TECHNISCHE UNIVERSITÄT DRESDEN

Angela Francke • Sven Lißner

Big Data in Bicycle Traffic

A user-oriented guide to the use of smartphone-generated bicycle traffic data

Publishing details

Big Data in Bicycle Traffic
A user-oriented guide to the
use of smartphone-generated
bicycle traffic data

Dresden,
November 2017

Published by: Technische Universität Dresden, Chair of Transport Ecology and
Chair of Traffic and Transportation Psychology

Authors: Dipl.-Ing. Sven Lißner,
Dipl.-Verk.wirtsch. Angela Francke

Contact: verkehrsoekologie@tu-dresden.de

Layout: Lisa-Marie Schaefer

Translation: Helen Grützner
Ian McGarry
Till Becker

Supported by:

Federal Ministry
of Transport and
Digital Infrastructure

on the basis of a decision
by the German Bundestag

Prefatory remarks
by the Federal Ministry of Transport and Digital Infrastructure

The present guide builds on two extremely positive trends:

1. Cycling is fun, healthy, easy on the wallet and environmentally friendly. Moreover, the bicycle is an extremely flexible and efficient means of transport. Cycling thus enhances not only the quality of life of cyclists themselves, but also the quality of life of a city as a whole.
2. The onward march of the digital revolution is opening up new possibilities for data collection and analysis, new participative data contents and the reorganization of existing processes by means of digital data.

Why not combine these two positive trends and use the obvious advantages of digitalization for local authority cycling planning?

Good cycling infrastructure – a prerequisite for a high share of cycling in the modal split – has to be exactly where cyclists need it. It is amazing, but it is a fact, that so far, many local authority transport planners have not had any robust data on the precise routes taken by cyclists. This is where digitalization can help.

In the research project entitled „Smartphone-generated behavioural data in cycling“, data of cycling movements were thoroughly examined and used as a model for cycling planning in Dresden. The findings confirmed that the movement data not only provide an interesting insight into the local cycling flows but can also help to identify the needs and expectations of cyclists and incorporate them into transport planning.

The present guide provides local authority planners and interested members of the public with an easy-to-use practical introduction to the topic of GPS data, and in doing so identifies potential for cycling planning and possible stumbling blocks in the use of these data. It encourages the use of digitalization in this sector as well and addresses key issues such as data protection or the challenges involved in interpreting the data.

Supported by:

Federal Ministry
of Transport and
Digital Infrastructure

on the basis of a decision by the German Bundestag

Abstract

For cycling to be attractive, the infrastructure must be of high quality. Due to the high level of resources required to record it locally, the available data on the volume of cycling traffic has to date been patchy. At the moment, the most reliable and usable numbers seem to be derived from permanently installed automatic cycling traffic counters, already used by many local authorities. One disadvantage of these is that the number of data collection points is generally far too low to cover the entirety of a city or other municipality in a way that achieves truly meaningful results. The effect of side roads on cycling traffic is therefore only incompletely assessed. Furthermore, there is usually no data at all on other parameters, such as waiting times, route choices and cyclists' speed. This gap might in future be filled by methods such as GPS route data, as is now possible by today's widespread use of smartphones and the relevant tracking apps. The results of the project presented in this guide have been supported by the BMVI [Federal Ministry of Transport and Digital Infrastructure] within the framework of its 2020 National Cycling Plan. This research project seeks to investigate the usability of user data generated using a smartphone app for bicycle traffic planning by local authorities.

In summary, it can be stated that, taking into account the factors described in this guide, GPS data are usable for bicycle traffic planning within certain limitations. (The GPS data evaluated in this case were provided by Strava Inc.) Nowadays it is already possible to assess where, when and how cyclists are moving around across the entire network. The data generated by the smartphone app could be most useful to local authorities as a supplement to existing permanent traffic counters. However, there are a few aspects that need to be considered when evaluating and interpreting the data, such as the rather fitness-oriented context of the routes surveyed in the examples examined. Moreover, some of the data is still provided as database or GIS files, although some online templates that are easier to use are being set up, and some can already be used in a basic initial form. This means that evaluation and interpretation still require specialist expertise as well as human resources. However, the need for these is expected to reduce in the future with the further development of web interfaces and supporting evaluation templates. For this to work, developers need to collaborate with local authorities to work out what parameters are needed as well as the most suitable formats. This research project carried out an approach to extrapolating cycling traffic volumes from random samples of GPS data over the whole network. This was also successfully verified in another municipality. Further research is still nevertheless required in the future, as well as adaptation to the needs of different localities.

Evidence for the usability of GPS data in practice still needs to be acquired in the near future. The cities of Dresden, Leipzig and Mainz could be taken as examples for this, as they have all already taken their first steps in the use of GPS data in planning for and supporting cycling. These steps make sense in the light of the increasing digitisation of traffic and transport and the growing amount of data available as a result – despite the limitations on these data to date – so that administrative bodies can start early in building up the appropriate skills among their staff. The use of GPS data would yield benefits for bicycle traffic planning in the long run. In addition, the active involvement of cyclists opens up new possibilities in communication and citizen participation – even without requiring specialist knowledge. This guide delivers a practical introduction to the topic, giving a comprehensive overview of the opportunities, obstacles and potential offered by GPS data.

Index

1 Introduction

In the future, new, smart technology will be indispensable to traffic planning in an ever more digitised world. Transport research stands on the cusp of a shift towards more citizen participation via (automated) capturing of transport habits. This guide provides an introduction to the topic of GPS data, which could open the way to new, digital data collection methods for bicycle traffic planning.

1.1 For whom is this guide intended?

In the project 'Smartphone-generated behavioural data in cycling traffic', TU Dresden's traffic ecology and traffic psychology research groups have been working together to assess the usefulness of voluntarily generated cycling behavioural data (GPS data) in general cycling planning.

Target group

This guide provides practical instruction for the use of GPS data in bicycle traffic planning. No matter whether you have already worked with GPS data or not, this guide can be used either to get into the topic or as a practical toolbox for quality assessment. More detailed project results can be found in the project report, which is available for download from the NRVP [National Cycling Plan] website*.

Structure of the guide

The chapters of this guide are independent from each other. The use of only specific chapters is therefore possible. The guide includes the whole process of GPS-data-based cycling traffic planning, starting with a theoretical overview of the topic and sources of GPS-data. Chapter 2 introduces characteristics, possibilities and limits of GPS-data and chapter 3 covers application scenarios. A comprehensive example of an application in chapter 4 makes it possible to trace a planning process, to transfer that knowledge to the own community and to develop an idea of how to use the data for own purposes. At last, criteria for the usability and interpretation of the data help to prevent wrong conclusions and to use the full potential of the data (Chapter 5). A comprehensive detailed knowledge part, a glossary (keywords are marked in dark blue) and a set of frequently asked questions and their answers (FAQ) in the appendix enable efficient working with the guide.

1.2 Overview: GPS data in bicycle traffic

Data in bicycle traffic

For motorised private transport, valuable data from various sources is already available. Conventional data in bicycle traffic planning, however, is based on complex and active data collection by counting devices, traffic observations and surveys. The patterns of movement of the road users therefore remain only snapshots that can seldom be connected to one another. Growing digital communications and the increasing use of GPS-enabled devices now make it possible to look at patterns of movement of road users in an aggregated manner and to derive demand-oriented bicycle traffic planning.

Within the scope of this project, working with GPS data has initially demonstrated positive results in the following areas:

- ✓ *Trend for smartphone-based transport control by the users*
- ✓ *Availability of the data*
- ✓ *Applicable as a basis for different planning purposes*
- ✓ *Scientific examination of usability yields positive results*

* https://nationaler-radverkehrsplan.de/de/praxis/mit-smartphones-generierte-verhaltensdaten-im

Smartphone usage

The number of smartphone users has been growing steadily for years. In 2016, 49 million Germans already owned a smartphone, according to Destatis [Federal Statistical Office]. Nearly all of these smartphones come equipped with GPS functionality, often used for navigation connected to map applications. Over the past few years, smartphone applications for cycling, which use and evaluate these functions, have also entered the market. They offer sport-oriented users in particular the chance to track and save their training routes, movement patterns and speed. Users gain a comprehensive overview of their performance and are able to compare their results with others. The number of users of these apps is also increasing and most app operators have transferred the rights of use to themselves from the users. They can therefore offer the collected data in anonymised form on the open market. These data sets can offer interesting insights for local authorities and others. An overview of the purchase of data can be found in section 2.3.

Data availablity

Data providers and third-party providers are increasingly developing customer-oriented services in terms of data preparation and analysis. In our experience a detailed profile of requirements and a GIS road network of the city are sufficient for the vendor to provide data analysis according to individual requests. Therefore the focus of this guide rests primarily on the ordering process and use of GPS data in bicycle traffic planning and not on the technical analysis of raw data. How the data can be acquired is covered in Chapter 2.

Advantages of GPS data

The basic features of user-generated GPS data have several significant advantages over conventional data sources, such as traffic counts and surveys. Especially in the parameters traffic volume, origin-destination matrices and speed the data is of good quality and is therefore usable in many central areas of demand-driven bicycle traffic planning. Chapter 3 covers the application areas network planning, facilities planning, prioritising measures as well as evaluation and monitoring and explains the corresponding parameters with regard to GPS data. Finally, ideas and examples of the public use of GPS data-based planning in terms of citizen participation and public relations are introduced.

Potential of data usage

Within the framework of the research project, purchasable GPS data from a (sport-oriented) app for smartphones from Strava Inc. were used as the basis of a validation process. The data was compared with empirical traffic data (counting devices, speed measurement) in the pilot municipality of Dresden. In the course of the project, not only were the data examined for their generalisability in terms of user structure, movement patterns and network coverage, but also practical suggestions for applications in bicycle traffic planning were formulated and integrated into this guide (Chapter 4). After a comprehensive scientific evaluation a general recommendation for GPS-based bicycle traffic planning can be made, provided that the data are interpreted in a deliberate and careful manner. Chapter 5 covers recommendations on that.

Planning practice with GPS data

This list offers an overview of the opportunities and challenges arising during work with GPS data, which are given in more detail in Chapter 2. A detailed breakdown of the individual points and case studies can be found in Chapter 4.

Opportunities

Data acquisition

- Data is generated in real time
- Filterable according to different criteria
- Manageable costs
- Replacement of high-maintenance automatic counting devices
- Acquisition of socio-demographic user data

Data structure and fields of application

- Quick overview of number of cyclists in main and side roads
- Before and after evaluation of measures
- Mapping the traffic volume of cyclists in the whole city
- Origin-destination matrices between polygons
- Computing the waiting times at intersections
- Average speed of cyclists at links in the road network
- Routing of tracks within wider corridors

Challenges

- Systematic bias in the data by focussing on app users
- People without a suitable smartphone are systematically excluded from the data sample
- Scepticism concerning data acquisition in terms of privacy protection
- Limitations to data usage due to legal restrictions, particularly those concerning privacy protection

What are GPS data? 2

The term 'GPS data' covers global, satellite-based positioning data. Their approval in the year 2000 meant that navigation systems in particular have achieved great significance. With the rising number of GPS receivers, as well as simple recording software, these data can deliver a precise picture of people's daily movements.

2.1 General

Possibilities for use

A growing proportion of people own a smartphone and use it to help them in their daily transport needs. The possibility of creating a picture of people's movements both individually and aggregated over a large group, is set to gain importance in future transport planning. Insights from transport surveys such as SrV and MiD can be cheaply supplemented on a yearly basis with such voluntarily compiled GPS data from a range of app vendors. In this guide, for the sake of readability, smartphone-generated GPS records will simply be designated as 'GPS data' from now on.

Background

Due to the willingness of users to expose their data, providers are starting to see marketable models for data sale and analysis. This development requires a critical scientific support and a validation of the data, to enable the development of a high quality basis for planning. Within an extensive research project, the data offered by the company Strava Inc. (at the point of project start the broadest data provider in the market, at this point 180 cities and municipalities worldwide use the data, status 08/2017) was exemplarily examined for their possibilities of usage.

Forms of data

What form do GPS data take?

GPS data are often recorded during sports activities for the purposes of training, or come as a side-effect of a navigation process. However, people have myriad other reasons for recording their movements, resulting in a wide range of users. The routes that are recorded are collected, anonymised, summarised and then, with the collaboration of the client (for instance a parish or local authority) projected into an agreed map. This means that the finished product is not usually the route data themselves, but a very detailed mapping of traffic volumes. These maps of bicycle traffic volumes on road segments appear quite similar to those created for motorised traffic (see Figure 1).

The data offered consists of longitude, latitude and a timestamp, usually at a frequency of 1 Hertz. Lower frequencies are possible and are determined by the different mobile devices, but are not necessarily useful for planning purposes. The timestamped locations can then be used to derive distances or speeds using simple calculations. For example, Strava differentiates between:

» Link data: Referencing traffic volumes on GIS element (link)
» Node data: Referencing traffic volumes and waiting times on GIS element (node)

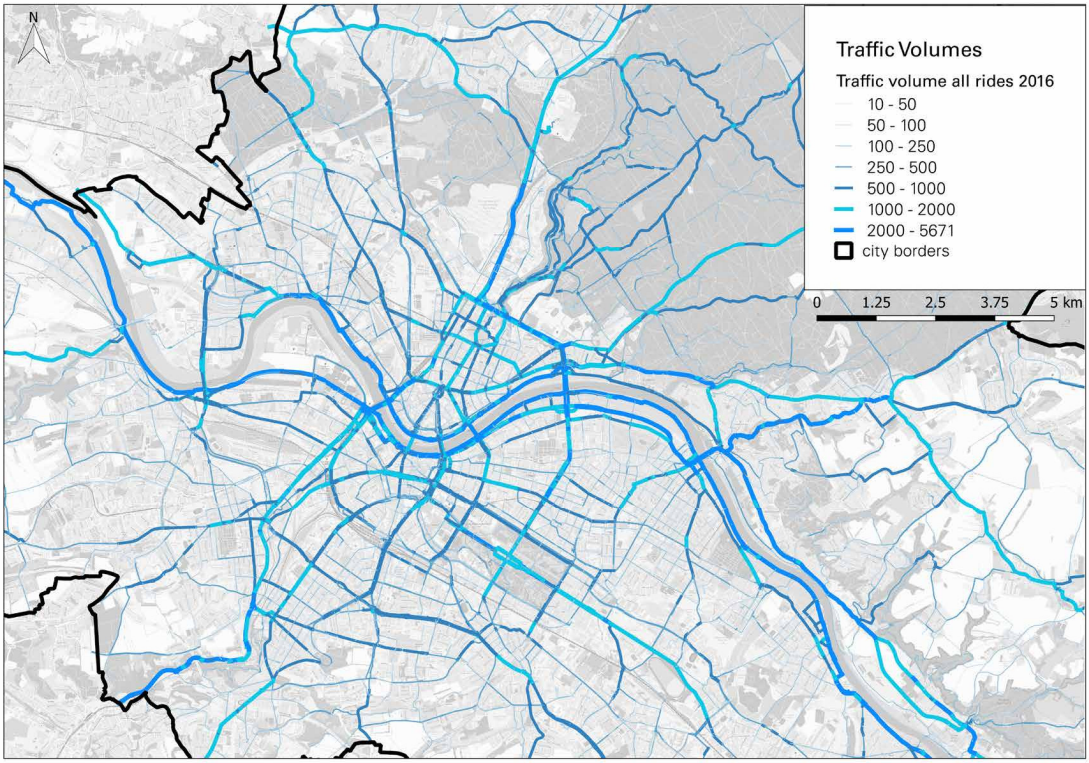


Figure 1: Bicycle Traffic Volume (BTV) of all Strava users in Dresden in the period from 06/2015 until 06/2016

Data protection law

Data protection plays an important role in smartphone-based data collection. In Germany, data protection is mainly regulated in the individual federal states by the data protection law in force in each state. This means that the area of use is defined by the processing of personal data by administrative bodies and other legal 'persons of public law' that are subject to the supervision of the federal state. The term 'personal data' is important here. Saxony's Data Protection Act defines it as follows: 'personal data are individual details concerning the personal or objective affairs of a specified or identifiable natural person'*. So the Data Protection Act is to be used when a dataset renders a natural person identifiable and when data can be traced to that person. Hence data providers generally provide anonymised datasets with which it is impossible for clients to trace back usage patterns to particular individuals.

Data anonymisation

How does anonymization work?

It is important to know that no route data are conveyed to the clients (in this case the municipalities). They are only given aggregated (summarised) data on the links of the GIS network used. Pieces of route information can also be passed on individually as basic data, with reference to the GIS links. This means that it would be possible, at least in theory, to reconstruct a route using the timestamp in lightly used networks. In order to avoid the threat of identifying a particular individual, the first and last 100 metres of the route are removed by the data provider before the data are handed over. However, in the vast majority of cases municipalities will buy ready-made map products to which the data protection law does not apply. For further information it is recommended to check the data protection law of the relevant federal state, usually to be found online.

Dealing with data protection concerns

Privacy concerns among users

It is not only the legal restrictions that raise concerns about data protection, going forward. Ever since large-scale data collection and integration began, data security has been an important and highly controversial issue in politics, business and society.

* Saxony's Data Protection Act of 25 August 2003 (SächsGVBl. p. 330), last amended by Article 17 of the Act of 29 April 2015 (SächsGVBl. S. 349).

Through smartphones that are not only kept close to the body, but can also record, save and transmit patterns of movement and usage, the debate reaches another dimension. It is not surprising that the use of GPS-data that were collected and resold for irrelevant purposes can cause criticism. Until now the routine for the handling of digitally generated personal data is missing. People might therefore not be fully aware of the effects and use of the data. Recent publications also show, that people are managing data security and new media in different ways. Trust, personal benefit and generational factors play an important role. A promising approach for participation of citizens is offered by the initiative Radwende, which will be discussed in chapter 3.

App user terms and conditions

Although users accept an application's terms and conditions of use during installation, many of them are unaware that the GPS data they generate can be sold on to third parties for completely different purposes. We must not lose sight of this when working with sensitive data. User surveys have also revealed that people are more willing to hand over information for research or planning purposes than for commercial use. So dealing with its acquisition and use transparently can have the effect of increasing trust.

2.2 What data are on offer?

GPS bicycle traffic data

The field of possible GPS data is certainly varied. The most commonly used solutions today are the data from app vendors with completely different user structures and data types, such as BikeCitizens or Strava, data from bicycle hire systems or data derived from local initiatives such as the ADFC [General German Cyclists' Club].

Smartphone-generated user data

Smartphone applications like Strava or BikeCitizens tend to generate route data that are saved in databanks together with the demographic details of the user derived from the application. These route data therefore contain sensitive information, such as the user's place of residence or workplace, which can also be connected with profile information such as name, age, gender and other freely given information. When passing on data to third parties, vendors are obliged to anonymise this information in accordance with the data protection laws and general conditions of business. In consequence, the buyer acquires data that have already been aggregated and do not allow any tracing back to the people that created the data. Anonymised demographic information such as gender and age are permitted to remain in the dataset. The data from global vendors of smartphone applications offer the largest range and number of possible users. Considerable differences can emerge within the user structure. The data are obtained second by second, saved at the end of the journey and transmitted to a server. The data can then be viewed by users on their smartphones and shared with others. This social factor feeds the user's motivation, in sporty applications such as Strava, to share the route just travelled with others or to keep a training journal. There are fundamental differences in background between sporting apps and those with different motivational models. The latter group includes apps that can find routes, such as Naviki or BikeCitizens, which are valuable to users for their navigational functions but also have other motivators such as social connections or reward systems. Well-known products at the time of writing (August 2017) include Strava, BikeCitizens and Scholz&Volkmer.

Who provides the data?

Data suppliers

When choosing a data supplier it is important to allow sufficient time for fundamental considerations on the user structures, data sources and motivational models that can be expected, as these have a direct influence on the data type, the sampling and the interpretation criteria.

As a rule, personalised user profiles in smartphone applications can give data suppliers very precise information on the user structure of their services. The sampling should be analysed on the basis of experience and of data available from other sources and assessed with regard to their representativeness, with due consideration of the way the data are going to be used. This does not necessarily mean that the user group has to correspond exactly with the broad range of cyclists in the municipality, as the case study in Chapter 4 aims to clarify. However, if there is a strong discrepancy in the user group, appropriate alterations to the data may have to be made in order to allow reliable interpretation.

App vendors

For example, our research project investigated the qualitative and quantitative reliability of the available GPS datasets from Strava. The Strava users, motivated by sport/fitness, differ from average cyclists, as they tend to belong more to the spectrum of ambitious 'power users'. This is predominantly reflected in their greater average journey length, frequency of bicycle use, and higher speed. Due to their strong motivation because of their sense of togetherness and control of their own performance, it should be expected that users will track their journeys carefully and in detail. During the course of this guide further pointers on this important aspect will be listed. The navigation-focused app vendor BikeCitizens, on the other hand, appeals more to everyday riders. These are closer to the average cyclist in their style of riding and their habits, but may be less intrinsically motivated to use the application to record every detail of every journey. Other app vendors are for example komoot, Naviki, and BikeMap which, however, up to this point according to the authors knowledge do not offer any data for planning purposes.

Local Apps

Besides the globally offered smartphone-applications there is a range of local alternatives. Some federal states offer bicycle route planning apps and also apps of transport associations (e.g. MVV or VVS) take bicycles into account. In addition, a range of local initiatives exists, including the initiative Radwende amongst others in Mainz and Wiesbaden, RADschlag in Düsseldorf, the BikeApp of the ADFC association Speyer or BBBike, which can offer a data basis. An individual data collection, as in the initiative Radwende, however requires large personnel efforts and know-how. Not only the programming of the app is a problem to deal with, but also the continuous acquisition of a sufficiently large sample. Also the data processing and the aggregated projection of the route data on a route network is rather complex. More details are presented in chapter 4.2. The additional differentiation of the resulting data and the assessment of their plausibility, as well as the processing for visualization requires knowledge on the application of databases such as SQL or PostgreSQL.

Bicycle hire systems as data suppliers

It is important to bear in mind that data based on smartphone use will obviously systematically exclude users without a smartphone. One further possible source of data with a completely different user structure could be bicycle hire systems. In some systems, chips built into the bicycles allow them to be tracked. Particular consideration must here be given to the recording intervals. Many hire systems record only the starting point and the location of the end of the journey. This delivers very little information on the routes chosen, if any, especially if the start and end points are the same. Also, it must be assumed that hire bicycles are mainly used in city centres or for strictly defined short routes within a series of designated paths. As a result, this will not give a picture of the regular daily routes of a monomodal cyclist. However, data from hire systems can be usefully added to other datasets. A highly dynamic development is currently occurring in this field. Vendors such as Mobike, YoBike, oBike and Gobeebike are falling over each other in German cities and focus on the evaluation of user data as well as on short-distance transport. How exactly these services will look is not yet predictable, but we must count on seeing more data sources in the near future.

In the future, apps with reward systems for kilometres ridden, such as Radbonus, might play a part in bicycle traffic planning. Either directly via their own data recording systems, or indirectly as a motivating element when using other tracking apps. It would be worth considering, and useful, to connect these together. Other stakeholders would include health insurance companies – the German health insurer AOK has already started to reward active modes of transport. Other potential stakeholders are the manufacturers or bicycle navigation systems such as Garmin, Teasi or Falk. These would tend to deliver data with a strong emphasis on leisure, but could give interesting insights in the fields of city marketing or tourism, as well as in rural areas.

Other potential data suppliers

How can data be procured? How should invitations to tender be made?

Obtaining data and invitations to tender

There are various ways of purchasing GPS data from vendors. Because the market is still so new, vendors are often also still in the test phase and have very individual methods of handling clients' requirements. The recommended system, developed in collaboration with local representatives for cycling matters and taking into account the structures present in many municipalities, is illustrated in Figure 2. Following this procedure keeps the buyer's staff requirements low and ensures that the preparation of the GPS data, a specialised task, is carried out by the supplier.

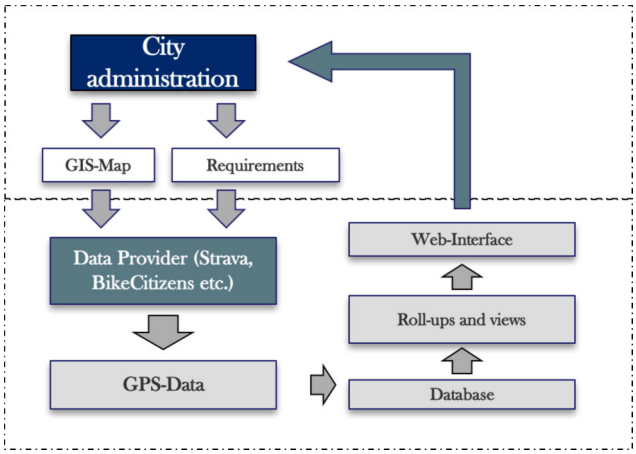


Figure 2: Recommended approach in data supply for the communal use

Advantages and disadvantages of various data types and Suppliers

Some suppliers are already set up to prepare the data in-house and to make them available in an easily accessible way. The local authority gives the supplier a suitable GIS map of the route networks (see Chapter 4) and the individual profiles of requirements, depending on the issues to be investigated. The vendor then moves on to the next step, which is to prepare the data completely and make them available online or in a data portal for further use (see Figure 3). Now the datasets can be accesses during the planning process and evaluated just as required. It is important when acquiring data to define the requirements clearly. If, at the beginning of the process, the decision is made to stick to pre-prepared data extracts and data views, then it must be ensured from the start that these will include all the necessary information. The route network also becomes very important in this. This absolutely must be checked in advance for any gaps in the network. More information on this can be found in Chapter 4.

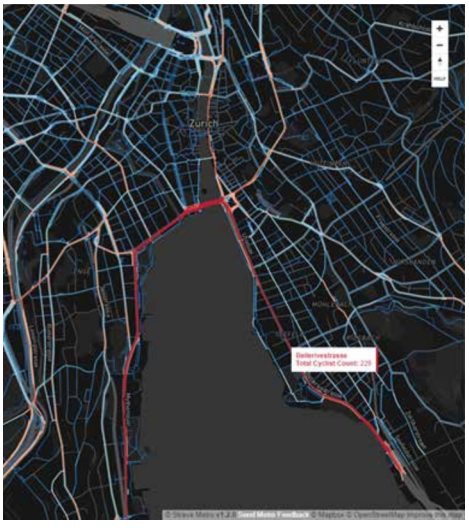


Figure 3: Exemplary representation of Strava Metro (source and copyright Strava Inc.)

When using data portals, a variety of approaches are possible. However, the display format chosen is always a card format, which uses buttons to show or hide information. This is then visible when the user clicks on the buttons, or via mouseover. Apart from showing the total number of journeys, Strava also offers details on everyday rides, user numbers and waiting times. It can also analyse origin-destination relationships.

Upon agreement with the data provider, it is also possible to obtain statistical data for a specified time frame. It is also possible for buyers to create their own aggregations or evaluations, but the depth of detail is limited. The output of the evaluation is released via an internet browser, so that no special software is needed by local authorities. It will not be possible to incorporate automated traffic counters in the near future. However, in the future it will be possible to show the routes starting and ending in originating traffic cells in aggregated form.

BikeCitizens go in a somewhat different direction with their Bike Analytics Tool (see Figure 4). In the foreground here are not only the macroscopic analysis, but onClick commands are also meaningfully linked to microscopic information. Here, various types of assessment can be made. Apart from bicycle traffic volumes, routes that contain a marked network element can be visualised too, as can individual diurnal variations. This form of analysis offers much deeper insights into the behavioural patterns of cyclists and allows detailed evaluations of individual links in the route network.

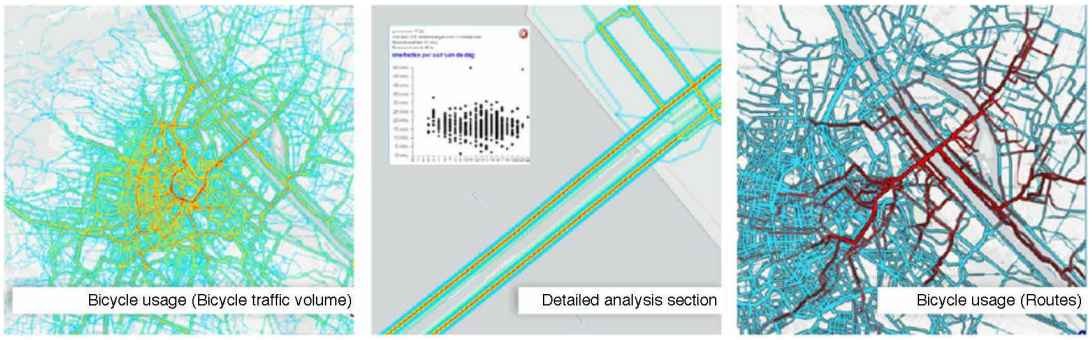


Figure 4: Exemplary representation of Bike Citizens Analytics (source and copyright BikeCitizens)

2.3 What parameters can be investigated?

Traffic volumes

Traffic volumes give the number of cyclists per unit of time. As only some cyclists participate in smartphone-based data acquisition, it is necessary to extrapolate the data to include the whole cycling community as a final step, in order to be able to draw conclusions that apply to all the cyclists in one area. An example of such an extrapolation process, validated using data from counting devices in the pilot municipality of Dresden, is given in Chapter 4.2, step 5.

Commuter / everyday journeys

Bearing in mind the usability of the data for bicycle traffic planning, it makes sense to look at the exact definition of commuter/everyday journeys by the vendor, as this may also include regularly recurring leisure- or sport-related trips. The proportion of purely fitness-related journeys should always be kept as low as possible within a dataset, or it should at least be possible to highlight and filter them effectively.

Origin-destination matrix

The complete journeys of users cannot be obtained - only aggregated data - for data protection reasons. This means that on the network link level only traffic volumes can be viewed, but not decisions on routes, nor origins and destinations. This restriction can be avoided by gathering the data from the supplier into blocks. These blocks might be postcode areas or districts of a city, traffic cells or block maps. While postcode areas and city districts give only a very rough idea, block maps are so finely detailed (50m or less) that anonymisation may no longer be effective. In cities with traffic models there are usually traffic cells that, depending on the quality of the resolution, are the size of a city district or somewhat smaller. These offer the possibility of setting up an origin-destination matrix for cyclists, however in different parts of the country they are at different resolutions or do not exist at all.

For that reason, the uniform European grid system for official statistics was used to create an origin-destination network within the framework of the research project. In addition, polygons were created as regular squares, with an edge length of 1000m. This process both satisfies data protection regulations and means the results from different regions are easier to compare. The exact methodology is detailed in the case study in section 4.2, step 4.

Speed

For every link in the route network there are two different speeds: one for each direction. The overall speed calculated should, when they are stated separately, correspond to the average value of both directions, weighted according to the volume of traffic. If the road slopes, large differences are to be expected between the speeds in different directions. As there is usually no obvious template within the basic maps to designate the direction of digitisation, it is not always easy to solve the problem of how to display all uphill and downhill trips. This could perhaps be solved by having separate graphic representations, made visible by mouseover.

Waiting times

GPS-based waiting times are suitable indicators for node point evaluation from a cyclist‘s point of view. Strava‘s data were not usable, however, without a lot of corrective work beforehand. Still, several interventions allowed at least a superficial assessment of node points using the proportion of unhindered rides through (see project report). The supplier BikeCitizens offers promising approaches in this regard. When the data is handed over it is important to note whether it is actually waiting times or ride-through times that are being provided. Solely georeferenced models for calculating waiting time usually give only the sum of waiting time and ride-through time. As in the case of the Strava data, this results in far too few unhindered ride-throughs. Ideally, the data provider should use the cyclist‘s speed and acceleration values as well as the georeferencing around the node point and the corresponding timestamps in order to calculate waiting times (see Figure 5).

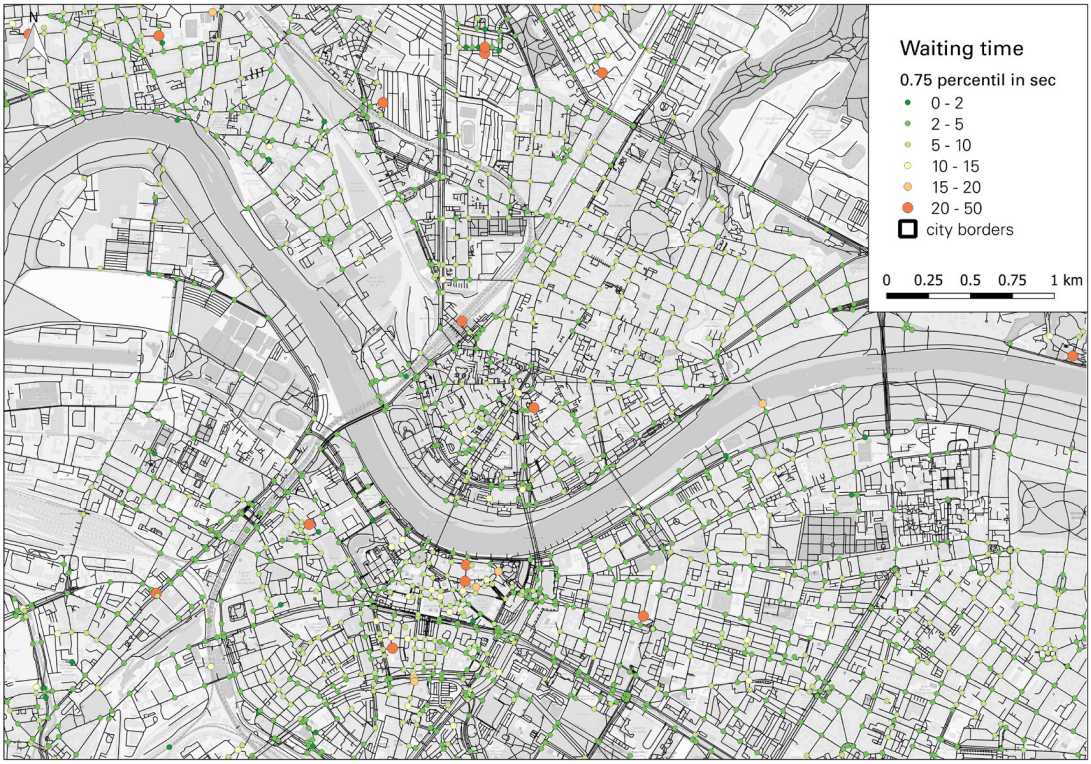


Figure 5: Waiting times of all Strava users in Dresden in the period from 06/2015 until 06/2016

3 What opportunities are offered by GPS data?

GPS data offer the opportunity to analyse where, when and how cyclists move around over the whole network. There could be various fields of application for the GPS data described in Chapter 2, such as network planning, facilities planning, setting priorities, monitoring and evaluation. This chapter gives an overview of these fields of application and suggests some accompanying measures (Chapter 3.6) which aim to involve citizens in the process of data collection and use.

3.1 Overview

GPS data offer a variety of areas of use that can currently only be insufficiently or with difficulty covered by conventional data sources. The various areas of use and their corresponding parameters in the GPS dataset are described below.

Potential uses

Overview of the potential uses of the three parameters traffic volume, origin-destination matrix and speed (cf. section 2.4) for bicycle traffic planning. A detailed description can be found in sections 3.2-3.5.

Table 1: Use cases for GPS data, parameters "traffic volumes", "Origin-destination" and "speed"

	Traffic volumes	Origin-destination matrix	Speed
Network planning	✓	✓	
Facilities planning		✓	✓
Priorisation of measures	✓	✓	
Monitoring & Evaluation	✓		✓

3.2 Network planning

GPS data can deliver valuable input for planning destination networks or for working out a bicycle traffic concept. They can show which paths cyclists choose, how great the demand is for particular routes, and where there is need for new infrastructure.

Gaps in the network and the side road network

Bicycle traffic flow in side roads is often particularly poorly represented by the data, as permanent counting devices usually lie on main routes and short-term counting data are not usually available on the required scale. Within bicycle traffic networks that have already been designed it is possible to discern gaps in the network and to close them, or to identify and resolve inadequate facilities on the basis of data from the city geoinformation systems. For these steps to take place, it is essential to know the bicycle traffic volumes on the network links. In future, rising user numbers are likely to allow the identification of capacity bottlenecks following extrapolation (for examples of extrapolation calculations, see Chapter 4). Here, route data can also give decisive information on detours taken and missing network elements. Many datasets that are available do not include route data, as they are lost during the aggregation of information on individual journeys that is necessary for data protection reasons. The fact that such evaluations are already possible despite this is shown by the services provided by companies like Bikeprint.nl and BikeCitizens.

Origin-destination relationships

The origin-destination matrices can initially be of help here. Origin-destination relationships, often in demand, offer information on a necessary expansion of connections and flag up places where the quality of facilities needs to be checked. Insufficiencies in infrastructure are mainly revealed when people need to cross line elements with high divisive effects, such as rivers or railway embankments. Origin-destination relationships can give information on the necessity of additional crossings.

Planning services

When planning networks, despite everything, planning services for places without any noticeable demand should still serve as the means of choice. This is because, as with almost every other form of data collection, non-users are barely represented, if at all.

3.3 Facilities planning

Speed as an indicator

Gaps in quality, such as defective surfaces or overburdened infrastructure, can be recognised due to lower speeds in comparable, level network elements. The parameter of speed can therefore give some initial findings on any potential for improvement in specific places. This allows potential for improvement to specific sections to be placed in the context of the network as a whole.

When speed is used as a parameter, it is important to be aware of the user group of the app in question. At least with apps with a sporting background the average speed needs to be reduced. To come into line with the speed of an everyday cyclist, the speed in Dresden has to be reduced by 5.5km/h (see Chapter 4).

Bicycle traffic volumes

Using particularly high and low values in bicycle traffic data, which are visible from the origin-destination matrix, the need for bicycle parks can be estimated or checked. The same applies to checking the quality of facilities along particularly well-frequented relationships. Origin-destination matrices can also be used to detect deficiencies in the number of bicycle parks.

Verifying the results

Further possibilities within facilities planning regard checking the success of completed projects. Here, the traffic volumes that occur can be used to visualise whether a particular measure is having the desired effect, or whether, for example, further accompanying informational and communication measures or signposting are needed.

3.4 Prioritisation of measures

Traffic volumes as an additional indicator

When setting priorities for the implementation of bicycle traffic measures, parameters such as traffic safety and any planned major remedial work are generally taken into account. These parameters can now be added to with information on the traffic volumes present. This helps to prioritise the use of cost-effective measures, particularly for the side road network. This includes adding protective road markings, approving one-way cycle lanes in the opposite direction to the other traffic, and setting speed restrictions for motorised vehicles. Decisions on the use of bicycle traffic concepts that are already in place can also be assisted by the use of traffic volume data. They can also be used to help with prioritising certain measures, along with other parameters, such as mitigating accident black spots. For areas with high origin and destination traffic, it is possible to prioritise the implementation of measures that concern bicycle parking.

The effect of measures over time

The use of GPS data in addition to the tools usually used in bicycle traffic planning can provide further clarity on a workable timescale for the measures, but should certainly not serve as the only indicator. Traffic safety should, of course, continue to be the most important parameter in carrying out any construction projects.

3.5 Monitoring and evaluation

Evaluation of measures

After the implementation of individual measures, it makes sense to carry out an evaluation. Bigger changes in traffic services generally result in the most visible reactions. For example, GPS data allow roadworks-related deflected traffic to be mapped, as well as an increase in traffic due to new or altered infrastructure.

Difference networks

This can be shown using difference networks, for example. Figure 6 shows an example of a difference map of the city of Dresden. While most roads show a significant increase in bicycle traffic, we can identify some individual sections where it has decreased. These sections point to alterations in behaviour among cyclists. In this way, looking at the entire network, we can identify possible deficiencies in the infrastructure or the effects of changes that have been made only in certain areas. However, it must always be borne in mind that the number of users of the app could also have gone up or down. It would be useful to normalise the data to one base year or to compare percentages. The most important thing here is a consistent route network, as comparisons are always best made using a clear, consistent index of network elements.

Alterations in speed

The speed of the cyclists can represent a further indicator for the monitoring of measures, insofar as the measure carried out was intended to improve surface quality or change the infrastructure on offer. Ideally, for instance, a separate cycle lane would lead to a much faster flow of traffic and therefore higher speeds than one that was combined with a pedestrian path. But combined cycle and pedestrian paths can also be checked for their workability. High speeds on such combined paths would point to the arrangement being unworkable or at least to the need for additional regulatory measures.

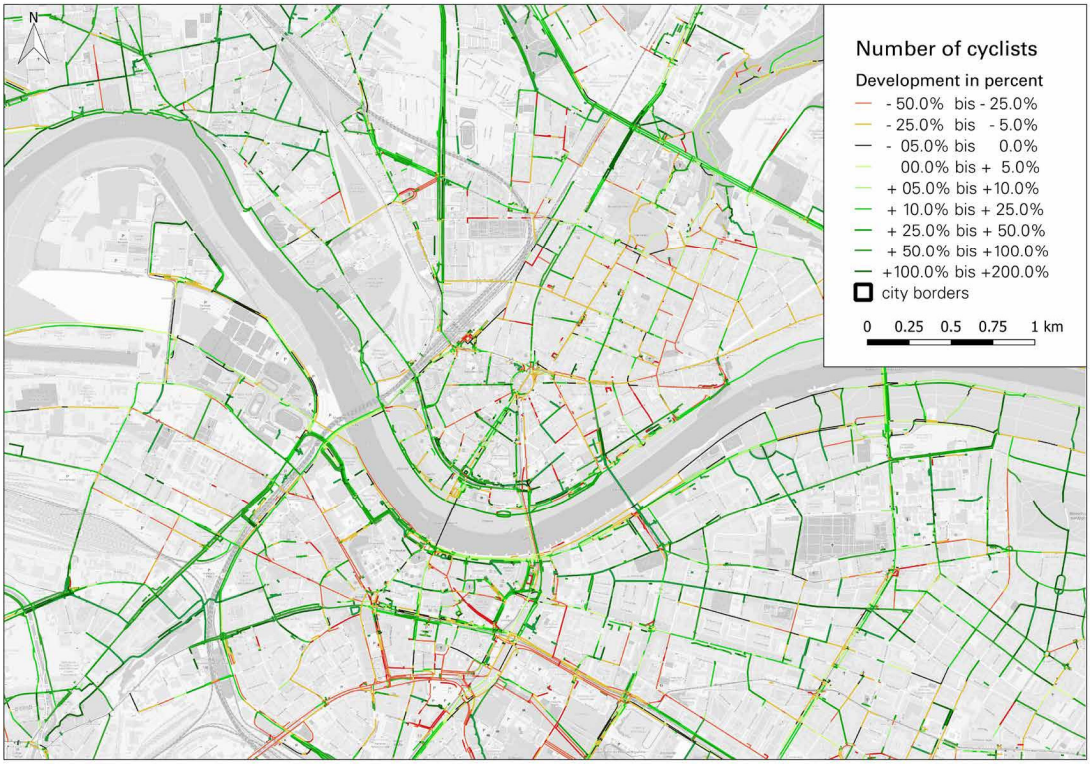


Figure 6: Difference network of Bicycle Traffic Volume, percentage development based on the change from 01/06/2015 to 01/06/2016

3.6 How could accompanying measures look?

Participation and public relations

GPS data have a high potential for public relations and internal marketing due to their interactive character and their close reflection of reality. Conventional planning tools are either hardly noticed by residents (traffic counting, speed measurement) or offer no direct objective added value (surveys of residents etc.).

However, as GPS data generation only takes place because of active cyclists, it would make sense to communicate this to the outside world. If cyclists can be actively involved in it, the use of GPS data will gain important added value in the long term. A means of direct communication between citizens and planning offices, integrated into the app, would reduce the barriers that town council meetings or displays of plans can build. It would also permit less expert residents to participate. Unfortunately, this advantage hardly has any effect on the data supplier Strava (investigated as an example in the research project) due to the scant possibility of interaction between the app operator and the users. This would make it hard for local authorities to approach users in practice. On the other hand, other competitors on the market offer much greater scope for citizen participation as far as design and interaction are concerned, and this will surely be expanded and/or covered by future providers. There are a few examples of cities in which citizens are involved in the generation of GPS data right from the start. The local authorities themselves offer them apps that record the routes travelled by the users and transmit them to the city authorities. This shows how personal identification with the municipality can be supported.

Example:
The Radwende-App

An example of this is the app Radwende from the agency Scholz&Volkmer in the cities of Mainz and Wiesbaden. This solution is very much influenced by the locality, and is not very suitable for collecting and preparing data on a global scale at present due to its lack of scalability. Yet it should be highlighted as an extremely positive example of data collection and communication due to its outstanding accompanying media campaign during the data collection phase. Among other things, the cycle routes that had been recorded were raised to an art form and beautifully displayed. This was particularly the case during collaborations with local cultural establishments, such as the Staatstheater Mainz. This meant that a target group could be reached that would not normally have been interested in using the app. In addition, the Mainz city authorities implemented cyclist-friendly measures in return for specified numbers of kilometres reached by people using the app. There were additional incentives in the form of discounts at the local retailer which could be claimed after riding a certain distance. All in all, the campaign had a very strong external effect, particularly because of the vivid depiction of the flow of bicycle traffic, and it reached even target groups that are hard to get at using conventional means of communication. The evaluation of the data, however, has not yet got further than the creation of a heatmap. Altogether, in both cities combined, 93,000km and around 16,000 journeys were recorded by users. This can be taken as a clear success for the concept.

Example:
ADFC Speyer RADschlag

Another example is the Bike Track App by the ADFC Speyer, which attracts citizens through possible improvements in the bicycle traffic infrastructure due to their app use, as well as RADschlag in Düsseldorf. Also RADschlag offers an interaction with the city of Düsseldorf as the app operator is possible, for example to report deficits in the infrastructure.

Support for internal communication

As the examples of Mainz and Wiesbaden show, one great advantage of GPS bicycle traffic data lies in the clarity of the datasets. In combination with GIS networks and street maps, it lends itself to the possibility of illustrating a variety of issues and of presenting them in a way that is understandable to the layperson. This means that, going beyond public relations, the datasets could be used for well-founded and reality-

based internal communication in administration and politics. Visualisations clarify what needs to be dealt with in traffic planning and deliver supporting arguments for upcoming decision-making processes.

3.7 Possible weaknesses

Precision of GPS data

When considering facilities and network planning, it is not yet possible at present to relate users to a particular infrastructural element on which someone has travelled. For instance, there is not yet enough evidence to state definitively whether cyclists should ride in mixed traffic or on approved pedestrian pathways. Help might soon arrive in the form of the new global navigation satellite system Galileo, which promises higher precision. With Galileo, which is set to be used as the new standard in future, the current precision level of 6.8m on a horizontal plane can be yet further increased. This means it will also be even more useful for planning. The new generation of smartphonesuse a mixture of all receivable satellite systems. This means that even in bad reception conditions a high level of precision can be achieved.

Sample composition

GPS data are currently still very homogeneous datasets. There are no deeper studies on their users to date. This is where the RadVerS project, sponsored as part of the 2020 National Cycling Plan, comes in. This places research on the user groups in the foreground, which should result in a reduction in this gap in our knowledge. Moreover, the topics of non-users and systematic exclusion of user groups also need further research. Evidence will be needed in the near future on how a representative sample can be created.

Resources required for data collection

Finally, it is necessary to note that the high motivation associated with sports apps like Strava due to their competitive nature and to their value as training aids may not necessarily be the case with other app concepts. In order to convince users to use apps actively in the long term, appropriate incentives must be created. This will require more resources, a fact that should be considered early on by interested municipalities and integrated into informational and communication campaigns.

4 Case study
Dresden as a pilot municipality

The following chapter uses a case study to discuss data acquisition, evaluation and interpretation, in order to alert planners to the possible pitfalls and indicate possible strategies for handling such data. The data verification using Dresden as a pilot municipality was done both qualitatively, via map-matching of bicycle traffic data that had been measured in other ways, and qualitatively using a user survey. Bias due to the fitness-oriented target group can be offset by mathematical and qualitative methods.

4.1 Project framework

Project content

For the ‚GPS data in bicycle traffic' project, the Chairs for Transport Ecology and Traffic and Transportation Psychology at the Technische Universität (TU) Dresden have conducted a comprehensive analysis of the GPS dataset for the city of Dresden, as an example. The approach used and relevant results are detailed in this chapter. They may help further in answering the question of whether the GPS data on offer can be useful for a municipality, and if so, how. The results also convey a picture of the limitations and peculiarities that can come up in the practical applicaiton and interpretation of the datasets. Detailed results and examples of evaluations from the project can be found in the appendix under ‚Detailed Findings'.

Dataset and cost

The data presented here were supplied by the supplier Strava, chosen from among several providers. Two different datasets were used, covering a period of 18 months. The dataset consists of 70,500 journeys by around 3,200 users in the pilot municipality of Dresden (approx. 550,000 inhabitants). Pricing methods differ between suppliers. Strava, for instance, calculates the final price at the time depending on the number of users per year. Strava currently charges about €1 per user. BikeCitizens follows an alternative model, which is agreed with the municipalities concerned, and, for example, integrates campaigns in collaboration with retailers. An annual fee of a low to medium five-figure sum can be assumed here.

4.2 Case study – Procedure

Step 1: Can I use GPS data for my municipality?

Aims and guidelines

A comprehensive analysis was carried out on the pilot municipality of Dresden. This analysis answers as many questions on bicycle traffic planning as possible, and makes it possible to evaluate the data in their entirety. The main thing that was investigated in this respect was issues surrounding validation. As described in Chapter 3, the further possibilities for bicycle traffic planning are many and varied. It is thus necessary to set targets at the beginning of the process. It may be necessary to abide by the relevant procurement guidelines when purchasing the data.

Definition of the customer's aims

What questions need to be answered?

Before data are purchased, the buyer must be perfectly clear on what exact questions these data are intended to answer. These questions should be defined from the start. In addition, it is necessary to note the limitations and possibilities of the method described in this guide. As already explained in Chapter 2, the dataset cannot be obtained in its entirety for data protection reasons. The complete dataset always remains with the supplier, who prepares the data for its customers. During this process, for the purposes of anonymisation the first and last 100m of the route are deleted, and aggregated

datasets are created out of the individually recorded journeys.

Profile of requirements of the data

In the case study given in the pilot municipality of Dresden, the data offered on the market were to undergo critical examination in order to uncover their potential for bicycle traffic planning. The first step was to define the following requirements for the data's use in bicycle traffic planning:

- Bicycle traffic volume
- Cycling speeds
- Sociodemographic information on app users
- Computing waiting times at intersections
- Origin-destination matrices
- Possibility of dividing the data into time slices so as to compare time variation curves using counting devices.

Apart from these minimum requirements, additional criteria that contributed to the choice of supplier were extensive coverage by the GPS data, user numbers and, finally, the price. Section 3.1 gives guidance as to what data are needed for what kinds of analysis. Although the detailed data would make comprehensive evaluations possible, there are some restrictions involved with the sampling and the data collection process (see section 2.2).

Step 2: What suppliers are there?

Choice of supplier

The offerings of the various suppliers are assessed on the basis of the requirements defined in step 1. While at the time of the analysis (February 2016) only a few providers had user data on sale, a continually growing number of stakeholders can be seen on the market offering a wide range of data in a number of different forms. The datasets offered by Strava were the only ones that appeared suitable, in both scope and quality, for bicycle traffic planning in Dresden. That is why the decision was made to buy two datasets:

Dataset 1: Consisting of 439,570 journeys made by 19,615 users between October 2014 and June 2016 in the German federal states of Saxony, Berlin and Brandenburg. This dataset permits us to compare the results across regions.

Dataset 2: Consisting of 70,500 journeys made by 3,200 users between June 2015 and June 2016 in Dresden. This dataset forms the basis of the case study presented here.

Purchasing data

In order to check in advance the suitability of the data to answer the questions defined by the customer, it makes sense to ask the chosen supplier for a sample dataset. This allows the potential buyer, even before the planned purchase, to carry out a rough analysis to compare the sample dataset with the expected values based on past experience, thus checking for general plausibility and any distortions of the data.

App user profile

Who uses the app?

This section describes the users in Dresden in order to give an initial picture of the possible class of users. While some apps work with more social motivators and therefore show a somewhat broader class of users, Strava works exclusively via sporty and competitive incentives. Consequently we have some reservations about administrative bodies using this data, as they appear to originate mainly from young male cyclists who ride at high speeds. However, the age distribution in Dresden has turned out better than initially conjectured, and represents users from all age groups (see Figure 7). The reservations expressed by the municipalities are therefore only partially justified: male users who ride at high speeds are indeed over-represented. It does therefore appear

necessary to check that the data is representative, as well as simply analysing them, in order to make them more meaningful (see section 4.2, step 5). In the project, this was achieved by carrying out a survey of the Strava users.

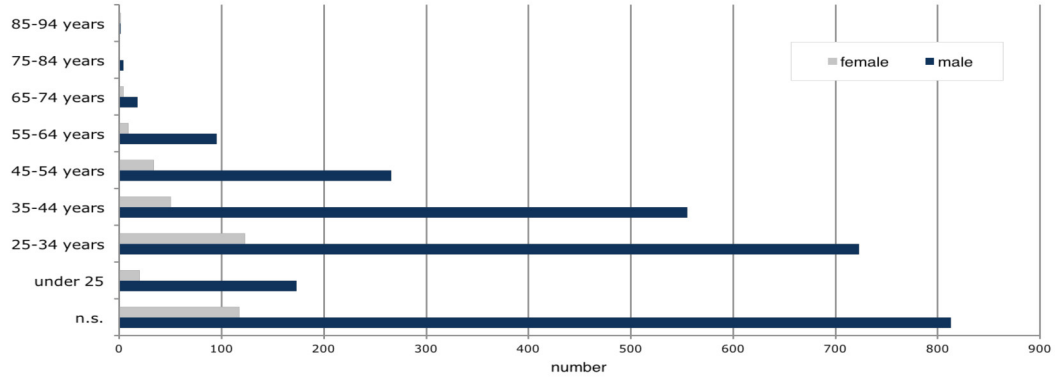


Figure 7: Age distribution of Strava users in the time period between 01/2016 and 06/2016 (N = 3,016)

Influence on Speeds

This form of motivation has an effect on the cycling performance of the users. A high number of intrinsically motivated long-distance trips was recorded (Ø 22 trips/year at Ø20km in length). The other side of the coin is the higher speeds achieved due to the high number of cyclists of above-average fitness. This lies an average of 5.5km/h over the average cyclist‘s speed. (see Detailed Findings VI). Within the dataset, the purely fitness-motivated trips can be distinguished from the everyday trips to a certain extent, but even the everyday journeys are influenced by this user group.

Influence of route choice behaviour

However, looking at the coverage of the network, this fact did not seem to have any particular influence on the routes chosen for everyday trips – instead it must be presumed that this user group is simply particularly well informed as cyclists. When interpreting the data, however, things like cycling events must be taken into consideration. This should always be brought to the data provider‘s attention, as they can otherwise produce spikes in cycling trips that would otherwise be hard to explain in retrospect.

User survey

In addition, a user survey was carried out, on the basis of which the user behaviour depicted in the data was compared with subjective estimates by the users themselves. Overall, the results correspond to the data described below. Though the motivation and speed of the Strava users are elevated, the choice of routes and route length show hardly any differences during everyday rides (see the project report for a detailed view of the user survey).

Step 3: How does the provider supply the data?

Data provision

The data purchased from the app provider Strava are based on GPS points, which are allocated to a GIS-based route network for the purposes of analysis. Strava has at its disposal all the GPS positions, stamped with the user, time and journey. Strava allocates these to a route network by means of a map-matching algorithm. The exact algorithm is not known to the project team. However, as during the evaluation a high number of duplicates was noted in areas with high network link densities, it can be assumed that a simple point-to-point or point-to-curve approach was used. An exact check of the data supplied as well as a statistical test for duplications is therefore always to be recommended, whoever the supplier.

Checking for duplicates

Figure 8 shows cyclists projected several times on different parallel links. Only the outer ones of these are actually usable by bicycle traffic. In the centre, cyclists are projected on a main traffic artery as well as in a tunnel.

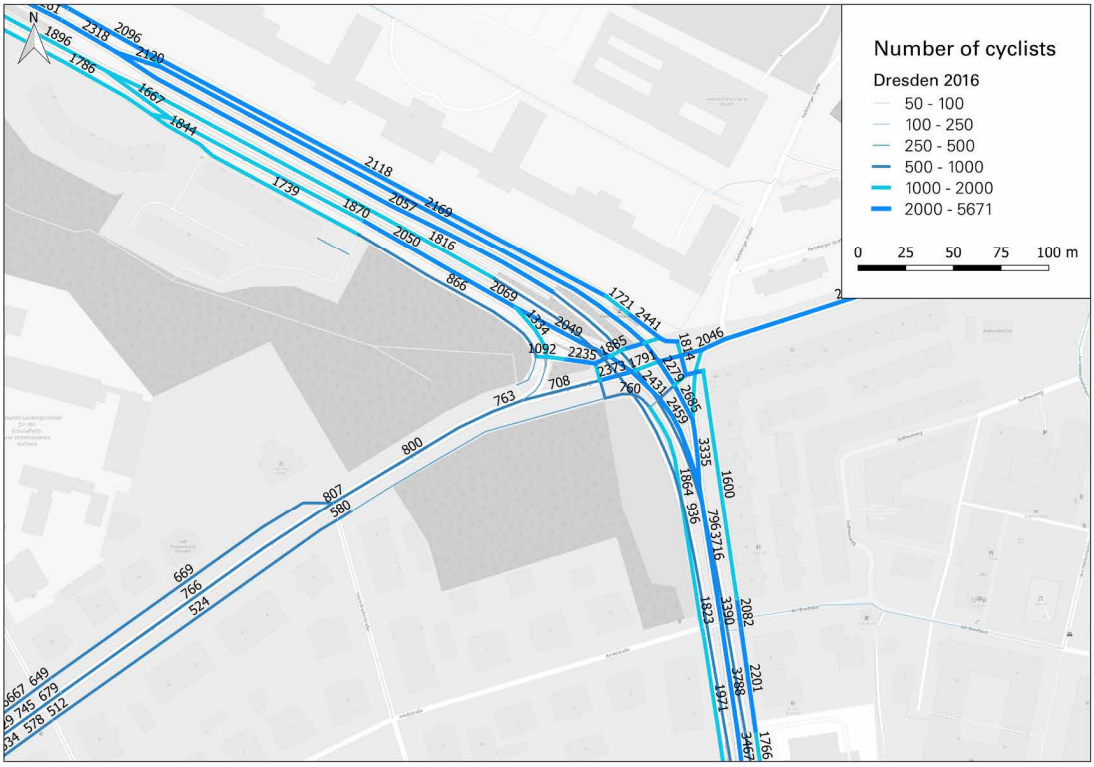


Figure 8: Visualization of duplicates in the Bicycle Traffic Volume (BTV, OSM road network in the city of Dresden, 2017

Which route network should be used?

Selection of a basemap

The route network upon which the map-matching is based can be chosen freely by the customer and is, in fact, generally provided by customers themselves. First of all, then, the project team had to sort out which route network of the pilot municipality Dresden was to be conveyed to the data provider Strava for map-matching; this network would later also form the basis of the later analysis (see Table 2).

In the early phase of the project it was decided to use two different route networks as the basis for the map-matching. One was OpenStreetMap (OSM), which was chosen to permit cross-regional comparisons within the three federal states of Saxony, Brandenburg and Berlin. The other was the extended street node networks provided by the city administration, which are ultimately linked to Germany's Authorative Topographic-Cartographic Information System (ATKIS®) and may contain various additional information relating to the infrastructure. It also offers the possibility that the city authorities might be more able to use the maps for their own planning purposes at a later date. The city's networks could be used without too much additional processing,

as every transport route consists of a link with a variety of attributes, such as bicycle traffic infrastructure, number of traffic lanes in a road, and speed limits.

The detailed OSM basemap was harder to use in some respects. In order to provide base data for the map-matching on the basis of OSM, a suitable route network had to be put together. For this purpose, the first step was to use route networks with the most complete datasets possible from OSM. These were exported from Metro Extracts from an online map portal. The map exports are very similar in their completeness to the raw data from OSM, so they offer wide-ranging base data. Problems arise when links run one above the other, or very close to each other in parallel, as the journeys allocated to them can jump from one link to another, as depicted in Figure 8. The analysis of cross-sections is also made more difficult during the evaluation of the data. To avoid this the networks were simplified so that only individual links were shown for each route. The exact method used can be taken from Detailed Findings I.

Table 2: Comparison of the route networks that were considered in Dresden

Suppliers	Advantages	Disadvantages
openstreetmap.org	Wide-ranging datasets Cross-regional comparability	Exports can be made of extracts only, xml format
mapzen.com	Wide-ranging datasets	Exports can only be made of selected cities
geofabrik.de	Reduced number of links	Limited range of data
ESKN 5*	Used by city authorities	Sometimes not many links (bridges)

** Extended road node network provided by the city authorities.*

Content of delivered data

What does the supplier deliver?

After the basemap has been handed over, the supplier generally aggregates the desired data and delivers it back referenced to the base network. The Strava data examined during the project came in the form of two packages of data. There is one dataset referenced to the links of the initial GIS network (see Detailed Findings II) and one dataset referenced to the nodes (see Detailed Findings III). This is carried out by simple indexing of each dataset. The data are provided exclusively referenced to the relevant network elements. Connecting routes cannot directly be deduced from them, so as to comply with data protection laws. From these basic data, various forms of aggregation, known as ‚roll-ups' can be generated according to the wishes of the customer. These are usually offered according to month, time of day or season. For example, the time-stamps can be used to aggregate all the data from January to March of a given year. These data can either be integrated into a city's online presence by the buyers themselves, or made available on an interactive platform by the supplier.

Step 4: What does the data reveal?

The basemap provided the starting point for the explorative data analysis. Firstly, traffic volumes and speeds were examined. Overall here, existing data is of relatively high significance.

Network coverage

The general network coverage by the app users is highly satisfactory, at least for the city of Dresden. Figure 12 illustrates this nicely with the blue lines closely distributed around the whole area of Dresden. It should be added that in this form of presentation no distinction was made between different purposes of trips, such as commuting, sport or leisure trips. The main routes within the city however in an east-west direction, as well as in north-south direction, still become visible.

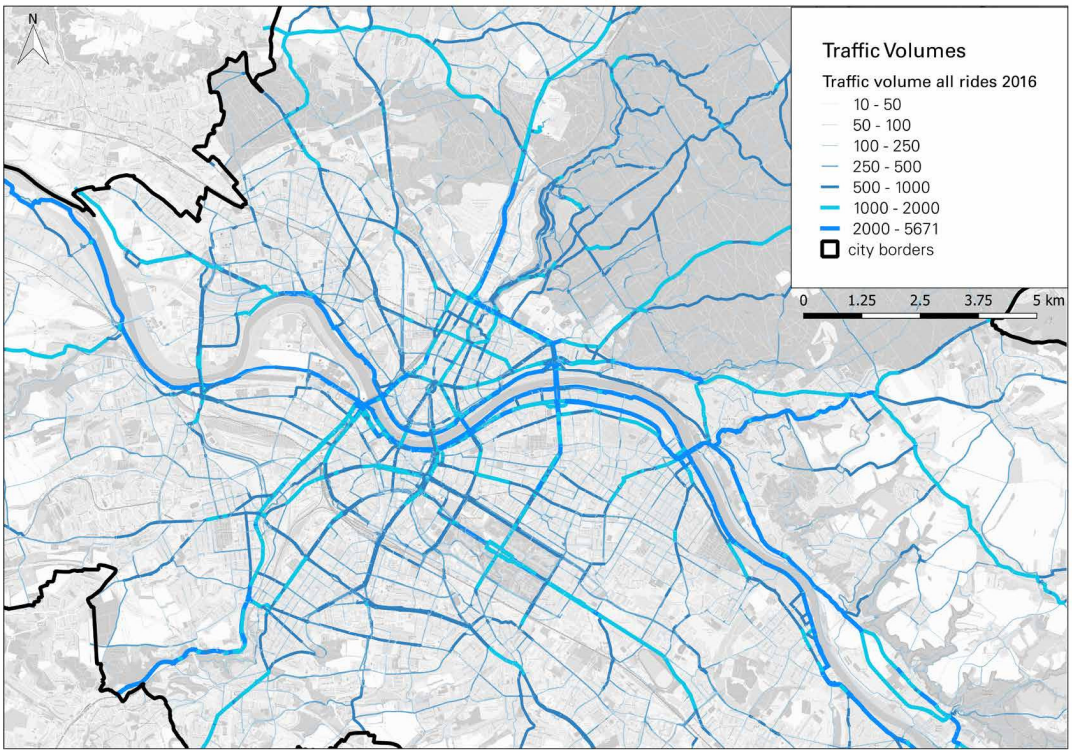


Figure 9: Bicycle Traffic Volume (BTV) of all Strava users for Dresden during the period June 2015 to June 2016

Investigation of individual network elements

A closer look reveals the specific possibilities available to the planners (see Figure 10). For example, the centre of the figure below contains a shared pedestrian- and cycle path through to a green space (Weißeritz greenbelt) that has virtually no node points, while the outer, far more busy links each represent shared walking and cycling routes in heavy traffic on major roads. It appears that these are used much more intensively than are places with less traffic, with their far higher amenity value. The case study here is not only represented by routes to and from work, for which the argument that they represent the shortest distance is a very strong one, but also for leisure trips. In line with this, the filter ‚Everyday trips' has been omitted. For city or cycling traffic planners, the following questions arise as part of this type of evaluation: is the current provision or offering attractive for cyclists in its present form? The quality of this offering is nevertheless very high. It is a little-used and very quiet route in green surroundings with ample seating and playgrounds that should make visitors want to spend time there. This aspect is thus probably not the reason for the low numbers of people coming here. Another question comes to mind regarding the visibility of the cycle lane: can cyclists actually see this offering?

In the example considered, this aspect of a lack of visibility is probably worth examining. The green space as such is not signposted for cyclists and is somewhat

Figure 10: Map excerpt of bicycle traffic volume (BTV) of all Strava users in the period June 2015 to June 2016, which shows the use of the Weißeritz greenbelt, as well as photos of this location

hidden away behind a railway underpass (Figure 10, see photos). Firstly, the possibility of improved signage should be determined in this type of case. The number of users can then be compared the following year to determine whether these measures have had an impact.

Consideration of purposes for trip

Commuter trips or everyday trips

Figure 11 shows the routes marked by Strava as those used by commuters or other people on everyday trips (see Detailed Findings IV). At first glance, a reduction in the amount of traffic is discernible compared to Figure 9 across the overall map. So, for example, the proportion of daily trips made on segments of the Elbe cycle path is only 50%. Decreases in numbers of trips are also noted in the north-eastern part of the city forest. The theory of above-average numbers of sport-related trips can thus, at least for this area, be confirmed. Since the routes include both paved and unpaved variants, both road bike and mountain bike users appear to use the app in equal numbers. It is therefore likely that both road and mountain bike riders use the app.

Connections in urban areas

While, due to the routes used, off-road/mountain bike sport activity can be localised relatively simply and then also excluded from the map displays, this is not as straightforward in the case of road biking. At this point the question arises of whether this is absolutely necessary. The filtered daily trips also offer planners opportunities for gaining insights (details on the definition of daily trips can be found in Detailed Findings IV). For example, in the case presented, in Figure 12 we can see the section of the Elbe cycle path that is marked in Figure 11. This is used to a discernibly lesser

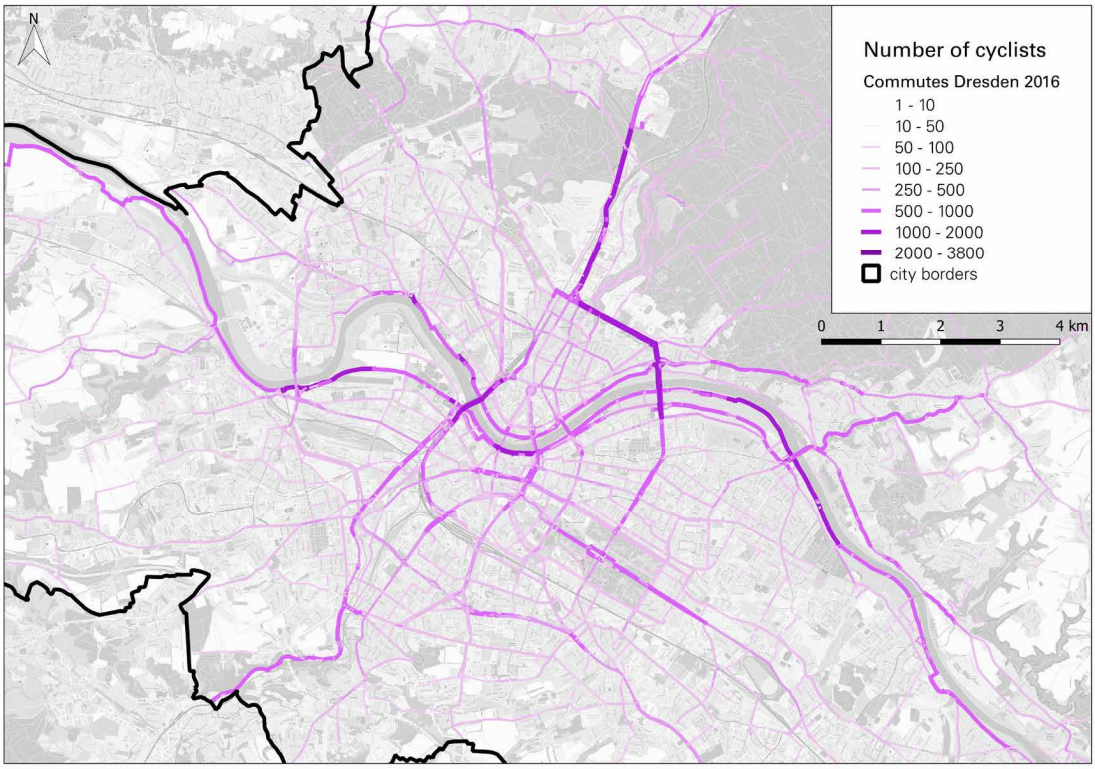
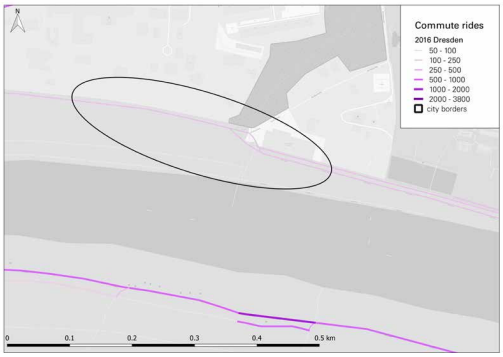


Figure 11: Trips by Strava users, designated as ‚commute'

degree than the opposite side and the subsequent sections. The connecting function here is absolutely the same. However, it is a section with a significantly lower surface quality attributable to existing preservation order constraints. This example is of course a very familiar one for local planners, but excellently illustrates the possibilities offered by a dataset based on recorded behaviour of cyclists. Due to the clarity of the



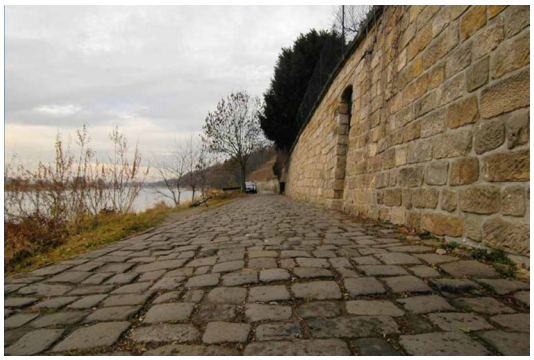


Figure 12: Map excerpt of bicycle traffic volume (BTV) of all Strava users in the period June 2015 to June 2016, which shows the use of the Körnerweg route, as well as a photo of the surface quality.

presentation, this can also back up arguments in discussion.

Origin-destination matrices

The origin-destination relationships can be seen in two different degrees of detail. So, consideration of all the data is possible as well as a look at direct origin-destination links of individual areas. Because the origin-destination matrices cannot be directly derived from the original dataset (see section 2.5), the data processing of the origin-destination matrices is shown by way of example in Dresden in Detailed Findings V.

Originating traffic

The source traffic of all trips within the city limits of Dresden marked by Strava as ‚commuting' are shown in Figure 13. For this, in order to also capture incoming commuters in terms of location, traffic within the origin-destination matrix was examined which started or ended in the city of Dresden. Clearly visible is the strong focus on the city centre, located in the centre of the map, with a high density of

residential development. Just north of the centre emerges the city district 'Outer new city'. This identifies itself as Dresden's trendy and nightlife district. The predominantly young population here would suggest that there is also a higher proportion of cyclists. This thesis can be confirmed on the basis of originating traffic. Other strong sources in the north of Dresden, respectively in the north-central area (shown in purple) are congruent with the location of large employers (Infineon, with 2,000, and Globalfoundries, with 3,700 employees). In the city centre and south of it are other significant sources of bicycle traffic in the form of the campus of the Technical University of Dresden and the nearby student accommodation. Equally striking is also the special contribution made by the city forest. The Dresdner Heide (Dresden Heath) constitutes a major source of originating traffic in the north-east of the city. Major sources of traffic generally tend to be more in the city centre than in the urban

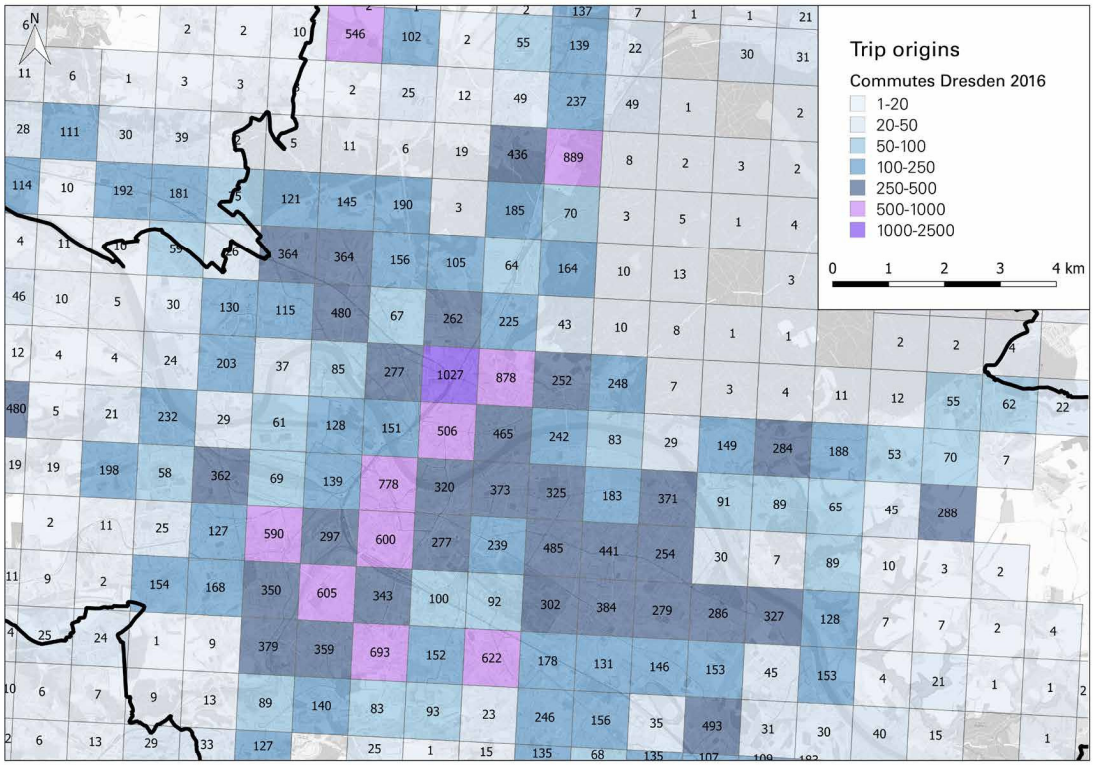


Figure 13: Bicycle Traffic Volume (BTV) of all Strava users for the city of Dresden during the period June 2015 to June 2016

Critical interpretation of data

environs. This is logical due to the higher population figures within the central city district as well as the concentration of retail infrastructure and jobs.
Originating traffic cells for commuters are mainly found in the towns of Radeberg, Pirna and Radebeul just outside Dresden. However, the absolute figures recorded for the commuters should generally be interpreted with caution, since figures under 200 trips taken could easily have been taken by one very active person on their own.

Critical examination of usage corridors

Direct origin-destination relationships enable, on one hand, the identification of certain, possibly preferable, usage corridors. Second, they provide a way to identify particularly active cyclists as sole sources based on the absolute number of routes. This is not necessarily possible because data on cyclists are not available, and it is theoretically also possible that several less active cyclists also often use the same route. As a result of the – compared with the urban population (> 550,000 adults) – comparatively small sample of Strava users (approx. 3,000) the likelihood of there being several less active users is however relatively low.

Figure 14 shows all recorded origin-destination routes for a selected neighbourhood in Dresden. First to note is the variety of different other targets and sources suggest a very

heterogeneous mix of users and a variety of active users, while, secondly, the influence of one very active user is very clearly identified between the Laubegast part of south-east Dresden and the outer part of the city's newer area. On this route alone, about 450 trips were made during the period of observation. Statistically at least, this figure cannot be explained by the number of inhabitants compared to other parts of the city. One possible explanation would therefore be a very regular commuter, or a regularly scheduled training session or an active gathering of cyclists. This problem area acts as an example for all procedures based on public participation. Very active users can lead

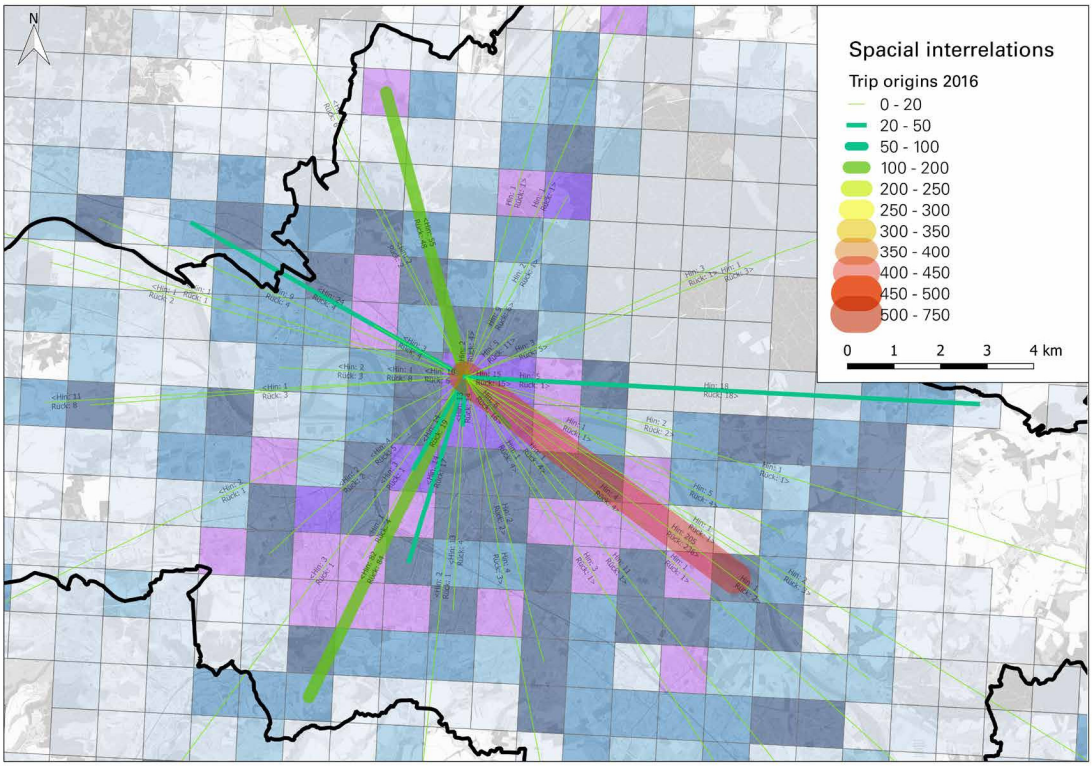


Figure 14: Route relationships used in a selected neighbourhood in Dresden in the period June 2015 to June 2016

to distortions. On the vendor side, what is necessary is to encourage normalising the user numbers or alternatively limiting the observation period, if necessary.

Speeds

Average speeds

The speeds shown in Figure 15 represent the average of all trips recorded from 05/2015 to 12/2015 on the respective network sections. It should be noted here that for each link in the system, of course separated according to direction, two different speeds are available. During the underlying data processing, these speeds are calculated only ‚going in' and ‚going counter to' the direction of digitalisation of the corresponding links in the GIS basemaps. The facts presented correspond to the median value of both directions weighted via the frequency of use. An increased cycle traffic speed, as compared to measurement of travel time, can generally be noted for the city centre area and also the Elbe cycle lane. This gives rise to an enhanced review of the real speeds travelled at in these areas. Fundamentally however, the speeds can be interpreted relatively. Areas with lower speeds along a strip of road indicate obstacles or potential areas for improvement in the workability of the infrastructure used.

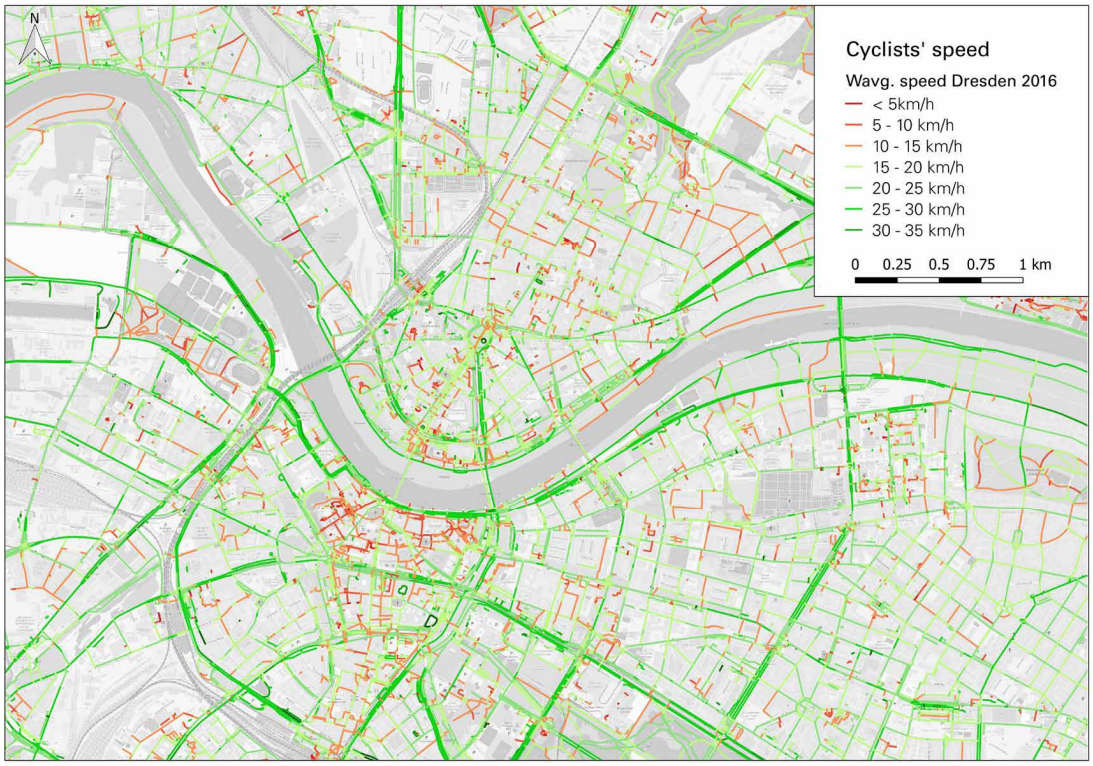


Figure 15: Average speeds of all Strava users across both digitalisation directions of the link during the period June 2015 to June 2016

Step 5: Are the data representative?

Investigation of individual parameters for representative-ness

To assess the representativeness of the data, a comprehensive examination for each parameter was performed. The context here is the random sampling of the Dresden Strava dataset featuring around 3,200 people and 70,500 trips, which, compared with the total number of cyclists in Dresden, is relatively low and which at least suggests a socio-demographic distortion due to the app's user structure. The exact procedure, as well as methods of dealing with systematic errors, are described in Detailed Findings VI and VII. The estimates are based on an extensive scientific validation between GPS data of the company Strava and data from corresponding established field measurements in the pilot city of Dresden (see project report). Here, the representativeness of the origin-destination matrices, waiting times and commuter trips has been investigated. This is not a practicable option for smaller municipalities in this depth, but their own empirical data should be included in the data analysis.

Online portals for analysis

Restrictions

The prerequisites for the use of GPS data depend on the vendor used and the desired depth of the evaluation. Strava, as well as other companies such as BikeCitizens, offer access to online portals with pre-defined evaluations and presentations. For this purpose the user simply requires Internet access at his or her workplace. This type of technical implementation is attractive especially for smaller municipalities and those with fewer resources, but should be the goal for all municipalities. This requires concrete agreement with data vendors in advance, particularly with regard to the requested method of visualisation.

However, should buyers wish to carry out their own, more in-depthevaluations, the minimum requirement is GIS software, the related expert personnel, as well as the opportunity to work with geo-databases. As a rule, the necessary technical requirements and human resource capacity are present at least in larger municipalities even if in most cases not usually embodied in a single person or an administrative unit.

Summary of the findings of the representative review

» *Traffic volume*:	The average daily load curve of cyclists can be roughly mapped as well as the average daily traffic. An extrapolation of the traffic volume by means of a constant factor calibration of data from permanent counter devices is possible.
» *Speed*:	The speeds are on average around 5.5 km/h higher than comparable readings. A reduction by 5.5 km/h is therefore recommended. Relative changes are nevertheless also visible without this step.
» *Origin-destination matrices:*	The distribution of trips correlates with the inhabitant figures; however, exclusion effects are observable
» *Waiting time*:	The latencies are not usable without massive corrective intervention.
» *Commuter trips / everyday trips*	The algorithm reliably detects everyday trips where these are not performed as a circular route in the form of a route chain and where the individual sections are not cached. These are very long however, compared to the usual route distance. The proportion of daily trips in the Strava dataset thus tends to be overrated.

Step 6: Has the question been answered?

Prerequisite for data evaluation

Various questions with regard to the target network planning and the analysis of the previous offering, as well as the prioritising of measures, can be answered based on traffic volumes, speeds and waiting times. To get high bicycle traffic quality, dips in speed, for example caused by barriers such as steps, must be avoided, the traffic volumes on the desired routes must be increased, and waiting times at node points have to be reduced. Therefore, these are also the parameters that should form the focus of a data analysis:

Where do cyclists ride (not ride)?
Where do cyclists ride slower than usual?
Where do cyclists wait for a very long time?

The data analysis carried out should provide information on the above issues. The resulting answers can be specifically transferred to action plans.

Summary in 6 steps

All steps for practical application

Step 1	Define question: There is a question at the beginning of the process. Examples of possible topics are contained in Chapter 3.
Step 2	Vendor selection: An overview of the vendors can be found in Chapter 2. a. Reference of a sample dataset: A sample dataset is useful for assessing the quality of data. An analysis of distortions is thus already possible at an early stage of the project. b. Analysis of user group: The user group represents a critical aspect of the random sampling. Depending on the selected vendor, this can vary extensively and must therefore definitely be taken into account.
Step 3	Provision of data: Decision for a route network
Step 4	Findings of data analysis: Data preparation, exploratory data analysis
Step 5	Representative assessment and adaptation of data: A representative assessment of the parameters taken into account is necessary in order to eliminate systematic errors. This must be observed in particular for distorted samples (e.g. sporty, male app users).
Step 6	Answering the question

5 Verifying the results

The following chapter uses a case study to discuss data acquisition, evaluation and interpretation, in order to alert planners to the possible pitfalls and indicate possible strategies for handling such data. The data verification using Dresden as a pilot municipality was done both qualitatively, via map-matching of bicycle traffic data that had been measured in other ways, and qualitatively using a user survey. Bias due to the fitness-oriented target group can be offset by mathematical and qualitative methods.

5.1 When is the data set usable?

User numbers

As a general rule, the dataset is usable if it provides meaningful, qualitative coverage of the network. This can equally well be the case with a few very active users or with many more sporadic users. If in doubt, it is important to check this with the data supplier in advance. To obtain a meaningful data aggregation buyers should choose a fairly long period, such as year. In the city of Dresden, for instance, for the application set out in this guide (see Chapter 4), very good coverage of the network was recorded with 3,000 users per year, and the correlation with the permanent counting devices was at a very high level. This number of users corresponds to a mere 0.68% of the city's residents, but an average of 22 trips a year were recorded per user. That results in 70,500 journeys within the city limits, around 50% of which were everyday rides.

Cycling performance by users

As a general rule: The less sporty the users' backgrounds, the broader the final picture of the general population. Nevertheless, the number of users opposes the aim for an exclusive evaluation of everyday cyclists. Sporty, active cyclists achieve high-performance rides both when riding for fitness and when simply travelling from A to B. With this in mind, a larger dataset with a higher proportion of fitness-motivated users can deliver more information than a dataset that consists solely of everyday cyclists but contains noticeably fewer trips.

Restrictions on interpretation

Apart from user numbers, when acquiring data it is also important to ensure a good mix of ages and genders. If a homogeneous distribution or concentrations are unavoidable, this always has to be allowed for when interpreting the data. There are no ideal datasets available on the market today. Therefore buyers will always have to accept limitations in one area or another. The important thing here is to evaluate these limitations carefully and to allow for them in the interpretation of the data.

5.2 What can go wrong?

Scope of sampling

There are two main potential sources of error. Firstly, it is possible that the dataset is too small for the buyer's own municipality. Secondly, a dataset such as that offered by Strava may be much less representative of the buyer's own municipality than it is of others. For example, though a dataset consisting largely of fit, young, male users may be at least partially representative of a university town such as Dresden, it would be less meaningful for a city with a significantly higher proportion of older residents. The connection with the GIS data used in a city can also bring problems. Map-matching is a central part of the data handling process, but it can lead to duplication if overly There are two main potential sources of error. Firstly, it is possible that the dataset is too small for the buyer's own municipality. Secondly, a dataset such as that offered by Strava may be much less representative of the buyer's own municipality than it is of others. For example, though a dataset consisting largely of fit, young, male users may

be at least partially representative of a university town such as Dresden, it would be less meaningful for a city with a significantly higher proportion of older residents. The connection with the GIS data used in a city can also bring problems. Map-matching is a central part of the data handling process, but it can lead to duplication if overly detailed basemaps are used. On the other hand, too few digitised links (such as over a bridge) can also lead to a large number of trips not being projected onto the basemap (see section 4.2, step 3). This generally depends on the data provider and it is imperative to clarify this with the provider in advance.

Staff resources

Essentially, it is possible to obtain a variety of depths when it comes to evaluating the data - almost all vendors offer evaluation via an analysis platform on differing terms. Some are designed to be more detailed than others. It is possible, as with Strava, to get minute-by-minute data for individual network links. This allows the buyer to create as many individual evaluations as they like. However, this does take up resources and is actually only useful for municipalities with sufficient staff on hand. It would be a pity if the 'buried treasure' of data could not be uncovered because the obstacles facing its evaluation were too high. Base data is easy to buy, but it is mainly the way the purchaser works with it that turns it into such a valuable tool for municipal bicycle traffic planning. This work takes time, but it quickly pays dividends in the form of data and conclusions that can be used in planning.

6 Outlook: Future developments in the market

Dealing with errors

The data set collected by Strava is, as mentioned at the beginning of this guide, probably a comparatively demanding application. The field of mobile mass data is expected to develop in the future. Its currently rather homogenous random sampling runs counter to the established methods of data collection in traffic planning. Nonetheless, it does provide a large sample size at very low cost in comparison to other methods. Strava's database has two further weaknesses, apart from the user group to which it appeals. Firstly, the results of the waiting time analysis at node points cannot be seen as satisfactory as they stand. However, the vendor is working on this, as well as on differentiated map-matching processes that would work even in more complex areas of the network, thus solving the problem of cyclists being counted twice. This would mean that we would see fewer instances of misleadingly high numbers of cyclists being recorded in areas with a high density of infrastructure. There is also great potential here in other functions such as the option to visualise journeys from a specified starting area combined together.

Alternatives to Strava

The project on which this guide is based aimed to analyse this data set. It is strongly recommended that we remain open to further developments and vendors and regularly check the development of this market (see Chapter 3). For example, the Radwende app from Scholz&Volkmer in particular (see Chapter 3) offers the chance to reach a broad target group and the necessary network coverage via its accompanying campaign. This guide also only represents a snapshot of what is available at the time of publication. In the future it could be expanded by data from other vendors covering sport, navigation, bicycle hire, bonus- and municipality-based apps and other stakeholders. It must be borne in mind that it is not worth using 'ready-made' GPS data if the user density in the area in question is not high enough, so that network coverage cannot be guaranteed. For cases such as this, however, an accompanying campaign can be launched with the collaboration of the vendor, so as to achieve the user numbers required. This kind of service is currently offered by BikeCitizens, for instance: this company is already far advanced in the field of analysis and evaluation. With the publication of cycling analysis tools, waiting times, traffic densities and speeds can be analysed down to individual junctions and segments of road, as well as alternative routes and detours between the origin and destination points. In addition, data collection and evaluation functions are intended to be added to the STADTRADELN campaign in the future. A research project in this topic within the framework of the BMVI's mFUND research initiative has been running since July 2017, aimed at automating data collection and evaluation for all participating municipalities. The first results are due in 2019.

Annex

In the following Annex further technical details are summarized, technical terms will be explained and frequently asked questions will be answered.

Detailed knowledge I: Prevention of overlapping links

To avoid overlapping links it was tried to filter the links in a way, so that every connection is only represented by one link. At the same time it was tried to avoid that routes are not represented by any link anymore. This could occur for example at separate bicycle ways or pathways, which play an important role as gap closures for the non-motorized traffic. Also links that are certainly not passable for cyclists were removed to avoid mismatches. The filter was developed though a definition query in ArcGIS. Attributes which are pointing to public transport or highways were excluded. After further tests with the generated road network, advantages and disadvantages were revealed (see Figure 16 and 17). The network is rich in content, but at the first allocation of traffic to the links, conflicts could still occur. As a consequence the decision was made for the road network export of geofabrik.de. This network is simplified in a way, so that it usually contains only one link per route, which results in a loss of information. Via the identical variable "OSM-ID" missing information can be added later. In the course of the project this was however not necessary. On a city level it is recommended to invest some time in the generation of a net and to adjust it manually if necessary. Table 2 shows a comparison of the analyzed path networks, does however not claim to present a comprehensive overview, but only aims to exemplify the diverse data basis.

Figure 16: Links in GIS based on Mapzen.com

Figure 17: Links in GIS based on Geofabrik.de. The reduced links in comparison to Mapzen.com (Figure 16) are clearly visible

Links in which the bicycle path is very close and parallel to the road network remain an unresolved issue. Because of the inaccuracy of the GPS-data it cannot be distinguished whether the road or the bicycle path was used. Possible wrong assignments are likely due to the inaccurate GPS-position-data. Further investigations are necessary to find an adequate solution for this problem.

Detailed knowledge II: The networks link data

The sample dataset in table 3 shows the time stamp, the amount of users, the users' activity and the time needed. The Edge-ID is explicitly referenced on the indices of the links within the GIS-net basis.

	edge_id integer	year integer	day integer	hour integer	minute integer	athlete_count integer	rev_athlete_count integer	activity_count integer	rev_activity_count integer	total_activity_count integer	activity_time double precision	rev_activity_time double precision	commute_count integer
1	6291272	2014	305	13	53	1	0	1	0	1	3.34325994805218		0
2	6291300	2015	169	18	34	1	0	1	0	1	1.16634166515055		0
3	9749260	2015	316	15	30	1	0	1	0	1	1.23008924058472		0
4	8582488	2014	305	12	51	0	1	0	1	1		1.41511103287362	0
5	9749263	2015	103	16	31	0	1	0	1	1		2.13238128748964	1
6	9749279	2015	237	8	2	1	0	1	0	1	0.354159852233221		0
7	9749282	2015	135	10	8	0	1	0	1	1		12.7332161530339	1
8	6291466	2015	123	11	53	1	0	1	0	1	1.13122306296737		0
9	6291499	2014	315	8	44	0	1	0	1	1		6.20219514092036	0
10	6291548	2015	104	6	28	0	1	0	1	1		1.27923814209955	1

Table 3: Sample dataset SQL for the GIS-net links, Strava 2016

The columns include the following parameters:

athlete_count	Number of cyclists in the direction of digitalisation of the link
rev_athlete_count	Number of cyclists in the opposite direction of digitalisation of the link
activity_count	Number of activities in the direction of digitalisation of the link, one activity can include several users
rev_activity_count	Number of activities opposed to the digitalisation of the link, one activity can include several users
total_activity_count	Total number of activities in both directions
activity_time	Time needed to pass the link in the direction of digitalisation, in connection to the length of the net-element the speed can be estimated
commute_count	Number of trips identified as commutes, either 0 = sport or leisure trip or >0 <= total_activity_count, then commute

Detailed knowledge III: The networks node data

The node data is treated similarly as the link data. Also for this dataset different aggregations are possible depending on the time stamp and the other parameters. Therefore, the index of the node of the GIS-basemap is connected to the indexed parameter node_id.

	node_id integer	year integer	day integer	hour integer	minute integer	athletes integer	activities integer	median_wait integer	max_wait integer	min_wait integer	commute_count integer
1	2752	2015	157	12	9	1	1	1	1	1	1
2	2752	2015	165	21	8	1	1	0	0	0	0
3	2752	2015	181	12	9	1	1	0	0	0	0
4	8254	2015	164	8	35	1	1	4	4	4	0
5	2752	2015	183	19	15	1	1	1	1	1	0
6	11005	2015	158	16	36	1	1	0	0	0	0
7	8254	2015	174	21	21	1	1	4	4	4	0
8	2752	2015	185	11	28	1	1	2	2	2	0
9	11005	2015	159	10	10	1	1	8	8	8	1
10	8254	2015	183	15	48	1	1	3	3	3	1

Table 4: Sample dataset SQL for the nodes of the GIS-net, Strava 2016

As presented in table 4, for the nodes the individual crossings and their waiting times are recorded. The calculation of those parameters and the problems related are covered in more detail in the project report. For nodes, in general the assumption of the node can be seen as a restrictive element. A distinction of commutes and sport and leisure trips is possible, but not mandatory. The parameters are similar to the parameters in the link dataset. The only exception is the waiting time instead of the crossing time for one link element. Different values for median_ wait, max_wait and min_wait are only expectable, if a group of cyclists, in which all are using the app, is crossing the node at the same time. Due to the low distribution of users of the app Strava among the population of cyclists, this is a rare event.

Detailed knowledge IV: How does Strava identify commuter trips?

Strava identifies "commutes" mostly through a "point-to-point"-matching-method. Frequently cycled relations between origin and destination are classified as regular and therefore as a commuting or everyday trip (Strava LLC, 2016). This method identifies about 98% of the commuting trips that are marked as those by the users and a large quantity of trips that are not marked by the users, but still fulfill these criteria. This approach does not meet the formal European definition of a commuting trip in the sense of a regular trip between place of residence and work- or training place. Also included are regularly occurring trips, such as popular training routes of race cyclists or stages of cycle races. It seems like the company tries to classify as many everyday trips as possible and to accept the possible 'noise' in favor of a higher number of commuter trips. This should be considered, when interpreting the data of the commuter trips.

Detailed knowledge V: Data basis for the origin-destination-matrices provided by Strava

In this project the standardized European grid system for official statistics was used. Regular squares (polygons) with a length of 1000m were built and the Strava data from the corresponding area was then projected onto this network. This data is corresponding with the standard of the new statistical polygon network. The projected data on the polygons have the structure pictured in table 5.

	polygon_id integer	year integer	day integer	hour integer	minute integer	commute integer	dest_polygon_id integer	intersected_polygons integer[]
1	339081	2015	191	7	30	1	344108	{339528,339970,340403,340404,340824,341243,341658,342069,342477,342478,342888,343297,343703,339081,339082,339527}
2	340399	2015	223	16	58	1	341238	{340399,340819,341238}
3	340403	2015	224	16	5	0	340403	{340404,340405,340825}
4	341648	2015	328	17	42	0	341648	{341648,342059,342467,342058,341647}

Table 5: Schematic presentation of the dataset polygons_ride, SQL-database extract

The columns include the following parameters:

Polygon_id	Start polygon of the cycled route, clear value for polygon designation, consecutive numbering within the project
year, day, hour, minute	columns representing the time stamp
commute	1 = commuter trip, 0 = leisure/sport trip
dest_polygon_id	Destination polygon of the cycled route, clear value for polygon designation, consecutive numbering within the project
Intersected_polygons	Labels the polygons crossed between start and destination

Start- and destination-polygon can have the same value if it was a round trip. Besides presenting the most common origins and destinations of cyclists, length of trip and the used corridor can be estimated. Thereby the link length of a polygon is multiplied with the number of crossed polygons. The length of the trip might be overestimated in some cases, but it still can be approximated in

Detailed knowledge VI:
Approaches for an analysis of representativeness – 1. Traffic volumes

The analysis of representativeness compared counting data in Dresden with the Strava-data. This included data from permanent counting devices, as well as data from manual short time counts.

Whereas the counting data is available for every hour, the realized trips in the Strava dataset are distributed to a lower ratio for the particular net elements. It can be the case, that for several hours of a day no cyclists are recorded. This makes it necessary to look at the distribution of cyclists over a longer time period. Because the dataset was available for a longer time period, the road network of the OpenStreetMap-map was used (see chapter 4.2, step 3). The basis for the analysis are the traffic intensities of all cyclists of a link, that are exact to the minute.
In the course of the data preparation, the following steps took place:

1. *Coding of a time- and date-variable*
2. *Combination of the cyclists of more than one parallel links and exclusion of duplicates with the criterion "several cases per minute on two parallel links"*
3. *Aggregation of data from minutes to hours*
4. *Addition of the value 0 for hours without any cyclist*

The comparison between GPS-based data and the permanent counting devices is done for exactly corresponding time periods. The whole time period, for which data of the permanent counting devices are available is included in the analysis.

At first the results of the comparison of the permanent counting devices will be presented. Data from six permanent counting devices in the state capital Dresden in the time period between 01/09/2015 and 31/05/2016 were compared with the GPS-data. The daily traffic flow based on the permanent counting devices is depicted in Figure 18. The traffic volume per day is between 400 and 2000 cyclists. Differences can be explained with the function of the respective road in the network. The sample size in the Strava dataset is smaller. The sum of cyclists in the nine months period is between 500 and 5500 cyclists. The resulting daily traffic flow is depicted in Figure 19.

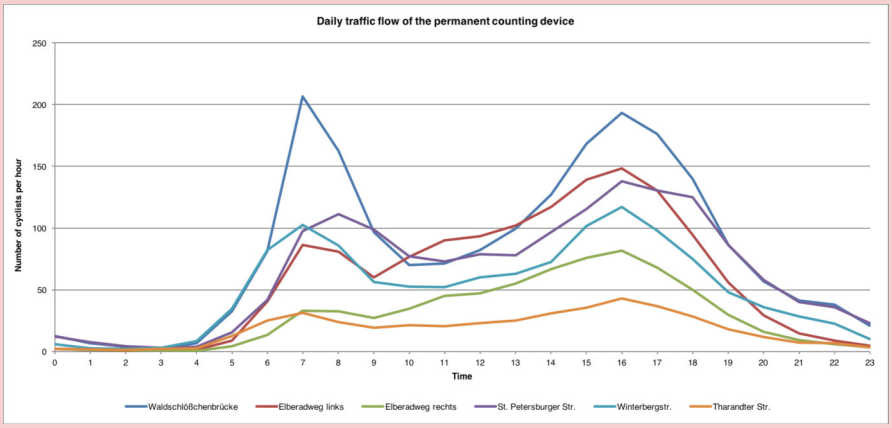


Figure 18: Daily traffic flow of the permanent counting devices in Dresden in the time period between 09/2015 and 05/2016

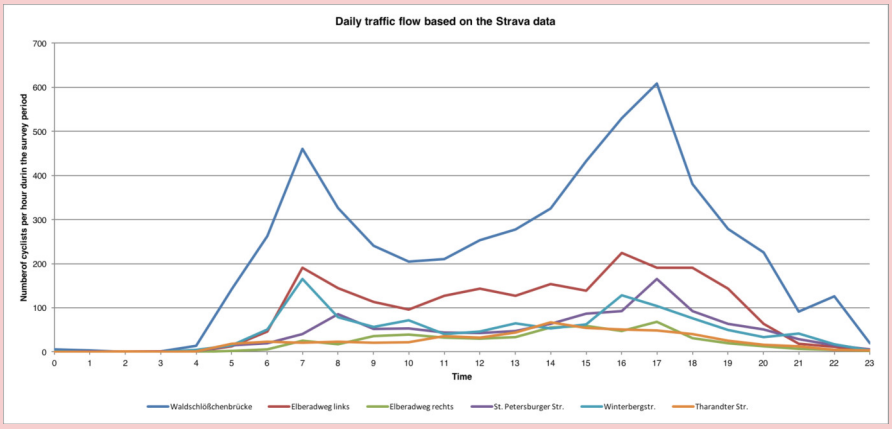


Figure 19: Daily traffic flow based on the Strava data in Dresden in the time period between 09/2015 and 05/2016

Detailed knowledge VI continued: Approaches for an analysis of representativeness – 1. Traffic volumes

The projection of the Strava-data was done with two different approaches. Departing from the time period considered in the analysis, the projection is based on the projection of Strava-data for a single, average day to the values of the permanent counting device data for a single, average day. At first, with the help of several linear regressions the sample of the traffic volumes was projected on the respective permanent counting devices in Dresden. Afterwards, out of the regression equations a single regression for the whole city of Dresden was derived (see Figure 20). The regression uses the Strava-data as output data (x) and derives the daily values f(x). The regression equation is f(x) = 0.5087 * x + 16.9719.

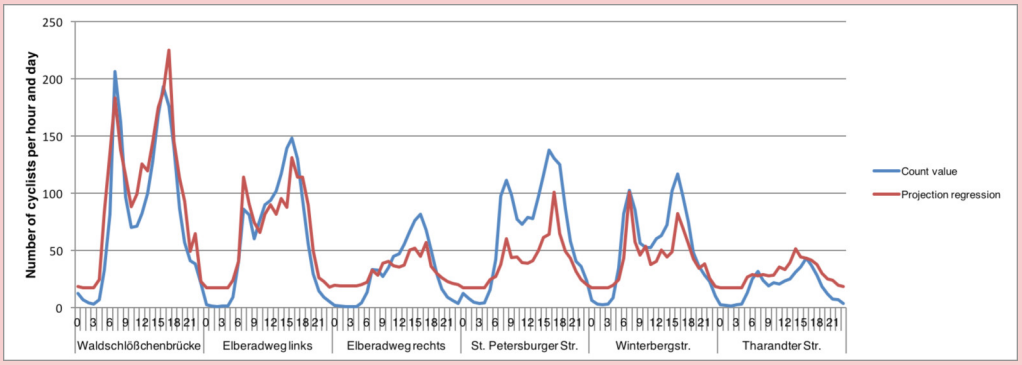


Figure 20: Projection of the Strava-data on the level of the permanent counting devices in Dresden for the average daily traffic flow, based on a linear regression

The second approach is the use of a constant factor. The factor is based on the means of the quotient of the values of the permanent counting devices and those of Strava on an hourly basis (see Figure 21). The arithmetic mean is 0.00371 and this value is used for the prognosis in Figure 24. The median is 0.00306, minimum and maximum are 0 and 0.1449 and the standard deviation is 0.00262. The projection of annual values to daily values results in a quotient of 1.3. The bicycle traffic volume of one year in the Strava dataset can therefore be scaled down with the factor 0.76 on the average daily traffic volume (DTV). The same method results in the factor 0.77 for the Berlin permanent counting devices. Noticeable is the deviation at the site Waldschlösschenbrücke (Waldschlösschen bridge), when the mean is used for projection. The projected traffic volume is in sum about twice as high as recorded at the permanent counting site. A possible source of error may lie within the map matching process. The bridge consists of four digitalized axes. With the search for duplicates the data was already reduced to half of the initial cases, however the methodology at parallel axes seems to entail elements of uncertainty. Because of the proximity to a junction the possibility of turning traffic cannot be completely excluded either in the Strava dataset. Another reason could be the values of the permanent counting device. Those are plausible on the daily and annual traffic flow, but a systematic underrecording cannot be disregarded, as the bridge can also be accessed without crossing the counting device through a staircase from the Elberadweg (Elbe cycling track).

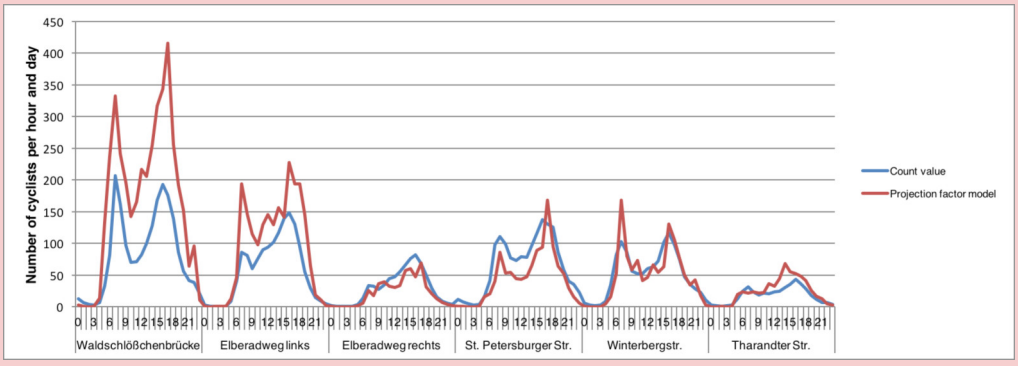


Figure 21: Projection for the permanent counting devices based on a constant factor, derived from the quotient of the values from the permanent counting devices and those of Strava on an hourly basis

While evaluating the methods for projection the following aspects should be considered. Within the linear regression of the form
$f(x) = ax + b$
there is always a constant value (here: b = 16.9719). Based on this projection a minimum traffic volume of about 17 cyclists per hour is expected, even on roads without any Strava-users. During night times this effect becomes clearly visible in Figure 23, comparing the difference between the red and the blue lines. This is hard to explain. Statistical parameters for forecasting reliability are a quantitative evaluation of the forecasting reliability of the bicycle traffic main net in Dresden (see Table 6). Overall, the projection via the constant factors performs better at the statistical parameters, because the mean absolute error and the relative error have lower values.

Table 6: Statistical values for the comparison of the projections based on a linear regression and a constant factor in Dresden

Statistical value	**Linear regression**	**Constant factor**
Relative error	0.363	0.023
Correlation	0.869	0.866
Explained variation (r^2)	0.76	0.76

Additionally, the chair of Traffic Ecology of the TU Dresden performed manual traffic counts in the time period from 22.05.2015 until 12.06.2015, Tuesday till Thursday from 2pm until 6pm. The sites were selected with regard to the importance of traffic, parallel routes to the main roads and suspicion of high frequency of sport cyclists.

The ratio of counted cyclists to Strava users in the same time period is 7309 : 7. It can be concluded that only 0.096 % of the recorded cyclists within the four hour observation periods used the Strava app. Statistically sample and population therefore are not the same. There are different reasons for the deviations between the two data sets. Whereas the Strava-data is based on a whole year period, the temporary manual counts were done in May and June under good weather conditions. The high numbers in the counting data might be due to a seasonal effect.

To conclude, at this point a constant factor, calibrated on the mean hourly values of permanent counting devices, is favoured for a projection of the data onto the population in the network. This method seems to represent the actual configuration in the cycling network best. Especially in the subnet deviation is still to be expected.

The use of temporary manual counting data as a comparison, increases the range of factors for the individual counting sites, resulting in an inaccurate mean projection factor and a lower forecasting reliability. The same is true for data based on a traffic model. This was tested with data in the city of Leipzig. Statistically there are two independent samples, even though the data in the traffic model in itself are dependent secondary data. Based on that a useful projection was not possible.

Detailed knowledge VI continued: Approaches for an analysis of representativeness – 2. Speeds

Because of the rather sport-oriented, male app users a difference in speed between the Strava-sample and the population 'cyclists in Dresden' was expected. This difference was confirmed via extensive comparative measurements (n = 1,000, measurement at five sections across the whole city of Dresden). To derive a useful sample size, the comparative values within the Strava-dataset were only adjusted for the time slices, not for the concrete date to the measurement specifics. This is due to the large sample size: a statistical comparison needs same sample sizes. Also the choice of a low form of disaggregation influences the use in the communal praxis. In this case this would be a simple mean in a certain observation period. It needs to be stated, that the values of the Strava-sample are not measured values per se, but estimated values, derived form the time needed to pass the according link in the GIS-net.

The GPS-data on average shows a higher speed as in the comparative measurements. Table 7 depicts this trend with the means of the speed in the Strava-sample and in the comparative measurements. The difference differs on a small scale. There are outliers in the sections Zellescher Weg R2 and Grundstraße R2. The section Zellescher Weg has the lowest measured average speeds. This is likely due to the high frequency of students, that are realising many short origin and destination traffics at the section, as the campus and the university library are nearby. The section Grundstraße is a route with a steep slope (length overall 2.9km, 124 metres high, average slope 4.2%, max. slope 7%). Determining for the interpretation of the data is not only the mean of the driven speed, but also the distribution. The distribution especially across even road segments is similarly normally distributed (see Figure 23).

Table 7: Means and standard deviations at the examined sections based on the GPS-data (Strava) and comparative measurements

Section	Mean value (in km/h)		Standard deviation (in km/h)	
	Strava	Measurement	Strava	Measurement
Elberadweg WSB R1	26.07	21.17	4.18	4.72
Elberadweg WSB R2	26.25	22.08	5.12	4.70
Elberadweg ALB R1	26.26	21.70	5.06	3.72
Elberadweg ALB R2	26.29	20.20	4.60	3.41
Zellescher Weg R1	26.33	21.56	5.73	5.00
Zellescher Weg R2	27.00	19.34	5.74	4.20
Chemnitzer Straße R1	29.25	23.56	6.07	5.06
Chemnitzer Straße R2	21.14	21.30	5.59	4.46
Grundstraße R1	41.89	33.68	8.41	5.55
Grundstraße R2	17.17	16.13	4.38	4.44

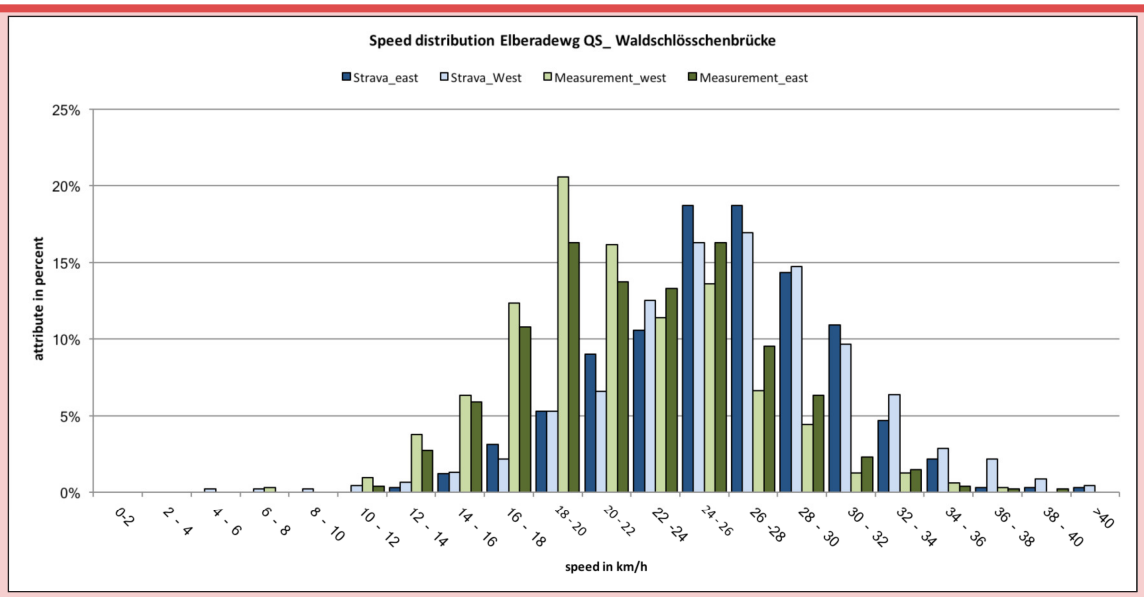


Figure 22: Average daily speed diurnal variation times and lines based on the Strava-data in Dresden in the time period 09/2015 - 05/2016

Based on the aggregated distributions for all measurement sections a mean deviation of 5.5 km/h across all sections was determined (see Figure 22). The datasets also show a high correlation in statistical analyses. For practical planning, the speed in the app-data can roughly be reduced by 5.5 km/h to derive the speed level of average cyclists in Dresden, achieving a higher explanatory power of the data (see Figure 23).

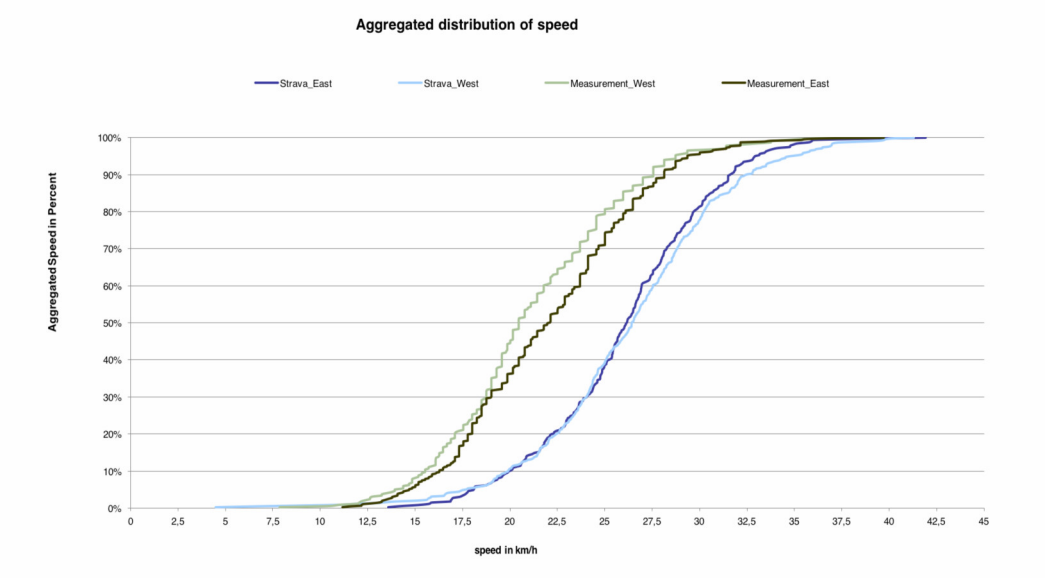


Figure 23: Aggregated speed distribution of the cyclists in Dresden in the main net, section: Elberadweg (n = 316/473)

A normally distributed sample allows for an analysis of regression of independent size. These requirements are met in the main net, as well as in the subnet for even measurement sections. A regression analysis is possible. This approach was chosen, because a correction-constant (-5.5 km/h) can point to the direction of correction, but not necessarily delivers a high quality of results for all sections. In this case the sample of app-users is the independent value and the measurements should be estimated via the regression equation in the future. A simple linear regression was calculated. For the purpose of simplification all other parameters were assumed to be equal and therefore negligible. Further research in this area may improve the quality of results. This would however require extensive field experiments that were not pursued in the frame of this research project.

The regression calculation resulted in a Pearson correlation of 0.922 and a determination coefficient of R2 = 0.85. This means, that 85% of the variation of the real speeds can be explained with this linear model. The accuracy of the model is therefore high. In this case, based on app generated GPS-data, the actual cycling speeds of cyclists in even areas in the city can be predicted.

Detailed knowledge VI continued: Approaches for an analysis of representativeness – 3. Origin-destination-matrices

In the previous sections origin-destination-matrices were often cited as a possible form of presentation. In addition, these data may be influenced and distorted by a heterogeneous sample. Therefore, at first the regional distribution of the originating traffic is validated with the population of the corresponding originating traffic-cells (see Figure 24). The population data is based on the census of 2011. This approach is limited as originating traffic can for instance also start from the workplace. Further analysis is therefore recommended. For this study the workplace data was however not available. With this limitation the statistical analysis of the originating traffic data is only partly valid, but delivers a first overview of the quality of the data.

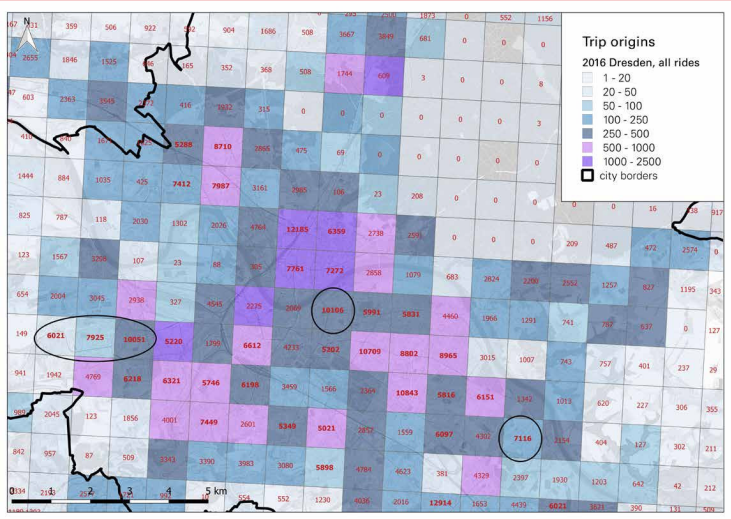


Figure 24: Comparison of the originating traffic (every-day traffic) with the population in the city of Dresden (source: census 2011)

Some originating traffic cells and the number of inhabitants seem to be highly connected. This finding is however not generalizable for all originating traffic cells. Some originating traffic cells with a high number of inhabitants and a rather low social background display a lower number of trips compared to the general findings. An influence by the homogeneous user group can be suspected here, resulting in a social exclusion of inhabitants with a low level of education or low income. At this point sociological studies are recommended.

The statistical analysis results at basic level in a high positive correlation of the values number of inhabitants and start_activity_count, which describes the number of journeys. The whole data set was analyzed in the time period 01/2015 until 06/2016. It can therefore be concluded that a connection between the values exists.

The value output for the cities Dresden, Chemnitz, Berlin and Leipzig in Table 5 shows, that this correlation (r) is locally differently distributed and stable at around 0.7.

Table 8: Correlation of the values origin of trip and population for the examined cities

Location	Slope	Correlation r	Coefficient of determination (r^2)	Adjusted r^2	Covariance	Number of datasets
Dresden	0.103	0.689	0.474	0.472	583.818	394
Chemnitz	0.050	0.672	0.452	0.448	153.053	276
Leipzig	0.0617	0.702	0.494	0.491	407.306	361
Berlin	0.060	0.678	0.460	0.458	1,110.467	1,017
Gesamt	0.052	0.660	0.435	0.435	159.337	377,048

The correlation is however strongly influenced through areas with a low number of inhabitants and journeys. A filtering for journeys greater than 100 and number of inhabitants greater than 0, leads to a correlation of 0.6. In this case still a positive connection exists however.

Detailed knowledge VII:
Further possibilities of evaluation of the datasets

The autonomous work with the GPS-data enables the users to create own visualizations, such as differential nets, with a PostGIS-extension for PostgreSQL-databases. Those visualizations can then continuously be integrated into the online-topic-map (see Figure 25). Also the plausibility check of the obtained data is possible.
An alternative is the contracting of the data preparation to an external engineering office. Especially the formats of the results have to be defined carefully beforehand in this case. The data delivery also takes place according to predefined criteria – Strava for example offers monthly data, as well as retrospective data for a whole year. From the authors' point of view, a monthly dataset is at this point not meaningful enough for many purposes, for example the quick verification of an applied measure, because of the low numbers of users compared to the general population.

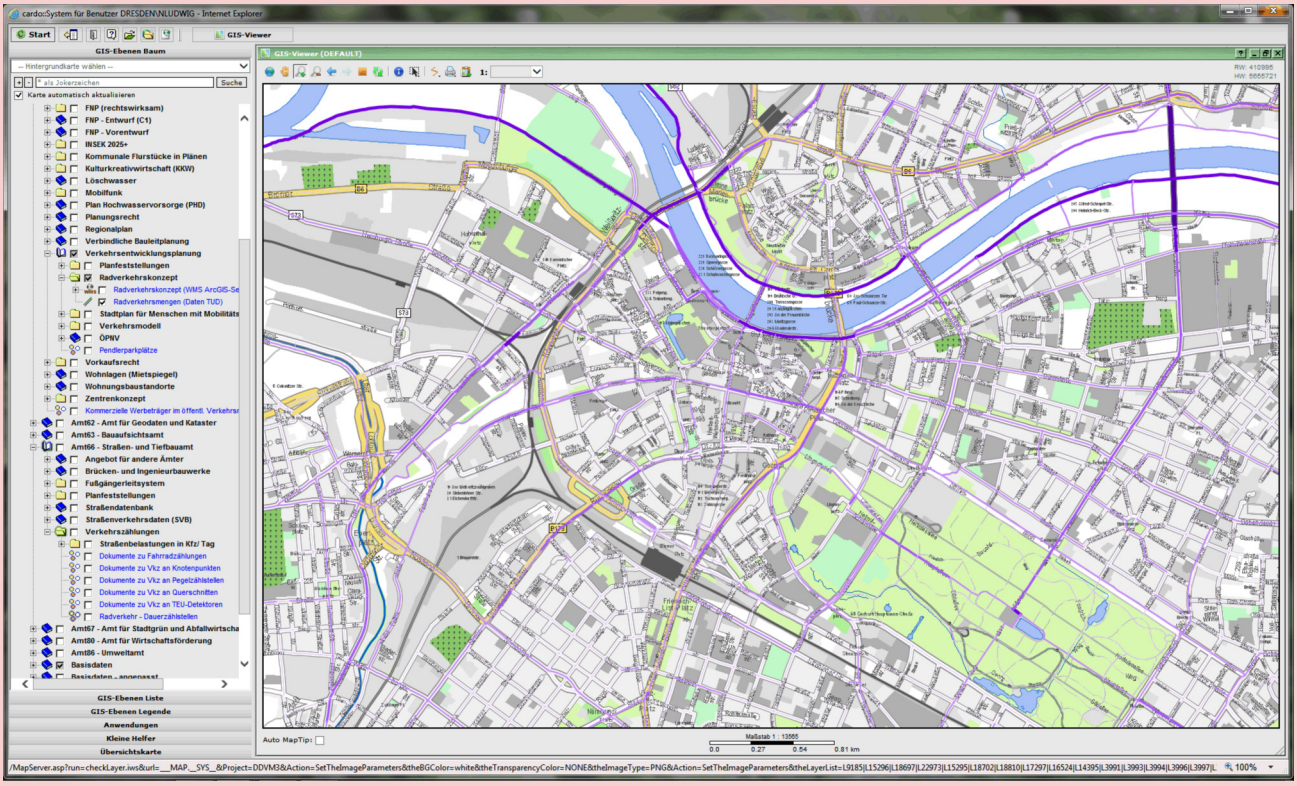


Figure 25: Online-topic-map for the city of Dresden (source: city of Dresden)

App	Apps or applications, are application programmes for smartphones- In this guide the focus is on applications, that enable the recording and storage of GPS-data.
Databases (SQL, PostgreSQL)	Relational databases serve the purpose of data management in computer systems and are based on a table based database model. Structured Query Language (SQL) is a database language for the definition of data structures and for the editing and request of data files. PostgreSQL is a free database management system, that is mostly conform with the SQL-standard. It has an extension for the management of geo-data (PostGIS).
Deflected traffic	Traffics that because of a measure or an event in the road network which increases the resistance or the travelling time, do not take place on the usual route, but take an in fact inferior or undesirable alternative route.
Differential nets	Differential nets describe a visualization in geo information systems, which show the same parameter in a net, for example at two different times. This visualization is usually used to depict developments.
GIS	Geographic Information Systems (GIS) are information systems for the collection, editing and organisation, as well as the analysis of data with a territorial dimension. Also data of the road network is for example filed and maintained in such information systems.
GIS-element	In a road link network, the network graphs are made of the elements link and node. A presentation of traffic flows needs a consistent and gap-free network. Single elements can carry different attributes, such as allowed maximum speed or rights of priority.
GPS-data	GPS-data are a synonym for position data determined via the Global Positioning System (GPS) and other services like GLONASS or Galileo. A global system of navigation satellites is used to determine the position. The satellites continuously transmit their position and the exact time. Specialized receivers can calculate the position and speed thereafter. The data include the longitude, latitude and the time. At the moment a horizontal accuracy of 7.8m is guaranteed.
Link data	See also GIS-element. Link data summarizes all attributes of a link in the network. This can include for example the speed limit, the road classification, the availability of bicycle traffic installations and cycling traffic volumes.
Node data	See also GIS-element. Node data summarize all attributes of a node in the network. An attribute is for example the type of crossing roads or the resulting link type. Also waiting times for traffic flows can be attributed.
Mapmatching	Mapmatching is a method that compares the position derived through localization, for example via GPS, with data on a map. In this project mapmatching is the comparison of GPS-data and a GIS-network and their assignment.
Metro Extracts	Metro Extracts are a product of the company Strava. It is an online-platform offering data on traffic volumes, waiting times and other informations for the communal use.

MiD	Mobility in Germany is a nation-wide survey of households on the traffic behaviour of citizens. The survey takes place every five years. The results of the most recent survey are expected to be available in 2018.
MIT	Motorised individual transport describes cars, light commercial vehicles, motorcycles, mopeds or other comparable motorised means of transport, owned by a private person and predominantly used for private mobility.
Subnet	The subnet in cycling traffic is not necessarily the subnet or comparable classifications in categories of the network as for the MIT, but elements of the network with secondary importance in the network of cycling traffic. Those can be side streets, but not necessarily.
OSM	OpenStreetMap is a free project that is based on an openly accessible database of all free geo information. With these data the maps used in this project can for example be produced. The use is free of charge.
Point-to-Curve Matching	Point-to-Curve-Matching describes a method of assignment of measured position data to a map, using the shortest distance of the position point to a curve (link).
Point-to-Point	See also Map-matching, Point-to-Point matching describes a method of mapmatching that assigns GPS-data-points to points of the available transport network. For a network graph, this can be nodes, start or end of a link or generated points along a link.
Origin-destination-matrices	Origin-destination-matrices describe parameters that are realised with a change of place from an origin to a destination. This may be traffic volumes, different types of transport or temporal expenditures.
Origin-destination-relationships	See also origin-destination-matrices. Origin-destination-relationships describe the local connection between an origin (start) and a destination (end) of a change of location. That is not necessarily a track.
Reliability	Reliability is a measure of formal accuracy of the measurements.
Representativeness	Representativeness (of a random sample) is the characteristic of specific types of data collection, that allow for conclusions for a population drawn from a small sample. The projection of age, level of education and marital status need to resemble the official statistics in their characteristics if personal data is used.
Route data	Route data describe the sum of all GPS-points and their corresponding speeds in a usually linear sequence, starting from the source, over a change of location to the destination. Route data can also include information on the reason for the trip, the person or other information.
SrV	The system of representative traffic surveys of the TU Dresden is a traffic data collection in the city traffic taking place every five years. The last period of data collection was in 2013.
Communal geo information systems (AKTIS)	The official topographical cartographical information system is built for the purpose of digitally collecting the results of topographical mapping and official topographical maps of the land surveying offices and the Federal Agency for Cartography and Geodesy.

Average daily traffic	The average daily traffic describes the distribution of the hourly traffic volumes of a modus at a section. During work days peak times are in the morning (around 7:00 – 9:00 o'clock) and in the afternoon (15:00 – 17:30 o'clock).
Validation	The validation describes the verification of the user input or statistical data on the suitability for specific purposes.
Traffic cells	A traffic cell is a theoretical unit of space representing a part of a city or a community. Traffic cells can be used as a unit of reference for modelling. In the process data such as ownership of a car, the population, the number of jobs or others is assigned to the cells.
Time slices	Classification of a day into time units. In this project usually hourly time slices were used.

What kinds of GPS data is there?
GPS-data are position data and can therefore depict the movement of cyclists through space. Possible data sources are tracking apps for smartphones (for example GPS-tracker), sport apps (for example Strava) or navigation apps (for example BikeCitizens). The frequency of data recording is important. Some bike rental systems also have a recording system for GPS-data. The intervals of recording are however large (usually only the start and destination). This makes the understanding of the used routes harder or even impossible (start = destination).

What does the data contain?
After the matching to the used GIS-net the GPS-data includes usually a data set for the nodes of the net and a data set for the links. The data set for the links includes traffic volumes and speeds, the nodes' data set additionally contains waiting times. Moreover, origin-destination-matrices can be calculated. For this, a grid needs to be delivered additionally to the GPS-data set. The data can be individually prepared for the cities depending on the provider. At least distinctions for day of the week or time slices are deliverable.

What can be accomplished with which data?
There are different kinds of data, based on different user groups or types of measurement (see chapter 2.2). Data based on bike rental systems offer a diverse user group, but also limited information on the purposes of the trip, or rather represent the use within trip chains, which complicates the interpretation. Data based on navigation apps can be seen as more representative. The reasons lie within the somewhat more heterogeneous sample (for example a slightly higher proportion of women) and a less sport oriented user group. This considerably increases the comparability with the average cyclist. The target group is also large, even though their motivation needs to be sustained over a long time period through well designed campaigns. This is not the case for the sport app users. Those users are mostly male and show a more sportive cycling behaviour. Adjustments are therefore necessary. Because of the intrinsic motivation of the users, even without accompanying campaigns a high number of participants can be achieved resulting in a good network coverage.

What needs to be considered when working with GPS-data?
There are many aspects to be considered when working with GPS-data. The most important aspects are the selection and the verification of the GIS-net. Especially when planning a continuous data supply, the same network basis is needed, because the comparison within the time periods otherwise becomes complicated. Additionally, the goals of the data acquisition and analysis and the existing basis should be clear beforehand, so that the most suitable offer can be chosen. When a strong cycling community is present a sport oriented data set can be chosen over a navigation based approach for example. For further information see also the question: "For which purposes can the data be used?".

Is the data representative?
The classical form of a statistical representativeness within the population of cyclists is not achievable with GPS-data. The reason lies within missing relevant target groups such as older aged cyclists or children. Still under certain aspects and with some corrections a data set generated through users of sportive apps can deliver a good picture of the cyclists traffic behaviour. In the end, the issue of a missing representativeness based on the population of cyclists is also a point of critique with permanent counting devices.

For which purposes can the data be used?
The data is suitable for the following aspects in cycling traffic planning:
- Target grid planning (with limitations)
- Planning of measures
- Evaluation of cycling measures

- Marketing
- Visualisation of cyclists

See also chapter 3 in this guide.

Which problems concerning privacy might come up?

Data security is in most cases fulfilled by the data vendors. The prominent offers do not deliver individual route data, but aggregated traffic volumes on the nodes and links of the GIS-net that was provided. Additionally, before the aggregation in the course of the mapmatching the data vendors cut off the first and last 100m of the tracks in the first step of data processing to complicate a reference to the single user. The aggregated final product therefore makes a tracing of single users in the final data set impossible. The data protection law of the federal states is formally not applicable any more.

How to cope with the free input of data, for example age or gender through the users, or how can it be ensured that the data stems from cyclists?

Comparable to household surveys the principle of trust in a large sample applies here. There is no guarantee for the correctness of the data. There are however also no advantages for the users through incorrect data input. There will still be users deliberately providing wrong data. The Strava-data set of the city of Dresden included a 90-year-old sportive cyclist. The correctness of this input is doubtful, however considering the amount of several thousand users, it is not a serious issue. The same applies for a divergent choice of mode. The data can also be recorded through a use of different modes. This becomes apparent in the course of the mapmatching (for example at railways) and a maximal average speed can also be defined together with the data vendor.

Who are the app-users?

Overall, besides the smartphone ownership, there are no barriers of use. The user group of the tested apps (BikeCitizens and Strava) was still dominated by men, which was more pronounced at Strava than at BikeCitizens. Furthermore, the age categories are distributed similarly, with a slightly more pronounced group of 20- to 40-year-olds. Looking at the cycling traffic behaviour, a clear difference because of the diverging motivation of the users can be expected. The sportive users cycle much more and faster than the average cyclists.

Can I buy the data?

Yes, at the moment three commercial vendors are known of: The sport app vendor Strava, BikeCitizens and Scholz&Volkmer. The modalities for the data purchase are different. Whereas Strava at this point bases their pricing per user, BikeCitizens and Scholz&Volkmer base their pricing on the requested offer. This might include an accompanying campaign, like a BikeBenefit program, navigation or a special feature in the data analysis.

Do I get single routes and how many do I get?

Because of privacy concerns route data is not delivered. Only aggregated data based on all available routes are delivered. Moreover, there is the possibility to obtain an extended origin-destination-matrix. The matrix allows for the analysis of route corridors. More information on this can be found in the subchapter "privacy". The number of route data is depending on the app distribution on the community. Key figures in the city of Dresden and the vendor Strava are about 3200 users with around 75200 trips per year. The number of users and trips is not influenceable in this case, opposing to the product of BikeCitizens. Here, a collaboration with the community is intended to gain a user basis.