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 in 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, taking the OSM data as the base line. 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 OSM 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 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 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.

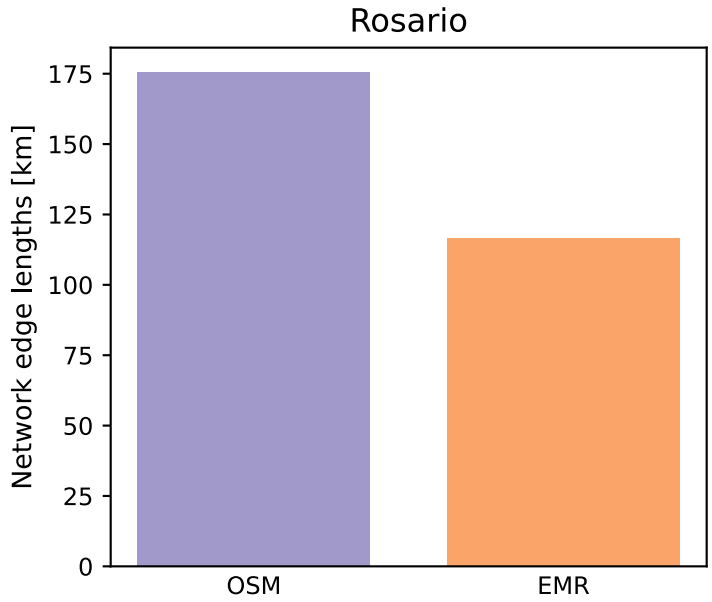
For example, a 100 meter long **bi**directional path (geometric length: 100m) contributes with 200 meters of bicycle infrastructure (infrastructure length: 200m).

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, missing 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: 175.49 km
Length of the reference data set: 116.71 km


The reference data set is 58.78 km shorter than the OSM data set.
The reference data set is 33.49% shorter than the OSM data set.
```

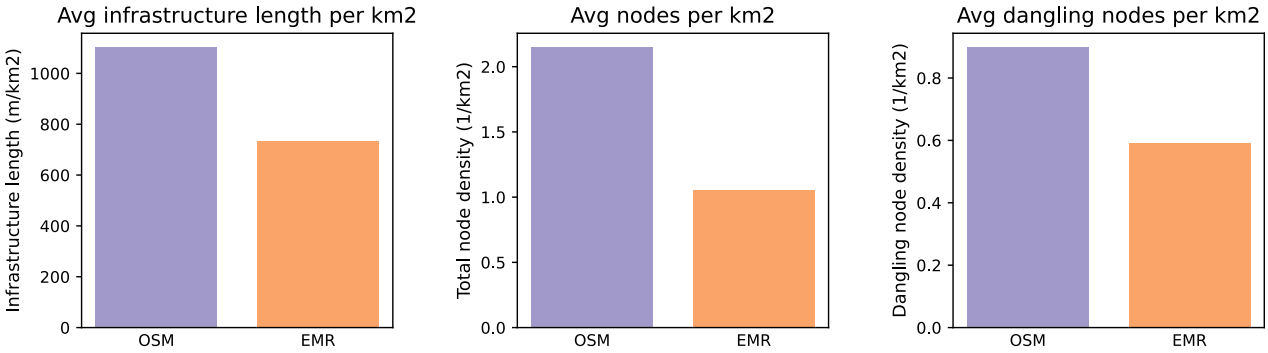


Network Density

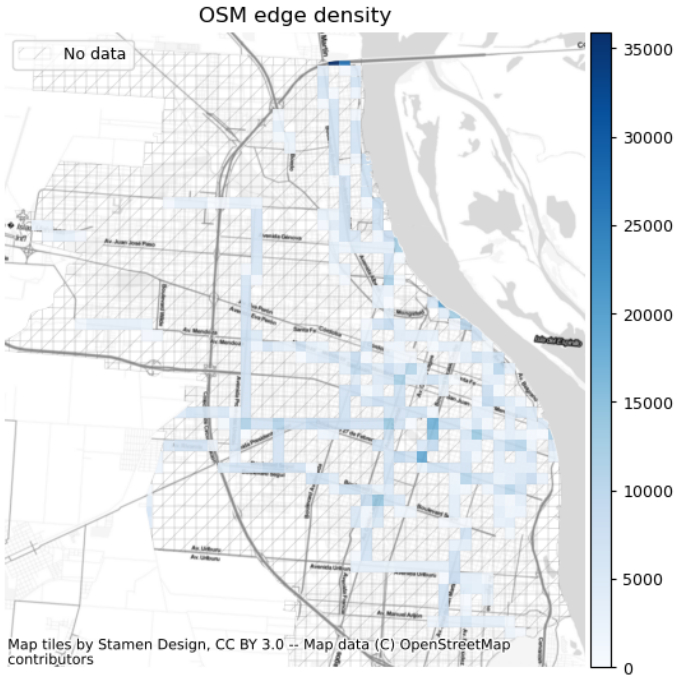
Global network densities

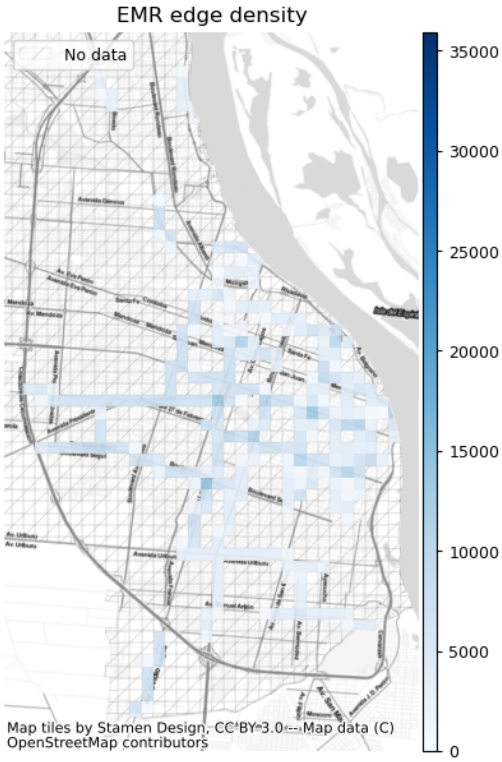
```
In the OSM data, there are:
 - 1102.58 meters of cycling infrastructure per km2.
 - 2.15 nodes in the cycling network per km2.
 - 0.90 dangling nodes in the cycling network per km2.


In the reference data, there are:
 - 733.29 meters of cycling infrastructure per km2.
 - 1.06 nodes in the cycling network per km2.
 - 0.59 dangling nodes in the cycling network per km2.
```

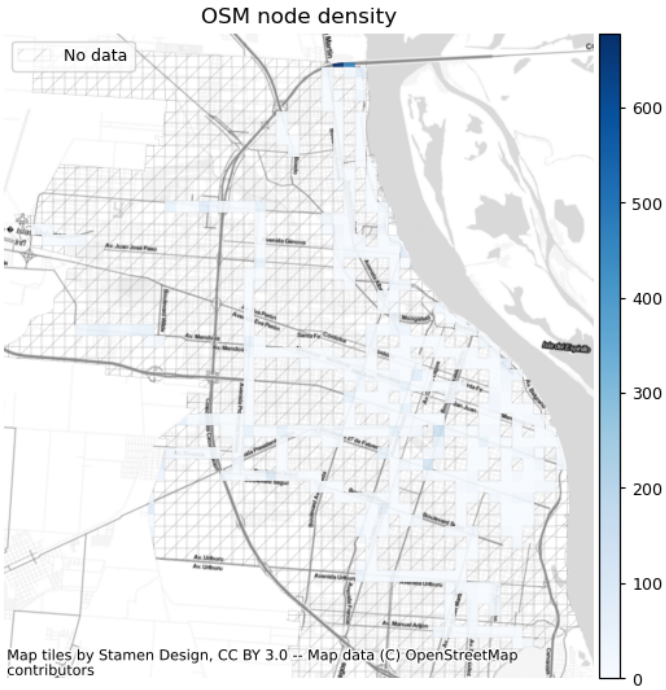
Global network densities (per km2)

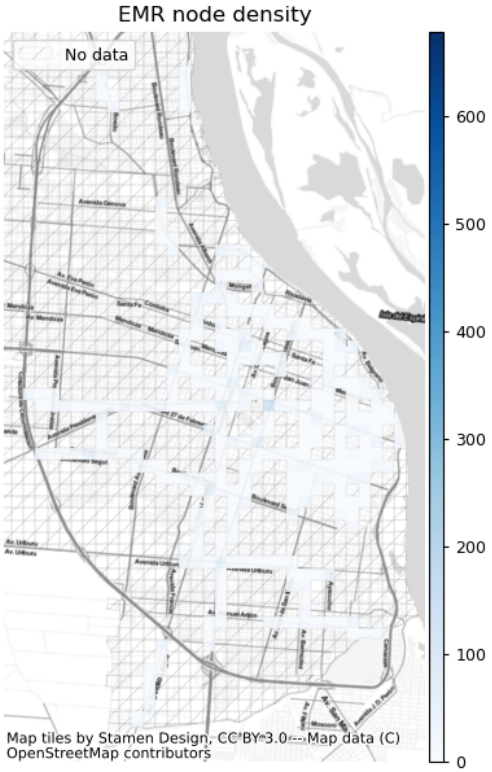


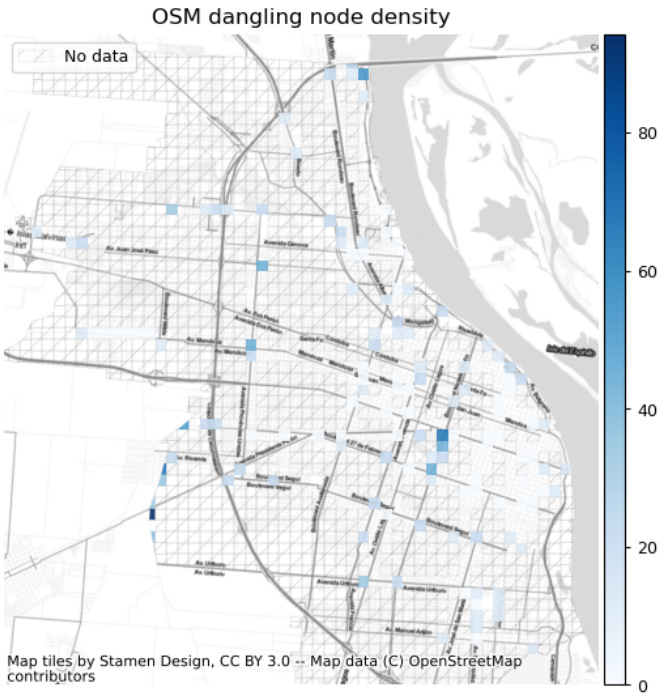
Local network densities

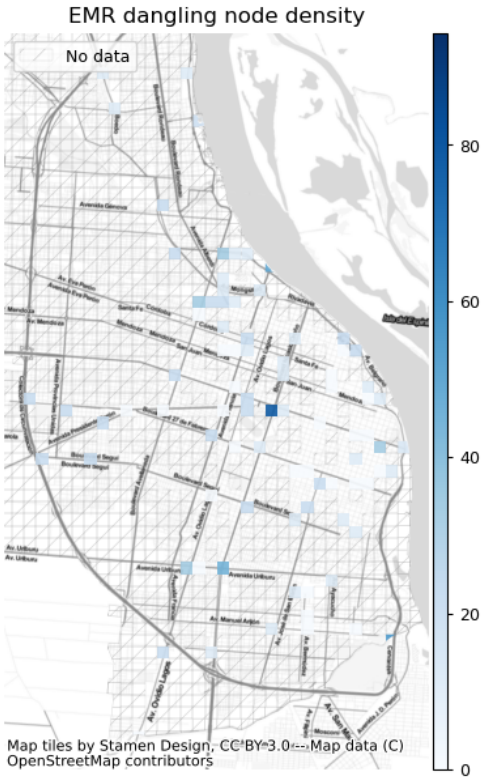








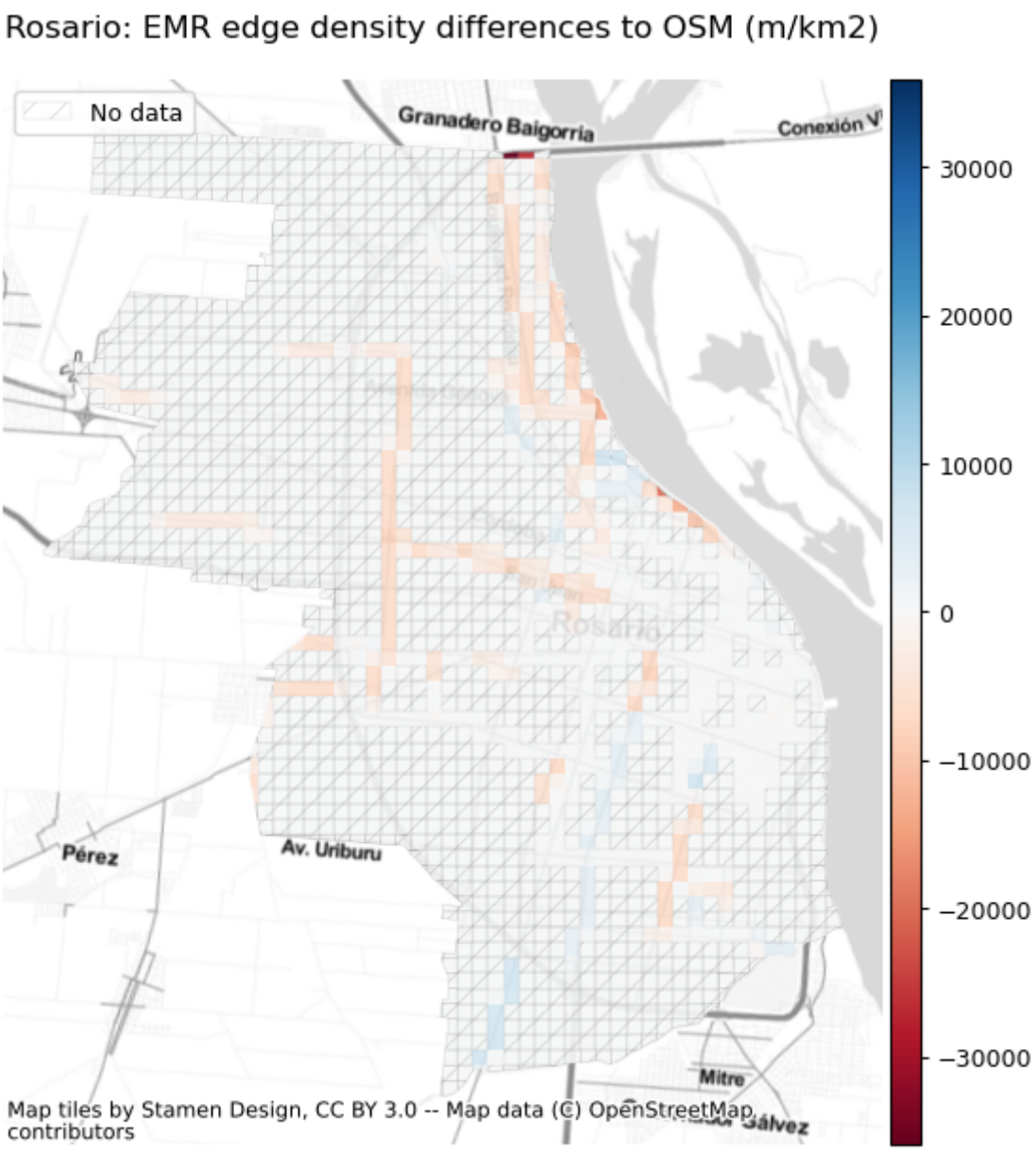


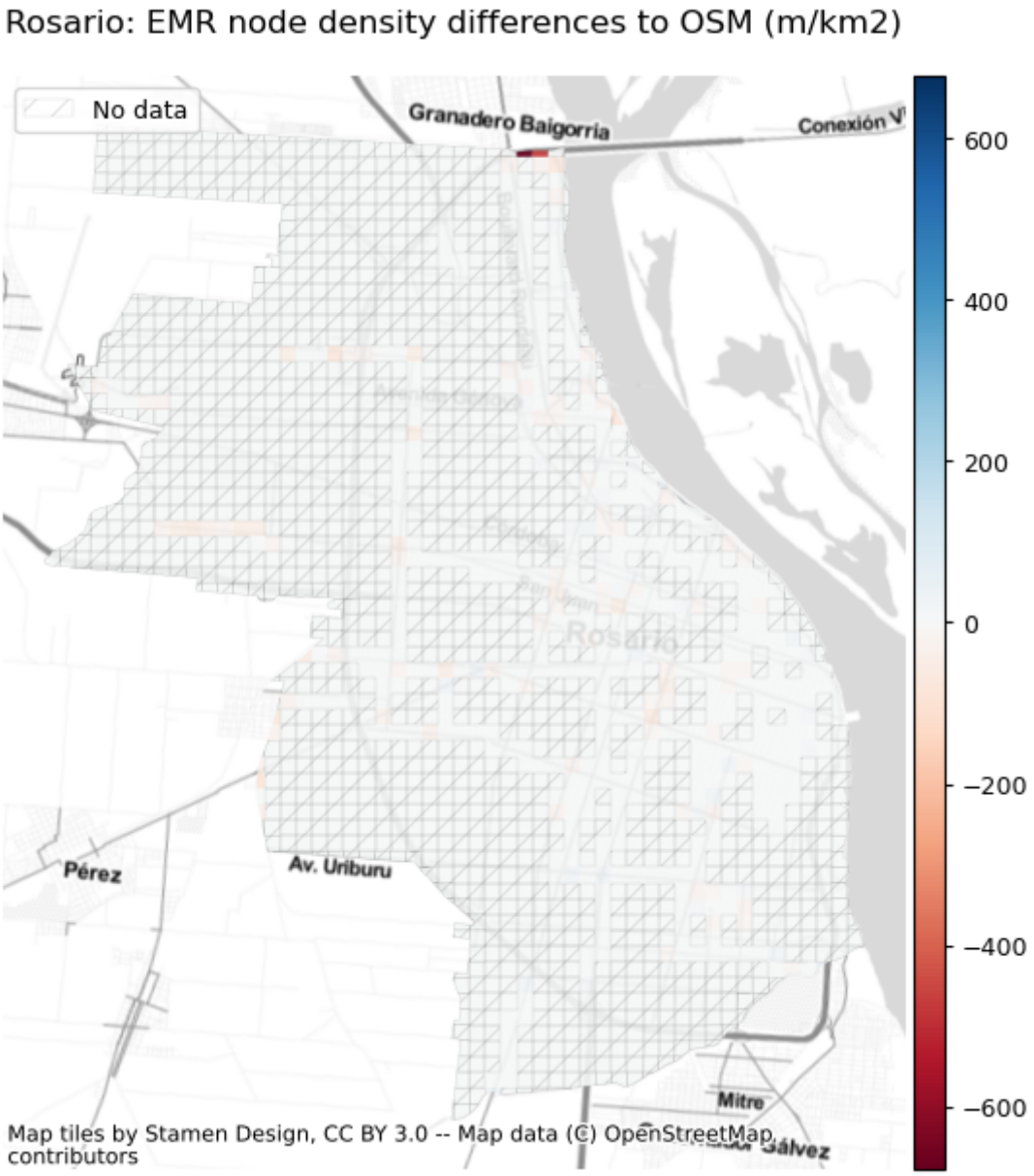


Local differences in network densities

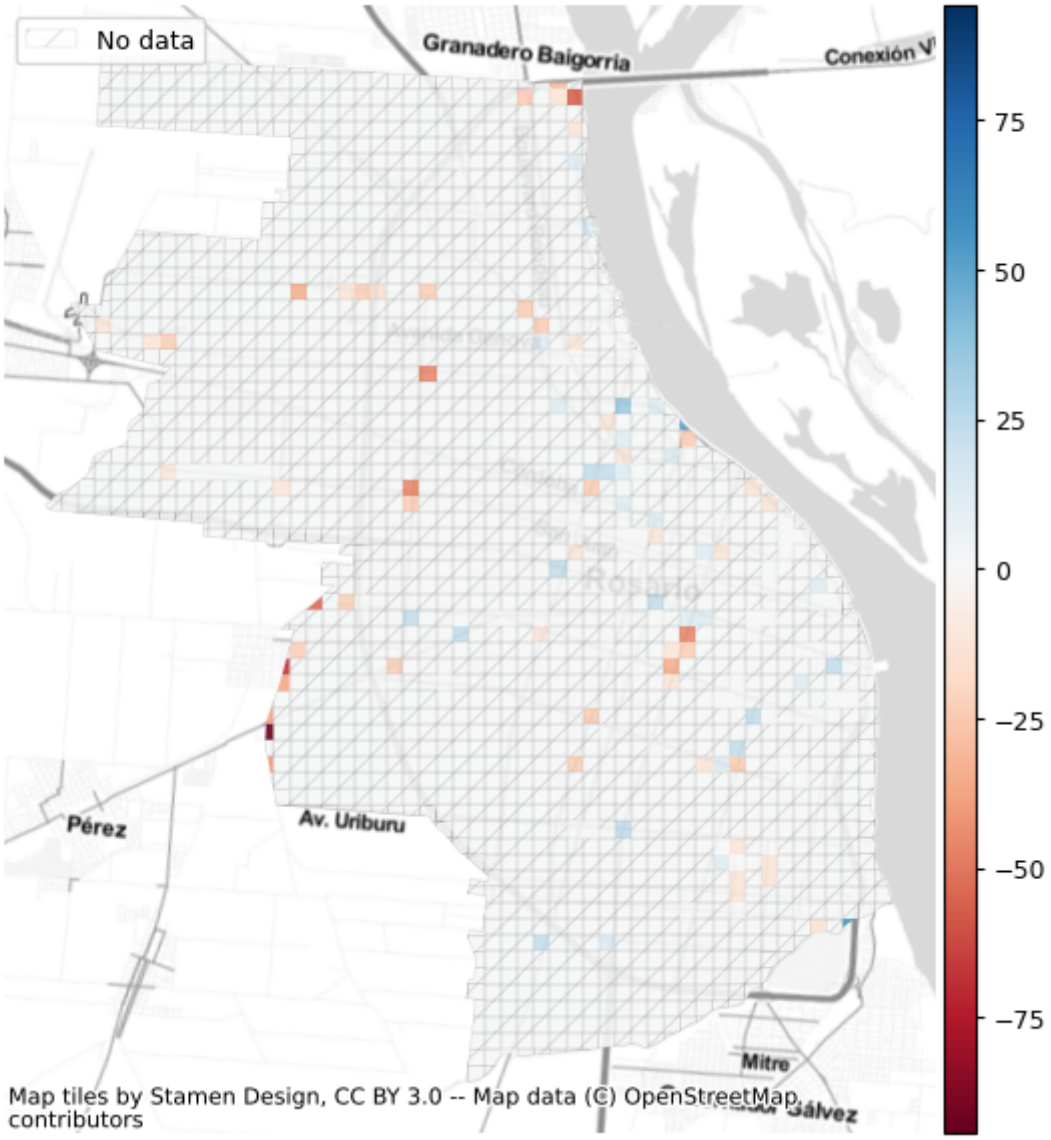
The densities in the OSM data are taken as base line for comparison, with absolute differences computed as `reference value` - `OSM value`. Hence, positive values indicate that the reference

density of the infrastructure type is higher than the OSM density; negative values indicate that the reference density is lower than the OSM density.



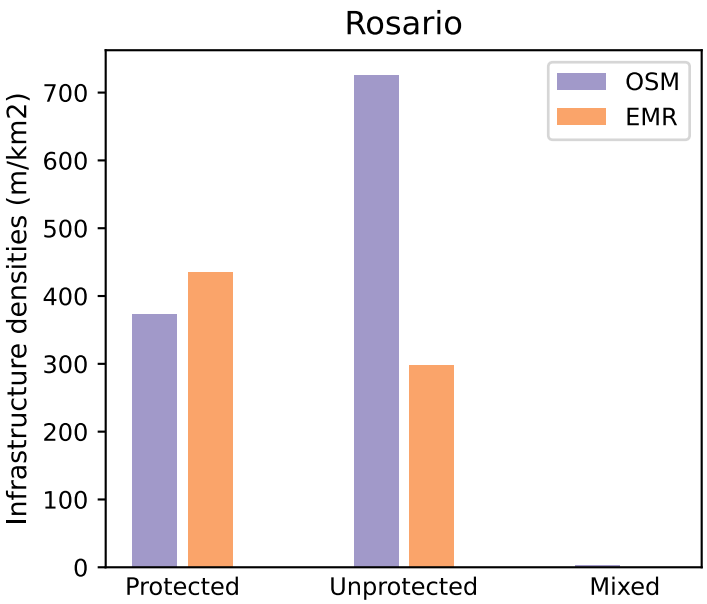

Rosario: EMR node density differences to OSM (m/km2)
No data
Granadero Baigorria
Conexión V
Rosario
Pérez
Av. Uriburu
Mitre
600
400
200
0
−200
−400
−600
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors

Rosario: EMR dangling node density differences to OSM (m/km2)



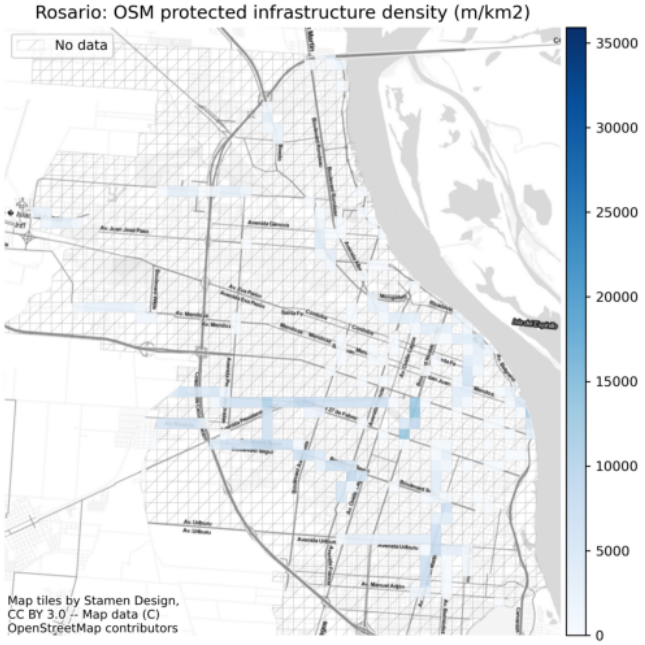
Densities of protected and unprotected bicycle infrastructure

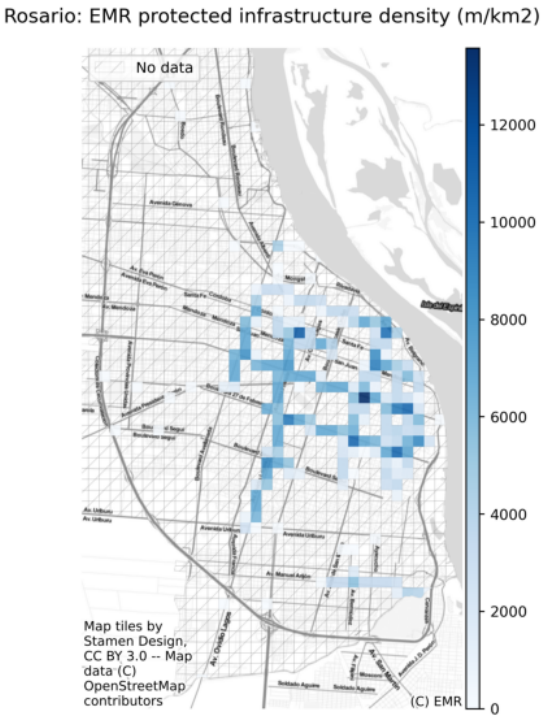
Global network densities for protected/unprotected infrastructure



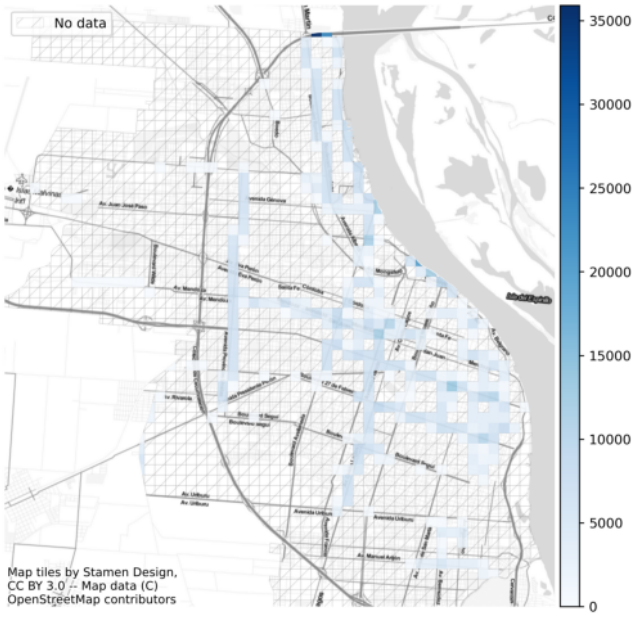
Local network densities for protected/unprotected infrastructure

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

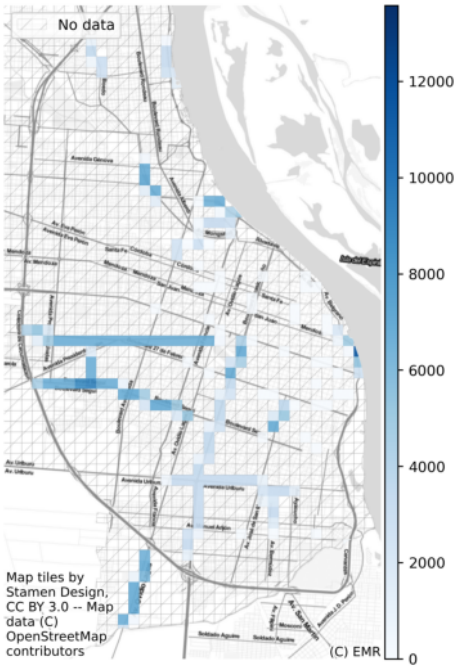




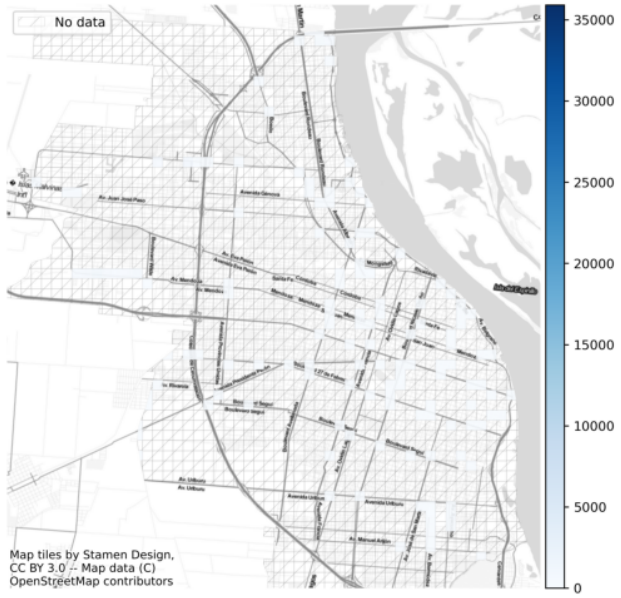
Rosario: OSM unprotected infrastructure density for (m/km2)



Rosario: EMR unprotected infrastructure density (m/km2)

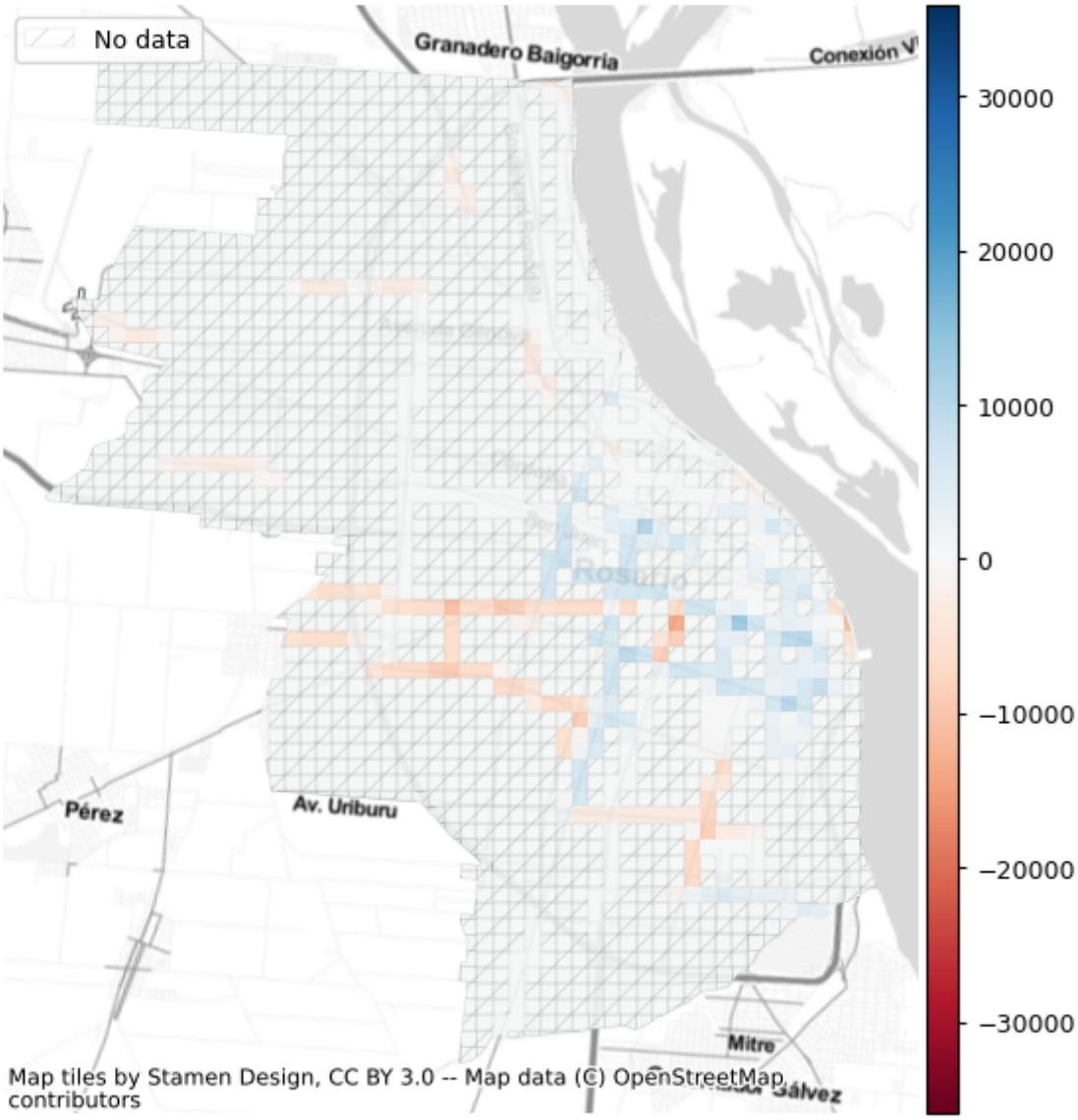


Rosario: OSM mixed protection infrastructure density (m/km2)

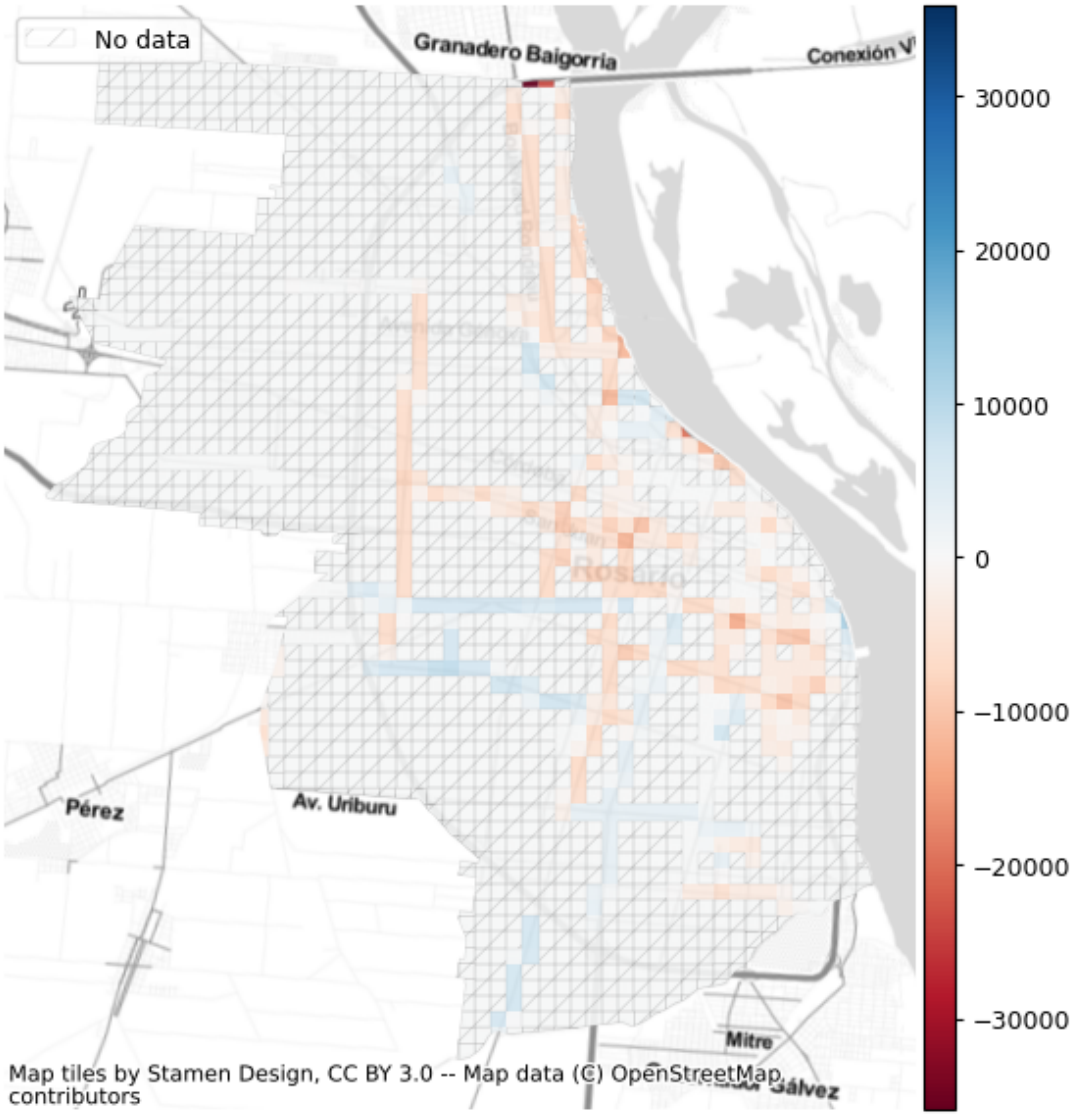


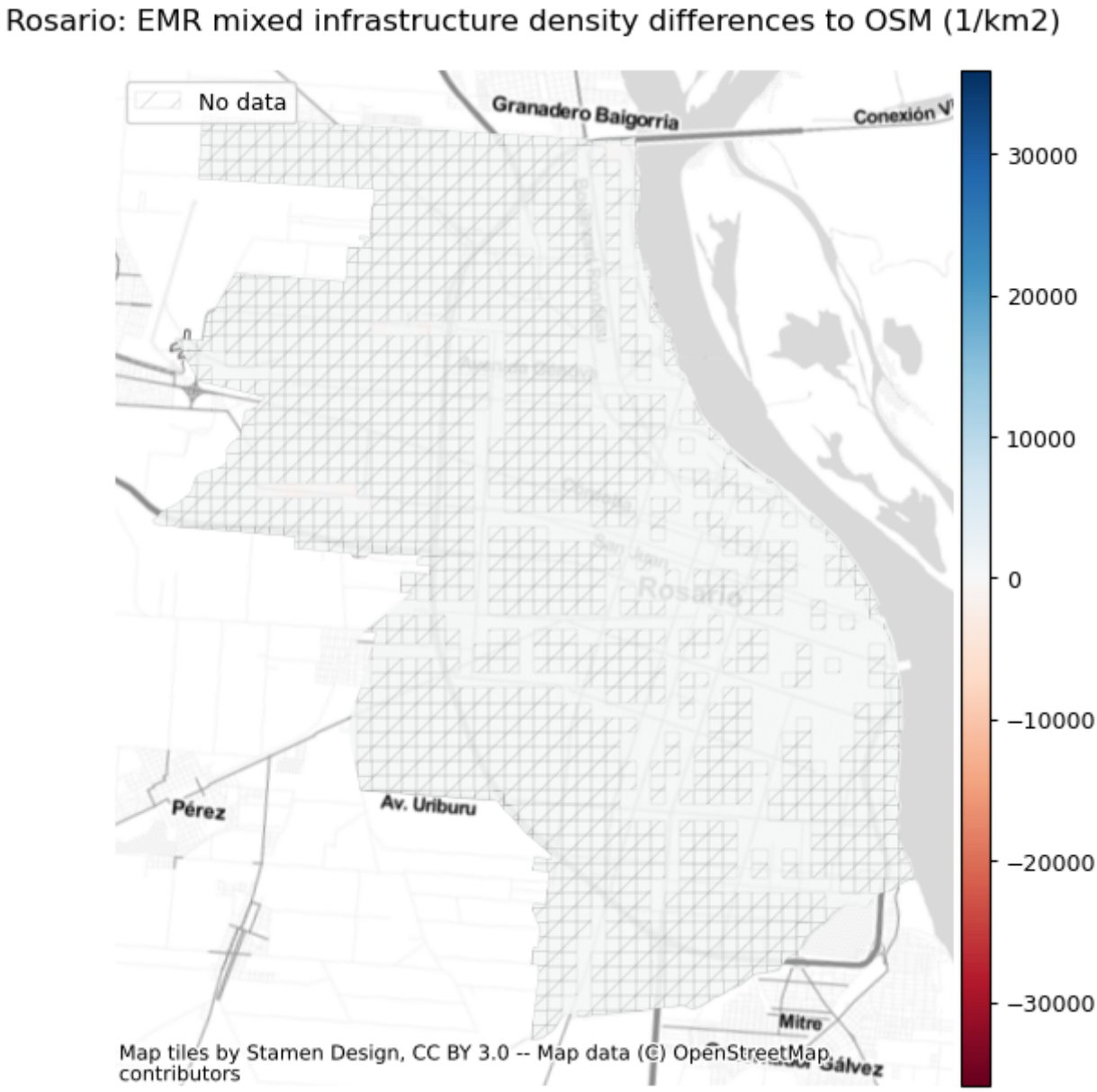
Differences in infrastructure type density

Rosario: EMR protected infrastructure density differences to OSM (m/km2)



Rosario: EMR unprotected infrastructure density differences to OSM (1/km2)





Network topology

After having compared data completeness, i.e. *how much* infrastructure is mapped, here we focus on differences in network *topology*, which gives 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.

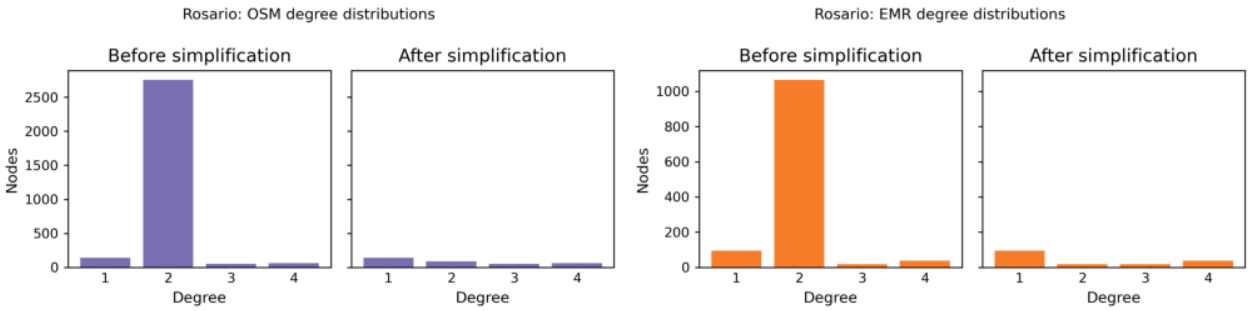
Simplification outcomes

```
Simplifying the OSM network decreased the number of edges by 89.9%.
Simplifying the OSM network decreased the number of nodes by 88.6%.
```

```
Simplifying the EMR network decreased the number of edges by 86.2%.
Simplifying the EMR network decreased the number of nodes by 86.2%.
```

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.

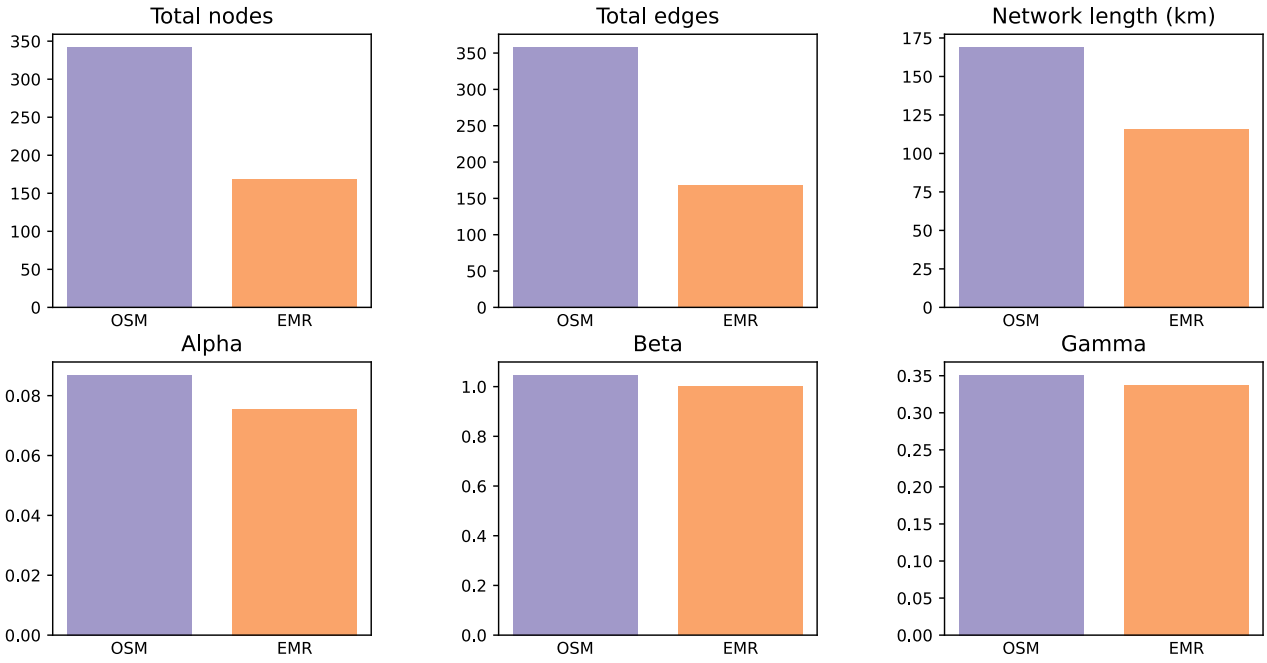
Interpretation

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 indication that the structure and topology of the two data sets are fundamentally different.

```
Alpha for the simplified OSM network: 0.09
Beta for the simplified OSM network: 1.05
Gamma for the simplified OSM network: 0.35
```

```
Alpha for the simplified EMR network: 0.08
Beta for the simplified EMR network: 1.00
Gamma for the simplified EMR network: 0.34
```



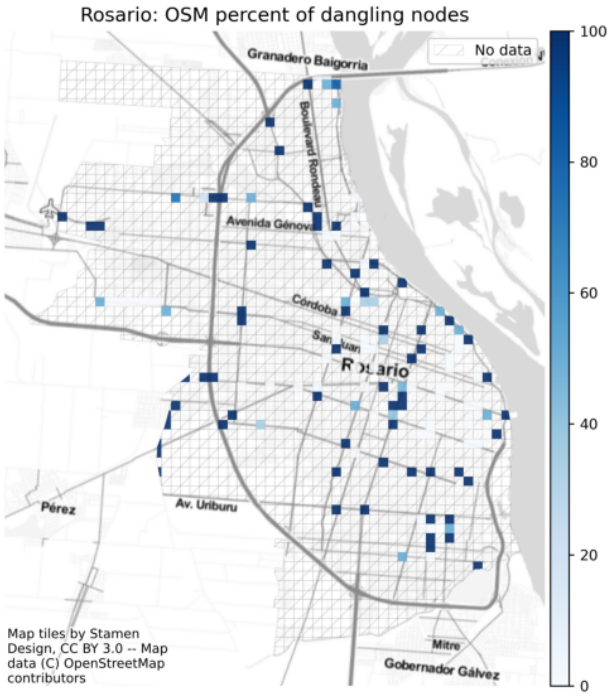
Dangling nodes

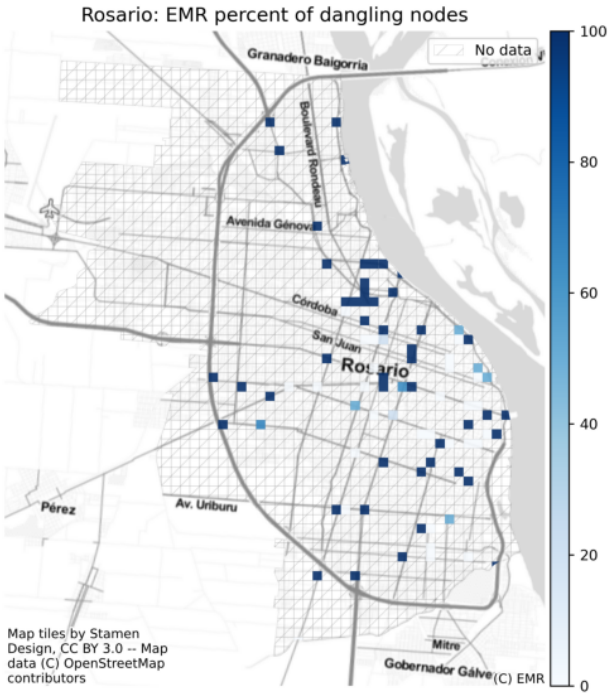
Dangling nodes in OSM & reference networks

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

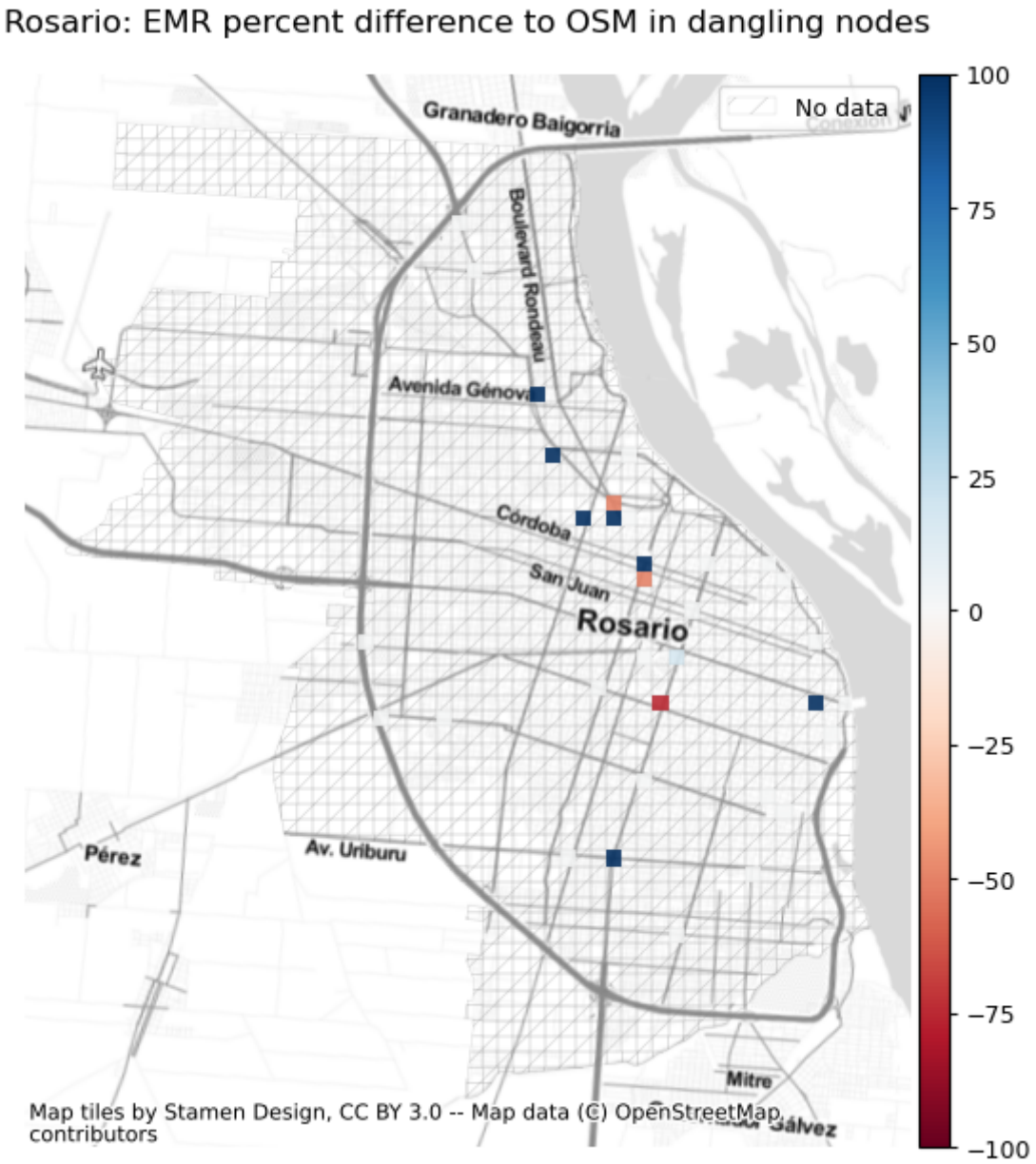
Local values for dangling nodes

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/Rosario/maps_interactive/overundershoots_3_3_compare.ht
ml
```

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.

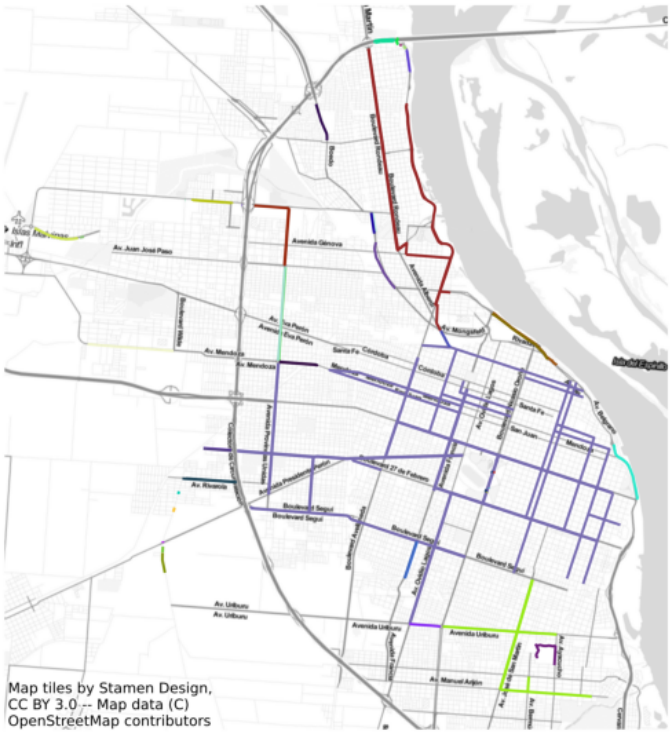
Interpretation

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 43 disconnected components.
The EMR network in the study area consists of 25 disconnected components.
```

Rosario: OSM disconnected components



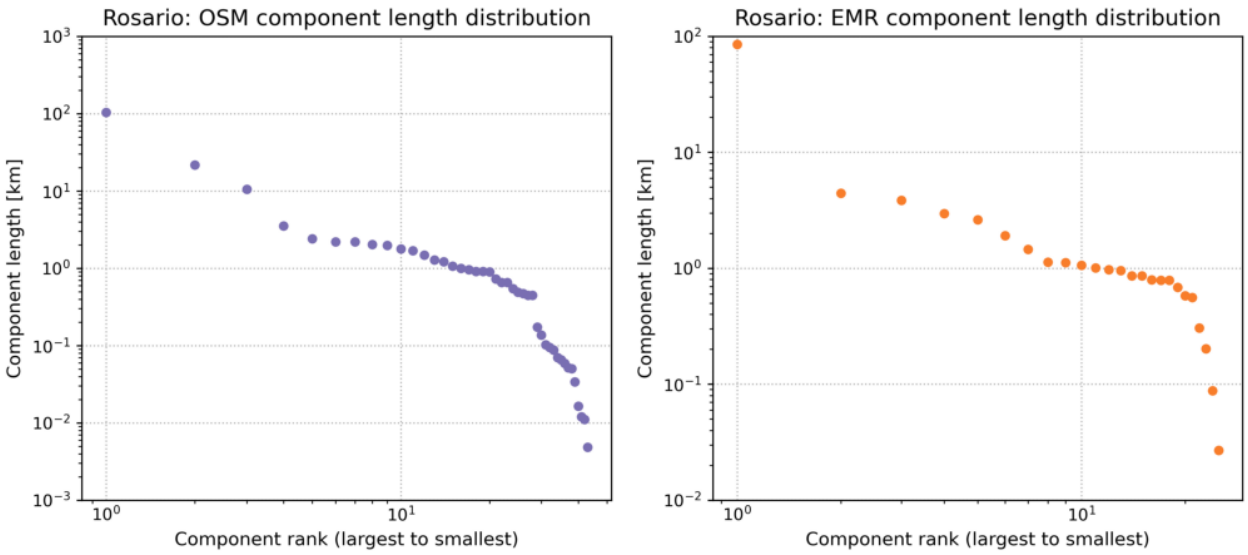
Rosario: EMR disconnected components

Component length distribution

The distribution of all network component lengths can be visualized in a so-called *Zipf plot*, which orders the lengths of each component by rank, showing the largest component's length on the left, then the second largest component's length, etc., until the smallest component's length on the right. When a Zipf plot follows a straight line in log-log scale, it means that there is a much higher chance to find small disconnected components than expected from traditional distributions (Clauset et al., 2009). This can mean that there has been no consolidation of the network, only piece-wise or random additions (Szell et al., 2022), or that the data itself suffers from many gaps and topology errors resulting in small disconnected components.

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 at the infrastructure level, most of the infrastructure has been connected to one large component, and that the data reflects this - i.e. the data is not suffering from gaps and missing links to a large extent. Bicycle networks might also be somewhere inbetween, with several large components as outliers.

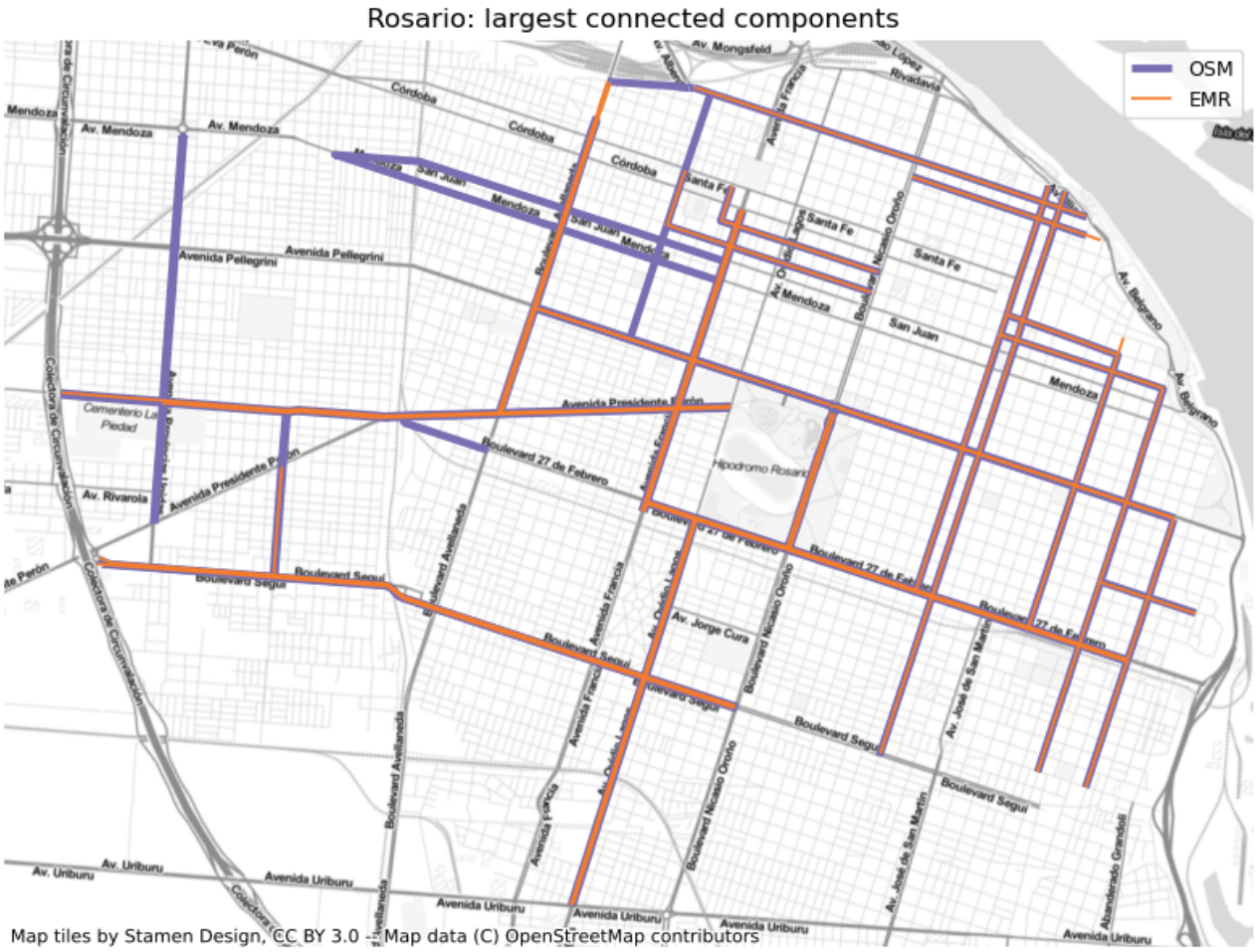
In case of a comparison over the same region, as shown below, if one data set shows a clear outlier in its largest connected component while the other does not, and if it is also at least as large, it can in general be interpreted as being more complete.



Largest connected component

```
The largest connected component in the OSM network contains 61.48% of the network length.
The largest connected component in the EMR network contains 74.01% of the network length.
```

Overlay of largest connected component in OSM and reference networks



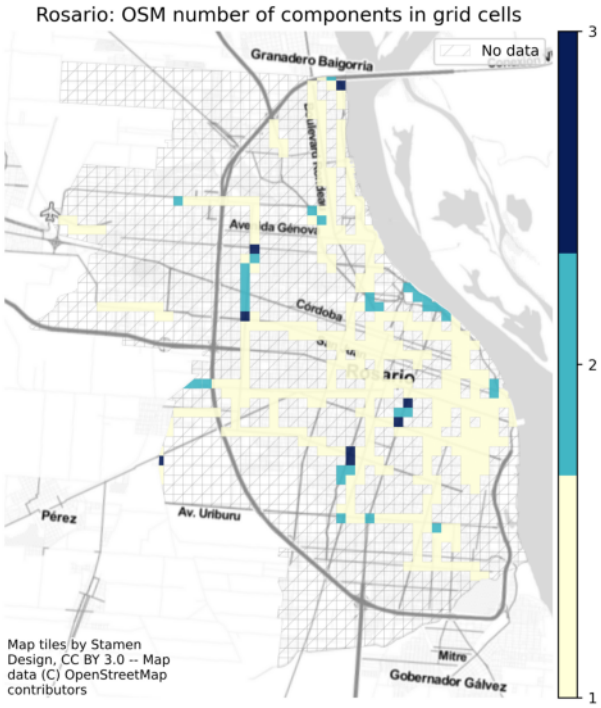
Missing links

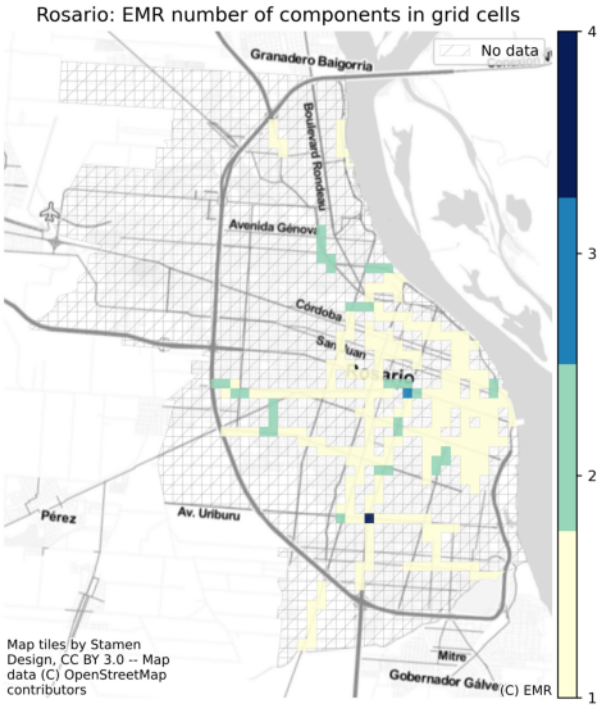
In the plot of potential missing links between components, all edges that are within the specified distance of an edge on another component are plotted. The gaps between disconnected edges are highlighted with a marker. The map thus highlights edges which, despite being in close proximity of each other,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

```
Interactive map saved at results/COMPARE/Rosario/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.

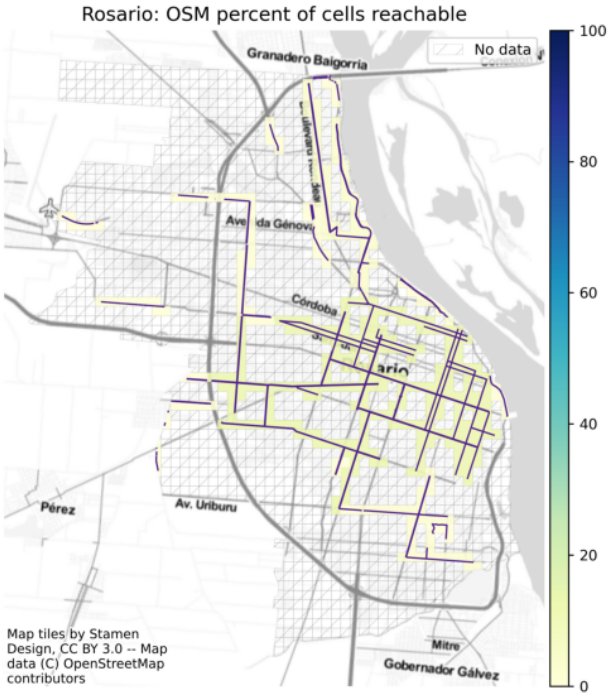


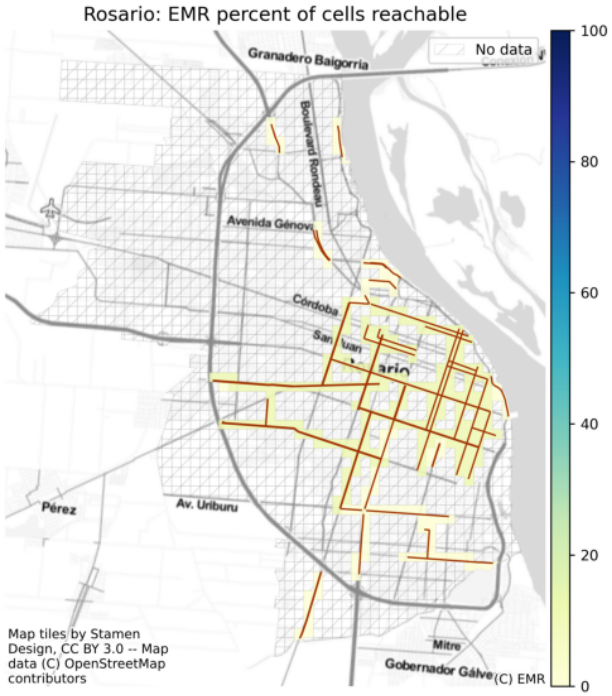


Component connectivity

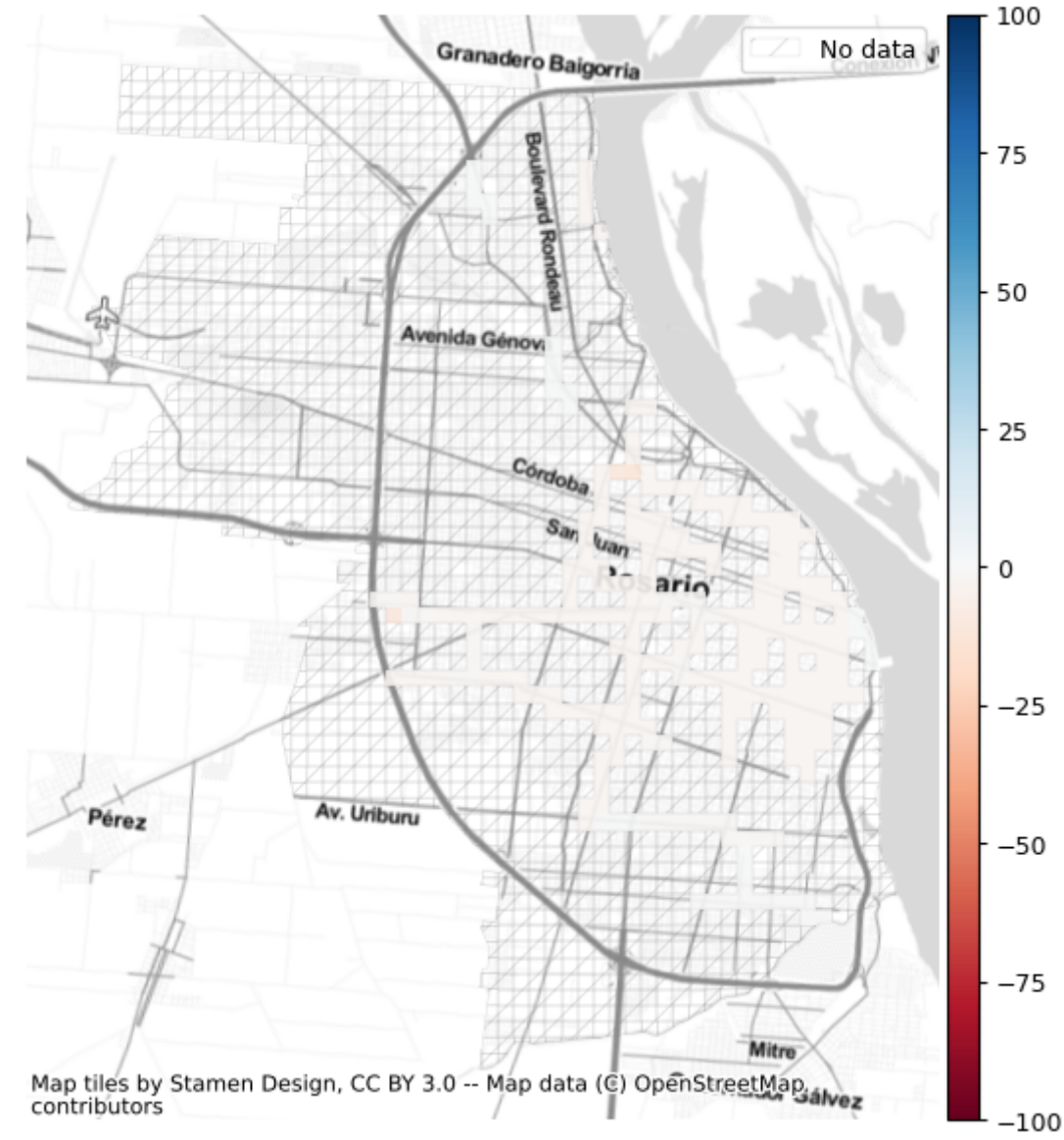
Here we visualize differences between how many cells can be reached from each cell. The metric is a crude measure for network connectivity but has the benefit of being computationally cheap and thus able to quickly highlight stark differences in network connectivity.

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.


Rosario: OSM percent of cells reachable
No data
Granadero Baigorria
Córdoba
Pérez
Av. Uriburu
Mitre
Gobernador Gálvez
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors
100
80
60
40
20
0


Rosario: EMR percent of cells reachable
No data
Granadero Baigorria
Boulevard Rondeau
Avenida Génova
Córdoba
Pérez
Av. Uriburu
Mitre
Gobernador Gálvez
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors
(C) EMR
100
80
60
40
20
0

Rosario: EMR difference to OSM in percent of cells reached



Summary

Extrinsic Quality Comparison		
	OSM	EMR
Total infrastructure length (km)	175	117
Protected bicycle infrastructure density (m/km2)	373	435
Unprotected bicycle infrastructure density (m/km2)	726	299
Mixed protection bicycle infrastructure density (m/km2)	3	0

Bicycle infrastructure density (m/km2)	1,103	733
Nodes	342	168
Dangling nodes	143	94
Nodes per km2	2	1
Dangling nodes per km2	1	1
Overshoots	1	1
Undershoots	0	19
Components	43	25
Length of largest component (km)	104	86
Largest component's share of network length	61%	74%
Component gaps	13	18
Alpha	0.09	0.08
Beta	1.05	1.00
Gamma	0.35	0.34