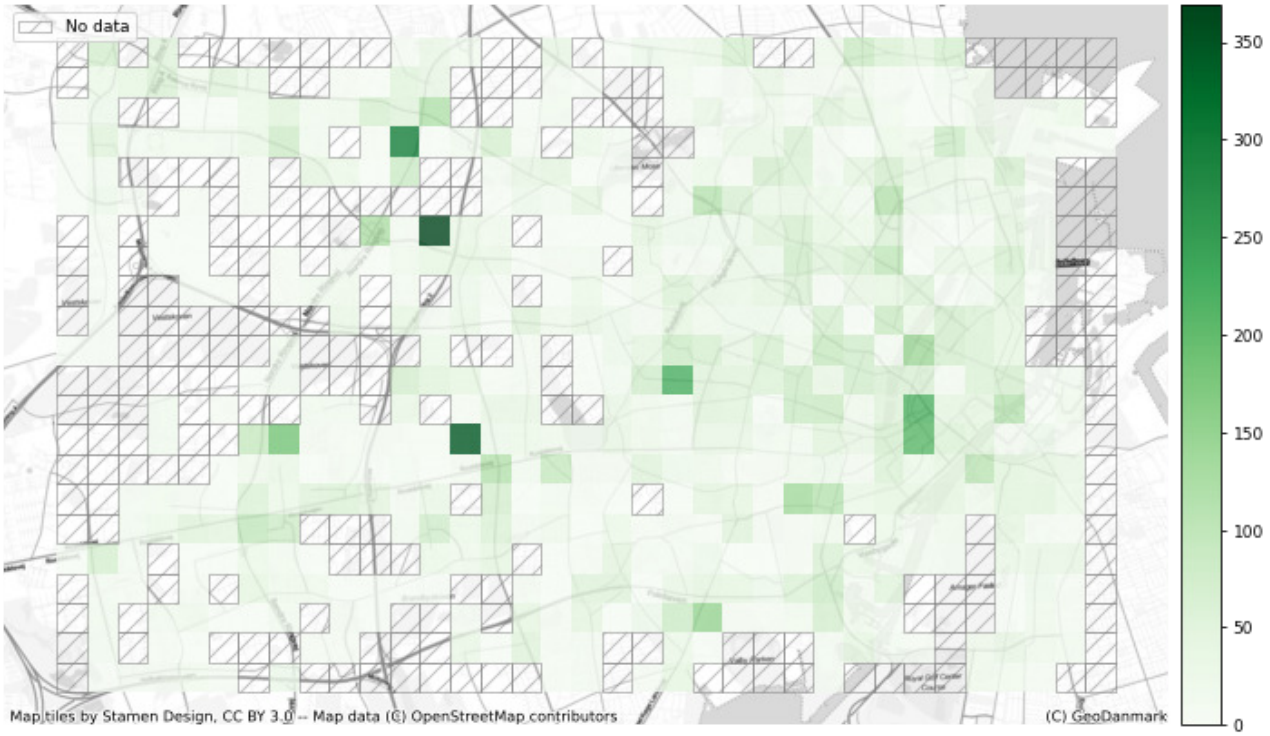
```
For the entire study area, there are:
- 3453.85 meters of bicycle infrastructure per km2.
- 22.74 nodes in the bicycle network per km2.
- 4.80 dangling nodes in the bicycle network per km2.
- 2998.80 meters of protected bicycle infrastructure per km2.
- 455.05 meters of unprotected bicycle infrastructure per km2.
- 0.00 meters of mixed protection bicycle infrastructure per km2.
```

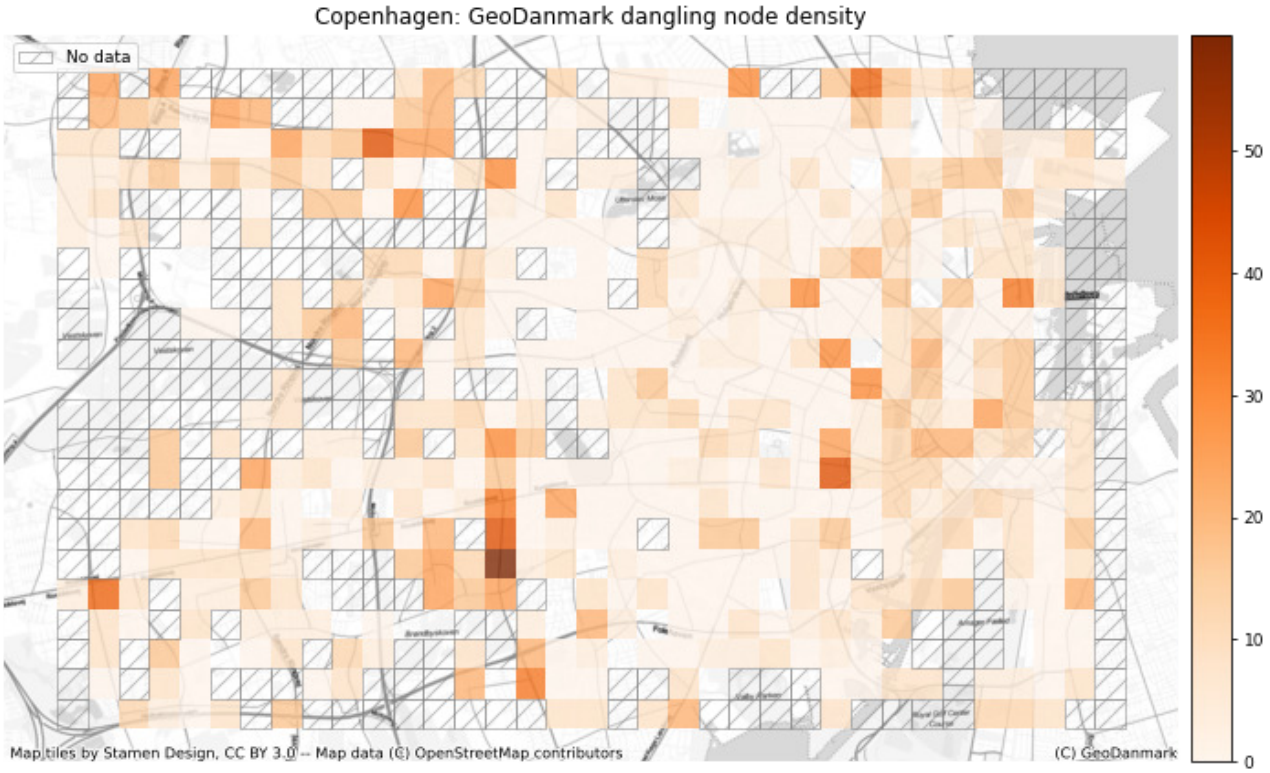
Local network density

Copenhagen: GeoDanmark edge density



Copenhagen: GeoDanmark node density

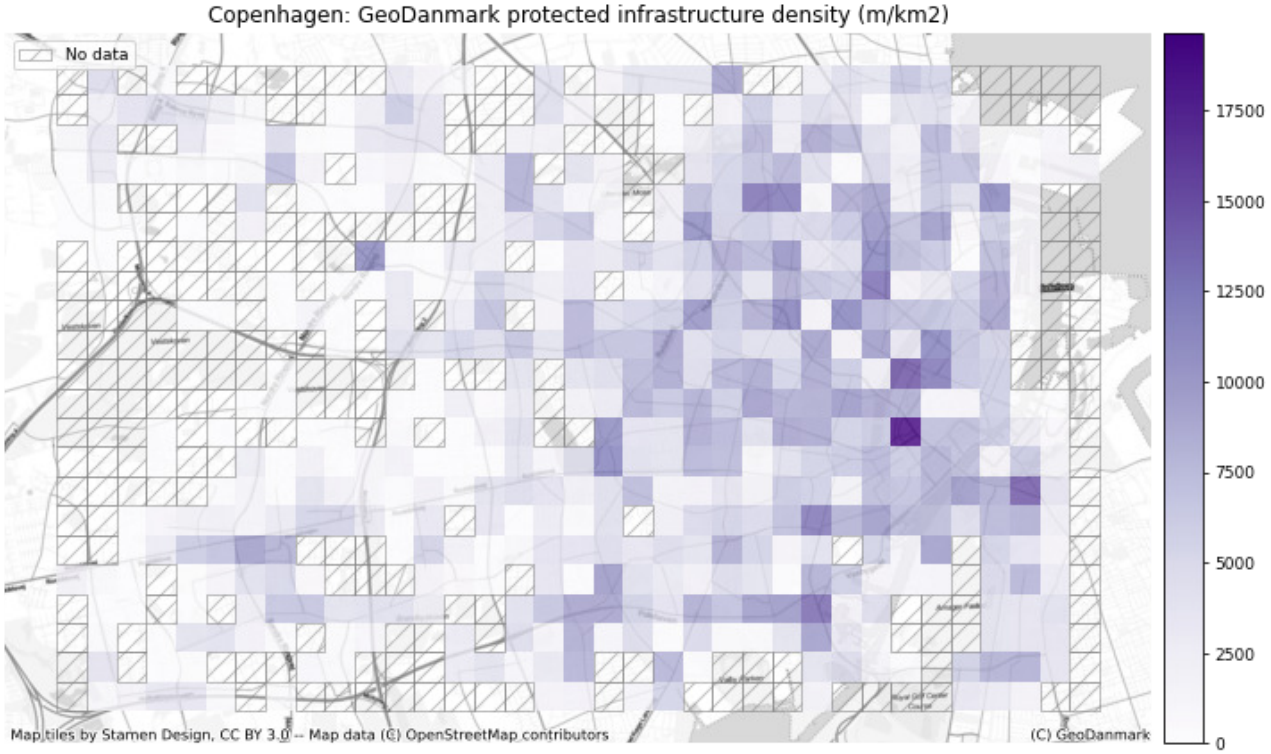


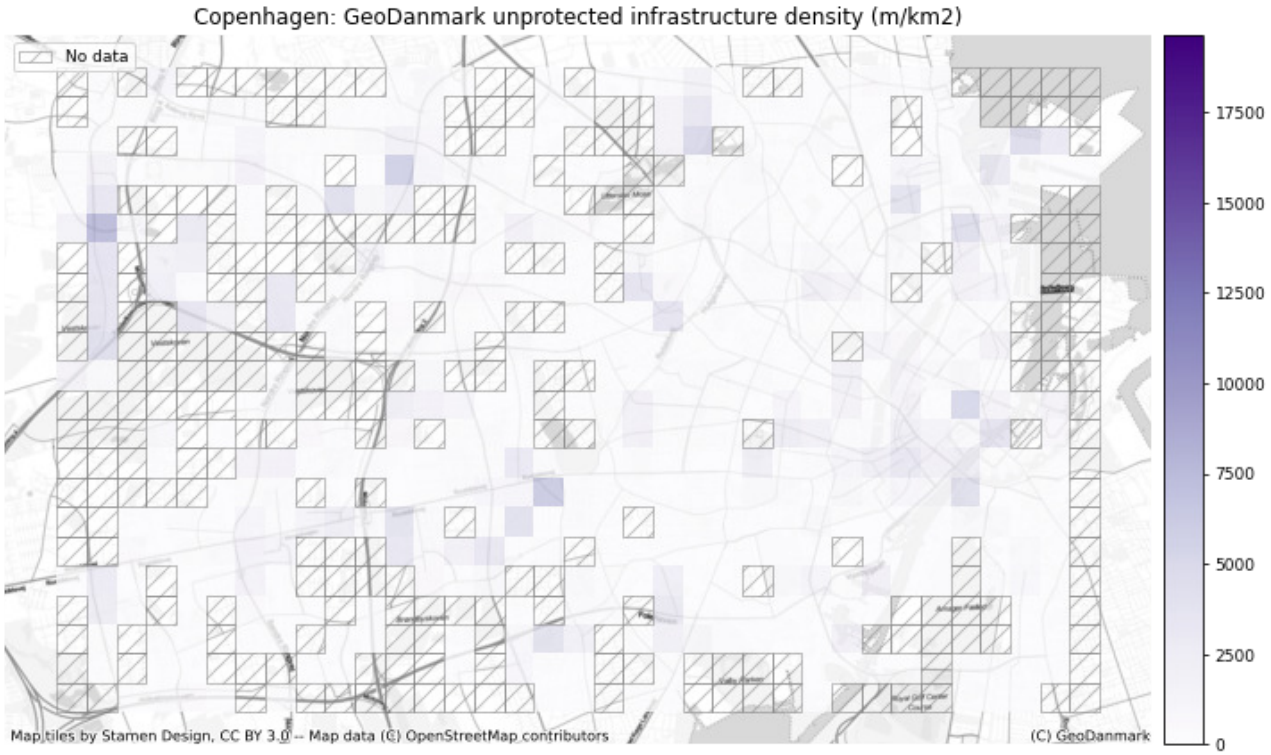


Densities of protected and unprotected infrastructure

In BikeDNA, *protected infrastructure* refers to all bicycle infrastructure which is either separated from car traffic by for example an elevated curb, bollards, or other physical barriers, or for cycle tracks that are not adjacent to a street.

Unprotected infrastructure are all other types of lanes that are dedicated for bicyclists, but which only are separated by car traffic by e.g., a painted line on the street.





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.

Simplification outcome

To compare the structure and true ratio between nodes and edges in the network, a simplified network representation which only includes nodes at endpoints and intersections was created in notebook `1b` by removing all interstitial nodes.

Comparing the degree distribution for the networks before and after simplification is a quick sanity check for the simplification routine. Typically, the vast 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.

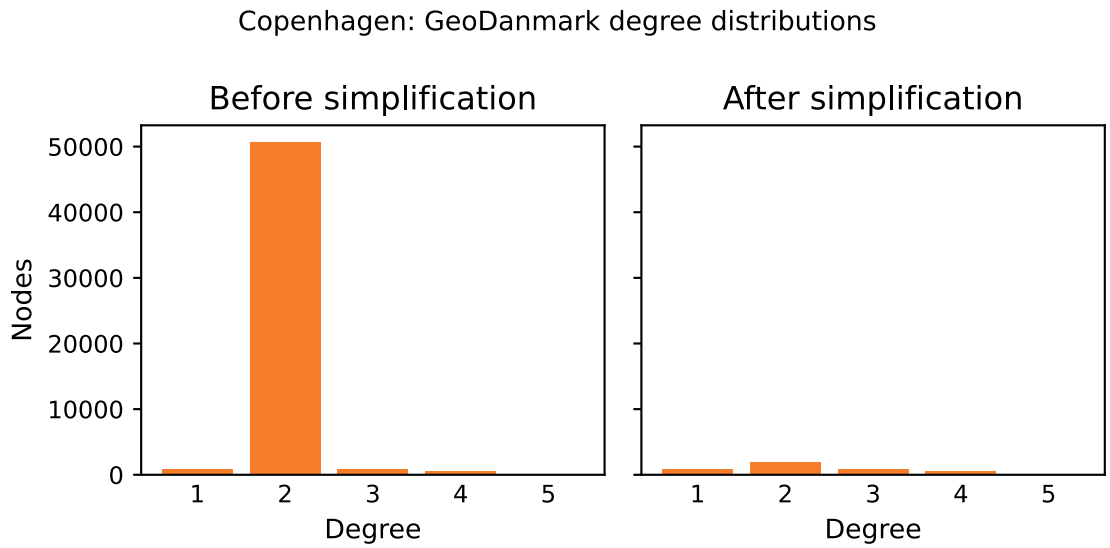
Method

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.

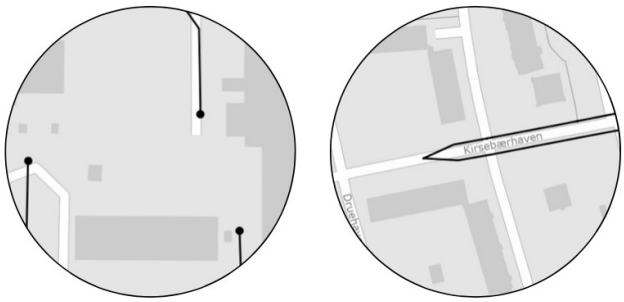
```
Simplifying the network decreased the number of edges with 91.2% and the number of nodes with 9
2.2%.
```



Dangling nodes

Dangling nodes are nodes of degree one, i.e. they 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 (see 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 containing many dead-ends can indicate digitization errors and problems with edge over/undershoots.



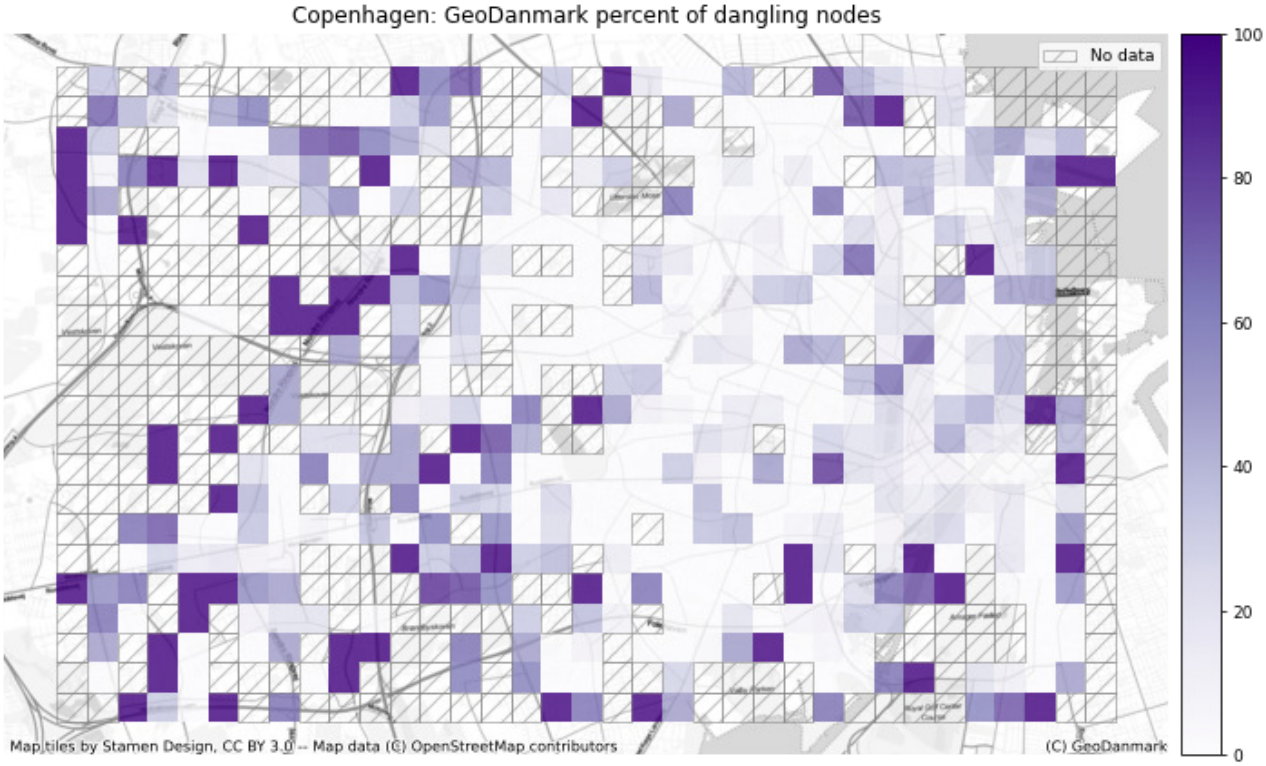
Left: Dangling nodes occur where road features end. Right: However, when separate features are joined at the end, there will be no dangling nodes.

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 color, all other nodes are shown in black.

Interpretation

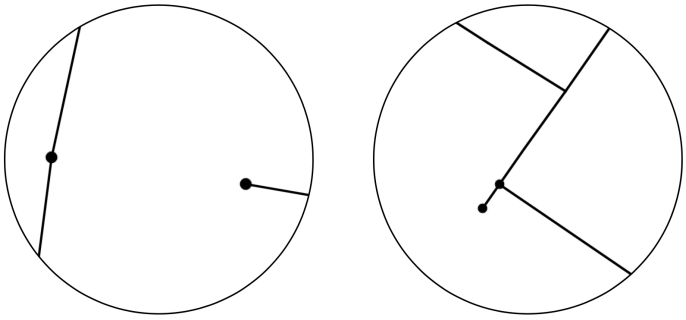
We recommend a visual analysis in order to interpret the spatial distribution of dangling nodes, with particular attention to areas of high dangling node density. It is important to understand where dangling nodes come from: are they actual dead-ends or digitization errors (e.g., over/undershoots)? A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




```
Interactive map saved at results/REFERENCE/cph_geodk/maps_interactive/folium_danglingmap_referen
ce.html
```

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 undershoot occurs when two features are supposed to meet, but instead are just in close proximity to each other. An overshoot occurs when two features meet and one of them extends beyond the other. See the image below for an illustration of an undershoot (left) and an overshoot (right). For a more detailed explanation of over/undershoots, see the GIS Lounge website.



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

Method

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.

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.

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

Interpretation

Under/overshoots 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.

```
21 potential overshoots were identified with a length tolerance of 3 m.
11 potential undershoots were identified with a length tolerance of 3 m.


Interactive map saved at results/REFERENCE/cph_geodk/maps_interactive/overundershoots_3_3_refere
nce.html
```

Network 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

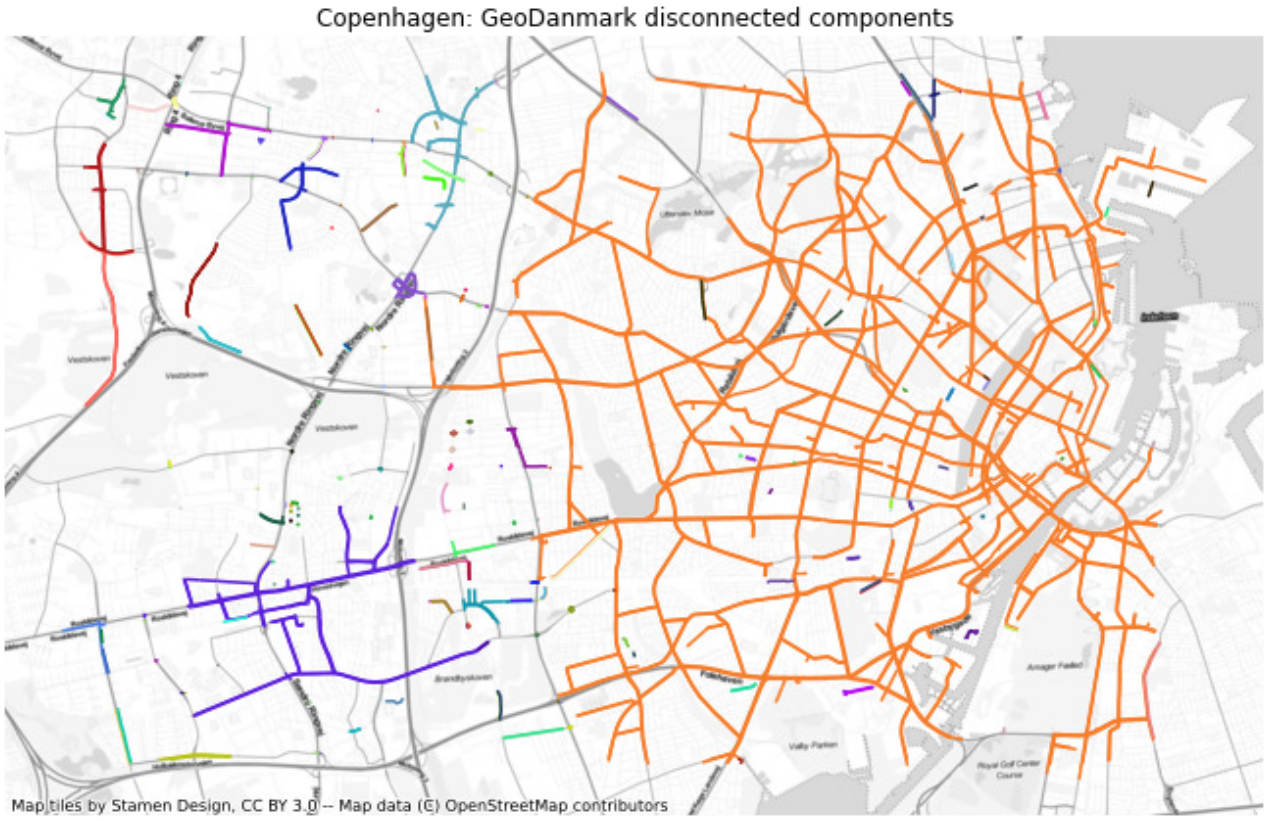
First, with the help of `return_components`, a list of all (disconnected) components of the network is obtained. The total number of components is printed and all components are plotted in different colors

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.

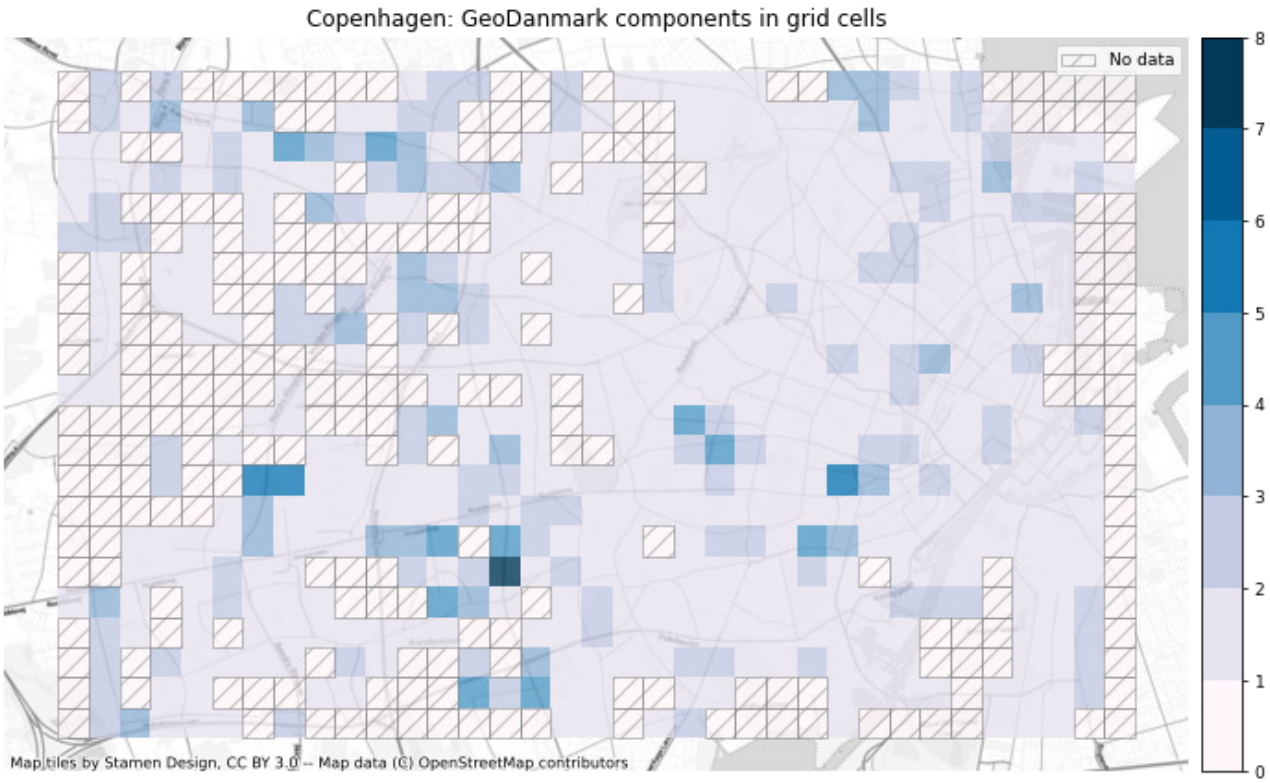
Interpretation

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.

```
The network in the study area has 204 disconnected components.
```



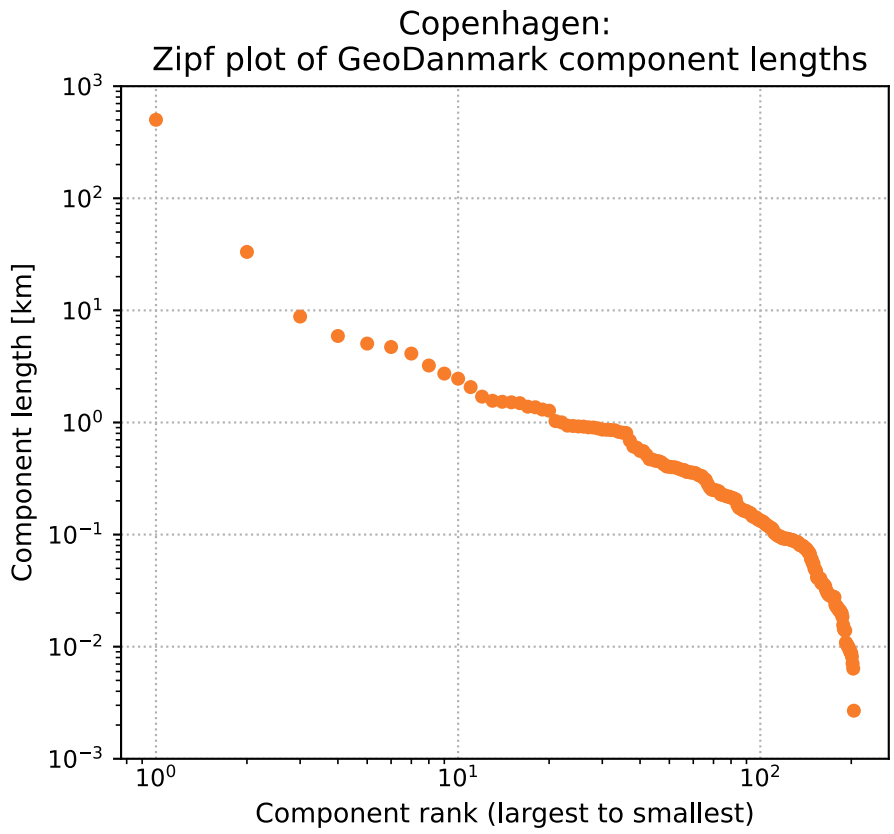
Components per grid cell



Component size distribution

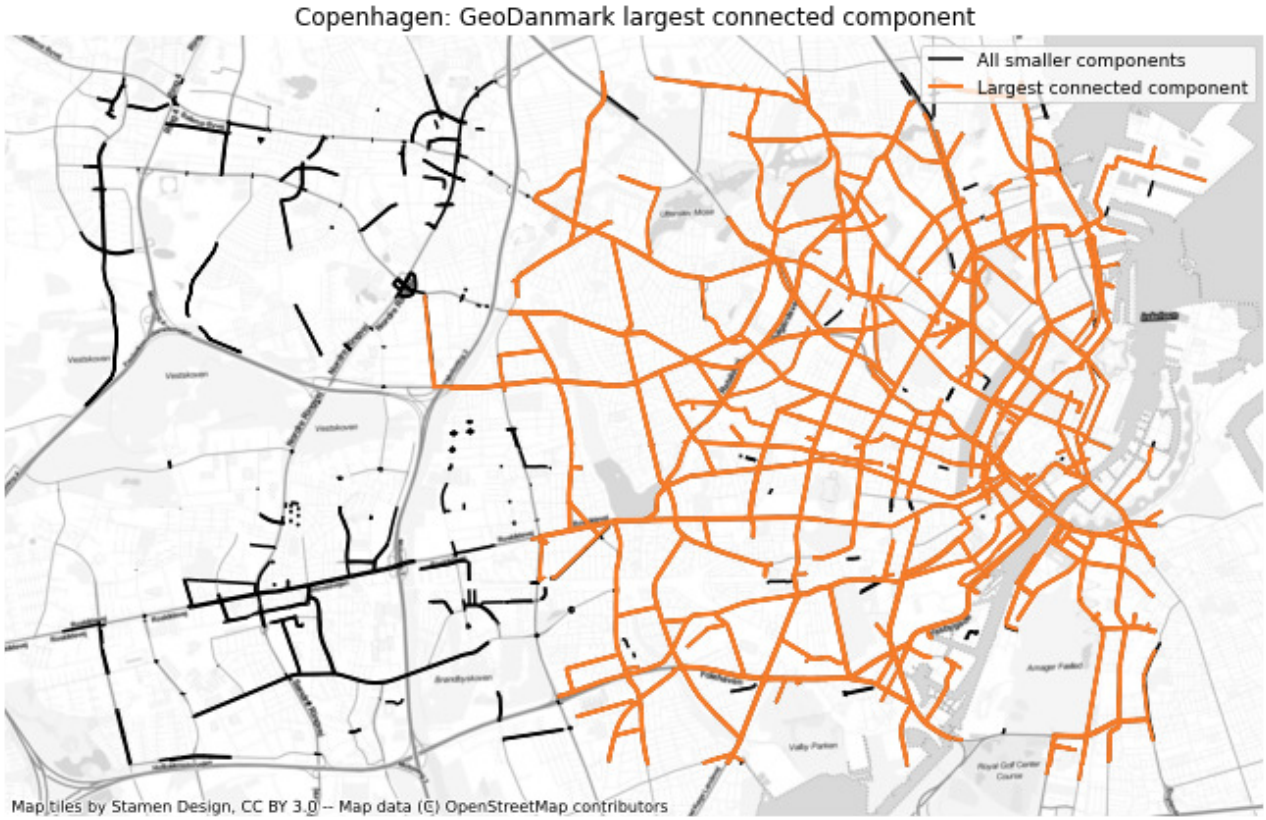
Many empirical distributions are skewed and often follow a power law, i.e. a straight line in a log-log plot, due to natural processes such as multiplicative network growth (Clauset et al., 2009). The network component size distribution (where size is length) can be visualized with a so-called Zipf plot, which plots the frequency of a component versus its rank (from largest to smallest). When a Zipf plot follows a straight line in log-log scale, it means that there is an much higher chance to find small disconnected components than expected by a distribution from an exponential family (like a normal distribution). This can mean that there has been no consolidation of the network, only piece-wise or random additions (Szell et al., 2022).

However, it can also happen that the largest connected component (the leftmost marker in the plot at rank 10^0) is a clear outlier, while the rest of the plot follows a different shape. This can mean that a consolidation *has* taken place, and that either a central planner has deliberately targeted to connect the network, or that the data are of high enough quality to have overcome many gaps.



The largest connected component contains 80.04% of the network length.

Largest connected component



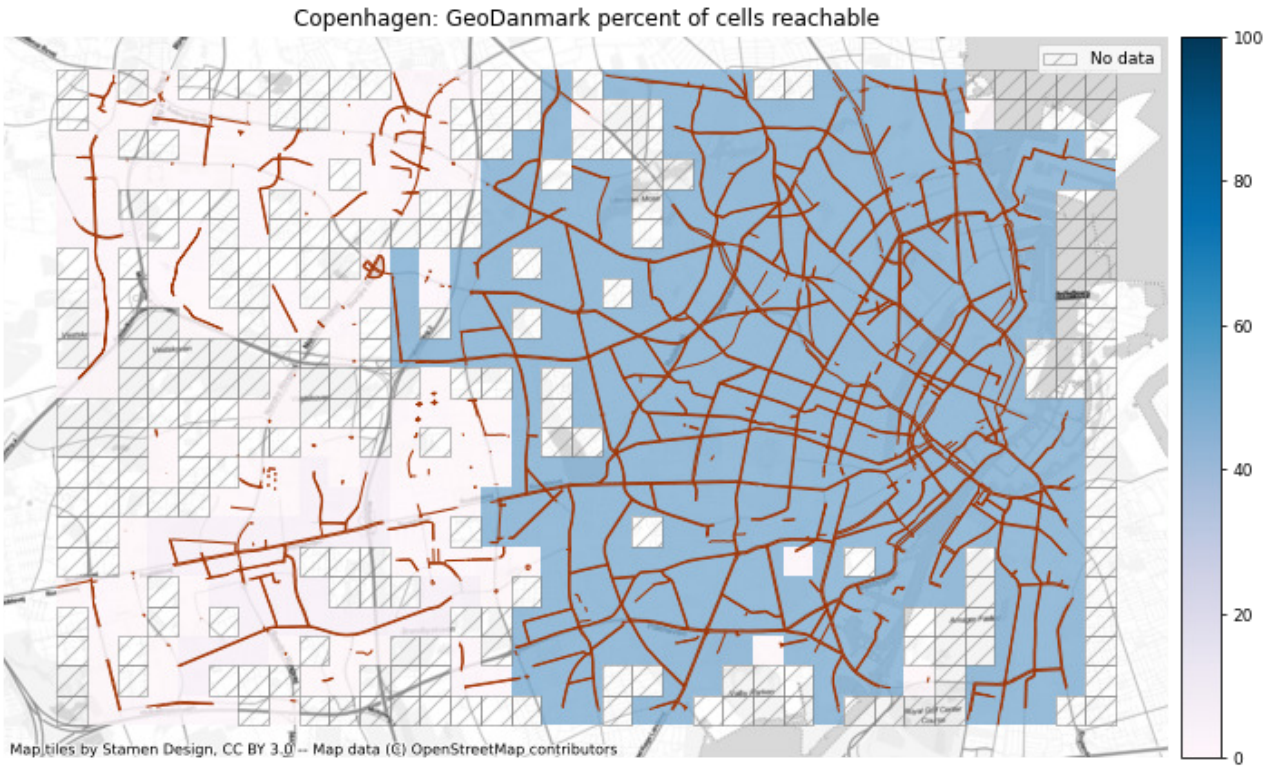
Missing links

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.

```
Interactive map saved at results/REFERENCE/cph_geodk/maps_interactive/component_gaps_10_referenc
e.html
```

Component connectivity

Here we visualize 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 quickly highlight stark differences in network connectivity.



Summary

Intrinsic Quality Metrics - GeoDanmark Data

Total infrastructure length (km)	626
Protected bicycle infrastructure density (m/km2)	2,999
Unprotected bicycle infrastructure density (m/km2)	455
Mixed protection bicycle infrastructure density (m/km2)	0
Bicycle infrastructure density (m/km2)	3,454
Nodes	4,125
Dangling nodes	870
Nodes per km2	23
Dangling nodes per km2	5
Overshoots	21
Undershoots	11

Components	204
Length of largest component (km)	501
Largest component's share of network length	80%
Component gaps	52

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

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

This analysis is based on comparing the reference data set to OSM and highlighting how and where they differ, both in terms of *how much* bicycle infrastructure is mapped in the two data sets, and of *how* the infrastructure is mapped, pinpointing differences in network structure.

All differences are computed for the reference data in relation to OSM. For example, the difference in network density is computed by calculating reference density minus OSM density. Hence, positive difference values (over 0) indicate how much higher the reference value is; negative difference values (below 0) indicate how much lower the reference value is. Accordingly, if differences are given in percent, the reference value is taken to be the total value (100%).

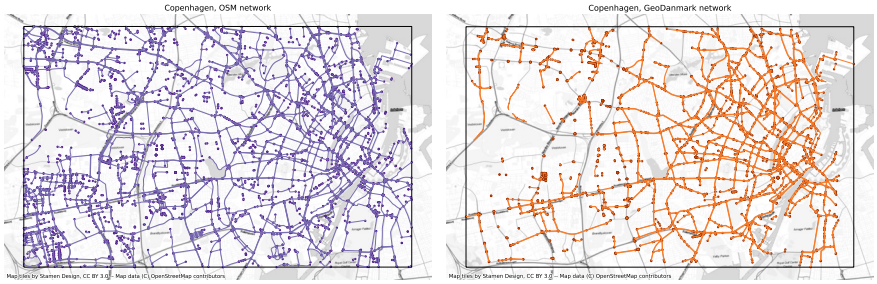
While the analysis is based on a comparison, it makes no a priori assumptions about which data set is better. The same goes for the identified differences: BikeDNA does not allow an automatic conclusion as to which data set is of better quality, but instead requires the user to interpret the meaning of the differences found, e.g., whether differing features are results of errors of omission or commission, and which data set is more correct. However, many low values can be an indication that the reference data is of lower completeness or quality than the OSM data.

The goal is that the identified differences can be used to both assess the quality of the reference and OSM data sets, and to support the decision of which data set 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

OSM versus reference network



Data completeness

This section compares the OSM and reference data sets in terms of data completeness. The goal is to identify whether one data set 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 km2) 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.

Computing gridded local density differences as a measure of data quality has also been applied by e.g. Haklay (2010).

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).

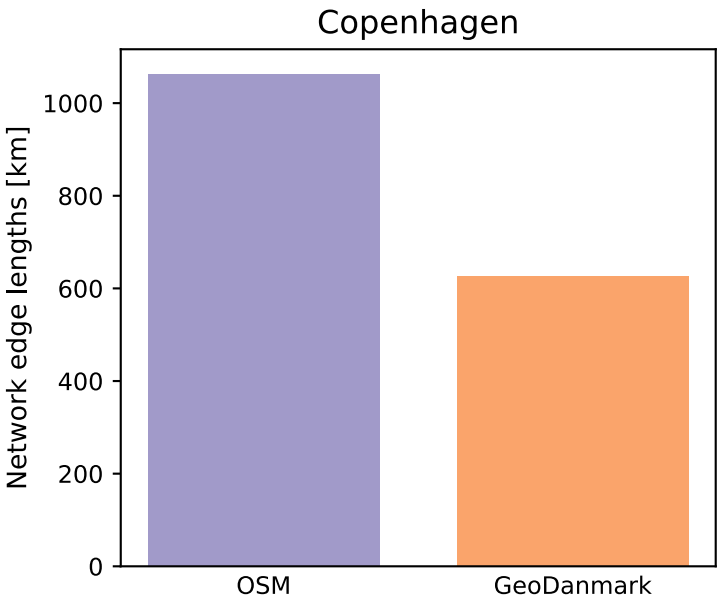
Interpretation

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 can indicate unmapped grid cell features in the reference data set, or that a street mistakenly has been tagged as bicycle infrastructure in OSM.

Network length

```
Length of the OSM data set: 1063.18 km
Length of the reference data set: 626.48 km
```

```
The OSM data set is 436.70 km longer than the reference data set.
The OSM data set is 41.08% longer than the reference data set.
```

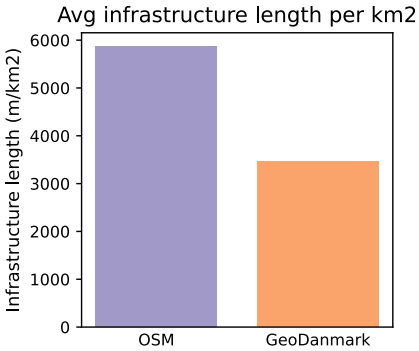


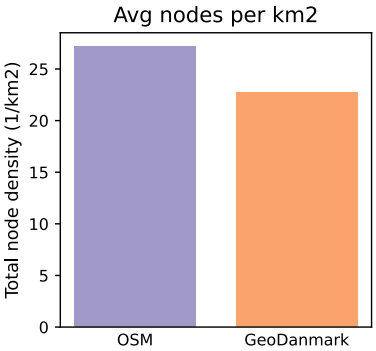
Network Density

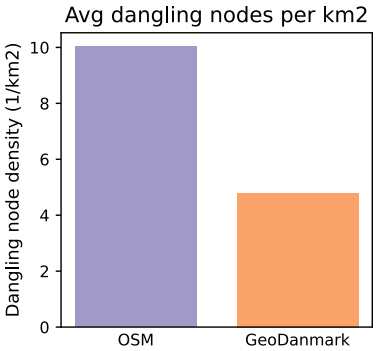
Global network densities

```
In the OSM data, there are:
 - 5861.46 meters of cycling infrastructure per km2.
 - 27.15 nodes in the cycling network per km2.
 - 10.02 dangling nodes in the cycling network per km2.


In the reference data, there are:
 - 3453.85 meters of cycling infrastructure per km2.
 - 22.74 nodes in the cycling network per km2.
 - 4.80 dangling nodes in the cycling network per km2.
```

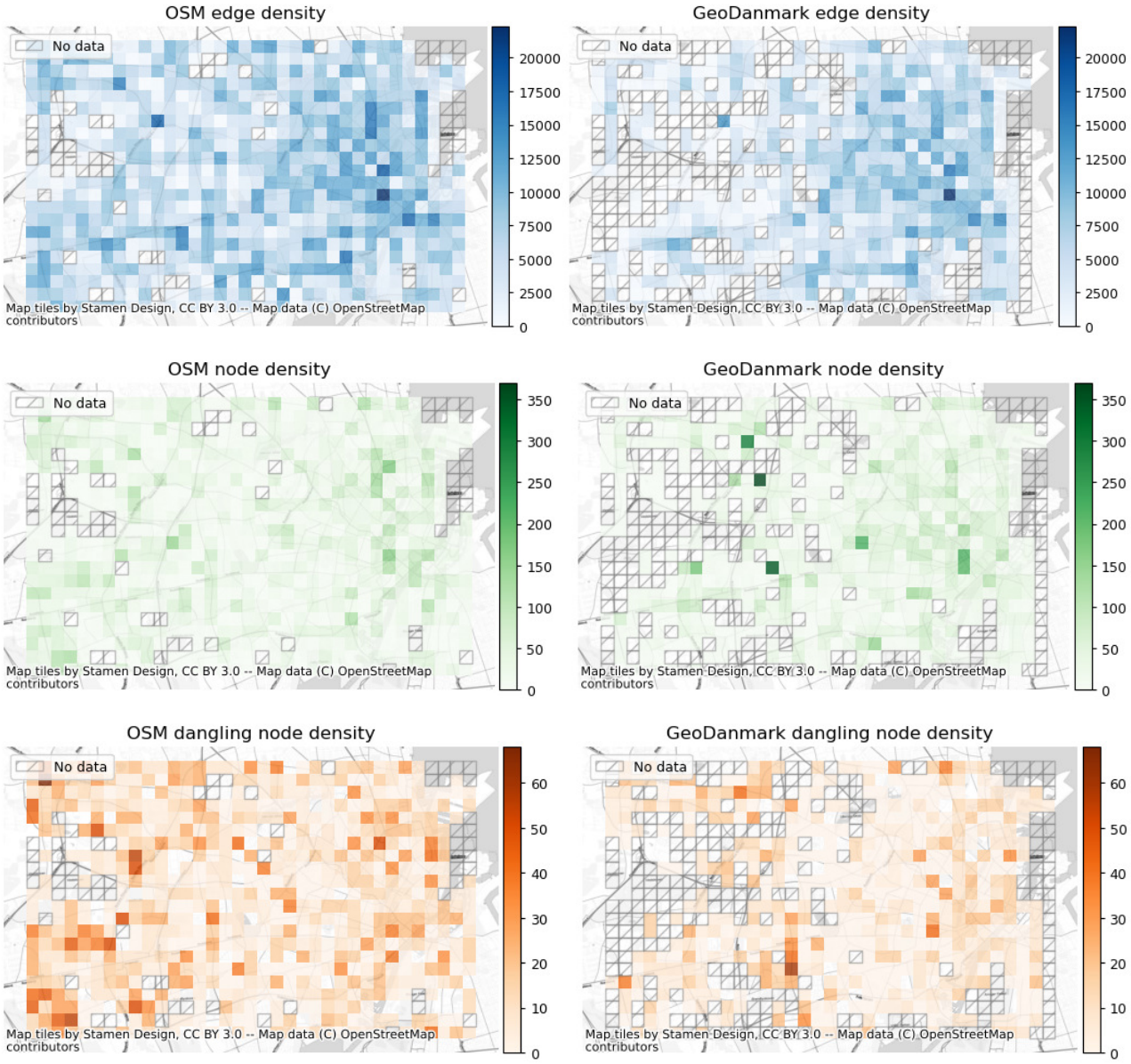
Global network densities (per km2)







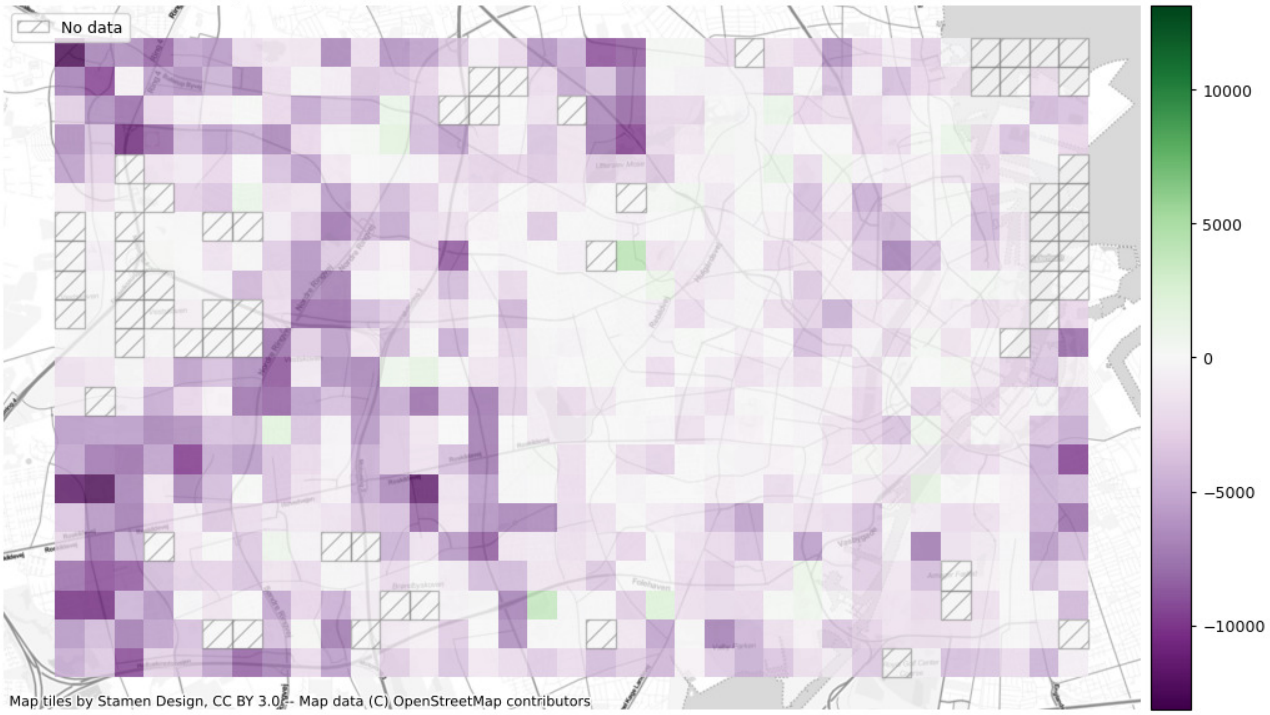
Local network densities



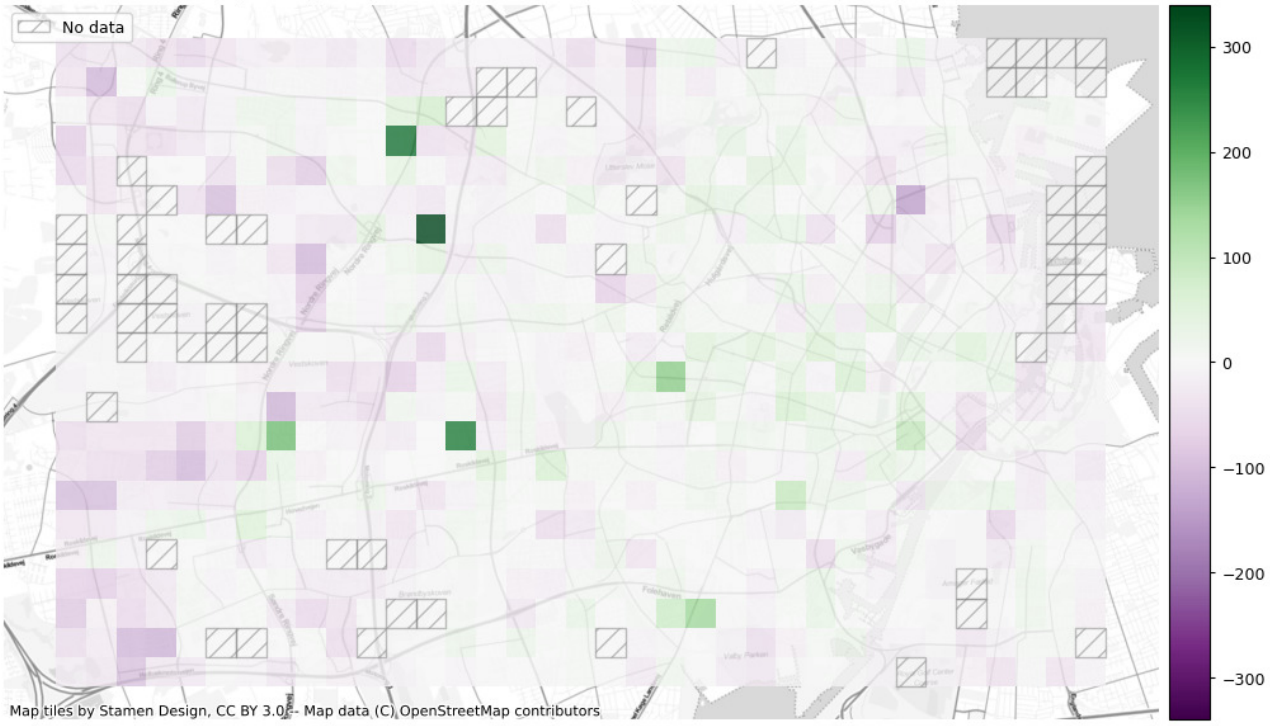
Local differences in network densities

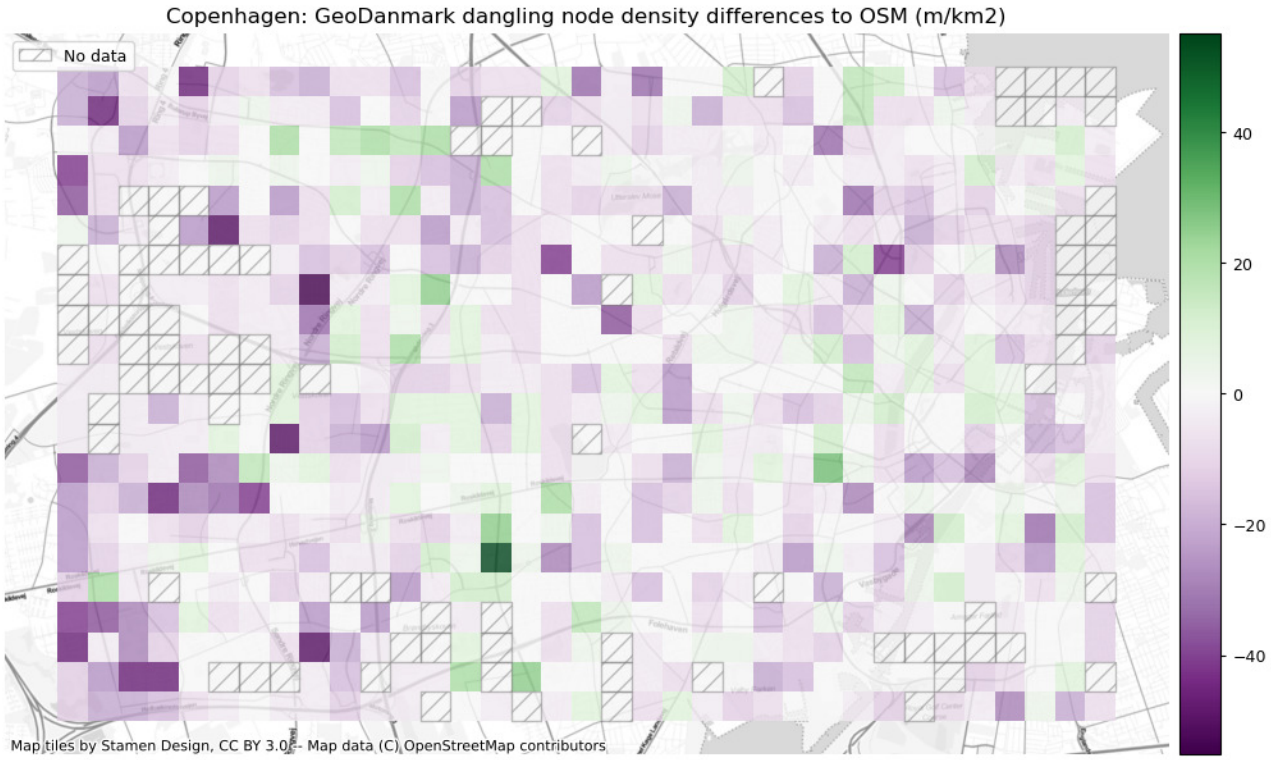
The densities in the OSM data set are taken as base line for comparison. Hence, positive values indicate that the OSM density of the infrastructure type is higher than the reference density; negative values indicate that the OSM density is lower than the reference density.

Copenhagen: GeoDanmark edge density differences to OSM (m/km2)



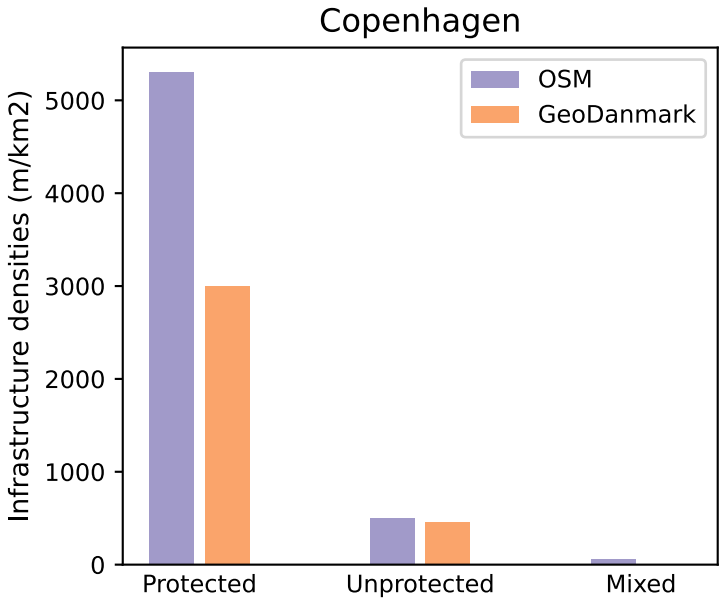
Copenhagen: GeoDanmark node density differences to OSM (m/km2)





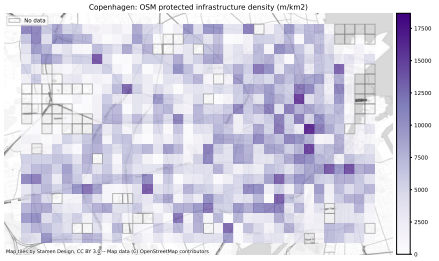
Densities of protected and unprotected bicycle infrastructure

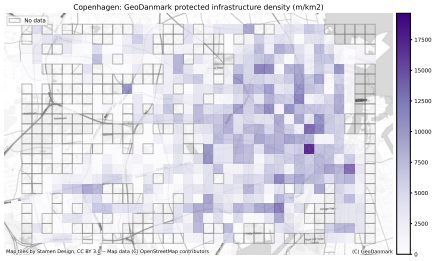
Global network densities for protected/unprotected infrastructure

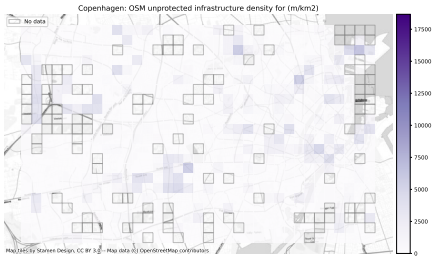


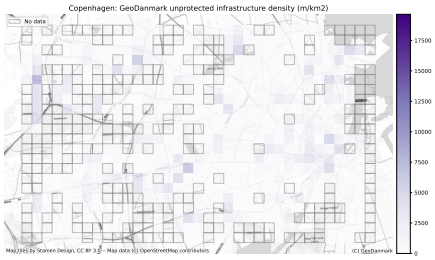
Local network densities for protected/unprotected infrastructure

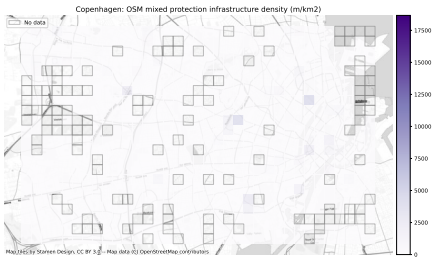
The densities in the OSM data set are taken as base line for comparison. Hence, positive values indicate that the OSM density of the infrastructure type is higher than the reference density; negative values indicate that the OSM density is lower than the reference density.





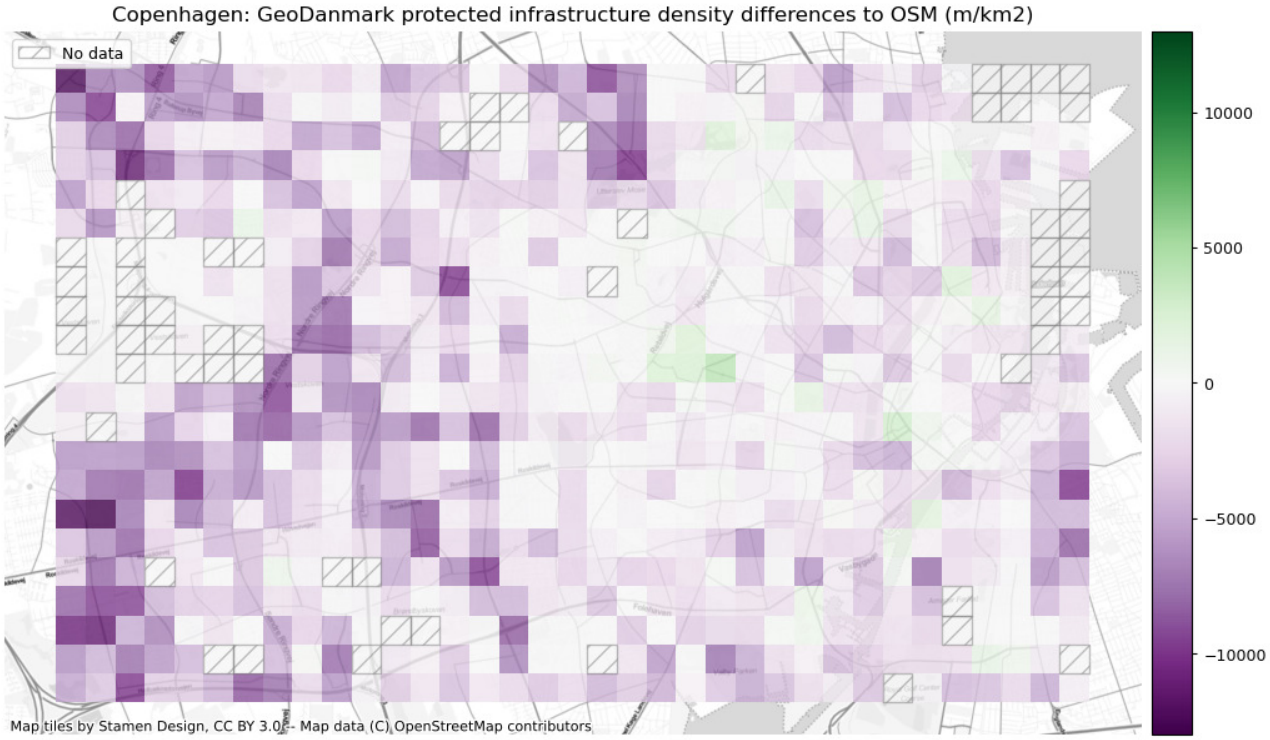


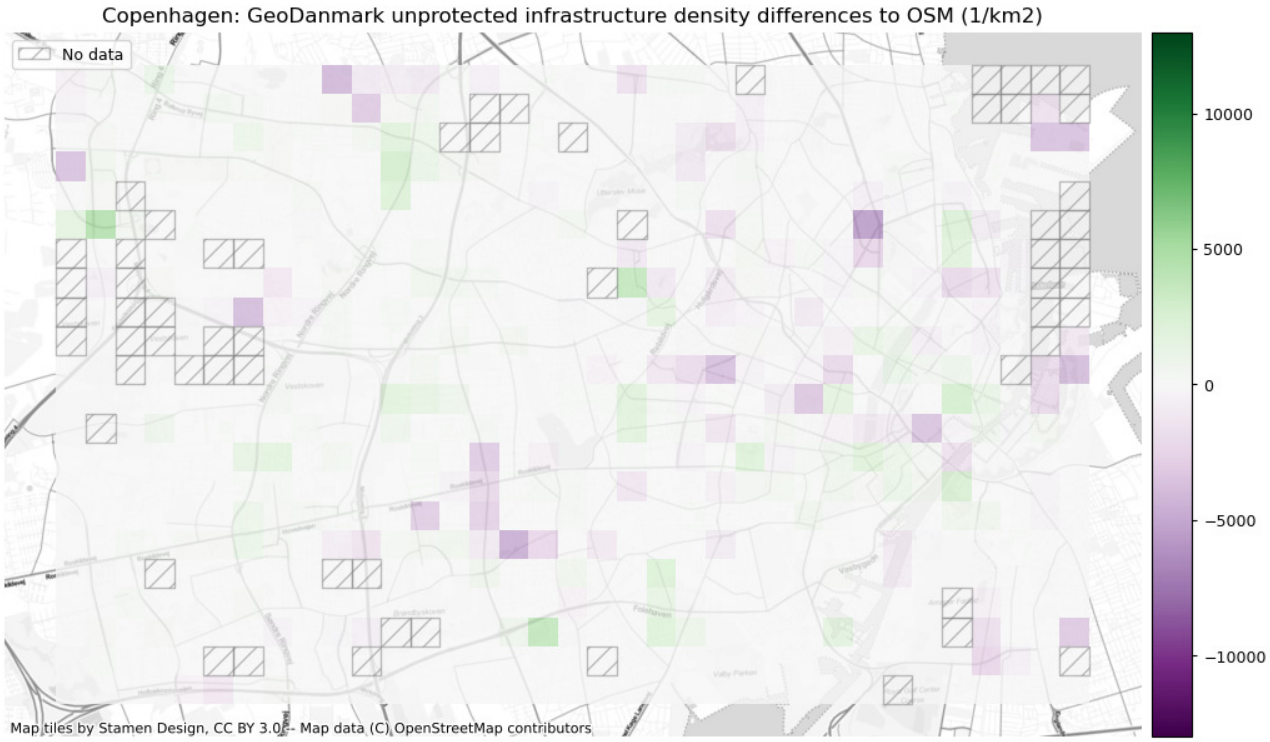


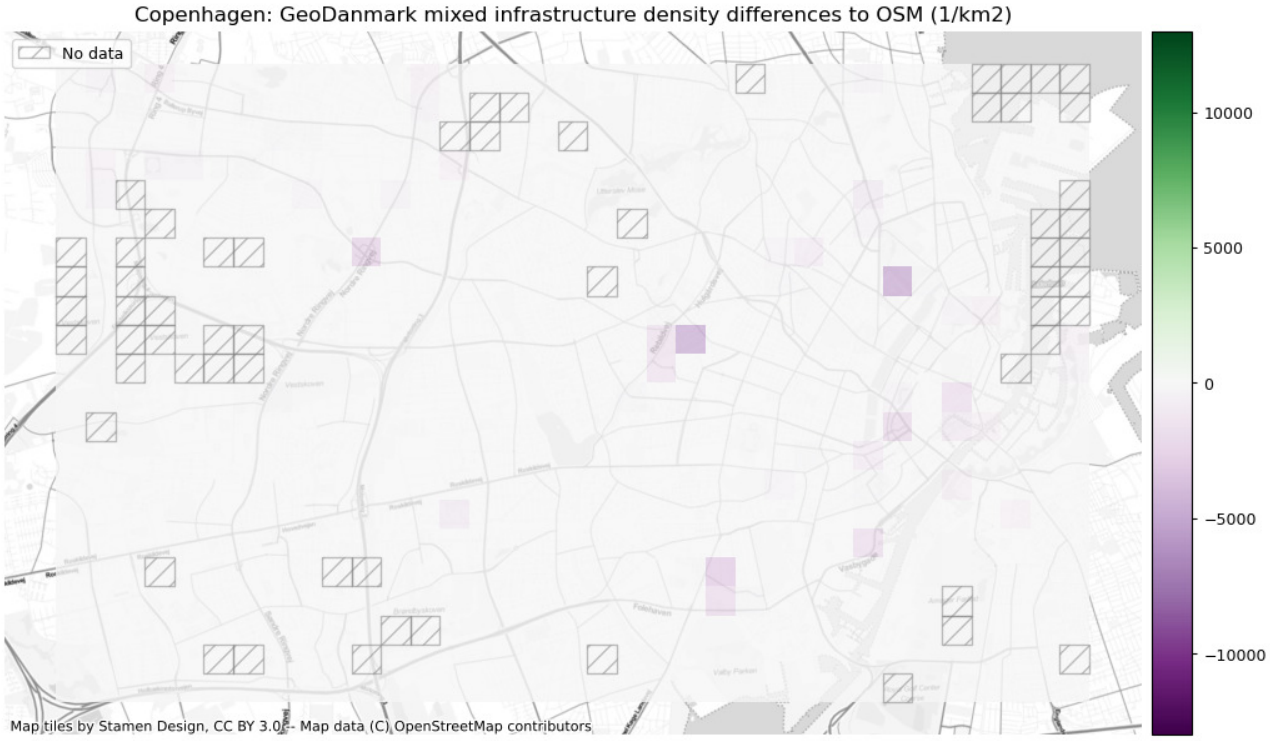


Differences in infrastructure type density

No infrastructure is mapped as mixed protected/unprotected in the GeoDanmark data.







Network topology

After having compared data completeness, i.e. *how much* infrastructure is mapped, here we focus on differences in network *topology*, which can give some information about *how* the infrastructure is mapped in both data sets. Here we also analyze the extent to which network edges are connected to

one or more other edges, or if they end in a dangling node. The extent to which edges are properly connected to adjacent edges are important for, for example, analyzes of accessibility and routing.

When working with data on bicycle networks, a data set without gaps between actually connected network elements is preferred - while of course reflecting the real conditions. Identifying the dangling nodes in a network is 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 data sets are plotted. Then, the local percentage of dangling nodes out of all nodes in each data set is plotted separately. Finally, we show the local difference in the percent of dangling nodes.

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 data set but not the other, this indicates a problem with the data quality. There either is a missing connection in the data, or 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 data set with another data set can help identify whether it is a question of data quality or the quality of the actual infrastructure. Systematic 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 are valid, data sets created with the former method are more suited for routing-based analysis.

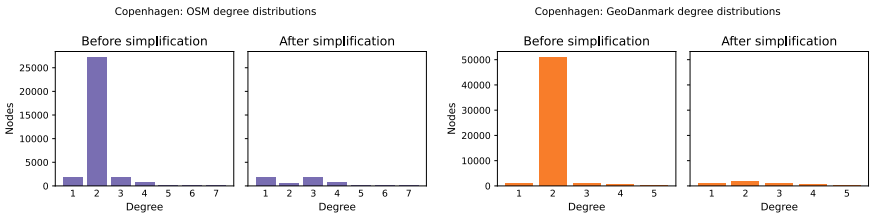
Overshoots will often be less consequential for analysis, but a high number of overshoots will introduce false dangling nodes and distort measures for network structure based on e.g., node degree or the ratio between nodes and edges.

Simplification outcomes

```
Simplifying the OSM network decreased the number of edges by 88.9%.
Simplifying the OSM network decreased the number of nodes by 84.3%.


Simplifying the GeoDanmark network decreased the number of edges by 91.2%.
Simplifying the GeoDanmark network decreased the number of nodes by 92.2%.
```

Node degree distribution

```
Note that the two figures below have different y-axis scales.
```



Alpha, beta, and 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 data set. 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 drawn 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, i.e. it is a tree. 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.

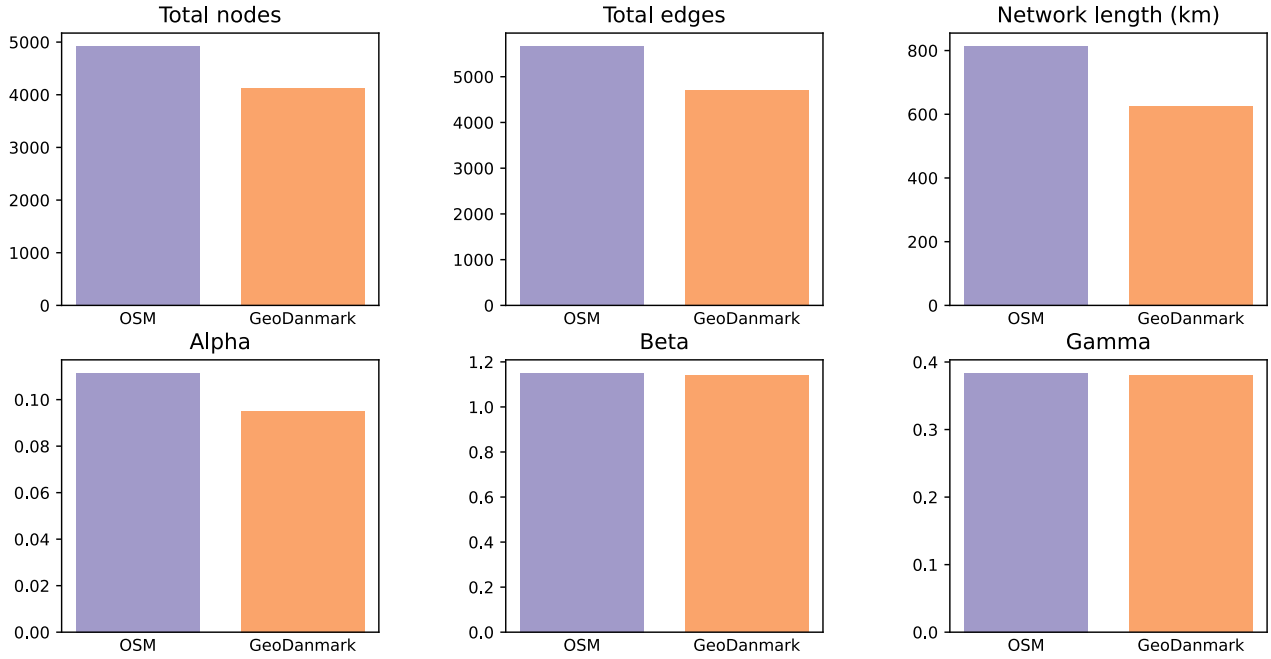
Interpretation

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 drawn 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 that one simply includes more features than the other. However, if the networks have roughly

the same total geometric length, but the values from alpha, beta and gamma differ, this can be an

```
Alpha for the simplified OSM network: 0.11
Beta for the simplified OSM network: 1.15
Gamma for the simplified OSM network: 0.38
```

```
Alpha for the simplified GeoDanmark network: 0.10
Beta for the simplified GeoDanmark network: 1.14
Gamma for the simplified GeoDanmark network: 0.38
```



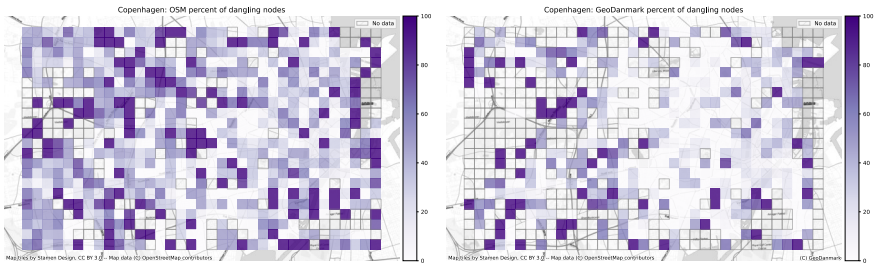
Dangling nodes

Dangling nodes in OSM & reference networks

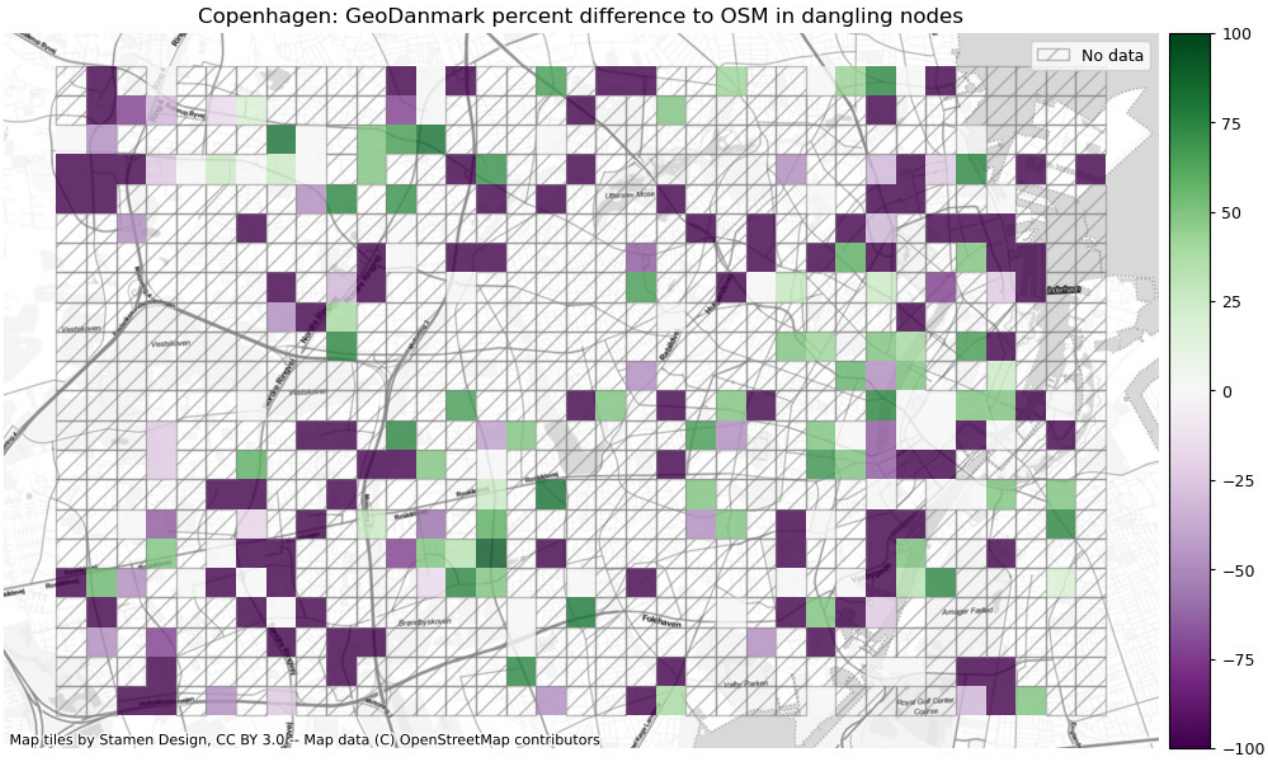
```
Interactive map saved at results/COMPARE/cph_geodk/maps_interactive/danglingmap_compare.html
```

Local values for dangling nodes

Dangling nodes as percentage of all nodes



Local differences in dangling nodes percentages



Under/overshoots

Over and undershoots in OSM and reference networks

```
Interactive map saved at results/COMPARE/cph_geodk/maps_interactive/overundershoots_3_3_compare.
html
```

Network components

This section takes a close look at the network component characteristics for 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

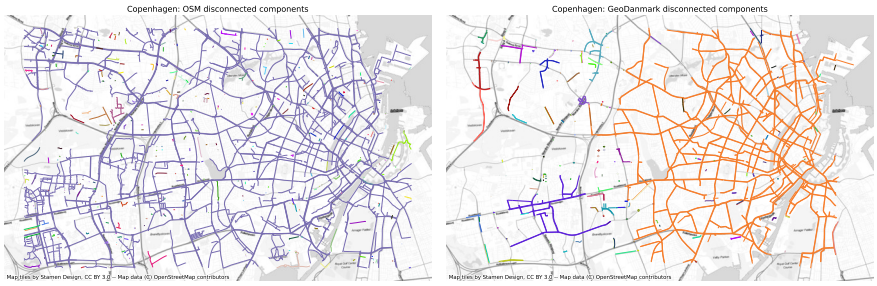
To compare the number and pattern of disconnected components in OSM and reference data, all component results from the intrinsic analyses are juxtaposed and two new plots showing respectively components gaps for OSM and reference data and the difference in component connectivity are produced.

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 bicycle infrastructure. Comparing disconnected components in two data sets enables a more accurate assessment of whether a disconnected component is a data or a planning issue.

Disconnected components

```
The OSM network in the study area consists of 352 disconnected components.
The GeoDanmark network in the study area consists of 204 disconnected components.
```

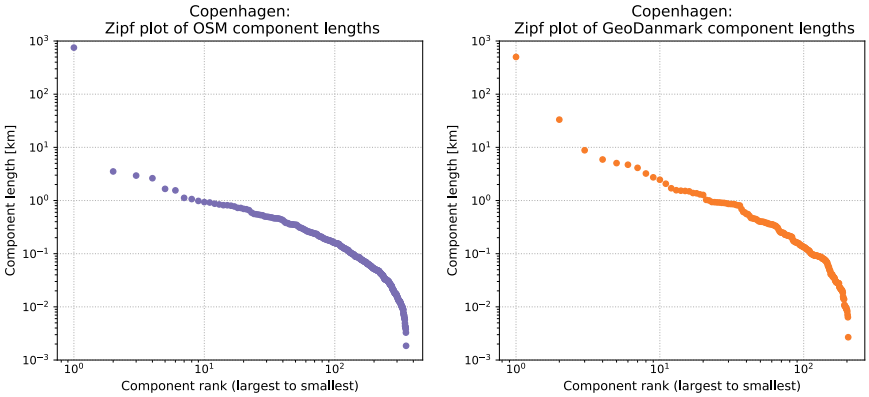


Component size distribution

Many empirical distributions are skewed and often follow a power law, i.e. a straight line in a log-log plot, due to natural processes such as multiplicative network growth (Clauset et al., 2009). The network component size distribution (where size is length) can be visualized with a so-called Zipf plot, which plots the frequency of a component versus its rank (from largest to smallest). When a Zipf plot follows a straight line in log-log scale, it means that there is an much higher chance to find small disconnected components than expected by a distribution from an exponential family (like a normal distribution). This can mean that there has been no consolidation of the network, only piece-wise or random additions (Szell et al., 2022).

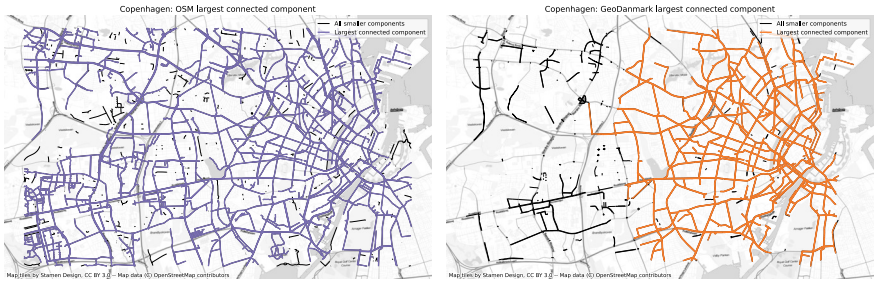
However, it can also happen that the largest connected component (the leftmost marker in the plot at rank 10^0) is a clear outlier, while the rest of the plot follows a different shape. This can mean that a consolidation *has* taken place, and that either a central planner has deliberately targeted to connect the network, or that the data are of high enough quality to have overcome many gaps.

In case of a comparison over the same region, as shown below, if one data set shows a clear outlier in its largest connected component while the other does not, and if it is also at least as large, it can in general be interpreted as being more complete. This issue of incompleteness can stem from different effects, for example a data set being merged together from different data providers such as different municipalities or regions that stop data collection at their boundaries and that do not connect their network data with their neighbors.

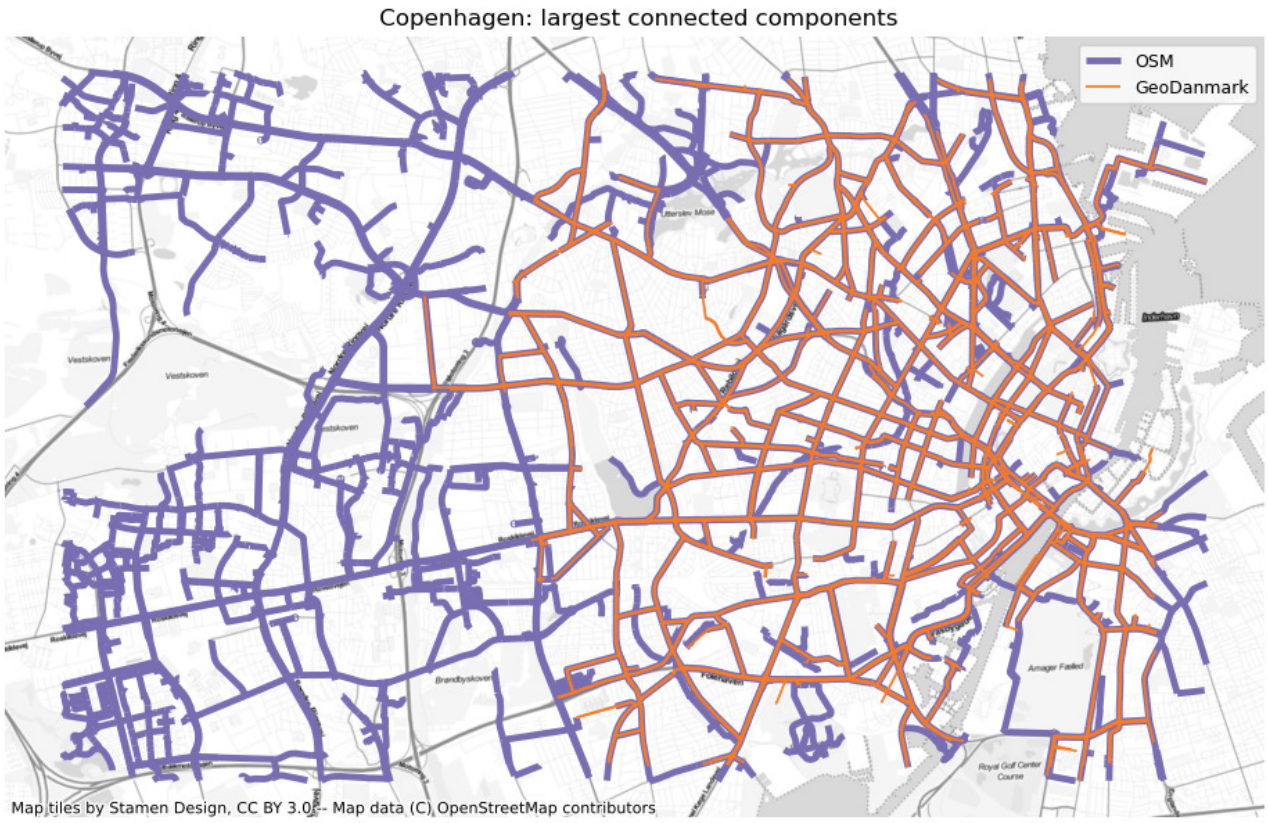


Largest connected component

```
The largest connected component in the OSM network contains 92.30% of the network length.
The largest connected component in the GeoDanmark network contains 80.04% of the network length.
```



Overlay of largest connected component in OSM and reference networks



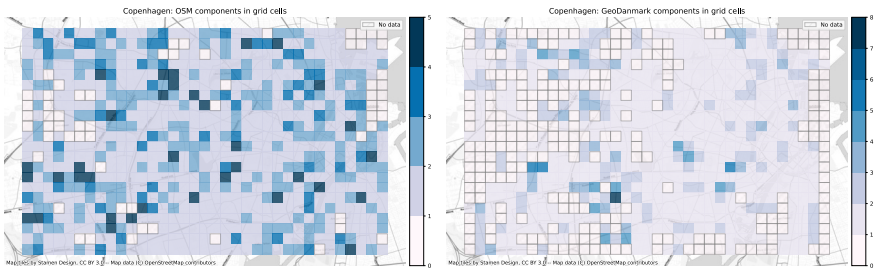
Missing links

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.

```
Interactive map saved at results/COMPARE/cph_geodk/maps_interactive/component_gaps_compare.html
```

Components per grid cell

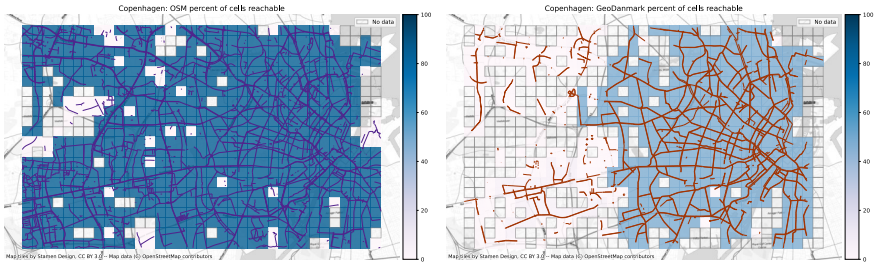
The plots below show the number of components intersecting a grid cell. A high number of components in a grid cell is generally an indication of poor network connectivity - either due to fragmented infrastructure or because of problems with the data quality.

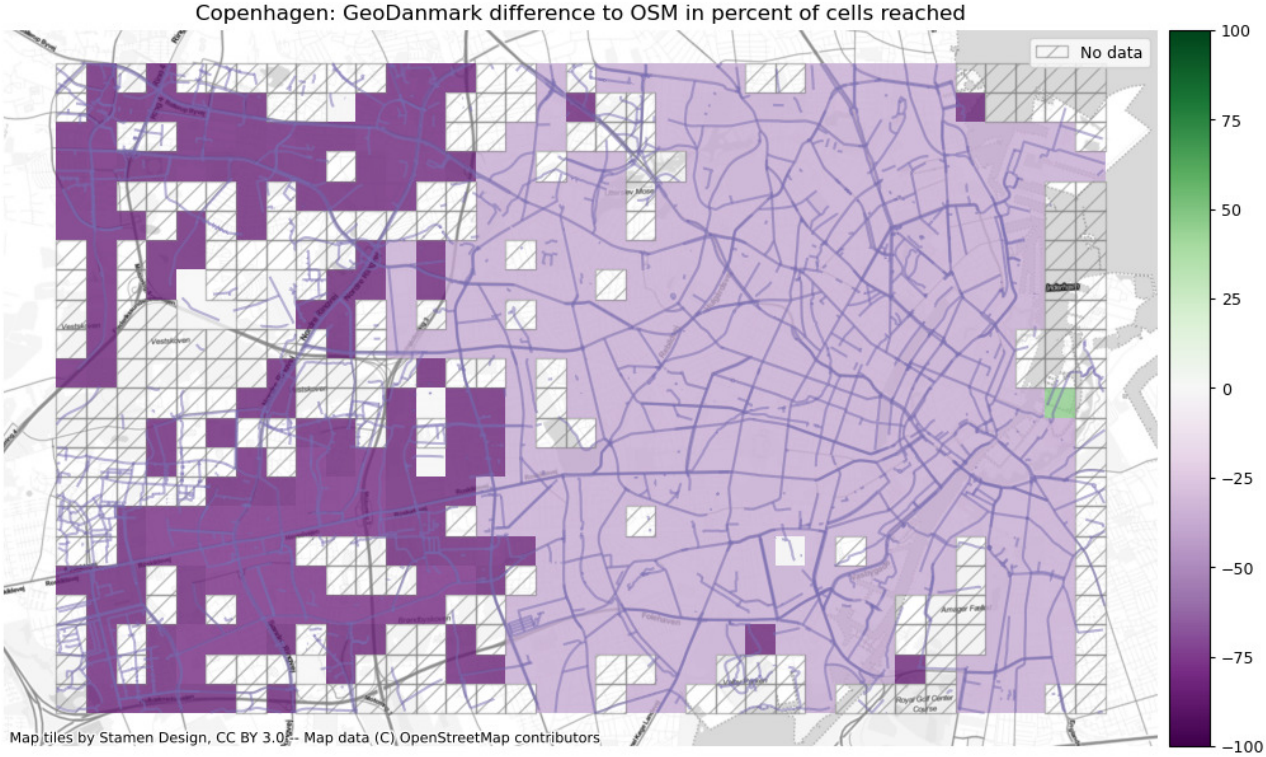


Component connectivity

Here we visualize differences between how many cells can be reached from each cell. The metric is a crude measure for network connectivity but has the benefit of being computationally cheap and thus able to quickly highlight stark differences in network connectivity.

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





Summary

Extrinsic Quality Comparison		
	OSM	**GeoDanmark**

Total infrastructure length (km)	1,063	626
Protected bicycle infrastructure density (m/km2)	5,303	2,999
Unprotected bicycle infrastructure density (m/km2)	499	455
Mixed protection bicycle infrastructure density (m/km2)	59	0
Bicycle infrastructure density (m/km2)	5,861	3,454
Nodes	4,925	4,125
Dangling nodes	1,818	870
Nodes per km2	27	23
Dangling nodes per km2	10	5
Overshoots	9	21
Undershoots	14	11
Components	352	204
Length of largest component (km)	752	501
Largest component's share of network length	92%	80%
Component gaps	91	52
Alpha	0.11	0.10
Beta	1.15	1.14
Gamma	0.38	0.38

3b. Feature Matching

The feature matching takes the reference data and attempts to identify corresponding features in the OSM data set. Feature matching is a necessary precondition to compare single features rather than feature characteristics on study area a grid cell level, as well as for merging two data sets.

Method

Matching features in two road data sets with each their way of digitizing features and a potential one-to-many relationship between edges (for example in the case where one data set only maps road center lines, while the other map the geometries of each bike lane) is not a trivial task.

The method used here converts all network edges to smaller segments of a uniform length before looking for a potential match between the reference and the OSM data. The matching is done on the basis of the buffered distance between objects, the angle, and the undirected Hausdorff distance, and is based on the works of Koukoletsos et al. (2012) and Will (2014).

Based on the matching results, the following values are computed:

- The number and length of matched and unmatched edges, in total and per grid cell
- A comparison of the attributes of the matched edges: Is their classification of cycling infrastructure as protected or unprotected the same?

Interpretation

It is important to visually explore the feature matching results, since the success rate of the matching influences how the analysis of number of matches should be interpreted.

If the features in the two data sets have been digitized differently - e.g. if one data set has digitized bike tracks as mostly straight lines, while the other includes more winding tracks, the matching will fail. This is also the case if they are placed too far from each other. If it can be confirmed visually that the same features do exist in both data sets, a lack of matches indicates that the geometries in the two data sets are too different. If however it can be confirmed that most real corresponding features have been identified, a lack of matches in an area indicates errors of commission or omission.

Sections

Match features

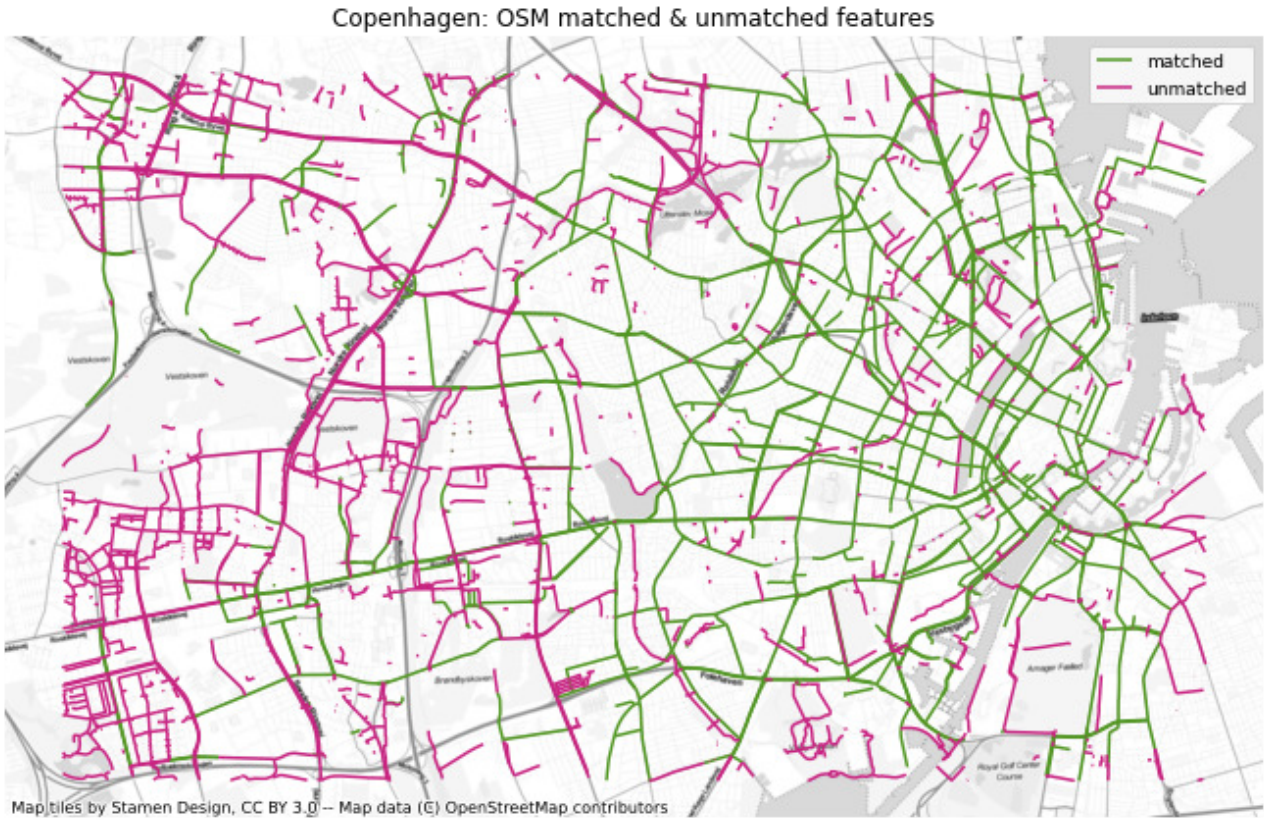
Run and plot feature matching

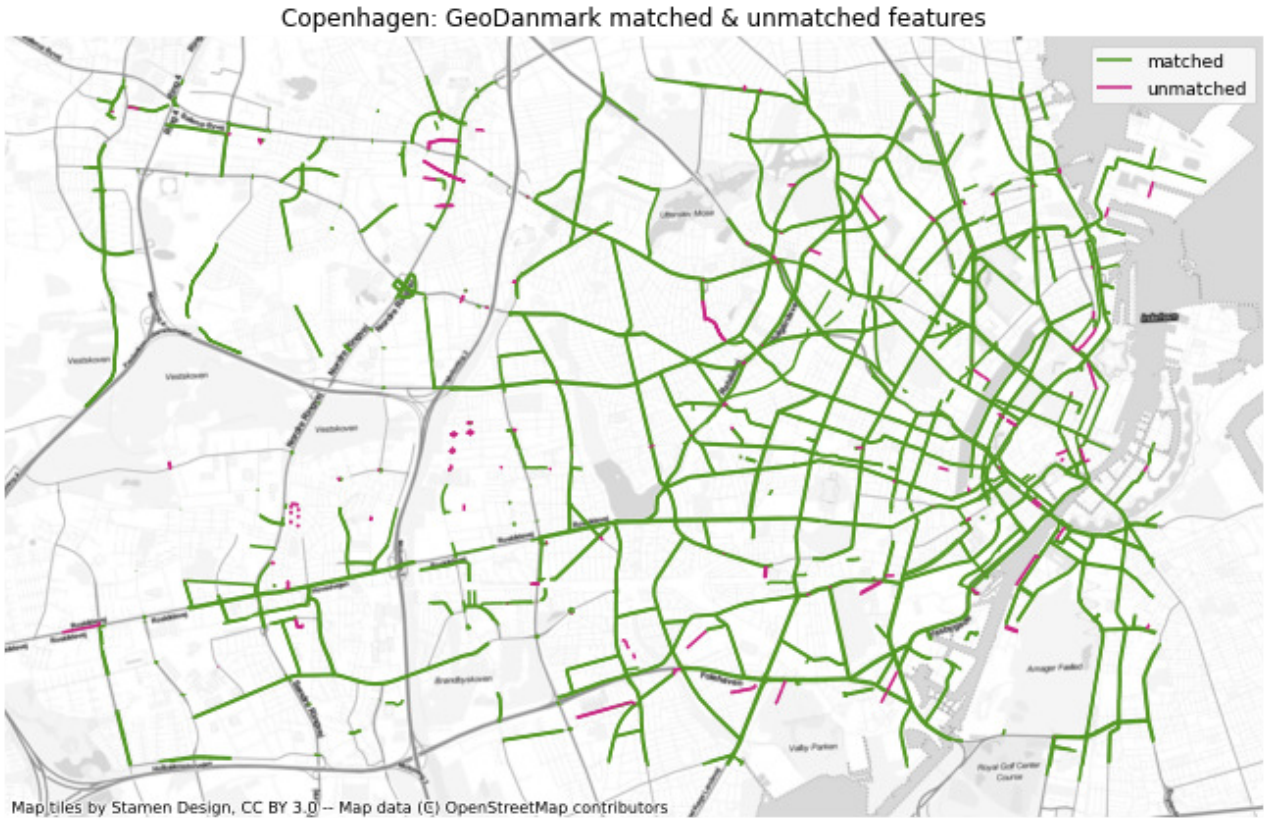
```
Segments created successfully!
```

```
Segment matching has already been performed. Loading existing segment matches, matched with a bu
ffer distance of 15 meters, a Hausdorff distance of 17 meters, and a max angle of 30 degrees.
```

```
Interactive map saved at results/COMPARE/cph_geodk/maps_interactive/segment_matches_15_17_30_com
pare.html
```

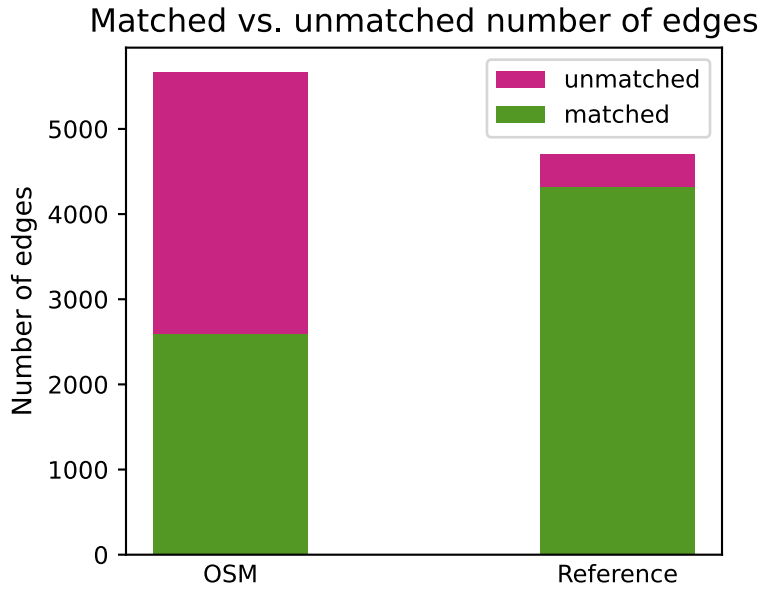
Matched and unmatched features

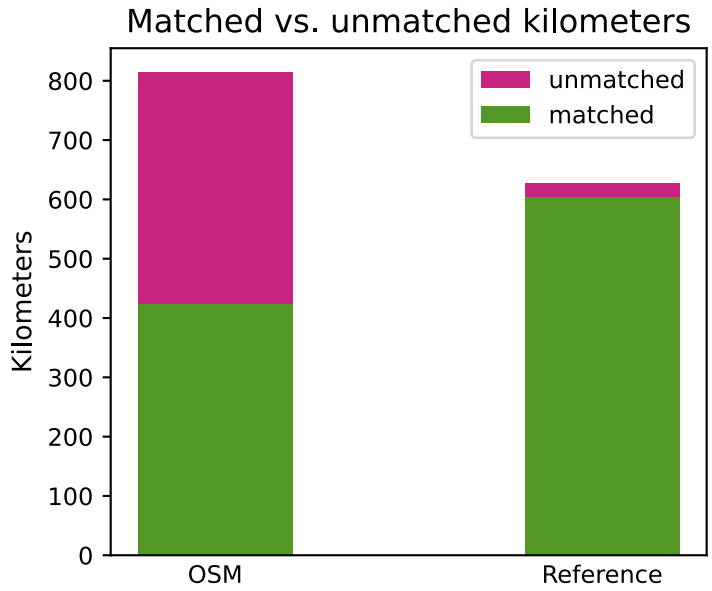




Feature matching summary

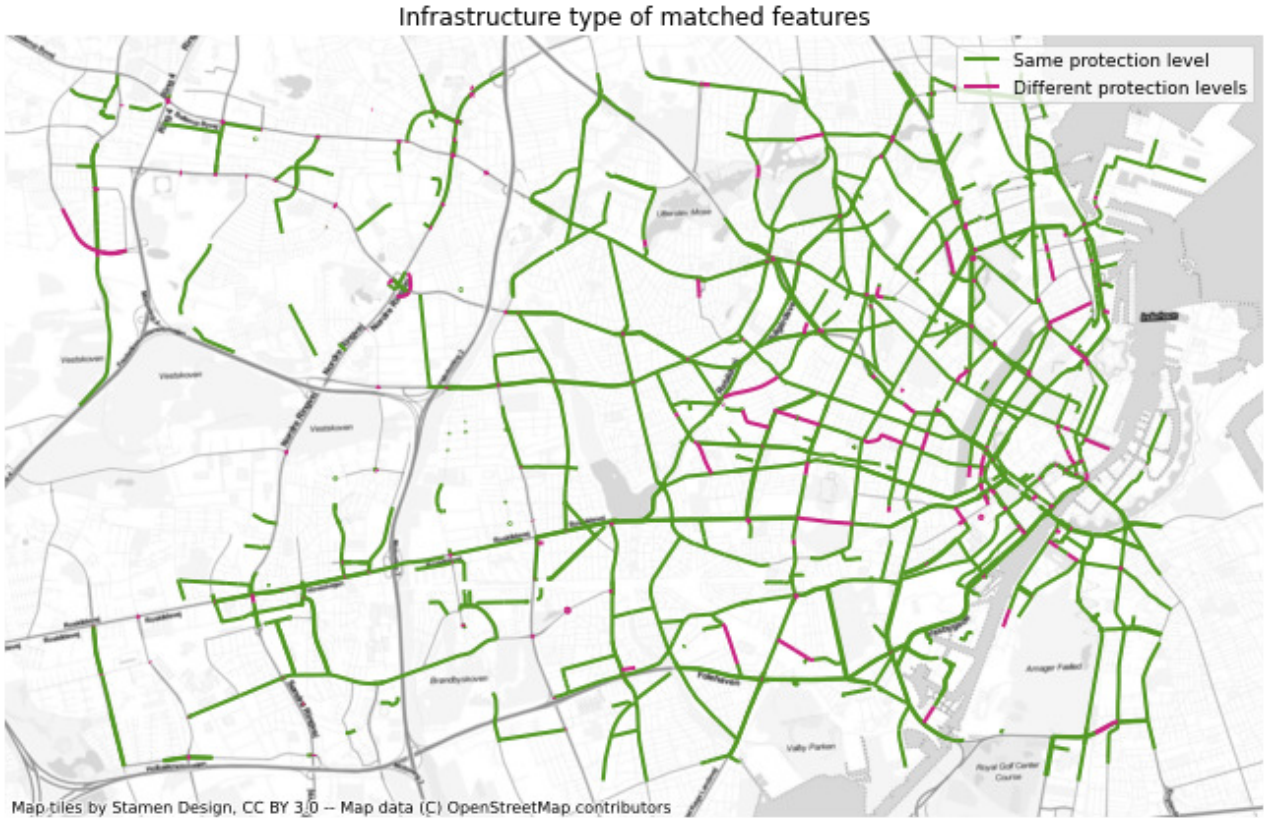
```
Edge count: 2588 of 5671 OSM edges (45.64%) were matched with a reference edge.
Edge count: 4313 out of 4705 reference edges (91.67%) were matched with an OSM edge.
Length: 422.50 km out of 814.18 km of OSM edges (51.89%) were matched with a reference edge.
Length: 603.15 km out of 626.48 km of reference edges (96.28%) were matched with an OSM edge.
```





Analyze feature matching results

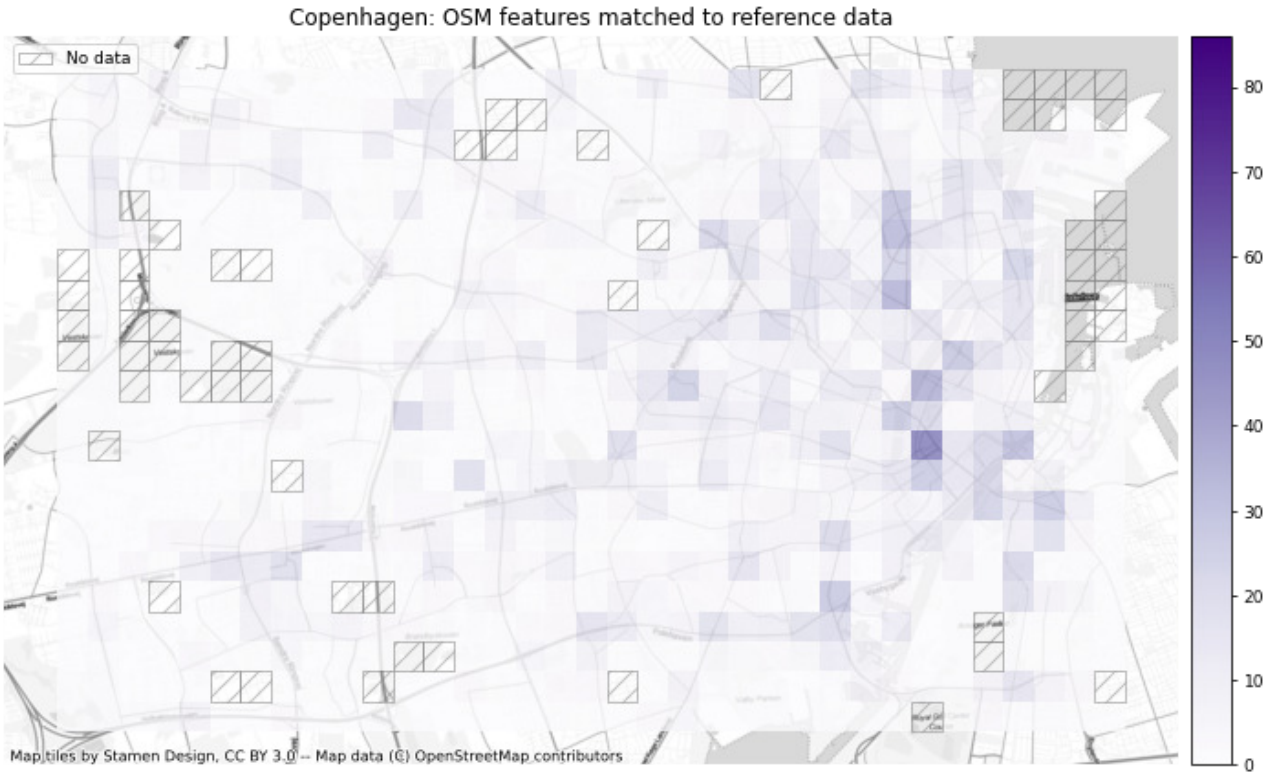
Matched features by infrastructure type



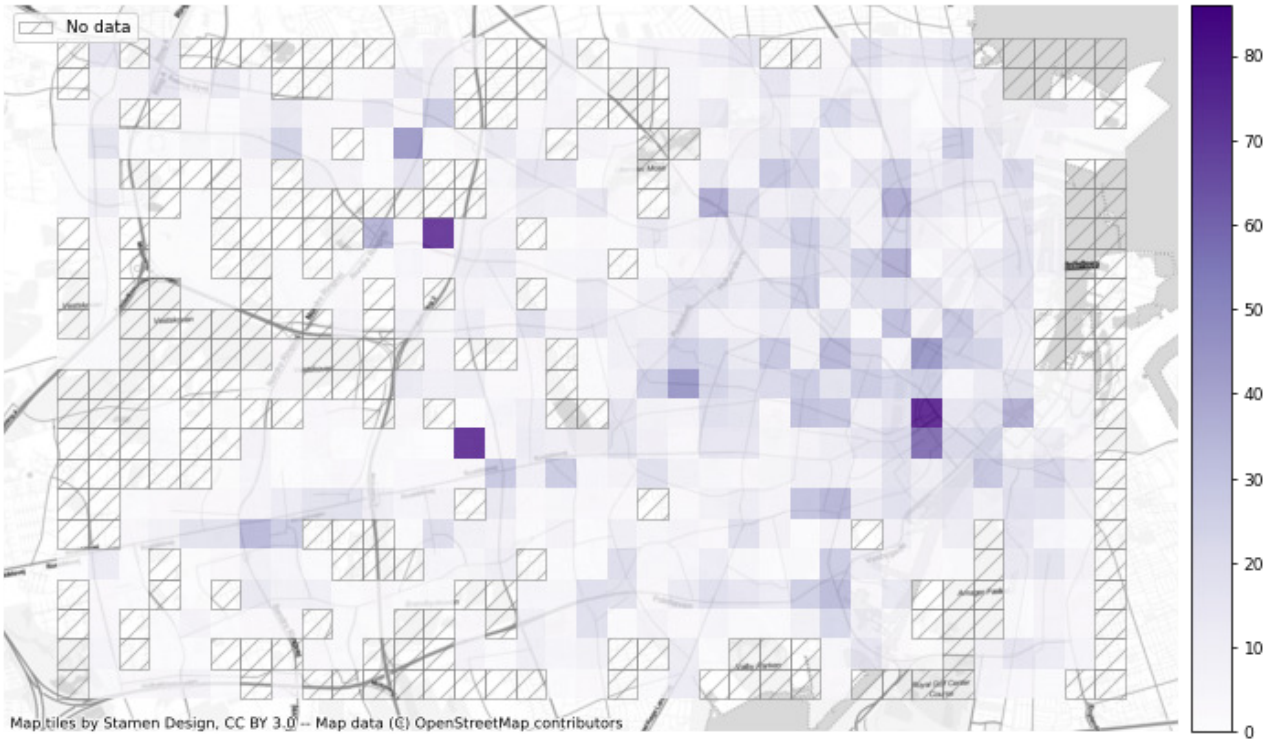
Feature matching success

In the plots below, the count, percent, and length of matched and unmatched features in each data set are summarized.

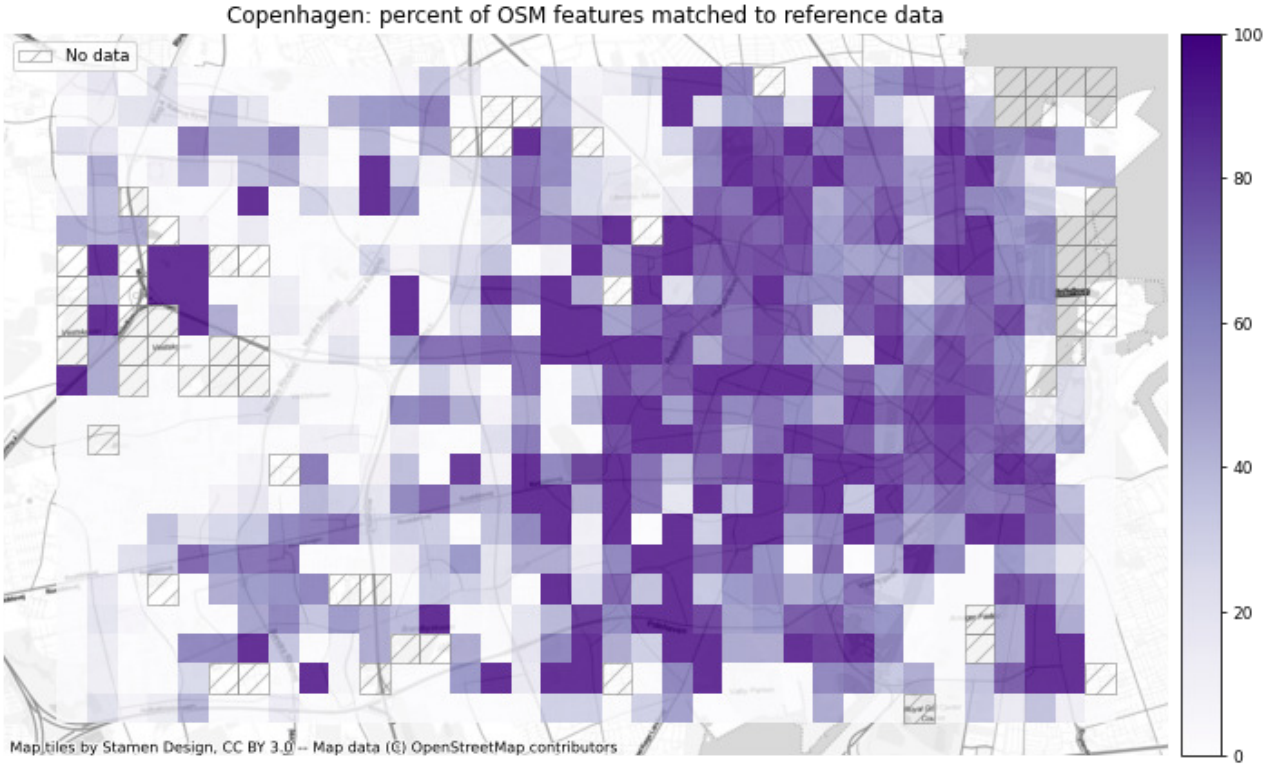
Warning
The number of matched features in one data set in a grid cell does not necessarily reflect the number of matched features in the other data set, since an edge can be matched to a corresponding edge in another cell. Moreover, the local count refers to edges intersected with the grid cell. For example, a long bike lane crossing 3 cells will thus be counted as matched in 3 different cells. This does not change the relative distribution of matched/unmatched features, but it does entail that the overall summary of matched/unmatched features above uses a different total count of edges than the plots below.



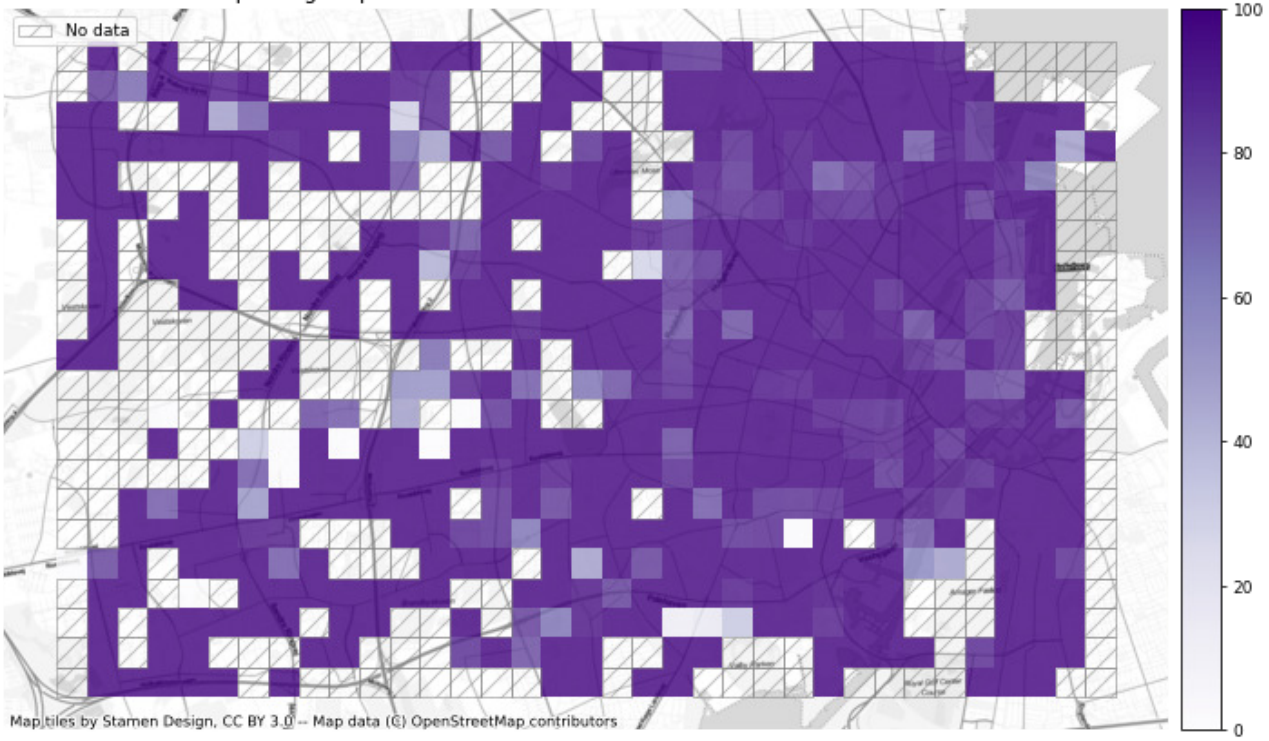
Copenhagen: GeoDanmark features matched to OSM data



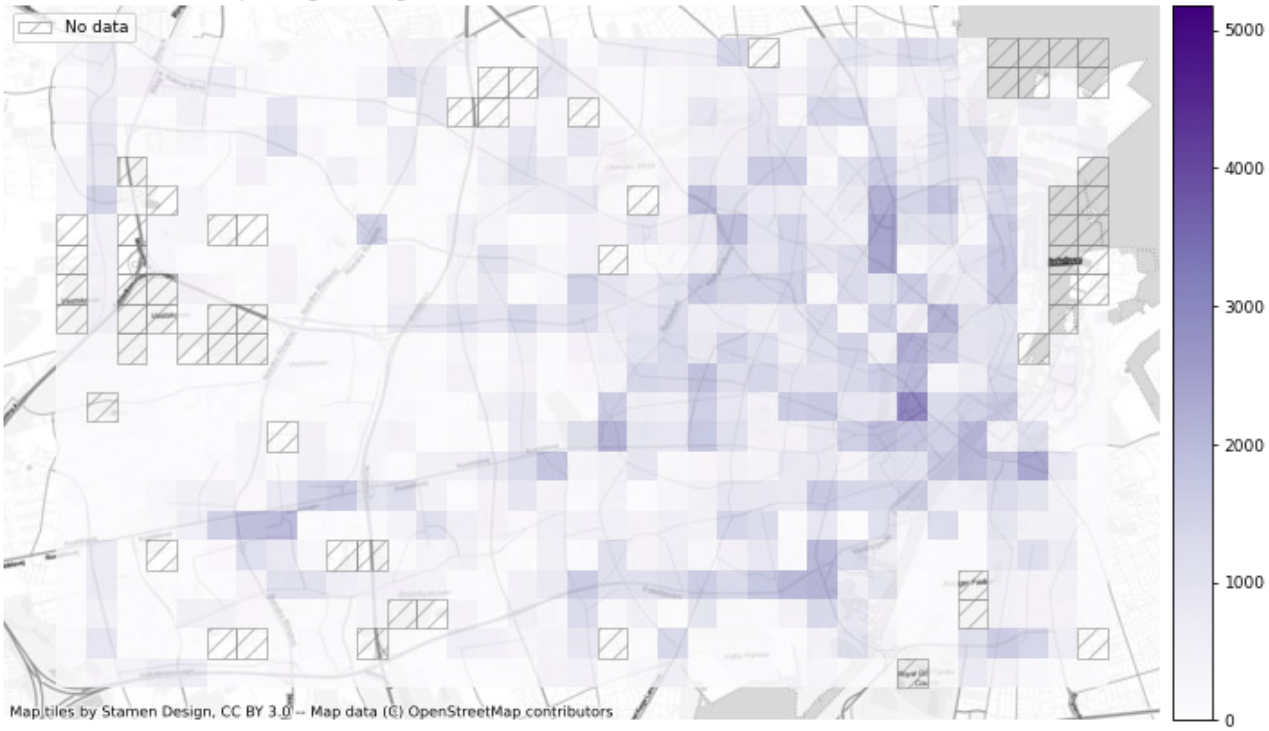
Copenhagen: percent of OSM features matched to reference data



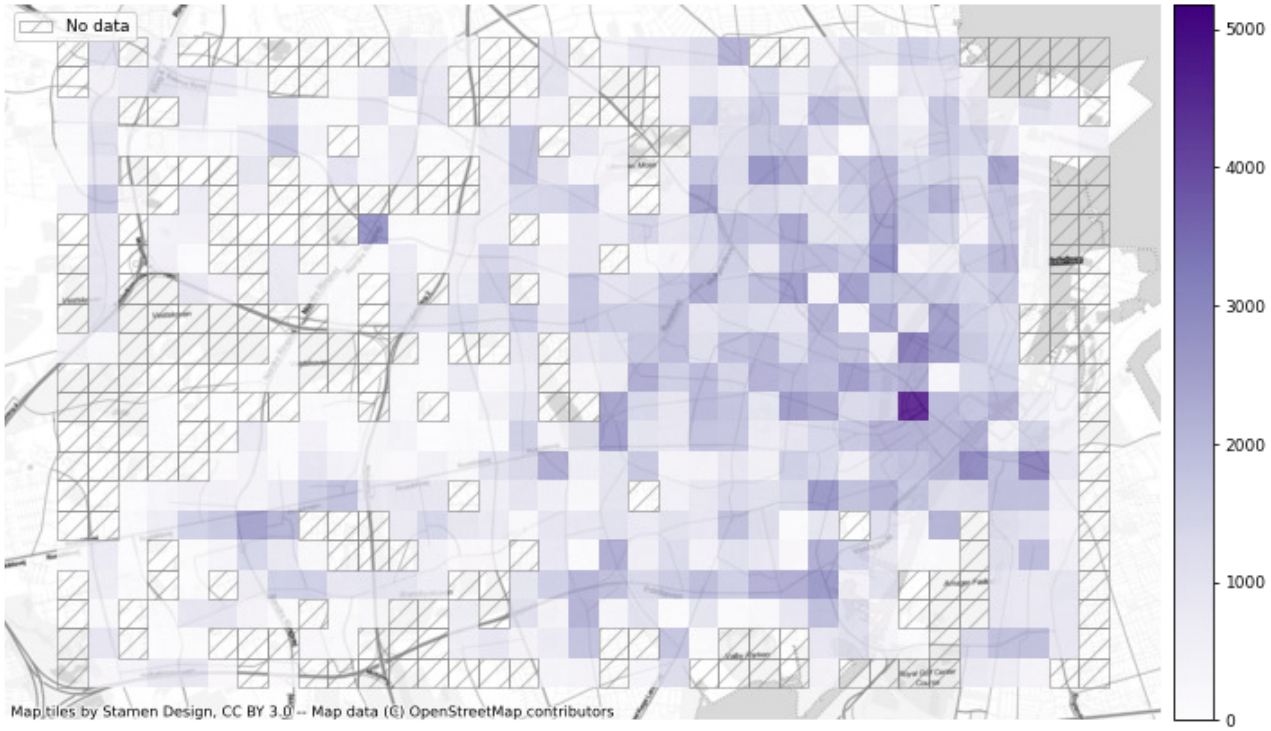
Copenhagen: percent of GeoDanmark features matched to OSM data



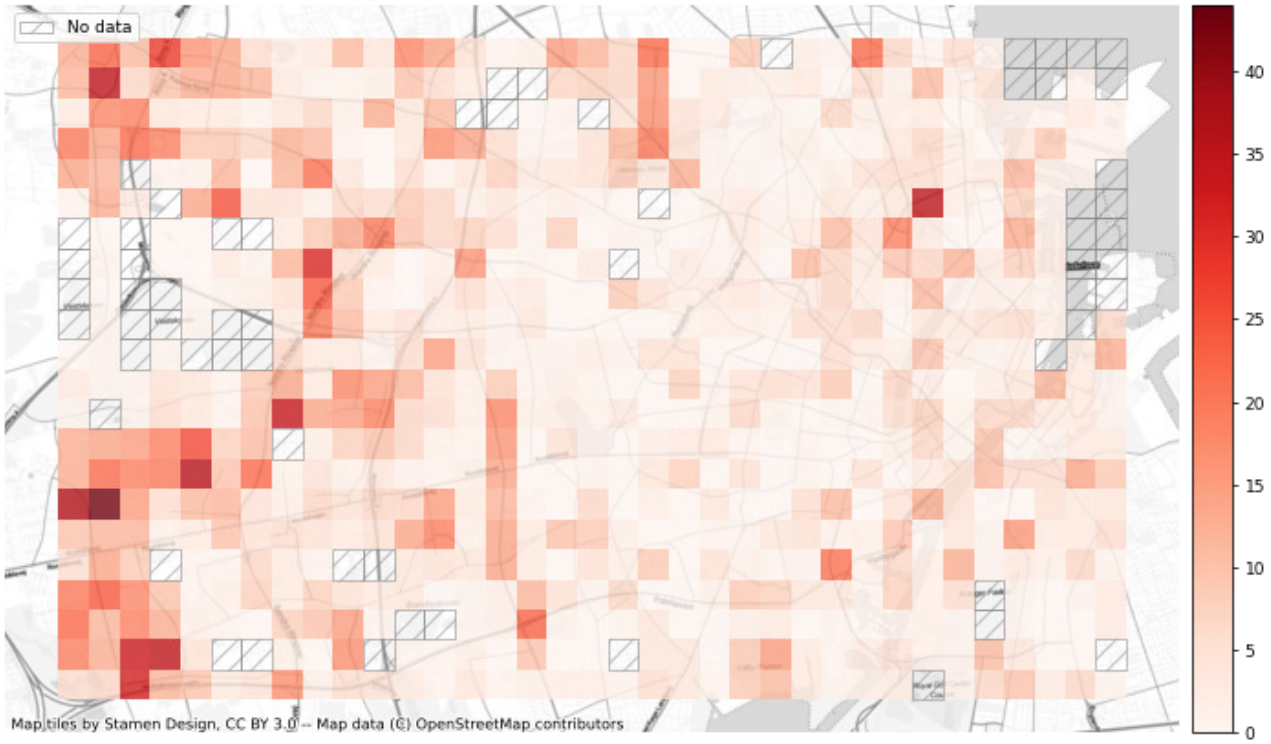
Copenhagen: length of OSM features matched to reference data (m)



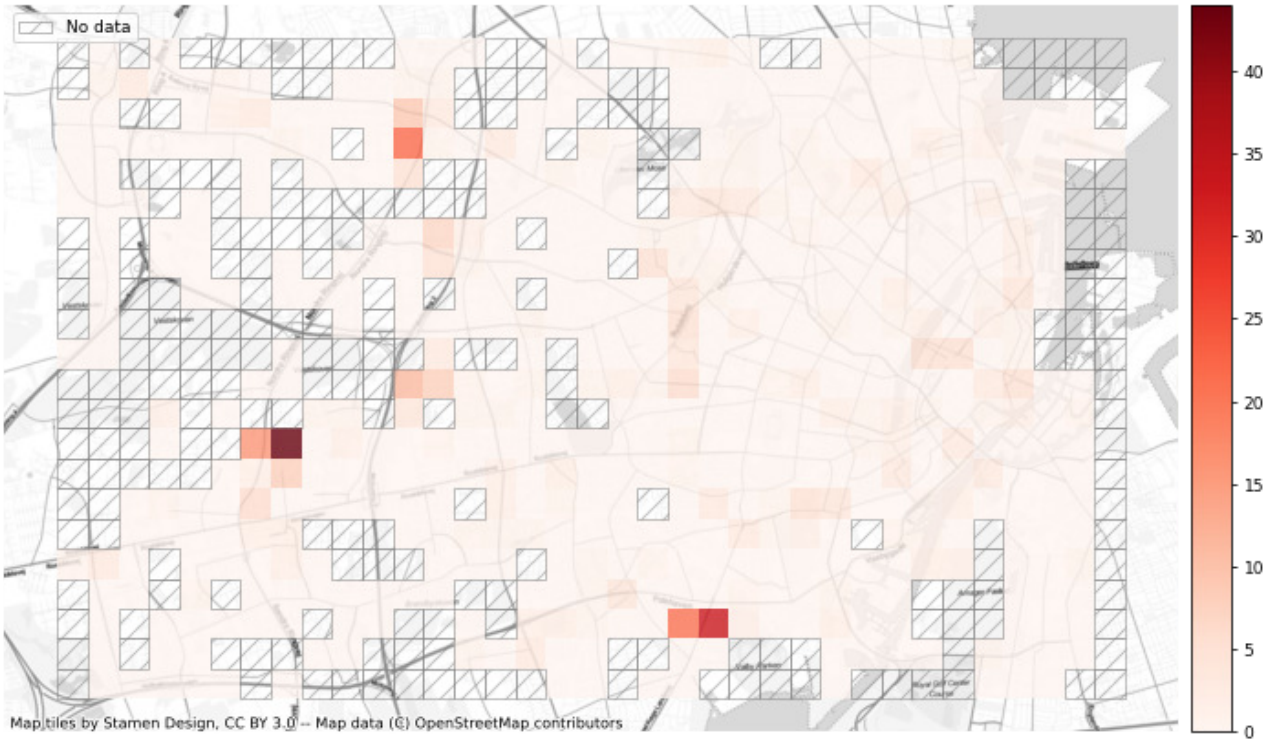
Copenhagen: length of GeoDanmark features matched to OSM data (m)



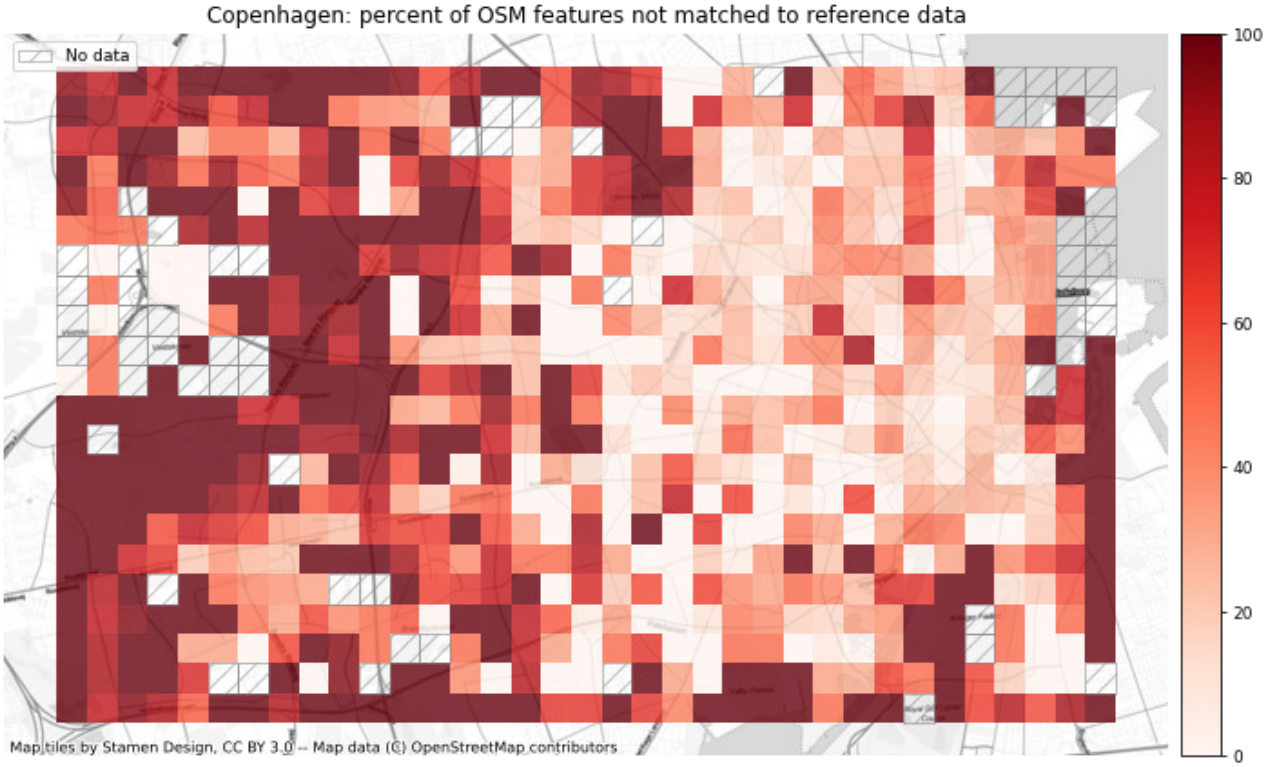
Copenhagen: OSM features not matched to reference data



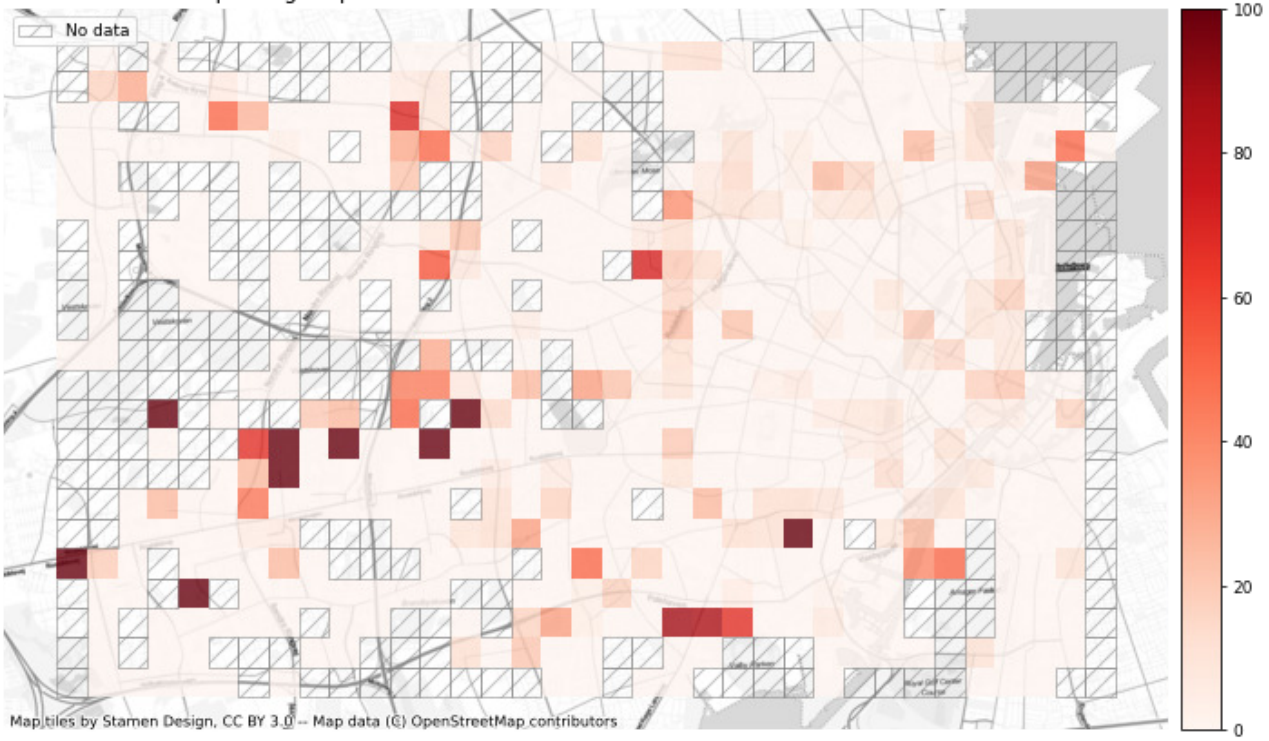
Copenhagen: GeoDanmark features not matched to OSM data



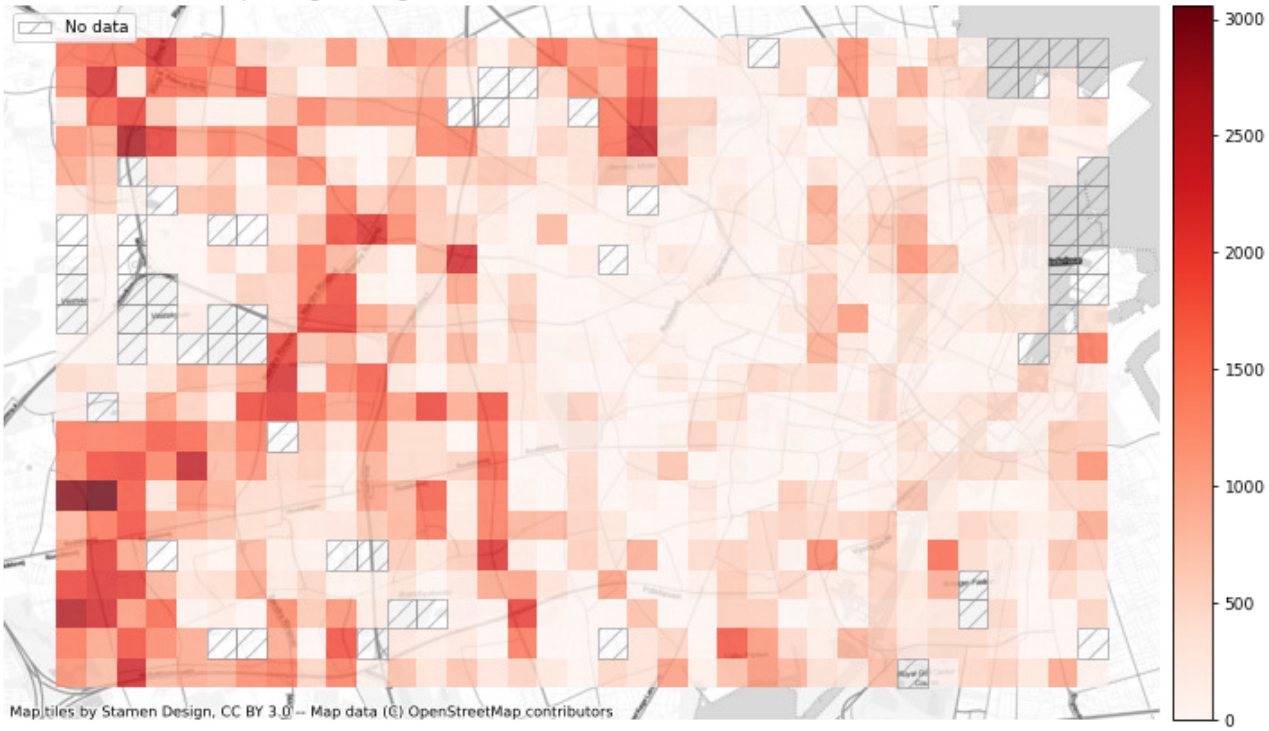
Copenhagen: percent of OSM features not matched to reference data

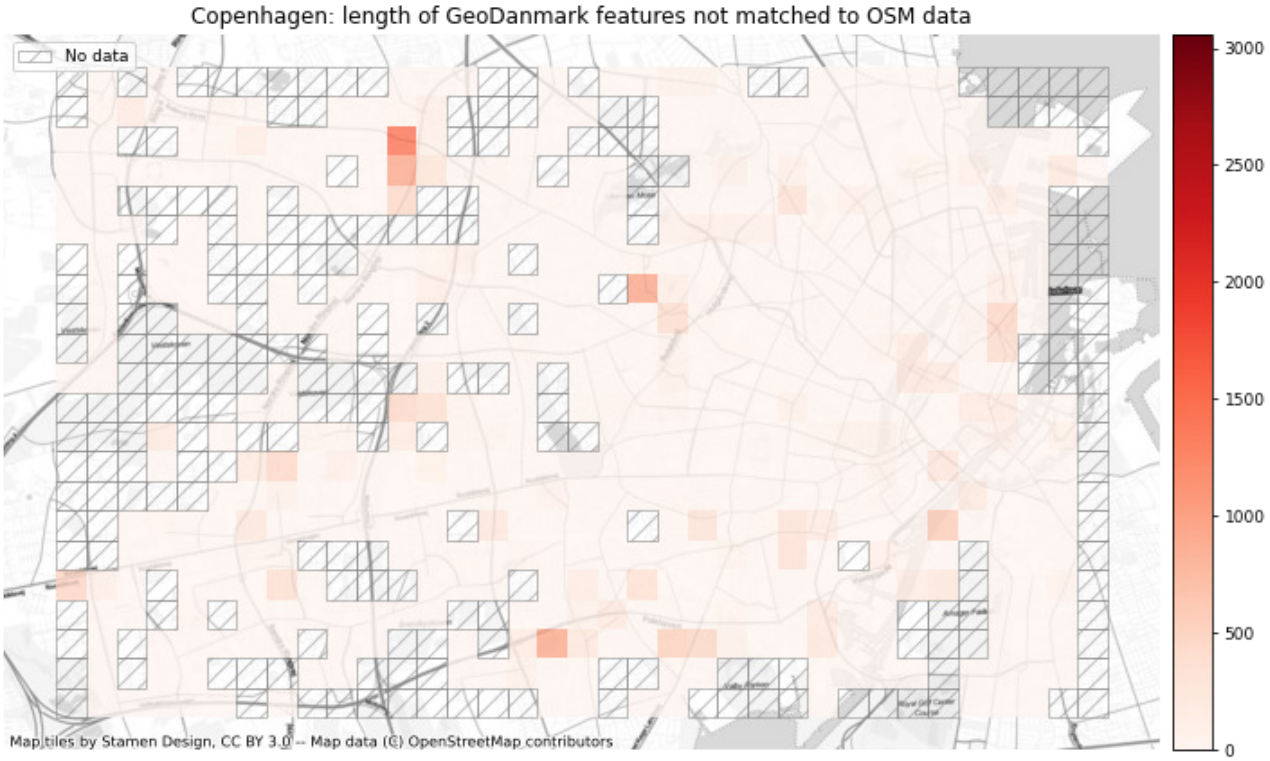


Copenhagen: percent of GeoDanmark features not matched to OSM data



Copenhagen: length of OSM features not matched to reference data





Summary

Feature Matching Results		
	OSM	GeoDanmark
Count of matched edges	2,588	4,313
Percent matched edges	46%	92%
Length of matched edges (km)	423	603
Percent of matched network length	52	96
Local min of % matched edges	4%	12%
Local max of % matched edges	100%	100%
Local average of % matched edges	62%	95%

Appendix A: config.yml

This notebook shows the content of `config.yml`, i.e. the parameters that were used to run the analysis.

```
{'area_name': 'Copenhagen',
 'reference_name': 'GeoDanmark',
 'study_area': 'cph_geodk',
 'study_crs': 'EPSG:25832',
 'plot_resolution': 'high',
 'bicycle_infrastructure_queries': {'A': "highway == 'cycleway'",
                                    'B': 'cycleway in '
                                         "['lane','track','opposite_lane','opposite_track','shar
ed_lane','designated','crossing']",
                                    'C': 'cycleway_left in '
                                         "['lane','track','opposite_lane','opposite_track','shar
ed_lane','designated','crossing']",
                                    'D': 'cycleway_right in '
                                         "['lane','track','opposite_lane','opposite_track','shar
ed_lane','designated','crossing']",
                                    'E': 'cycleway_both in '
                                         "['lane','track','opposite_lane','opposite_track','shar
ed_lane','designated','crossing']"},
 'osm_bicycle_infrastructure_type': {'protected': ["highway == 'cycleway'",
                                                   'cycleway in '
                                                   "['track','opposite_track']",
                                                   'cycleway_left in '
                                                   "['track','opposite_track']",
                                                   'cycleway_right in '
                                                   "['track','opposite_track']",
                                                   'cycleway_both in '
                                                   "['track','opposite_track']"],
                                     'unprotected': ['cycleway in '
                                                     "['lane','opposite_lane','shared_lane','cro
ssing']",
                                                     'cycleway_left in '
                                                     "['lane','opposite_lane','shared_lane','cro
ssing']",
                                                     'cycleway_right in '
                                                     "['lane','opposite_lane','shared_lane','cro
ssing']",
                                                     'cycleway_both in '
                                                     "['lane','opposite_lane','shared_lane','cro
ssing']"],
                                     'unknown': ["cycleway in ['designated']",
                                                 'cycleway_left in '
                                                 "['designated']",
                                                 'cycleway_right in '
                                                 "['designated']",
                                                 'cycleway_both in '
                                                 "['designated']"]},
 'osm_way_tags': ['access',
                  'barrier',
                  'bridge',
                  'bicycle',
                  'bicycle_road',
                  'crossing',
```

```
                  'cycleway',
                  'cycleway:left',
                  'cycleway:right',
                  'cycleway:both',
                  'cycleway:buffer',
                  'cycleway:left:buffer',
                  'cycleway:right:buffer',
                  'cycleway:both:buffer',
                  'cycleway:width',
                  'cycleway:left:width',
                  'cycleway:right:width',
                  'cycleway:both:width',
                  'cycleway:surface',
                  'foot',
                  'footway',
                  'highway',
                  'incline',
                  'junction',
                  'layer',
                  'lit',
                  'maxspeed',
                  'maxspeed:advisory',
                  'moped',
                  'moter_vehicle',
                  'motorcar',
                  'name',
                  'oneway',
                  'oneway:bicycle',
                  'osm_id',
                  'segregated',
                  'surface',
                  'tracktype',
                  'tunnel',
                  'width'],
 'existing_tag_analysis': {'surface': {'true_geometries': ['surface',
                                                           'cycleway_surface'],
                                       'centerline': ['cycleway_surface']},
                           'width': {'true_geometries': ['width',
                                                         'cycleway_width',
                                                         'cycleway_left_width',
                                                         'cycleway_right_width',
                                                         'cycleway_both_width'],
                                     'centerline': ['cycleway_width',
                                                    'cycleway_left_width',
                                                    'cycleway_right_width',
                                                    'cycleway_both_width']},
                           'speedlimit': {'all': ['maxspeed']},
                           'lit': {'all': ['lit']}},
 'incompatible_tags_analysis': {'bicycle_infrastructure': {'yes': [['bicycle',
                                                                    'no'],
                                                                   ['bicycle',
                                                                    'dismount'],
                                                                   ['car',
                                                                    'yes']]}},
 'grid_cell_size': 500,
 'reference_geometries': 'true_geometries',
 'bidirectional': False,
 'ref_bicycle_infrastructure_type': {'protected': ["vejklasse ==  'Cykelsti "
                                                   "langs vej'"],
                                     'unprotected': ["vejklasse == 'Cykelbane "
                                                     "langs vej'"]},
 'reference_id_col': 'fot_id'}
```