Appendix 2a – Data preparation: Network cleaning

Why network cleaning?

* OpenStreetMap (OSM) is a large and heterogeneous dataset that is being audited by the OSM community.
* Unfortunately, the immense amount of data and heterogeneity of the OSM contributors guarantee that the data is not ready to use.

Network cleaning

* A variety of errors can cause problems when using network data for an analysis.
* For example the coordinates of the endpoints of the lines that make up the network can be *almost* the same.
* A small deviation can cause the lines to be found unconnected by out analysis tools.
* Other common mistakes are duplicated entries or crossing lines that have no common node at the intersection an are thus not connected.
* On top of the usual pitfalls of network data, we have downloaded city-wide data in tiles of 2×2 km.
* Prior to the analysis, these tiles have to be joined together to create a network covering the entire city.
* On the edges of these tiles, there may be overlapping paths due to e.g. a street crossing through the border of one of our tiles.
* This street would be downloaded in the two adjacent tiles.
* For a consistent analysis, we have to make sure that most of these errors are being detected and corrected.

Functionality