Manuscript

BikeDNA: A tool for bicycle infrastructure data and network assessment

Ane Rahbek Vierø ⓘ, Anastassia Vybornova ⓘ and Michael Szell ⓘ
IT University of Copenhagen, Copenhagen, Denmark

Michael Szell ⓘ
ISI Foundation, Turin, Italy

Complexity Science Hub, Vienna, Austria

Abstract
Building high-quality bicycle networks requires knowledge of existing bicycle infrastructure. However, bicycle network data from governmental agencies or crowdsourced projects like OpenStreetMap often suffer from unknown, heterogeneous, or low quality, which hampers the green transition of human mobility. In particular, bicycle-specific data have peculiarities that require a tailor-made, reproducible quality assessment pipeline: For example, bicycle networks are much more fragmented than road networks, or are mapped with inconsistent data models. To fill this gap, we introduce BikeDNA, an open-source tool for reproducible quality assessment tailored to bicycle infrastructure data with a focus on network structure and connectivity. BikeDNA performs either a standalone analysis of one data set or a comparative analysis between OpenStreetMap and a reference data set, including feature matching. Data quality metrics are considered both globally for the entire study area and locally on grid cell level, thus exposing spatial variation in data quality. Interactive maps and HTML/PDF reports are generated to facilitate the visual exploration and communication of results. BikeDNA supports quality assessments of bicycle infrastructure data for a wide range of applications—from urban planning to OpenStreetMap data improvement or network research for sustainable mobility.

Corresponding author:
Ane Rahbek Vierø, IT University of Copenhagen, Rued Langgaards Vej 7, Copenhagen 2300, Denmark.
Email: anev@itu.dk

Introduction

Cities across the globe are striving to make their transportation systems more environmentally and socially sustainable. One of the most cost-effective ways to do so is to boost cycling, like in Paris or Bogotá (C40 Cities, 2019; City Of Paris 2021), and as recommended by the European Commission and the IPCC (EC, 2021; Jaramillo et al., 2022). However, achieving a substantial shift towards cycling remains a challenge. Adequate infrastructure is typically lacking, as is the *collection and provision of data* on bicycle infrastructure, which is necessary to harness the potential of data-driven bicycle network planning. Recent advances demonstrate that quantitative analyses of bicycle infrastructure on the network level can assist planning decisions and considerably improve the impact of planned investments (Natera Orozco et al., 2020; Olmos et al., 2020; Steinacker et al., 2022; Szell et al., 2022; Vybornova et al., 2022), but the necessary precondition of readily available and complete bicycle infrastructure data is typically not fulfilled. Moreover, even when such data are available, there is often little knowledge on data quality. This is true both for administrative data and for crowdsourced data such as OpenStreetMap (OSM) (Nelson et al., 2021; Rambøll 2022).

Further, because of the persistent political and financial constraints of bicycle network planning (Cox 2020), bicycle networks are qualitatively different from road and public transport networks. Networks of designated bicycle infrastructure consist of a much high number of disconnected components and have a much lower degree of connectivity than road networks for motorized transport—even in famously bicycle-friendly cities like Copenhagen (Natera Orozco et al., 2020). This high degree of fragmentation poses a challenge when applying standard Geographical Information System (GIS) tools to investigate network quality issues, since it becomes impossible to judge in an automated manner whether over/undershoots and disconnected components are due to poor data quality or an accurate representation of the (poorly connected) network.

To fill this gap and to help researchers, planners, and others in the field assess the quality of bicycle infrastructure data, we introduce BikeDNA (Bicycle Infrastructure Data and Network Assessment). BikeDNA is a computational tool written in Python that performs a reproducible quality assessment on one or two bicycle infrastructure data sets (see Figure S1 in the Supplementary Material). The default definition of "bicycle infrastructure" in BikeDNA is any part of the road network that is designated specifically for cycling, that is, protected bicycle infrastructure plus designated bicycle lanes; this definition can be modified by the user, for example, to include residential or low-traffic streets. BikeDNA can perform standalone analysis of either OSM or administrative data. If both OSM and administrative data are available, BikeDNA compares differences between the two data sets. To our knowledge, there are currently no other open-source tools for reproducible data quality assessment available that are tailored to the peculiar nature of bicycle infrastructure data.

Previous research on spatial data quality

While BikeDNA is designed to be compatible with both crowdsourced and administrative data on bicycle infrastructure, the quality assessment of OSM data is a central feature of the tool due to the widespread usage of OSM for bicycle network research (Ferster et al., 2020; Nelson et al., 2021). Much recent work on spatial data quality has moreover focused on the quality of crowdsourced or volunteered data, in attempts to build trust in data sets from non-official sources. Our review of previous work is therefore focused on quality assessment of crowdsourced road network data. For more comprehensive reviews of quality assessment methods for Volunteered Geographic Information (VGI), we refer to Fonte (2017), Senaratne et al. (2017), and Degrossi et al. (2018). For an overview of solutions for spatial data quality more generally, see Medeiros and Holanda (2019).

Intrinsic versus extrinsic analysis. A common way to classify spatial data quality assessment methods is as either "intrinsic" (standalone) or "extrinsic" (comparative)—a classification also used by BikeDNA. Intrinsic methods study inherent properties of the data set itself or how it was created, while extrinsic methods make use of an external reference data set for comparison. Early work on VGI data quality was primarily occupied with extrinsic evaluations with an emphasis on evaluating the completeness and spatial accuracy, using local metrics for the length of the mapped network in an area (Haklay 2010) as well as more advanced methods for feature matching (Koukoletsos et al., 2012). While data completeness and accuracy are relevant data quality metrics, they do not address topological quality, that is, how network elements are connected *structurally*, which, however, is crucial for many applications of road network data (see Figure 1). For this reason, extrinsic approaches were quickly expanded to also include road network topology, routing, and other network attributes (Girres and Touya 2010; Mondzech and Sester 2011; Neis et al., 2012; Zielstra and Hochmair 2012; Graser et al., 2014). Intrinsic evaluations of VGI data quality have been developed for situations where no reference data are available, or to avoid some of the more computationally intensive methods used in extrinsic comparisons. Methods for intrinsic evaluation analyze the spatial and structural dimensions of the data (Neis et al., 2012; Barron et al., 2014), as well as the
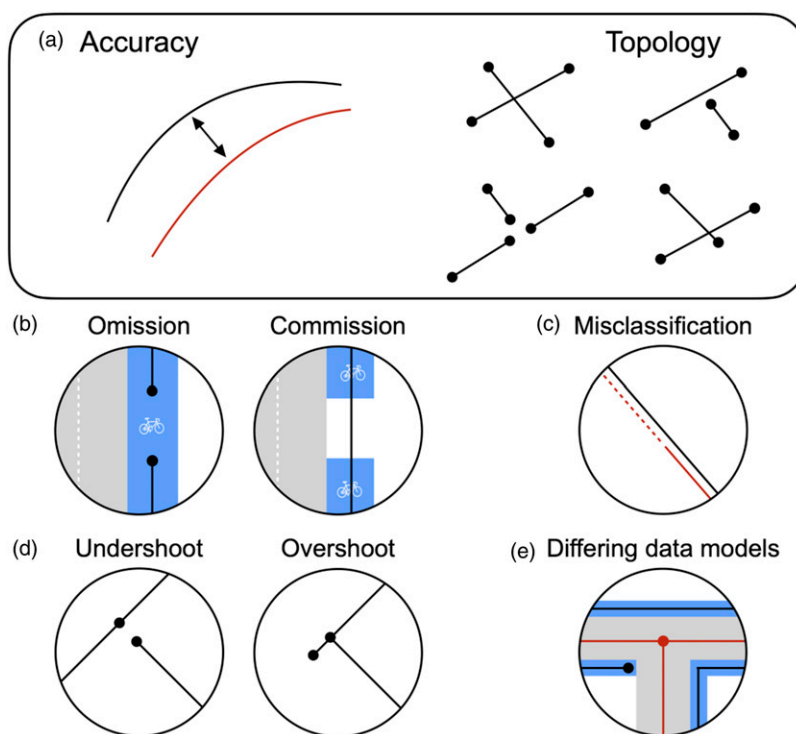

Figure 1. Known quality issues in bicycle infrastructure data. a) Different aspects of data quality assessment: Accuracy (left) versus topology (right). Early research on spatial data quality focused on accuracy, while BikeDNA puts a special emphasis on topology. Adapted from Haklay et al. (2010) and Neis et al. (2012). b) Errors of omission (left) or commission (right) result in differing data completeness. c) Misclassification leads to different types of bicycle infrastructure in two corresponding data sets (red vs black). d) Edge undershoot (left) and edge overshoot (right). e) Mapping of bicycle infrastructure to the centerline of the road (red) or on the side (black). In this case, the centerline mapping creates routable network data at the cost of accuracy.

number of contributors and edits (Keßler et al., 2011; Neis et al., 2013; Barron et al., 2014; Gröchenig et al., 2014). Extrinsic and intrinsic methods are often used in combination, sometimes in more elaborate frameworks for assessing spatial data quality (Barron et al., 2014; Ballatore and Zipf 2015), as it is the case for BikeDNA.

Quality standards and fitness for purpose. Another way to classify existing approaches to spatial data quality is to distinguish between the formulation of standards for spatial data quality versus the idea of "fitness for purpose." Following the former approach, for example, ISO 19157 defines spatial data quality as composed of: accuracy (positional, thematic, and temporal); data completeness; internal logical consistency; and finally, usability (Fonte 2017). While these are useful concepts that are used throughout the quality assessment in BikeDNA, the consistency implied by formal standards for data quality is difficult to apply to heterogeneous, volunteered data like OSM (Hashemi and Abbaspour 2015), and might be less relevant for some use cases. Therefore, Barron et al. (2014) suggest to instead evaluate OSM data based on fitness for purpose, meaning that the quality and suitability of a data set are evaluated based on whether the data fulfill the requirements for the individual use case. This is also the approach adopted in BikeDNA. Therefore, instead of referencing universal standards, BikeDNA is designed to help decide whether a data set is *good enough* for an intended use case.

Gaps in network-level bicycle data assessment and tools. Only few studies of quality assessments specifically of bicycle infrastructure data exist, with Hochmair et al. (2015) and Ferster et al. (2020) as the most prominent examples. Both studies compare OSM data on dedicated bicycle infrastructure to external reference data sets, assessing data completeness generally and for different types of bicycle infrastructure. This is of particular interest, since studies on the general completeness of OSM have shown that bicycle lanes and paths often are among the later features to be mapped (Neis et al., 2012; Barron et al., 2014). These studies provide important insights into the completeness of bicycle infrastructure data in OSM, but they do not sufficiently address network topology. This is a highly relevant research gap, since infrastructure data for pedestrians and cyclists are more prone to topological errors than data for motorized transport (Neis et al., 2012). Many of the assumptions used to evaluate road networks for car traffic, for example, interpreting disconnected components and small gaps between edges as errors, do furthermore not hold for bicycle networks which are often built in a fragmented and piece-wise manner (Natera Orozco et al., 2020; Szell et al., 2022).

   Overall, the quality of OSM data is well-studied, particularly when it comes to data on the car road network (Fonte 2017), and OSM data are often shown to be of high quality compared to other data sources (Neis et al., 2012; Zielstra and Hochmair 2012; Hochmair et al., 2015; Ferster et al., 2020; Zhang et al., 2021). However, data errors and inconsistencies are not randomly distributed in OSM, but instead correlate with lower population densities and socioeconomic variables; vary from country to country; and differ for different parts of the transportation network (Haklay 2010; Mondzech and Sester 2011; Ziemke et al., 2019; Ferster et al., 2020). Administrative data have also been shown to suffer from uneven data quality, for example, due to the role of local data maintainers and differing mapping practices (Hvingel and Jensen 2023). For this reason, quality assessment in one particular location cannot be generalized to other locations. Moreover, the quality of data on bicycle infrastructure often lags behind the quality of other road network data, both in terms of completeness and consistency. While previous research does not identify any single reason for the quality differences between different parts of the road network, Ferster et al. (2020) and Hvingel and Jensen (2023) point to differing mapping practices when looking at OSM and administrative data. Paths and other infrastructure for active mobility are furthermore often among the later features to be mapped in OSM (Neis et al., 2012), leaving less time for errors to be identified and corrected.

Although these findings have profound implications for bicycle research that uses OSM data, to date only few studies have assessed specifically the quality of bicycle infrastructure data in OSM. In short, the lack of studies and tools that assess how well bicycle infrastructure has been mapped is a considerable barrier to OSM data uptake and to data-driven bicycle network planning.

Filling the gap with BikeDNA

To address the lack of available tools tailored particularly to bicycle network quality assessment, we have developed BikeDNA. The tool contains a systematic and easy-to-follow quality assessment of bicycle infrastructure data. This is relevant for use cases such as bicycle routing (Murphy and Owen 2019), connectivity analysis (Natera Orozco et al., 2020; Vybornova et al., 2022), network quality assessment (Dill 2004), and accessibility studies (PeopleForBikes 2023). BikeDNA implements and visualizes a wide range of existing spatial VGI data quality metrics, but is tailored specifically to bicycle infrastructure data. The main features of BikeDNA are:

- Analysis of completeness, network density, and OSM tags
- Analysis of network topology and connectivity
- Feature matching between OSM and reference data
- Creation of detailed PDF and interactive HTML reports

BikeDNA thus addresses both data *completeness*, that is, how much information each data set contains, and data *consistency*. BikeDNA moreover accounts for different scenarios of data availability and can be applied in scenarios where only OSM, only reference data, or both are available. All data quality metrics, whenever relevant, are computed globally for the entire study area and locally for each cell in a grid covering the study area, which allows to detect local variations in data quality. The quality metrics are made accessible by extensive documentation and interpretation assistance in each step, and automated plotting of results with static and interactive maps. BikeDNA can help improve the quality of a data set by pinpointing both in which *features* and which *locations* gaps, errors, or inconsistencies exist. Potential users are researchers, planners, data maintainers, and everyone else who needs an indication of bicycle infrastructure data quality.

The rest of the paper is organized as follows: In the next section, we outline the concepts of evaluating bicycle infrastructure data quality. Then, we provide a descriptive overview of BikeDNA: First, we show how to use BikeDNA, then we provide example outputs of BikeDNA's quality assessment of OSM and the open government data "GeoDanmark" for a showcase area in Greater Copenhagen, Denmark. We conclude with a discussion of limitations and the need for future work on bicycle data quality. In addition, we outline possible improvements to BikeDNA and quality assessments more generally. For a description of the technical setup and analysis workflow, we refer to the Supplementary Material and Figure S1. BikeDNA makes extensive use of existing open-source Python libraries (Jordahl et al., 2021; Boeing 2017; Hagberg et al., 2008; Fleischmann 2019). The full description and technical specifications of BikeDNA are available on GitHub: https://github.com/anerv/BikeDNA.

Evaluating bicycle infrastructure data

In the following section, we introduce the analytical framework with an overview of the different scenarios of data availability and data sources considered by BikeDNA, describe known data quality issues in bicycle infrastructure data, and discuss how to perform and interpret quality assessments without ground truth data.

Data sources

BikeDNA considers three different scenarios of data availability for bicycle networks: either data are available from OSM; data are available from an administrative source, for example, a local municipality; or data are available from both. OSM is the primary data source for data used in bicycle research and routing applications. To our knowledge, administrative data are used primarily by governmental institutions. OSM and administrative data are not necessarily at odds, since OSM data sometimes are partly based on administrative data (Zielstra et al., 2013; Rambøll 2022).

Bicycle infrastructure data from OSM are in many cases of a similar or higher quality than administrative data when it comes to aspects such as data completeness (Hochmair et al., 2015; Ferster et al., 2020). This is partly due to a lack of government resources, as well as the unfavorable treatment of active mobility in data collection in many jurisdictions (Rambøll 2022). The inherently heterogeneous nature of VGI data, combined with the lack of metadata on data quality parameters, does, however, pose a barrier to data uptake in, for example, public transport planning (Hashemi and Abbaspour 2015). The lack of standardized quality assurance for crowdsourced data necessitates methods for local data quality assessment, as offered by BikeDNA.

Known quality issues

Within the fitness for purpose approach, it depends on the use case whether specific data errors are considered a problem. For example, spatial inaccuracies such as smaller displacements of objects are rarely an issue for network-based bicycle research, as long as the data topology is correctly represented (see Figure 1a). There are, however, several types of errors and inconsistencies which will be a problem for most data applications, and which we briefly explain below.

Errors of omission/commission. "Errors of omission" describe missing data, while "errors of commission" refer to features in the data set that should not be there (see Figure 1b). These error types appear in both OSM and administrative data sets on bicycle infrastructure (Hochmair et al., 2015; Ferster et al., 2020).

Misclassification. Misclassification happens when bicycle infrastructure is classified as something else, or when non-bicycle infrastructure is classified as bicycle infrastructure, resulting in errors of omission and commission, respectively (see Figure 1c). For example, unprotected bicycle infrastructure might be misclassified as protected, which has implications for example for Levels of Traffic Stress (Mekuria et al., 2012). Different types of misclassification errors have been documented in previous research on the quality of OSM bicycle infrastructure data (Hochmair et al., 2015; Ferster et al., 2020).

Topology errors. Topology errors occur where network edges are not properly connected at intersections, either because of a missing node at intersections, or in situations of "undershoots" (one or more edges are too short and thus do not connect at an intersection) or "overshoots" (one or more edges are too long at an intersection, which creates small edges with dangling nodes), as shown in Figure 1d. Within OSM, this type of problem is more prominent for bicycle infrastructure than for other types of the road network (Neis et al., 2012). Undershoots are a particularly common feature in networks of designated bicycle infrastructure, which typically are fragmented. Therefore, they have a high potential to identify real infrastructure gaps and not just data quality issues (Natera Orozco et al., 2020; Vybornova et al., 2022). When working with only a subset of the OSM network, such as that of designated bicycle infrastructure, topology errors can also emerge when OSM edges leading

up to an intersection have not been tagged as bicycle infrastructure (see section *Undershoots in subsets of road networks* and Figure S2 in the Supplementary Material).

Inconsistent mapping procedures. While not technically an error, differing mapping methods make it difficult to assess data completeness. Within some approaches, bicycle lanes are digitized as road center lines, regardless of whether only one or both sides of the road have a bicycle lane. Other approaches digitize bicycle lanes as their individual geometries, in which case bicycle lanes on two sides of the same road are represented separately in the data set (see Figure 1e for an illustration). This can significantly distort measurements of network density and network length and makes it complicated to compare data sets that are based on differing mapping approaches. Other studies have approached this challenge by only including one lane in length computations (Hochmair et al., 2015) or by generalizing parallel lines to the center line (Ferster et al., 2020).

Quality assessment without ground truth data

The absence of ground truth or authoritative data sets on bicycle infrastructure is a challenge—not only for bicycle planners and researchers, but for anyone attempting to assess data quality. Without validation data, data quality assessment cannot be fully automated. Moreover, given the fragmented, low-quality nature of many actual bicycle networks, it is often impossible to distinguish in an automated way whether a poor performance of a data quality metric is due to low *data* quality or due to low quality of the *infrastructure* itself. For these reasons, BikeDNA does not issue any final verdict about the quality of the analyzed data sets. Rather, it allows for exploring different characteristics of the data which, in combination with local knowledge, can help make a better-informed assessment of the data quality and potential limitations to the data usability. As pointed out by Brovelli et al. (2017) and Ferster et al. (2020), local knowledge of both on-the-ground conditions and OSM mapping norms is crucial for verification and interpretation of quality assessments. Accordingly, BikeDNA emphasizes the relevance of local knowledge for any kind of planning or research process.

What BikeDNA does

In this section, we present the main features and outputs of BikeDNA by the example of an area in Greater Copenhagen, using a local data set from GeoDanmark (GeoDanmark 2023) as the reference data (see Figures 2a and 2b). Detailed elaborations of methodology and interpretation of results are found in the analysis notebooks. All code and results can be found on the "GeoDanmark" branch on the GitHub repository.

Global and local analysis

BikeDNA conducts analysis on two levels: global and local. In global analysis steps, aggregated metrics are computed for the entire study area. Some results, such as the total network length, are only meaningful on a global scale. For some data inconsistencies, however, we do not only want to know *whether* they occur, but also *where* they happen. Their location is particularly relevant because of the spatial heterogeneity in locations of high and low data quality that often characterize crowdsourced data (Haklay 2010). Therefore, in the local analysis steps, BikeDNA computes metrics separately for each grid cell on a customizable square grid that covers the study area.
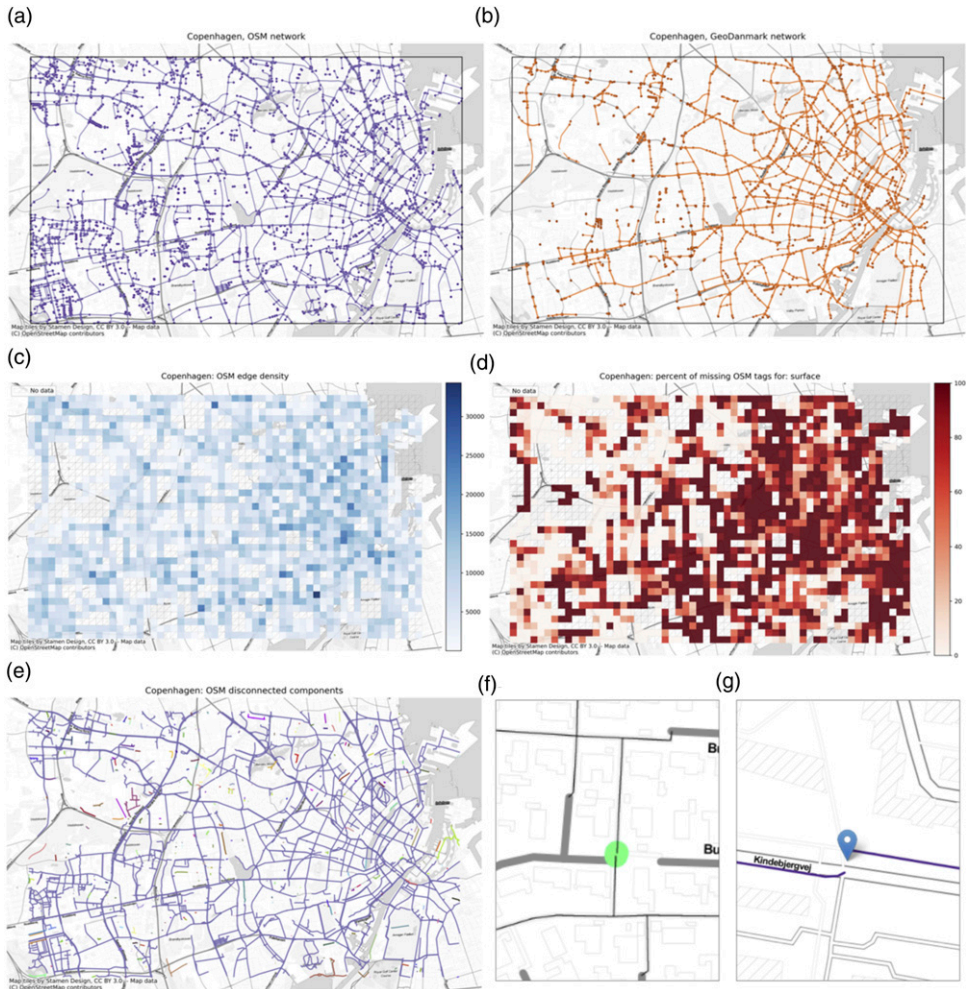
Figure 2. Input data for the showcase area from a) OSM and b) GeoDanmark. Results from intrinsic analysis of OSM and GeoDanmark data: c) OSM edge density. d) Missing OSM tags. e) Disconnected components. f) Example of an undershoot. g) Edges from two disconnected components with less than the specified distance threshold between them. Maps created with BikeDNA v.1.0.0.

Network density

The density of a transportation network is defined as the length of edges or number of nodes per square kilometer, which is the most basic descriptive statistic that can indicate data completeness. Comparing completeness between two bicycle infrastructure data sets based on network density is notoriously difficult due to differing mapping approaches and data models (see Figure 1e). OSM implements both methods for mapping bicycle infrastructure, as discussed in the section *Inconsistent mapping procedures* above. Therefore, BikeDNA computes the edge density based on the infrastructure length, *not* geometric edge length. For example, a 100-meter-long bidirectional path is counted as 200 m of bicycle infrastructure. This allows to compare data completeness between data sets with differing mapping approaches.

The analysis requires a simplified network in order not to count interstitial network nodes (i.e., nodes that do not represent intersections or dead-ends) in the computation of network density. For this reason, both OSM and reference networks are simplified using a modified OSMnx function (Sebastiao 2022) for network simplification, which keeps nodes only at intersections and dead-ends, as well as at locations where the value of important attributes changes.

Within the intrinsic analysis, network density values can indicate under- or over-mapped regions of the study area, if the density pattern strongly deviates from expected patterns (see Figure 2c). In the extrinsic analysis, contrasting network density values for OSM and the reference data set can be used for comparing data completeness (Haklay 2010).

Consistency of OSM tags

One important characteristic of OSM data on bicycle infrastructure is the highly heterogeneous distribution of tags (e.g., width, speed limits, street lights, or other characteristics of interest to cyclists), which poses a barrier to evaluations of, for example, bikeability (Wasserman et al., 2019). Likewise, the lack of restrictions on OSM tagging can lead to conflicting tags, which undermines the evaluation of bicycle conditions. For this reason, BikeDNA allows for checking the consistency of tags in OSM, tailored to OSM's data structure. BikeDNA conducts analysis of OSM tags in three ways: by identifying and visualizing where user-defined OSM tags are lacking information (see Figure 2d); by highlighting where edges are labeled with two or more tags defined as contradictory by the user; and finally, by visualizing tagging patterns, that is, the spatial variation in tags that are used to describe bicycle infrastructure in OSM. In addition, BikeDNA makes use of OSM tags to identify missing intersection nodes: when two edges intersect without having a node at the intersection and neither of them is tagged as bridge or tunnel, it is an indication of a topology error (Neis et al., 2012; Barron et al., 2014).

Network topology: Under/overshoots and dangling nodes

BikeDNA implements two methods for checking the consistency of network topology at intersections and dead-ends: analysis of undershoots and overshoots (see Figures 1 and 2f), and analysis of dangling nodes. Under/overshoots commonly occur due to errors in data digitization, but can also be an accurate representation of network conditions, for example when protected bicycle lanes end in intersections that do not provide protection for cyclists. The presence of over- and under-shoots skews the ratio of nodes and edges in a network, and thereby distorts network metric computation. Moreover, undershoots in bicycle infrastructure hinder correct routing for cyclists on the network. BikeDNA finds all over/undershoots within a user-defined distance threshold and displays them for further analysis.

Dangling nodes are nodes of degree one, that is, nodes that have only one single edge attached to them. Most infrastructure networks will naturally contain a number of dangling nodes at actual dead-ends or at the endpoints of certain features, for example, when a bicycle path ends in the middle of a road. However, they can also be a consequence of under/overshoots or data omissions. It is important to understand whether dangling nodes are caused by actual dead-ends or by digitization errors. BikeDNA therefore visualizes individual dangling nodes and their density on grid cell level to allow investigation of their spatial heterogeneity.

Network topology: Disconnected components

A network component is a maximal set of nodes which are linked by paths. In other words, all nodes within a component can reach each other, but they cannot reach any nodes in the rest of the network.

Most real-world bicycle infrastructure networks consist of many disconnected components (Natera Orozco et al., 2020) (see Figure 2e). Two disconnected components that are very close to each other can, however, be a sign of a real "missing link" (Vybornova et al., 2022) or of a digitizing error similar to an undershoot (see Figure 2g). BikeDNA identifies and visualizes all disconnected components of the input network and plots the distribution of all network component lengths on a Zipf plot, which ranks the lengths of all components by descending order. 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 exponential distributions (Clauset et al., 2009). This can mean that there has been no consolidation of the network and elements have been added only piecewise or randomly (Szell et al., 2022). However, it can also happen that the largest connected component (see the leftmost marker in Figure 3a at rank $10^0 = 1$) is a clear outlier, while the rest of the plot follows a different shape. This can mean that network consolidation has taken place. In case of a comparison over the same region, as shown in Figure 3, if one data set shows a clear outlier in its largest connected component while the other data set does not, it can be an indication that the first data set is more complete. In this particular case, the OSM data set (Figure 3a) is likely more complete than the reference data (Figure 3b).

Two disconnected components of bicycle infrastructure might of course be connected by other types of infrastructure, if the entire road network is considered. However, erroneous lack of connections between components of dedicated bicycle infrastructure poses a problem for bicycle routing and will lead to misleading or undesirable results when evaluating bicycle accessibility and Level of Traffic Stress (Murphy and Owen 2019; Wasserman et al., 2019). BikeDNA therefore also identifies potential missing connections between components by finding and highlighting edges from disconnected components that are within a user-defined distance threshold from each other. BikeDNA furthermore conducts a component connectivity analysis on grid cell level, visualizing differences in number of cells reachable from each cell. This measure is particularly useful for comparing and quantifying network reach and connectivity between two data sets (see Figures 3c and 3d).

Feature matching

Feature matching is the process of identifying which features from two different data sets correspond to the same real-life object. For example, the same road might be represented by slightly different geometries in two different data sets. Feature matching is necessary for a comparison of *individual features*, rather than feature densities. BikeDNA includes a feature matching algorithm which identifies corresponding edge segments between the reference and OSM data sets (see Figures 3g and 3h). The matching algorithm uses the undirected Hausdorff distance and angle between line segments to identify the best potential match from edges within a maximum search distance (Koukoletsos et al., 2012; Will 2014). For further details on the matching procedure, see section *Feature matching* and Figure S4 in the Supplementary Material.

Comparing OSM and reference data

If a reference data set is provided, BikeDNA compares and contrasts it with OSM data from the same area, highlighting how and where the two data sets differ, that is, both *how much* bicycle infrastructure is mapped in the two data sets and *how* the infrastructure is mapped. The comparative analysis makes no prior assumptions about which data set is of higher quality. BikeDNA does thus not lead to an automatic conclusion, but instead requires interpretation of the differences, for example, whether differing features are results of errors of omission or commission, and which data
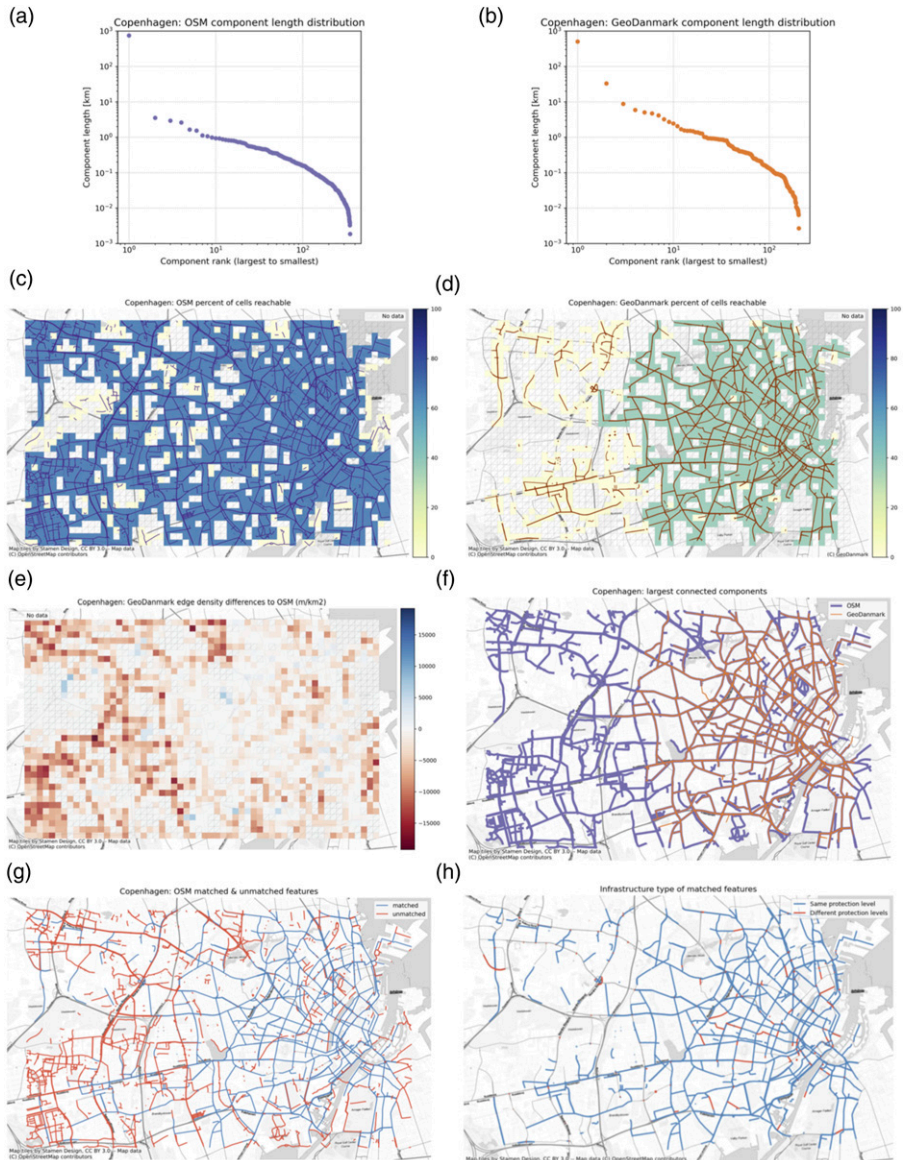
Figure 3. Extrinsic analysis based on a comparison of intrinsic results: Zipf plots of a) OSM and b) GeoDanmark component length distribution. Percent of cells reachable through c) the OSM and d) the reference networks. Extrinsic analysis of differences between OSM and reference data: e) Comparison of edge density per grid cell. f) Largest connected components. Feature matching: g) Matched (blue) and unmatched (red). h) Matched OSM features with same protection level in both data sets (blue) and differing protection levels (red). Maps created with BikeDNA v.1.0.0.

set is more fit for purpose. The differences between the data sets are computed and presented on a global and local level, and overlaid where visually adequate (see Figures 3e and 3f).

Discussion

In the following section, we sum up BikeDNA's findings for our showcase area in Greater Copenhagen, and present the general contributions of BikeDNA. We end the section with a discussion of BikeDNA's limitations and recommendations for future work.

Substantial problems with data quality persist

The results from the showcase area demonstrate that data quality issues can result in misleading representation of actual bicycle conditions, both in terms of the extent and the connectivity of the bicycle network. Despite of Denmark being known for its strong bicycle culture (Agervig and Ebert 2012), our analysis found several serious issues with bicycle infrastructure data quality, particularly in the administrative reference data set, with large areas being significantly under-mapped. Both OSM and the reference data were moreover suffering from missing links and undershoots resulting in data sets which are substantially more fragmented than the actual bicycle network. Importantly, the analysis indicated large spatial variations in missing or misleading data, which highlights the importance of localized data assessments.

BikeDNA: A tool for bicycle researchers, planners, and cyclists

BikeDNA provides fundamentally new insights into bicycle infrastructure data quality by enabling straightforward exploration of spatial heterogeneity in data quality, and by implementing network-based measures that go beyond quality indicators traditionally applied on bicycle infrastructure data. These novel features open up a plethora of use cases, such as:

- Urban and regional planning of bicycle infrastructure, where considering the network as a whole is particularly important (Szell et al., 2022).
- Improvement of OSM and administrative data.
- Transport research on active mobility and on multi-modal networks which require high data quality on all transport layers (Alessandretti et al., 2022).
- Improvements to bicycle network data used in tools for transport planning and research, such as "Propensity to Cycle" for transport planning (Lovelace et al., 2017), "A/B Street" for traffic simulation (Carlino et al., 2023), and People for Bike's Bicycle Network Analysis (PeopleForBikes 2023).
- Citizen science, by enabling citizens' contributions to reliable and quality-assessed data sets, along the lines of other projects supporting citizen data collection for more sustainable transport, such as OpenBikeSensor (2023) and StreetComplete (2023).

Ensuring high-quality data is particularly vital considering the recent growth in bicycle network research. Existing data should be scrutinized prior to their usage, to make sure that research results are not undermined by low data quality.

Limitations

Although we attempted to cover the main aspects of data quality relevant to bicycle networks, there are some limitations in the design of BikeDNA.

In terms of data modeling, BikeDNA makes use of an undirected, simplified network. This means that information about allowed travel directions and turn restrictions is not considered, movement is assumed in both directions on all edges, and different travel directions on the same road are not represented by separate edges. Therefore, the current state of the tool does not make use of routing on the network. For future iterations, it might be useful to include travel directions and the underlying road network for accurate path computations.

Another limitation touches upon the core purpose of the tool and the type of results it can produce: since we do not operate with one data set as ground truth against which another data set is evaluated, we cannot conclude where the error lies when differences are identified. For a successful application of BikeDNA, the user is thus expected to have some familiarity with OSM data structure and tagging conventions, but also to have enough knowledge of the study area to correctly interpret the results.

Lastly, we do not directly evaluate the positional accuracy of neither the OSM or the reference data—although a certain level of comparative positional accuracy can be deduced from the feature matching.

Future work

A lot of potential future work remains—not only for quality assessments of bicycle infrastructure data generally, but also for BikeDNA.

First, it is yet to be determined whether any of the analyzed metrics can serve as a more general predictor of data quality, based on correlations with other quality metrics. This would allow for a much faster and simpler assessment and remove the need for reference data comparison and more complex analytical tools. Second, the evaluation of bicycle infrastructure data quality in OSM and other data sources should be considered in connection with broader efforts to standardize the mapping of bicycle infrastructure. One challenge of comparing different data sets is translating between different typologies of bicycle infrastructure. Within BikeDNA, we have solved this ambiguity by only distinguishing between protected and unprotected bicycle infrastructure, but this approach could be refined with better classification schemes. To support better planning and research, more work is needed on how to best represent bicycle conditions in data.

For BikeDNA, a functionality for allowing an easy comparison of two non-OSM data sets remains to be implemented. In addition, for the intrinsic analysis of OSM data, a future inclusion of historical data on contributors and edits can offer a complementary perspective, since these metadata are potential indicators of data quality (Neis et al., 2012; Gröchenig et al., 2014). A future iteration of this work could aim at adjusting the BikeDNA pipeline to other types of networks, for example, pedestrian infrastructure or public transit. Likewise, BikeDNA's implementation of feature matching could be enhanced to produce a standalone tool for other network types. Lastly, BikeDNA could be re-implemented as a more user-friendly, interactive web app, to overcome the required Python notebook know-how. We hope to hear from the community about future case studies and will be grateful to receive suggestions for possible modifications and improvements to the current tool.

Conclusion

BikeDNA is an open-source tool for reproducible quality assessments of bicycle infrastructure data. BikeDNA allows for a customized evaluation based on the idea of "fitness of purpose" and comes with a configurable design that can be adapted to different scenarios of data availability, using both stand-alone and comparative methods. The tool computes a wide range of quality metrics and extends previous research on bicycle infrastructure quality by adding a focus on data topology and

structural properties. The provided example application of BikeDNA based on a comparison of OSM and Danish administrative data has demonstrated that BikeDNA can reveal important spatial variations, for example, in missing data and topology errors, and that quality assessments of bicycle infrastructure data are necessary, as demonstrated by the many inconsistencies and errors revealed by the example analysis. Some open questions remain, both for BikeDNA and more broadly, regarding the definition and collection of high-quality data on bicycle infrastructure. The absence of high-quality official data and recognized typologies for bicycle infrastructure reflects how active mobility modes have been historically neglected. For cities, regions, and countries around the world that strive to make use of recent advances in data-driven planning to expand their bicycle networks, better data on existing bicycle conditions are urgently needed.

Acknowledgments

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

ORCID iDs

Ane Rahbek Viero ⓘ https://orcid.org/0000-0002-2404-7271
Michael Szell ⓘ https://orcid.org/0000-0003-3022-2483
Anastassi Vybornova ⓘ https://orcid.org/0000-0001-6915-2561

Supplemental Material

Supplemental material for this article is available online.

References

Agervig CT and Ebert AK (2012) Cycling cultures in northern europe: from 'golden age' to 'renaissance. In: Parkin J (ed), *Cycling and Sustainability, Transport and Sustainability*. Bingley, UK: Emerald Group Publishing Limited, volume 1, pp. 23–58. ISBN. URL. DOI: 10.1108/S2044-9941(2012)0000001004

Alessandretti L, Natera Orozco LG, Saberi M, et al. (2022) Multimodal urban mobility and multilayer transport networks. In: *Environment and Planning B: Urban Analytics and City Science*. DOI: 10.1177/23998083221108190

Ballatore A and Zipf A (2015) A conceptual quality framework for volunteered geographic information. In: Fabrikant SI, Raubal M, Bertolotto M, et al. (eds), *Spatial Information Theory*. Cham: Springer International Publishing, volume 9368, pp. 89–107. ISBN. URL. DOI: 10.1007/978-3-319-23374-1_5 Series Title: Lecture Notes in Computer Science

Barron C, Neis P and Zipf A (2014) A comprehensive framework for intrinsic openstreetmap quality analysis: a comprehensive framework for intrinsic openstreetmap quality analysis. *Transactions in GIS* 18(6): 877–895. URL. DOI: 10.1111/tgis.12073

Boeing G (2017) OSMnx: new methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems* 65: 126–139. URL. DOI: 10.1016/j.compenvurbsys.2017.05.004

Brovelli MA, Minghini M, Molinari M, et al. (2017) Towards an automated comparison of openstreetmap with authoritative road datasets. *Transactions in GIS* 21(2): 191–206. URL. DOI: 10.1111/tgis.12182

C40 Cities (2019) Upgrade of the cycle network in bogotà dramatically increases bike trips. URL https://www.c40.org/case-studies/upgrade-of-the-cycle-network-in-bogota-dramatically-increases-bike-trips/

Carlino D, Li Y and Kirk M (2023) A/B Street. URL https://github.com/a-b-street/abstreet. Original-date: 2018-06-04T00

City Of Paris (2021) Un nouveau plan vélo pour une ville 100 % cyclable. URL https://www.paris.fr/pages/un-nouveau-plan-velo-pour-une-ville-100-cyclable-19554

Clauset A, Shalizi CR and Newman MEJ (2009) Power-law distributions in empirical data. *SIAM Review* 51(4): 661–703. URL. DOI: 10.1137/070710111

Cox P (2020) Theorising infrastructure: a politics of spaces and edges. In: *The Politics of Cycling Infrastructure* Bristol, UK: Policy Press. ISBN, pp. 15–34. URL https://bristoluniversitypressdigital.com/display/book/9781447345169/ch001.xml. Section: The Politics of Cycling Infrastructure

Degrossi LC, Porto de Albuquerque J, Santos Rocha Rd, et al. (2018) A taxonomy of quality assessment methods for volunteered and crowdsourced geographic information. *Transactions in GIS* 22(2): 542–560. URL. DOI: 10.1111/tgis.12329

Dill J (2004) Measuring network connectivity for bicycling and walking. In: *83rd Annual Meeting of the Transportation Research Board*. Washington, DC: Transportation Research Board, volume 83, p. 19.

EC (2021) The new eu urban mobility framework. technical report, european commission, brussels. URL https://transport.ec.europa.eu/system/files/2021-12/com_2021_811_the-new-eu-urban-mobility.pdf

Ferster C, Fischer J, Manaugh K, et al. (2020) Using openstreetmap to inventory bicycle infrastructure: a comparison with open data from cities. *International Journal of Sustainable Transportation* 14(1): 64–73. URL. DOI: 10.1080/15568318.2018.1519746

Fleischmann M (2019) momepy: urban morphology measuring toolkit. *Journal of Open Source Software* 4(43): 1807. URL. DOI: 10.21105/joss.01807 https://joss.theoj.org/papers/10.21105/joss.01807

Fonte C (2017) Assessing VGI Data Quality. In: *Mapping and the Citizen Sensor*: 137–163. URL. DOI: 10.5334/bbf https://www.ubiquitypress.com/site/chapters/10.5334/bbf.g/

GeoDanmark (2023) Danmarks geografi - geodanmark. URL https://dataforsyningen.dk/data/3563

Girres JF and Touya G (2010) Quality assessment of the french openstreetmap dataset: quality assessment of the french openstreetmap dataset. *Transactions in GIS* 14(4): 435–459. URL. DOI: 10.1111/j.1467-9671.2010.01203.x

Graser A, Straub M and Dragaschnig M (2014) Towards an open source analysis toolbox for street network comparison: indicators, tools and results of a comparison of osm and the official austrian reference graph. *Transactions in GIS* 18(4): pp. 510–526. Wiley-Blackwell, URL. DOI: 10.1111/tgis.12061 https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=97351436&site=ehost-live. Publisher

Gröchenig S, Brunauer R and Rehrl K (2014) Estimating completeness of vgi datasets by analyzing community activity over time periods. In: Huerta J, Schade S and Granell C (eds), *Connecting a Digital Europe Through Location and Place, Lecture Notes in Geoinformation and Cartography*. Cham: Springer International Publishing, pp. 3–18. ISBN. URL. DOI: 10.1007/978-3-319-03611-3_1

Hagberg AA, Schult DA and Swart PJ (2008) Exploring network structure, dynamics, and function using networkX. In: Varoquaux G, Vaught T and Millman J (eds). *Proceedings of the 7th Python in Science Conference*. Pasadena, CA USA: 7th Python in Science Conference, 11–15.

Haklay M (2010) How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and Planning B: Planning and Design* 37(4): 682–703. URL. DOI: 10.1068/b35097

Haklay MM, Basiouka S, Antoniou V, et al. (2010) How many volunteers does it take to map an area well? The validity of linus' law to volunteered geographic information. *The Cartographic Journal* 47(4): 315–322. URL. DOI: 10.1179/000870410X12911304958827

Hashemi P and Abbaspour RA (2015) Assessment of logical consistency in openstreetmap based on the spatial similarity concept. In: Jokar Arsanjani J, Zipf A, Mooney P, et al (eds), *OpenStreetMap in GIScience: Experiences, Research, and Applications, Lecture Notes in Geoinformation and Cartography*. Cham: Springer International Publishing, pp. 19–36. ISBN. URL. DOI: 10.1007/978-3-319-14280-7_2

Hochmair HH, Zielstra D and Neis P (2015) Assessing the completeness of bicycle trail and lane features in openstreetmap for the United States: completeness of bicycle features in openstreetmap. *Transactions in GIS* 19(1): 63–81. URL. DOI: 10.1111/tgis.12081

Hvingel L and Jensen T (2023) Gode cykeldata til alle. Trafik og veje URL https://www.kl.dk/media/53604/artikel_gode_cykeldata_trafik_og_veje_jan_2023_layout.pdf

Jaramillo P, Kahn Ribeiro S, Newman P et al (2022) *Transport Climate Change 2022: Mitigation of Climate Change*. Cambridge, UK: Intergovernmental Panel on Climate Change (IPCC), pp. 1049–1160. URL doi:. DOI: 10.1017/9781009157926.012

Jordahl K, Bossche JVd, Fleischmann M, et al (2021) geopandas/geopandas: v0.10.2. URL. DOI:10.5281/zenodo.5573592. https://zenodo.org/record/5573592

Keßler C, Trame J and Kauppinen T (2011) Tracking editing processes in volunteered geographic information: the case of openstreetmap In: *Proceedings of the Conference on Spatial Information Theory, Workshop: Identifying Objects, Processes and Events in Spatio-Temporally Distributed Data*. 7.

Koukoletsos T, Haklay M and Ellul C (2012) Assessing data completeness of vgi through an automated matching procedure for linear data. *Transactions in GIS* 16(4): 477–498. URL. DOI: 10.1111/j.1467-9671.2012.01304.x

Lovelace R, Goodman A, Aldred R, et al. (2017) The propensity to cycle tool: an open source online system for sustainable transport planning. *Journal of Transport and Land Use* 10(1). URL. DOI: 10.5198/jtlu.2016.862 https://www.jtlu.org/index.php/jtlu/article/view/862. Number: 1

Medeiros G and Holanda M (2019) Solutions for data quality in GIS and VGI: A Systematic Literature Review. In: Rocha A, Adeli H, Reis LP, et al. (eds), *New Knowledge in Information Systems and Technologies, Advances in Intelligent Systems and Computing*. Cham: Springer International Publishing, pp. 645–654. ISBN. DOI: 10.1007/978-3-030-16181-1_61.

Mekuria MC, Furth PG and Nixon H (2012) Low-stress bicycling and network connectivity. Technical Report 11-19, Mineta Transportation Institute. URL https://www.semanticscholar.org/paper/Low-Stress-Bicycling-and-Network-Connectivity-Mekuria-Furth/a50063c06112d3eb6aa752dfd362e1bdbc7f1c7e

Mondzech J and Sester M (2011) Quality analysis of openstreetmap data based on application needs. *Cartographica: The International Journal for Geographic Information and Geovisualization* 46(2): pp. 115–125. URL. University of Toronto Press. DOI: 10.3138/carto.46.2.115

Murphy B and Owen A (2019) Implementing low-stress bicycle routing in national accessibility evaluation. *Transportation Research Record* 2673(5): pp. 240–249. URL. Sage Publications Inc. DOI: 10.1177/0361198119837179

Natera Orozco LG, Battiston F, Iñiguez G, et al (2020) Data-driven strategies for optimal bicycle network growth. *Royal Society Open Science* 7(12): 201130. Royal Society. URL. DOI: 10.1098/rsos.201130

Neis P, Zielstra D and Zipf A (2012) The street network evolution of crowdsourced maps: openstreetmap in Germany 2007–2011. *Future Internet* 4(1) pp. 1–21. URL. Molecular Diversity Preservation International. DOI: 10.3390/fi4010001

Neis P, Zielstra D and Zipf A (2013) Comparison of volunteered geographic information data contributions and community development for selected world regions. *Future Internet* 5(2): pp. 282–300. URL. Multidisciplinary Digital Publishing Institute. DOI: 10.3390/fi5020282 https://www.mdpi.com/1999-5903/5/2/282. Number: 2 Publisher

Nelson T, Ferster C, Laberee K, et al. (2021) Crowdsourced data for bicycling research and practice. *Transport Reviews* 41(1): 97–114. URL. DOI: 10.1080/01441647.2020.1806943

Olmos LE, Tadeo MS, Vlachogiannis D, et al. (2020) A data science framework for planning the growth of bicycle infrastructures. *Transportation Research Part C: Emerging Technologies* 115: 102640. URL. DOI: 10.1016/j.trc.2020.102640 https://www.sciencedirect.com/science/article/pii/S0968090X19306436

OpenBikeSensor (2023) OpenBikeSensor. URL https://openbikesensor.org/en/

PeopleForBikes (2023) BNA bicycle network analysis. URL https://bna.peopleforbikes.org/#/

Rambøll (2022) Walking and cycling data. Practice, challenges, needs and gaps. URL https://ramboll.com/-/media/files/rgr/documents/markets/transport/walking-cycling-data-gaps-2022.pdf

Sebastiao C (2022) Modified version of OSMnx Simplification. GitHub URL https://github.com/anerv/NERDS_osmnx

Senaratne H, Mobasheri A, Ali AL, et al. (2017) A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science* 31(1): 139–167. URL. DOI: 10.1080/13658816.2016.1189556

Steinacker C, Storch DM, Timme M et al (2022) Demand-driven design of bicycle infrastructure networks for improved urban bikeability. *Nature Computational Science* 1: 655-664. URL. Nature Publishing GroupPublisher. DOI: 10.1038/s43588-022-00318-w https://www.nature.com/articles/s43588-022-00318-w.

StreetComplete (2023) StreetComplete. URL https://streetcomplete.app

Szell M, Mimar S, Perlman T, et al. (2022) Growing urban bicycle networks. *Scientific Reports* 12(1), p. 6765. URL. Nature Publishing Group. DOI: 10.1038/s41598-022-10783-y https://www.nature.com/articles/s41598-022-10783-y. Number: 1 Publisher

Vybornova A, Cunha T, Gühnemann A, et al. (2022) Automated detection of missing links in bicycle networks. Geographical Analysis n/a(n/a). DOI:10.1111/gean.12324. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/gean.12324

Wasserman D, Rixey A, Zhou XE, et al. (2019) Evaluating openstreetmap's performance potential for level of traffic stress analysis. *Transportation Research Record* 2673(4): pp. 284–294. URL. Sage Publications Inc. DOI: 10.1177/0361198119836772

Will J (2014) Development of an Automated Matching Algorithm to Assess the Quality of the OpenStreetMap Road Network: A Case Study in Göteborg, Sweden. PhD Thesis. Lund: Lund University. URL https://www.semanticscholar.org/paper/Development-of-an-automated-matching-algorithm-to-%3A-Will/b3b77d579077b967820630db56522bef31654f21

Zhang X, Wang T, Jiao D, et al. (2021) Detecting inconsistent information in crowd-sourced street networks based on parallel carriageways identification and the rule of symmetry. *ISPRS Journal of Photogrammetry and Remote Sensing* 175: 386–402. URL. DOI: 10.1016/j.isprsjprs.2021.03.014 https://www.sciencedirect.com/science/article/pii/S092427162100085X

Zielstra D, Hochmair H and Neis P (2013) Assessing the effect of data imports on the completeness of openstreetmap – a united states case study. *Transactions in GIS* 17: 315–334. DOI: 10.1111/tgis.12037.

Zielstra D and Hochmair HH (2012) Using free and proprietary data to compare shortest-path lengths for effective pedestrian routing in street networks. *Transportation Research Record* 2299(1): pp. 41–47. URL. Sage Publications Inc. DOI: 10.3141/2299

Ziemke D, Metzler S and Nagel K (2019) Bicycle traffic and its interaction with motorized traffic in an agent-based transport simulation framework. *Future Generation Computer Systems* 97: 30–40. URL. DOI: 10.1016/j.future.2018.11.005 https://www.sciencedirect.com/science/article/pii/S0167739X17320447