When Traffic Hits: A geospatial Big Data Analysis of Berlin's Road Network with Applications

Master's Thesis submitted

to

Prof. Dr. Vitaly Belik

Prof. Dr. Ottmar Edenhofer

Humboldt-Universität zu Berlin

School of Business and Economics

Institute for Statistics and Econometrics

by

Ben Thies

(604409)

in partial fulfillment of the requirements

for the degree of

Master of Science in Statistics

Berlin, September 30, 2021

Acknowledgements

I wish to thank Saskia & my family for everything.

I would also like to thank my colleagues at the Mercator Research Institute on Global Commons and Climate Change, especially Nicolas Koch, Alex(ander) Rohlf, and Nolan Ritter for the opportunity to work at such an inspiring place, including me in their team, and lending tremendous support while I wrote this thesis.

I am grateful to the *Studienstiftung des Deutschen Volkes*, which awarded me a scholarship in 2017 and has been supporting me since. It was a great privilege.

Finally, I would like to thank my advisors, Prof. Dr. Vitaly Belik (Freie Universität Berlin) and Prof. Dr. Ottmar Edenhofer (Technische Universität Berlin).

Abstract

Using comprehensive network and traffic data, we perform an analysis of the Berlin road network. Demonstrating the the usefulness of approaches from computational network science in economics, we use an OSM road network representation to identify potentially critical road segments and build a zero-inflated Poisson regression model to predict accident counts. Acording to our model, a 13% reduction in traffic, as observed in 2020 compared to 2017, leads to an unproportional 16% reduction of accidents, making up for 1.3M € in economic savings. Extrapolations of the data to all traffic accidents reveal highly precise estimates of the number of accidents as well as their associated economic cost.

Contents

List of Abbreviations

API	Application Programming Interface
COVID-19	Coronavirus Disease 2019
cp.	compare
e.g.	*exempli gratia* (for example)
GHS(L)	Global Human Settlement (Layer)
i.e.	*id est* (that is)
kJ	kilojoule
km(/h)	kilometres (per hour)
m	meter
μg/m^3	micro grams per cubic metre
OSM	OpenStreetMap
RSI	Road Segment Congestion Index
UC(D)	Urban Centre (Database)
UN	United Nations

List of Figures

List of Tables

1 Introduction

Real-world networks, especially spatial networks, have become a popular unit of study with many applications in fields as different as biology (Bullmore & Sporns, 2009; Mason & Verwoerd, 2007), sociology (Garton, Haythornthwaite, & Wellman, 1997; Mislove, Marcon, Gummadi, Druschel, & Bhattacharjee, 2007; Scott, 1988), computer science (Alderson, Li, Willinger, & Doyle, 2005; Zegura, Calvert, & Donahoo, 1997), or physics (Albert, Albert, & Nakarado, 2004). In urban geography, road and mobility networks emerge as a particularly palpable unit of analysis. As the field of (computational) network science has been growing steadily, much work has been undertaken to describe the mathematical properties of the latter, but only seldom was the focus of economic nature. Even research concerned with the topological and structural properties of city road networks rarely discuss economic implications. Yet one intricate property of road networks that is beyond their mathematical description is that they are designed to accommodate traffic.

Traffic affects nearly everyone, everyday, and so the economic interactions of road networks are numerous. One of the more obvious problems in this matter relates to congestion: can we learn from a network's mathematical properties where it is prone to congestion (Amézquita-López, Valdés-Atencio, & Angulo-García, 2021) or bad air quality? Can we track the drivers responsible for the most road traffic on certain roads (P. Wang, Hunter, Bayen, Schechtner, & González, 2012), given traffic data is available? Another problem of high traffic is an increased probability of traffic accidents. Here, as well, methods from network science can help us answering questions such as: can we identify critical road segments that are particularly prone to accidents?

In this work, we aim to expand on the last question. We draw from comprehensive traffic and network datasets to (1) describe Berlin's road network and its graph-theoretic properties, (2) use the spatial graph representation of Berlin to classify roads into a spectrum of potential congestion proneness, and (3) perform a pilot case study in which we relate the traffic on critical road segments to the number of traffic accidents, and calculate the economic savings resulting from the reduction in traffic caused indirectly by the COVID-19 pandemic.

Using high-resolution traffic and accident data and an open-source graph representation of Berlin's traffic network, we combine a computational network-theoretic approach with an applied economic analysis to exemplify their potential synergies. Notably, with this work, we contribute to the present state of research in the following ways: we

- use OSM (OpenStreetMap) data to systematically analyse Berlin's road network,

- provide an implementation of a road speed congestion index using OSM data, and
- demonstrate the usefulness of approaches from computational network science in supporting and enable certain economic analyses.

Moreover, all source code is released in a public GitHub repository[1], including easy-to-read notebooks with all calculations, data preparation steps, and code for figures, which can be readily adopted for the use in related projects answering different research questions.

Results indicate that computational network analysis can aid in answering economic questions: based on the traffic data at hand and a predictive zero-inflated Poisson regression model, we estimate that 127 traffic-relevant accidents were avoided due to less traffic on Berlin's roads, accumulating approximately 1.3M € in economic savings. Further analysis shows that simple extrapolation yields surprisingly precise results: when we estimate the total number of traffic accidents for 2020, our result underestimates the real number by only 2.86%, while our estimated cost for these accidents underestimates the real cost by 4.61%.

The remainder of this thesis is structured as follows: Section 2 provides an overview of the literature on spatial networks and mobility. Then, the data and technology used are introduced in Section 3. Subsequently, network-theoretic aspects are discussed in Section 4, where a primer in graph theory is given (Section 4.1), along with a descriptive and comparative analysis of the Berlin road network (Section 4.2). Section 5 presents the case study. We describe its methodology in Section 5.2, including the identification of critical road segments, and how we model accident frequency based on traffic volume. In Section 5.3, empirical results are discussed, section 6 concludes.

[1] `https://github.com/thiesben/thesis-when-traffic-hits`

2 Literature

2.1 Spatial and Road Networks

While the study of networks has a long history and was particularly pronounced in sociology and graph theory, modern network science emerged as late as the first decade of the 21st century. Network science combines quantitative aspects (especially graph theory, statistics and statistical physics), computational aspects (efficiency of algorithms, data management, software) and empiricism, as its primary focus is on application (Barabási & Pósfai, 2016). Here, we focus on a special kind of such networks, namely road networks as a sub-category of spatial networks.

Spatial networks today are found, e.g., in transportation, infrastructure, and biology, as in neural networks. Early comprehensive work in spatial networks that already discusses tools for analysis and possible models is provided by Haggett and Chorley (1969). Kansky (1965) introduced measures to characterize roads and highways, and Taaffe, Morrill, and Gould (1963) proposed a model to describe road network evolution in cities (Barthelemy, 2018). These studies were based mostly on graph theory and had their foundation in the discipline of geography. According to Barthelemy (2018), these early studies were constrained by the computational power of their time, and by not taking into account topological aspects of spatial networks. Thirty years later, when Watts and Strogatz (1998) published their seminal paper on small-world networks, and Barabási and Albert (1999) discovered scale-free networks, network analysis as a tool to describe all kinds of complex systems spread heavily across disciplines (Barabási & Pósfai, 2016; Barthelemy, 2018). The scale-free property is exhibited by functions $y = f(x)$ whose slope in the log-log plot is the same over all x, meaning that whatever range of x values one is looking at, one will always observe the same proportion of small to large y values. Scale-free networks are characterised by their degree distribution: most nodes will have very few links, while a small number of nodes will have a larger number of links (cp. Figure 7 (c)). Small-world networks have two important properties: first, they typically display high degrees of clustering and second, they have small characteristic path lengths (see Section 4.1)[2]. But despite these developments, the aspect of space had largely been ignored (Barthelemy, 2018). Due to the importance of space for real-world networks, especially because "there is a cost associated to the length of edges which in turn has dramatic effects on the topological structure of these networks" (Barthélemy, 2011,

[2]Hence the name by analogy with the small-world phenomenon, which postulates that, on average, anybody is separated from anybody else by only six people (Travers & Milgram, 1969)

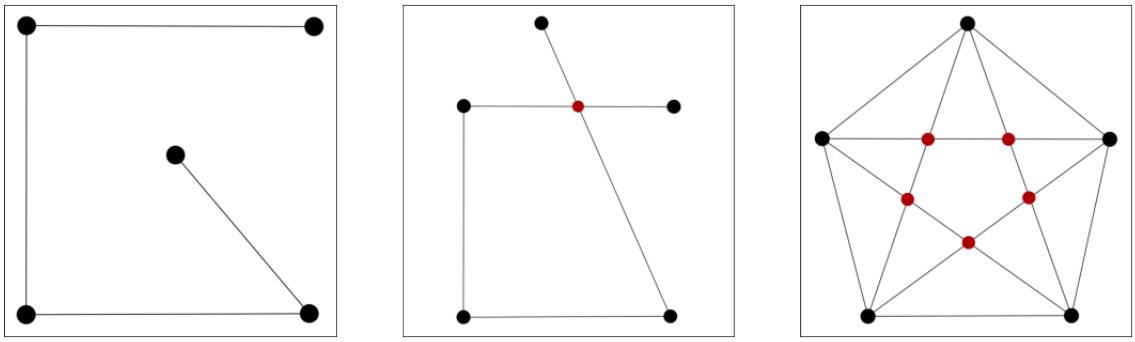
p. 1), the same author provides a comprehensive review of characteristics of spatial networks, important empirical observations, models of and processes on spatial networks.

Considering the field of road networks specifically, past research often focussed on typical properties and the structure of such networks. Crucitti, Latora, and Porta (2006) analyse the road networks of 1-square-mile street pattern samples from 18 planned and self-organised cities in terms of four different centrality measures and find that the networks of self-organised cities exhibit scale-free properties in the distribution of Betweenness Centrality (see Section 4.1 for a definition). Cardillo, Scellato, Latora, and Porta (2006) use a similar sample to study and compare basic properties of road networks in cities. They find that nodes (junctions) with more than five adjacent roads are rare, and that planned cities are different from self-organised ones in terms of their degree distribution and other measures. Jiang and Claramunt (2004) analyse the road network of three different cities in its functional view (*dual*[3] representation) to perform structural analyses using street connectivity, average path length and the network clustering coefficient, and show that those networks form small-world networks but do not exhibit the scale-free property. Jiang (2007) considers a larger sample of 40 US cities and finds, in contrast, that the dual representation of these networks display a scale-free property for street length and connectivity degree. More concisely, he finds that only 20% of all streets in a given network have a degree or path length greater than the average. Of these 20%, 1% are suitable to serve as a backbone of the whole network. Moreover, the author conjectures that 20% of the streets account for 80% of the traffic flow. This is solely based on network characteristics, not on traffic data. Lämmer, Gehlsen, and Helbing (2006) analyse the road networks of 20 German cities in their primal representation, provide descriptive statistics, and come to similar conclusions as Jiang (2007). Kalapala, Sanwalani, Clauset, and Moore (2006), too, observe scale-free degree distributions of road networks in the dual representation and additionally find that journeys across national road networks have a largely identical structure, independently of their length. Buhl et al. (2006) provide an analysis of topological patterns of 40 mostly European settlements and introduce a "meshedness coefficient" describing the structure of the graph (tree-like vs. planar). Connecting graph properties with economic activity, some authors find strong correlations between centrality measures and economic activity (Porta et al., 2012, 2009; Strano et al., 2007).

Most of these studies work with undirected planar street network models. While omitting

[3]In the dual representation of a road network, instead of edges representing road segments and nodes representing junctions (as in the *primal* representation), junctions are operationalised as edges that connect different streets.

Figure 1: Spatially and formally planar and non-planar graphs



Notes: Spatially and formally planar and non-planar graphs following Boeing (2020d). Formal planarity is explained in the main text. Spatial planarity is similar, but depends on the geometric embedding of the graph. For example, formally, the left and middle graph are the same (they can be drawn the exact same way), but they differ spatially, because in one graph, two edges overlap, whereas in the other they do not. Left: a spatially and formally planar graph. Middle: This graph is spatially non-planar, but formally planar, because it can be drawn without an edge intersection. Right: a spatially and formally non-planar graph. Red dots indicate the nodes that would have to be added to make the spatial representation planar.

the direction of edges is a simplification of the real circumstances with unclear implications, assuming planarity can have serious consequences for analysis. A planar graph is a graph that can be drawn without overlapping edges except where they intersect in joint nodes (Figure 1). There are some classical mathematical results for planar graphs, which many measures rely on (Barthélemy, 2011), and they offer computational simplicity and algorithmic tractability (Boeing, 2020d). But real road networks, with tunnels, over- and underpasses, are rarely planar. Boeing (2020d) examine the effects of imposing planarity on road networks, e.g., by adding artificial nodes at every line intersection of the non-planar graph. They introduce three measures describing the degree of planarity of a graph and the degree of network characteristic misrepresentation of a graphs planar version and conclude that imposing planarity breaks routing and network-based accessibility modelling, misrepresents connectivity and over- / underestimates intersection counts and average edge length, respectively.

Another criticism that many of the earlier studies face is that they were based on analyses of road networks of limited scale (e.g., one square mile grids), and it might be hard if not impossible to capture the properties of a city's road network by analysing a likely unrepresentative part of it.

A comparatively recent development has been the large-scale analysis of road networks made possible through open data becoming increasingly more available, especially since the advent of OpenStreetMap (OSM, see Section 3.3) and a corresponding API, providing broad access to street network and related data. Eventually, obtaining and analysing these data

became much easier following the introduction of `OSMNx` (Boeing, 2017), a python package designed for doing just this. Since then, a multitude of studies were conducted analysing large-scale road networks. For example, descriptive analyses were conducted for major cities in Ghana (Dumedah & Garsonu, 2021), every city, town, urbanized area and Zillow neighbourhood[4] of the US (Boeing, 2018), and every urban area of the world (Boeing, in press). `OSMNx` was also used to explore spatial order, urban fabric patterns, and street network orientation (Boeing, 2019b, 2021), as well as the evolution of street patterns on tract-level street networks in the US Boeing (2020c). Moreover, using OSM data and `OSMNx`, Shang et al. (2021) estimates the environmental impact of bike sharing in Beijing during the Covid-19 pandemic, Abdulla and Birgisson (2021) examine the robustness of road networks to flooding, and Boeing (2019a) analyse the circuity of walkable and drivable road networks and find that walkable routes allow for more direct routes than drivable ones in 40 US cities. This thesis will add to the existing research in that we use `OSMNx` and OSM to analyse the Berlin road network and identify potentially critical road segments. To establish a link with an economic application, we will use this identification method to predict the number of avoided accidents in 2020 due to reduced traffic.

2.2 Human Mobility on Road Networks

A closely related area is the science of human mobility. Here, we will highlight a few interesting results. After Marchetti (1994) discussed the idea of a fixed daily travel time budget, D. M. Levinson and Kumar (1994) and D. Levinson and Wu (2005) formulated the *rational locator hypothesis*, which states that individuals' travel times stay approximately steady over years, even if commuting distance and congestion changes. However, they object the idea of fixed personal travel time budgets, as the stability of travel time seems to depend on urban structure and geography. Drawing from the idea that mobility is essentially about energy, Kölbl and Helbing (2003) measure the energy consumption (in kJ/min, calculated based on the related O_2 consumption of the human body) of different modes and find that the average journey times for different modes of transport are inversely proportional to the energy consumption rates measured for the respective human physical activities. Consequently, given an energy budget, a mode that requires more energy is associated with less average travel time. Scaling the travel time distributions of the different modes, the authors find a universal functional relationship, pointing to a law of constant average energy consumption for the

[4] `https://www.zillow.com/research/data/`

physical activity of daily travel. Lastly, Schneider, Belik, Couronné, Smoreda, and González (2013) analyse daily human mobility patterns. Based on mobility data of different sources, the authors identify 17 unique daily mobility patterns (motifs) that make up up to 90% of the mobility patterns in the three datasets. Motifs essentially describe network patterns or subnetworks that appear more often than expected (Alon, 2007). In this case, a motif describes a movement profile encompassing all visited locations and the directed connections between those locations. For example, the leftmost motif in Figure 2 might describe the daily trips of a person who commutes to work and prefers to visit two other locations in just one trip, whereas the rightmost describes a commuter with only one daily destination. Barbosa et al. (2018) provide an in-depth discussion of mobility models and applications.

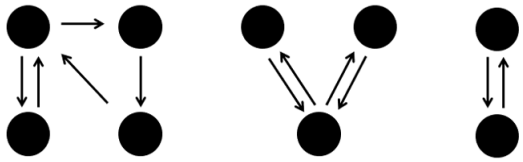


Figure 2: Sample mobility motifs according to Schneider et al. (2013)

3 Data

3.1 Trip Data

We rely on three main data sets. The first is provided by mobility data provider INRIX[5] and features trip data from approximately 7% of the traffic in Berlin in 2017. It contains data about a total of 34,209,113 unique automotive trips either originating in, ending in, or passing through Berlin, Germany, during the whole year 2017. *Trips* start as soon as the GPS device is turned on (e.g., when the car's engine is started or the app is turned on) and end when the device is turned off. Ten minutes without activity, that is, without a change in GPS coordinates or without transmission of coordinates, also mark the end of a trip[6].

The data set comprises information from (1) connected car data, (2) floating car data, (3) data from satellite navigation devices (Satnavs), and (4) commercial fleet data. For every trip, connected cars wirelessly transmit high-frequency GPS data. The term "floating car data" refers to similar data from mobile devices. Satnavs were the most widely used navigational devices until a few years ago. Commercial fleet data comes from, e.g., taxis, or heavy-duty vehicles such as trucks. The threat of sample selection bias is alleviated by relying on different data sources. For example, while younger drivers may be over-represented in (2), they may be under-represented in (3). However, the identification of individual drivers in the data set is virtually impossible, as several attempts at anonymisation have been made to adhere to data protection regulations. Therefore, assigning several trips to single drivers becomes a daunting task. This is a severe impediment, as certain questions regarding human mobility, transport related on a city level (e.g., concerning daily mobility energy consumption or mobility motif analyses as outlined in Section 2.2) are difficult to answer with these data. A viable approach seems to be the clustering of trips according to their geographical start and end points in order to identify recurring and consistent driving patterns (Koch, Ritter, Rohlf, & Thies, 2021).

Overall, the trip data consists of two components. On the one hand, there is an origin-destination data set with latitude / longitude coordinates of start and end points for each unique trip ID and several variables describing properties of the vehicle as well as summary statistics of the trip. On the other hand, there is a data set which records waypoints of each trip as timestamped in-between GPS markers. This data set also holds information on a

[5] `https://inrix.com/`

[6] For instance, there are trips in the data set with a travelled distance of 10 meters on a busy highway. This may result from congestion on this highway, as the data-transmitting car turns on and off frequently in the stop-and-go traffic.

Table 1: Variable descriptions of the trips and waypoints data set

	Variable Name	Type	Description
Trip Data	TripID	string	Unique trip identifier
	DeviceID	string	Unique device identifier
	StartDate	datetime	The trip's start date and time in UTC
	EndDate	datetime	The trip's end date and time in UTC
	StartLocLat	decimal	Latitude coordinate of the trip's start point in degrees
	StartLocLon	decimal	Longitude coordinate of the trip's start point in degrees
	EndLocLat	decimal	Latitude coordinate of the trip's end point in degrees
	EndLocLon	decimal	Longitude coordinate of the trip's end point in degrees
	DrivingProfile	integer	Vehicle class (consumer, taxi, delivery, for-hire truck)
	VehicleWeightClass	integer	Weight class (light- / medium- / heavy-duty)
	TripMeanSpeedKph	decimal	The trip's mean speed in km/h
	TripMaxSpeedKph	decimal	The trip's maximum speed in km/h
	TripDistanceMeters	decimal	Total distance of the trip in meters
	MovementType	integer	Indicates if the trip is moving or not
Waypoint Data	TripID	string	Unique trip identifier
	DeviceID	string	Unique device identifier
	WaypointSequence	integer	Order of the waypoint within the trip, incrementing from 1
	CaptureDate	date	The waypoint's capture date and time in UTC
	Lat	decimal	The waypoint's latitude coordinate in decimal degrees
	Lon	decimal	The waypoint's longitude coordinate in decimal degrees
	RawSpeed	decimal	The speed measured at the time of the waypoint
	RawSpeedMetric	string	The speed's unit of measurement

Notes: Unpopulated or deprecated variables have been excluded from the summary. TripID and DeviceID variables are present in both data sets.

vehicle's velocity. Table 1 presents variable descriptions.

Many different sources of data have been used in the literature to assess traffic and / or mobility patterns on routes of interest. Among them are surveys, mobile phone records, data from traffic detectors, data from GPS devices, and others. Surveys are useful to assess information about travel activities and purposes, as well as travel times. For example, the Chicago Metropolitan Agency for Planning provides data for two such surveys[7], and Liang, Zhao, Dong, and Xu (2013) use similar data for the Los Angeles area. Drawbacks of surveys include the comparatively low number of respondents and the narrow geographic and temporal scope. Call Detail Records (CRS) of mobile phones that contain information about time and the location of the cell tower routing the communication provide a much finer grained spatio-temporal resolution than surveys and have the advantage that mobile phones are commonly only used by their owners only. Such records were used, e.g., by Schneider et al. (2013) and P. Wang et al. (2012), even though this kind of data is not often available to researchers. Data from traffic detectors, such as freeway loop detectors, are the most common source of traffic flow data on freeways (Vanajakshi & Rilett, 2004). While their capability to measure

[7] `https://www.cmap.illinois.gov/data/transportation/travel-survey`

traffic flows and vehicle presence is used, for instance, for real-time traffic information boards, forecasting, and other applications, they come with several disadvantages. First, their locations are necessarily fixed, and they are almost exclusive to freeways. Thus, they do not cover different road types and city locations. They also suffer from equipment malfunctions. Barbosa et al. (2018) discuss additional sources of traffic and mobility data. Among all of these types of data, GPS is the one with the highest resolution and accuracy. However, one challenge of GPS devices is that they need a battery to function and that GPS coverage is unavailable in certain places, such as tunnels. Another typical drawback of GPS data is that if it is available, it is rather sparse in comparison to, e.g., mobile phone data (Barbosa et al., 2018). Fortunately, that does not apply in our case.

To use the data, not much prior cleaning is required. While applications which depend on reasonably defined individual private trips are well advised to exclude trips with lengths below a certain threshold and artifacts such as trips starting and ending at the same point in space or those made by commercial vehicles (Koch et al., 2021), we focus on traffic rather than trips, so we retain these observations. For further processing, we use all waypoints that fall into the administrative bounds of Berlin, or, respectively, in the Urban Centre of Berlin, which is defined in Section 3.3. Table 8 provides waypoint and trip statistics for the Berlin-Brandenburg region, Berlin in its administrative boundaries, and the Urban Centre of Berlin. Moreover, from the total sample of 34,209,113 trips, we select a sample of 20,000 (ca. 0.06%) trips and 631,525 corresponding waypoints to exemplify the precision of the waypoint data. Figure 3 (a) shows the driven routes of this sample within a bounding box encompassing Berlin. Following Bazzani, Giorgini, Rambaldi, Gallotti, and Giovannini (2010), in Figure 3 (b), the same sample is represented in a zoomed-in clipping of Berlin, coloured by speed categories.

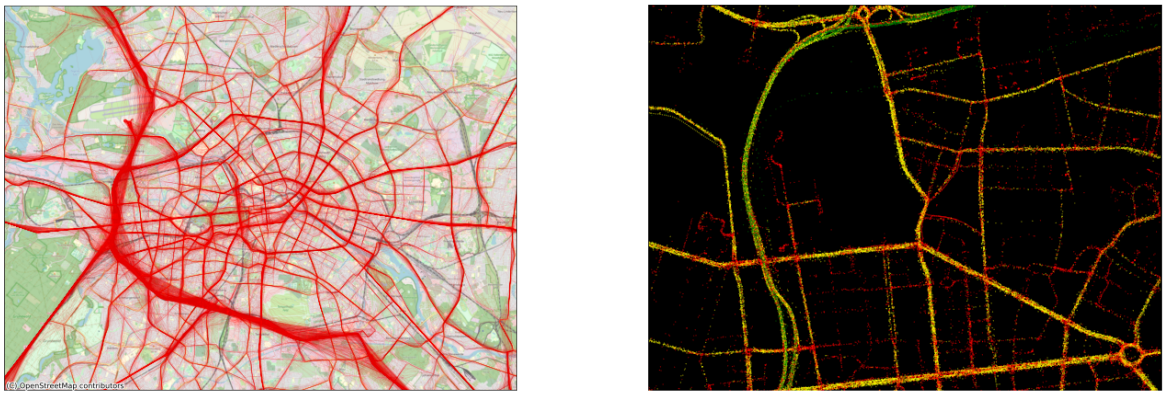
3.2 Accident Data

The second data set comprises information on accidents provided by INRIX in the form of a larger traffic incident data set for the year 2017. Incident types include construction work, accidents, closures and lane restrictions, weather conditions, and many others. INRIX uses several sources to retrieve these kind of data, especially official police records. Incidents are not recorded with high geographic precision, instead, they are mapped to the nearest junction, exit, or similar intersection point. Table 2 contains variables and variable descriptions[8].

[8]More details can be found at `http://docs.inrix.com/ra/incidents/`

Figure 3: The Berlin Waypoint Data

(a) **(b)**



Notes: (a) Traffic in and around Berlin. Lines are extrapolated from a sample of randomly selected waypoint sequences, the arterials (especially highways A9 and A10) are clearly distinguishable. (b) Single waypoints in a part of Berlin-Charlottenburg. Red dots correspond to a recorded current velocity of less than 30 km/h, yellow dots correspond to a velocity between 30 and 60 km/h, and green dots to a velocity of 60 km/h or more.

We limit the incident data to accidents which either (1) fall into the subcategory "accident" as classified by INRIX or (2) contain the word "accident" (without regard to capitalisation) in either the *full_description* or the *event_text* columns. The latter step is necessary because some incidents classified as accidents by the data provider had incongruous descriptions, for example "Lane closed due to maintenance work on A11 Southbound between 15 L304 Wandlitzer Chaussee and 16 L200 Schwanebecker Chaussee.". After filtering, 3,224 accidents remain in the dataset. The number of accidents is relatively small, because INRIX only records the traffic-relevant ones, i.e., which affect the traffic in some way. For instance, accidents in which a car damages another while parking, and which do not affect traffic in any way, may not appear in the accident data set, even if they are recorded by the police.

3.3 Road Network Data

The third data set is a preprocessed representation of Berlin's road network by Boeing (2020a, in press). We use this particular data because it provides an attempt at standardization for the possibly global comparison of road network models, which is much needed, given the many sources of road network data one can find in the literature. Boeing (in press) first derives urban area boundaries from the GHSL Urban Centre Database (UCD) [9]. The GHSL (Global Human Settlement Layer) project uses satellite imagery, census data and volunteered geographic information to produce knowledge primarily concerned with the human presence on

[9] `https://data.jrc.ec.europa.eu/dataset/53473144-b88c-44bc-b4a3-4583ed1f547e`, version 1.2

Table 2: Variables and variable descriptions for the incident data set

Variable Name	Type	Description
id	string	Unique incident ID
severity	integer	severity of impact, 0-4, 4 is highest
is_road_closure	string	"Yes" if incident causes a road closure, else "No"
category_ids	integer	Main Incident Category, e.g. "1:Construction"
sub_category_ids	integer	Incident Subcategories, e.g. "1:Roadworks"
full_description	string	Incident description
earliest_start_time_local_seen	datetime	Earliest local date and time seen for the incident
latest_end_time_local_seen	datetime	Latest local date and time seen for the incident
direction	string	Impacted road direction, e.g. "Northbound"
event_text	string	Brief description, e.g. "Road closed, construction"
location_type	string	Location type of the incident, e.g. "Point", "Linear"
length	decimal	Impact length of incident if location type is "Linear"
lat	float	Latitude coordinates of the incident in degrees
lon	float	Longitude coordinates of the incident in degrees
bidirectional	boolean	Indicates if the incident impacts both sides of a road
frcs	integer	List of functional road classes (e.g., FRC 3:"Arterials") impacted
causes	string	Cause of the incident, e.g. "2-Accident" for a road closure incident
effects	string	Effects of the incident, e.g. "8-Road closed" for an accident incident.
infos	string	Additional information text, e.g., "1277-drive with extreme caution"
alertCs	string	AlertCode, e.g. "1277-drive with extreme caution"
schedule	string	Days of week and local times the construction is active, if construction is scheduled

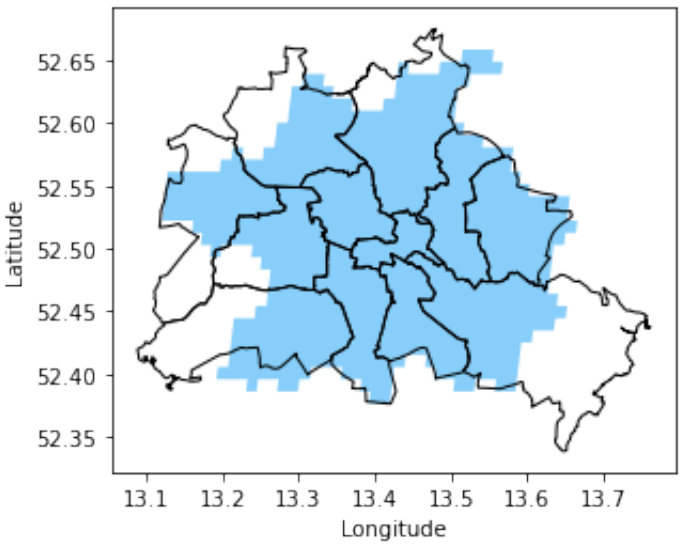
earth. The UCD comprises of so-called "urban centres", whose definition is not derived from national sources, but rather has been developed by the European Union, the Organisation for Economic Cooperation and Development, the World Bank, the UN Food and Agriculture Organization, and the UN Human Settlements Programme. (Florczyk et al., 2019) The "Urban Centres" (UC) are defined formally as

> "[...] the spatially-generalized high-density clusters of contiguous grid cells of 1 km^2 with a density of at least 1,500 inhabitants per km^2 of land surface or at least 50% built-up surface share per km^2 of land surface, and a minimum population of 50,000" (Florczyk et al., 2019, p.13).

Berlin's UC definition deviates from the official administrative boundaries (Figure 4).

The drivable street network of the UC Berlin is then queried from OpenStreetMap (OSM) and processed using `OSMNx` for Python (Boeing, 2017) (see Section 3.4). OpenStreetMap is an "open-source, collaborative, worldwide mapping project and database" (Boeing, in press, p.3), comprising map data, including information about addresses, speed limits, widths and lengths of roads, number of lanes, and points of interest. Its coverage is vast, and its almost

Figure 4: Administrative Boundaries of Berlin vs. Urban Centre of Berlin (blue)



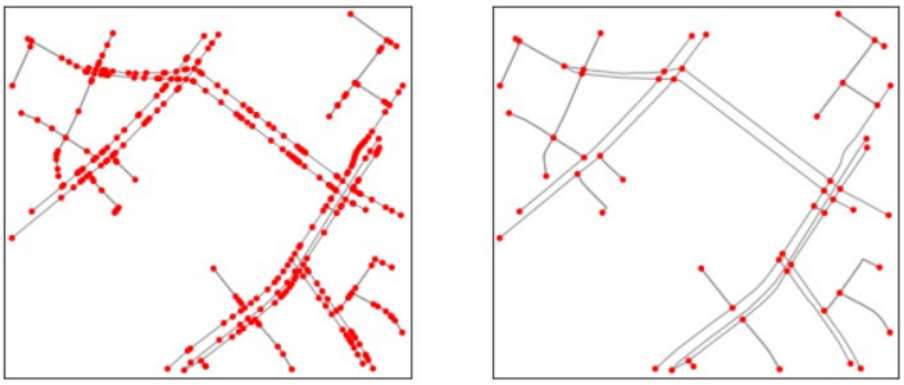
8 million contributors (OpenStreetMap, 2021) have managed to achieve 83% completeness worldwide in 2016, and complete street networks in 40% of countries (Barrington-Leigh & Millard-Ball, 2017). Even though the Berlin user base does not declare full completeness of their street network, as do many German cities in the project's "hall of fame"[10], Germany has a particularly active user base, and according to Seifert (2021), across all boroughs (*Bezirke*) in Berlin, streets are at least 96.5% complete. Therefore, OSM provides a free and good source of map information (Haklay, 2010; Ludwig, Voss, & Krause-Traudes, 2011; Neis, Zielstra, & Zipf, 2012).

The network is represented as non-planar directed primal multigraph[11]. Service roads and streets where motor traffic is forbidden are excluded. Moreover, the raw network that is retrieved from OSM is topologically simplified by `OSMNx` to enable meaningful network analyses, since raw OSM data comprises of nodes connecting straight-line edge segments. Subsequently, nodes represent intersections or dead-ends, and edges are the connecting road segments. Figure 5 shows a part of the Berlin network with and without simplification. To prevent adverse periphery effects originating from enforced artificial boundaries on the street network, `OSMNx` downloads a larger part of the graph and calculates node degrees before removing edges and nodes that fall outside the specified area (Boeing, in press).

Finally, information on node elevation is available from the Advanced Spaceborne Thermal Emission and Reflection Radiometer v2 (NASA/METI/AIST/Japan Spacesystems &

[10] `https://wiki.openstreetmap.org/wiki/Hall_of_Fame/Streets_complete)`
[11] A multigraph allows for more than one edge (road) between two nodes (junctions).

Figure 5: Berlin Alexanderplatz Network Representations



Notes: Raw OSM network data (left) and simplified road network data by `OSMNx` for Berlin, Alexanderplatz. Red dots signify nodes.

U.S./Japan ASTER Science Team, 2009) or the Shuttle Radar Topography Mission (SRTM, EROS Earth Resources Observation and Science Center (2017)). This dataset is available as `.graphml` file in the Harvard Dataset and includes information about 39 indicators (see Table 9 in the appendix) (Boeing, 2020a, 2020b, in press).

Waypoint data is combined with the network data as follows: first, the data is limited to the 850,587,925 waypoints within the Berlin UC area. For each individual waypoint, we use `OSMNx` to find the nearest edge in the road network. If the edge is farther than 50 meters away from the waypoint, we drop the observation. The threshold is arbitrary, but note that 90% of the waypoints were mapped onto the network with an accuracy of about 23 meters. note that the waypoint data has a precision of ca. 10m since they are mapped on a geographic grid defined by the fourth decimal in latitude / longitude coordinates. Moreover, precision also depends on the original signal and transmitting device. 22,167,933 ($\sim 2.6\%$) waypoints were removed using this criterion.

Comparing trip data within three different areas, the general trip statistics do not differ much (Table 8). Regarding trips that either start or end in the respective area, the Berlin UC has about 500,000 less trips than Berlin in its administrative boundaries. Average speed, distance and duration are virtually the same, as well as their maximum counterparts.

3.4 Technology

We processed all data using Python 3.8 in conjunction with Jupyter Notebooks which are also available as supplementary material to this thesis. Python (Van Rossum & Drake, 2009) is a high-level all-purpose programming language. It is highly flexible and more user-friendly

compared to compiled languages such as C++. However, it is slower than compiled languages, even though many popular packages make use of C and C++ for computational efficiency. Jupyter Notebooks (Thomas et al., 2016) offer an interactive interface for Python for a structured and reproducible workflow. We ran most computations on the high-performance cluster system at Potsdam Institute for Climate Impact Research Germany.

`OSMNx` (Boeing, 2017) is a Python package that builds on the popular network analysis package `NetworkX` (Hagberg, Schult, & Swart, 2008). It adds substantial functionality to include, amongst many others, spatial attributes in graphs. Its main contributions are that it allows automated downloading of street network data from OpenStreetMap, includes algorithms to correct network topology, enables saving and loading of these networks and, finally, provides tools to analyse street networks.

4 The Berlin Traffic Network

4.1 Networks in Graph Theory

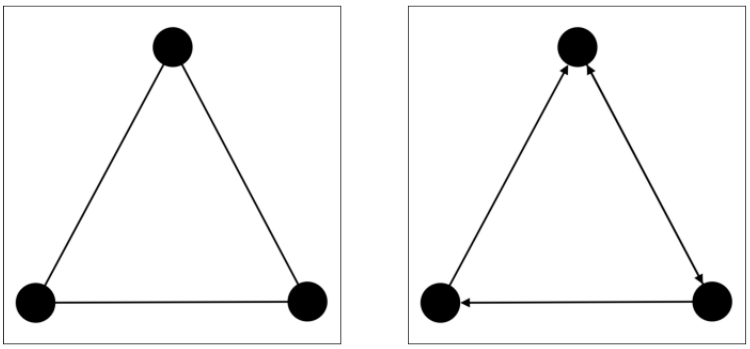
In the present thesis, we wish to discuss the Berlin traffic network's properties and use it to identify critical links that we will use for further economic analysis. Before doing so, we introduce basic aspects of graph theory. Additional characteristics are discussed in more detail in Barabási and Pósfai (2016). Here, we focus solely on the primal representation of the graph, which is also non-planar.

A graph representing a network usually consists of *nodes*, or *vertices*, and the connections between them, termed *edges*, *links*, or *arcs*. A graphs is called *directed* if all of its edges are directed, and they are called *undirected* if all of its edges are undirected. Technically, a graph can be both if some of the graphs edges are directional and some are bi-directional, which can be interpreted as undirected. More formally,

Definition 1. A *directed graph* (*digraph*) is a pair $\mathcal{G} = (V, E)$, where V is the set of nodes and $E \subseteq V \times V$ is the set of *directed* edges.
An *undirected graph* is a pair $\mathcal{G} = (V, E)$, where V is the set of vertices and $E \subseteq \binom{V}{2}$ is the set of *undirected* edges. In this notation, $\binom{V}{2}$ is the set of all 2-element subsets of V.

For example, the Facebook friend network is undirected, because if u is friends with v, v is automatically friends with u. The Twitter follower network is a directed network, since u can follow v while v must not necessarily follow u back. Road networks in the real world are usually directional, because even though most roads can probably be lawfully driven in both directions, there still exist one way streets. Note, that by the above definition, self-loops are not possible in undirected networks, since $\{u, u\} \notin \binom{V}{2}$.

Figure 6: An undirected (left) and a directed (right) graph with three nodes



Let N denote $|V|$, the total number of nodes in the network (also called *size* of the network), and L denote $|E|$, the total number of edges in the network. Nodes are usually labeled (e.g., u, v, w) or indexed ($u_i, i = 1, 2, ..., N$). Edges are not indexed or labelled, since they can be identified by the two nodes they connect, e.g., (u, v) is the link between u and v. Adjacent nodes (i.e. nodes that are connected by an edge) are called *neighbours*:

Definition 2. A node u is called *neighbour* of or *adjacent* to a node v if there is an edge $(u, v) \in E$ connecting them in $\mathcal{G}(V, E)$. If $\mathcal{G}$ is directed, u is an *in-neighbour* of v if $(u, v) \in E$ and an *out-neighbour* if $(v, u) \in E$. Edges are neighbours if they share a node.

Definition 3. The *neighbourhood* of a node $u \in V$ in an undirected graph $\mathcal{G}$ is defined as the set of all its neighbours: $N_{\mathcal{G}}(u) = \{v | \{u, v\} \in E\}$. In case $\mathcal{G}$ is directed, $N^{\text{out}}_{\mathcal{G}}(u) = \{v | (u, v) \in E\}$ denotes the *out-neighbourhood* while $N^{\text{in}}_{\mathcal{G}}(u) = \{v | (v, u) \in E\}$ denotes the *in-neighbourhood*.

Knowing about neighbourhoods, we can now define the *degree*:

Definition 4. The *degree* $k_{\mathcal{G}}(u)$ of a node $u \in V$ in an undirected graph $\mathcal{G}$ is defined as the size of its neighbourhood, $k_{\mathcal{G}}(u) = |N_{\mathcal{G}}(u)|$. Similarly, in a directed graph $\mathcal{G}$, we define the *in-degree* $k^{\text{in}}_{\mathcal{G}}(u) = \left|N^{\text{in}}_{\mathcal{G}}(u)\right|$ and *out-degree* $k^{\text{out}}_{\mathcal{G}}(u) = \left|N^{\text{out}}_{\mathcal{G}}(u)\right|$ of a node $u \in V$. The (in- / out-)degree for the i^{th} node in $\mathcal{G}$ is frequently written as k_i ($k^{\text{in}}_i / k^{\text{out}}_i$)

Note that in an undirected graph, the total number of links can be expressed as the sum of the node degrees

$$L = \frac{1}{2} \sum_{i=1}^{N} k_i, \tag{1}$$

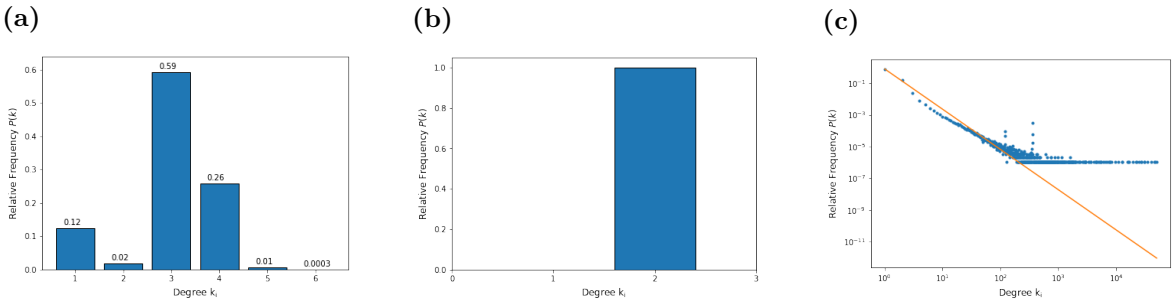
where the $\frac{1}{2}$ factor accounts for the fact that in the sum of node degrees, every edge is counted twice. In a directed graph, it holds that

$$L = \sum_{i=1}^{N} k^{\text{in}}_i = \sum_{i=1}^{N} k^{\text{out}}_i, \tag{2}$$

with the factor from Equation 1 missing, because the total degree of a node in a directed graph is $k_i = k^{\text{in}}_i + k^{\text{out}}_i$, so that each of the two summands separately account for one half of the total number of edges already. We can now calculate a graph's average degree for directed and undirected networks:

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2L}{N}, \tag{3}$$

Figure 7: Degree Distributions



Notes: Degree distributions of different networks. Distributions are usually discrete, hence bar diagrams often capture their basic nature. However, for distributions with many degrees, such as in (c), other displaying options are more suitable. (a) The degree distribution of the Berlin Traffic Network (see 4.2). (b) The degree distribution of the undirected version of the triangle (Figure 6). All nodes have degree 2. (c) The degree distribution of follower counts of influential Twitter users (blue dots). The x axis denotes the in-degree (follower count) and the y axis the probability to observe a user with the respective follower count. The data for the graph is taken from Münch, Thies, Puschmann, and Bruns (2021). The social network almost exerts the scale-free property (following a power law with exponent $\gamma \approx -2.54$, orange line), but it is not a perfect fit; probably due to the authors' sampling strategy.

$$\langle k^{\text{in}} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{\text{in}} = \langle k^{\text{out}} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{\text{out}} = \frac{L}{N}. \tag{4}$$

The average degree is an important statistic and is usually indicative of the networks underlying structure. The average degree can differ greatly across networks of different types, and we will see that road networks tend to have a small average degree. For an even better description of the network we can look at the *degree distribution* of the graph $\mathcal{G}$

$$P_{\mathcal{G}}(k) = \frac{N_k}{N}, \tag{5}$$

where N_k denotes the number of nodes with degree k. For directed networks, in- and out-degree distributions can be set up equivalently, and a joint probability distribution takes the form

$$P_{\mathcal{G}}(k^{\text{in}}, k^{\text{out}}) = \frac{N_{k^{\text{in}}} + N_{k^{\text{out}}}}{N}, \tag{6}$$

which is the probability of a node in $\mathcal{G}$ having in-degree k^{in} and out-degree k^{out}. For undirected networks, degree distributions are simply the normalized histogram of node degrees; the degree distribution of directed networks can be displayed either by plotting in-degree and out-degree distributions separately, adding up in- and out-degree to a total degree and only show the total degree distribution, or by using three-dimensional plot techniques such as surface plots. Degree distributions can differ heavily across networks, see Figure 7.

The degree distribution can tell something about the networks general structure and

about its robustness. Robustness in networks has its roots in percolation theory, and the primary question revolves around the impact of node failures on the integrity of a network. In other words, the question becomes what fraction of nodes (and which ones) have to be removed from a network to divide it in two or more disconnected *components* (see below).

Next comes the concept of *paths*. Other than in real-world applications, mathematically, physical distance is not a concept of particular interest in graphs and is replaced by *path length*.

Definition 5. A *path* is an ordered sequence of adjacent nodes, allowing for repeats. A *simple path* is a path without repeats.

Path length is then equal to the number of nodes crossed (or, the number of edges that the path contains). We are often interested in the (length of the) *shortest path* d_{uv} between two nodes u, v. The shortest path is the path with the fewest number of edges crossed. While $d_{uv} = d_{vu}$ holds for undirected networks, this is not necessarily the case for directed networks. In fact, the existence of a path d_{uv} does not imply the existence of a path d_{vu}. Path lengths are of special importance to *Betweenness Centrality*, a centrality measure for graphs. If there is a path between two nodes u and v, these nodes are called *connected*, otherwise they are *disconnected* and $d_{uv} = \infty$. This extends to whole graphs:

Definition 6. A graph is *connected* if all pairs of nodes in the graph are connected. The graph is *disconnected* if there is at least one disconnected pair of nodes.

If a graph is disconnected, there exist at least two different *components*, i.e., disconnected parts of the graph. Graphs can be equipped with weights, and if they are, we call them *weighted* and they are defined as follows:

Definition 7. A *weighted* graph is a triple $\mathcal{G} = (E, V, w)$, where $w : E \to W$ is a function mapping edges to their weights, and W is the set of possible weight values.

Many things can be used as weights. In a road network, it makes sense to use edge lengths (distances) or traffic density as weights.

There exist many different centrality measures for a graph that describe certain properties of the graph's nodes or edges. Three of them are *Closeness Centrality*, *Betweenness Centrality*, and *Straightness Centrality*. In this text, we adopt the definitions used by Crucitti et al. (2006).

Definition 8. *Closeness Centrality* is a measure of closeness of a node i to all other nodes $j \in \mathcal{G}$ and it is defined as

$$C_i^C = \frac{N-1}{\sum_{j \in \mathcal{G}; j \neq i} d_{ij}} \tag{7}$$

In its weighted version, each element in the denominator is multiplied by some weight. In the case of spatial networks, it is sensible to replace d_{ij} with the actual distance in units (e.g., meters).

Definition 9. *Betweenness Centrality* is a measure of how much a node i stands between other nodes $j \in \mathcal{G}$, indicated by the number of shortest path between other nodes that i is part of. It is defined as

$$C_i^B = \frac{1}{(N-1)(N-2)} \sum_{j,k \in \mathcal{G}; j \neq k \neq i} \frac{n_{jk}(i)}{n_{jk}},$$

where n_{jk} is the number of shortest paths between j and k and $n_{jk}(i)$ is the number of shortest paths between j and k that contain i. In its weighted version, the path lengths are modified. In the case of spatial networks, shortest paths are not defined by the number of traversed nodes but by the sum of distances between them.
Betweenness is also defined for edges:

$$C^B(e) = \sum_{j,k \in \mathcal{G}} \frac{n_{jk}(e)}{n_{jk}},$$

where $n_{jk}(e)$ is th number of shortest paths passing through edge e.

Definition 10. *Straightness centrality* is a measure of how much actual distances from any node to a node i differ from the straight (Euclidean) distances. It is defined as

$$C_i^S = \frac{1}{N-1} \sum_{j \in \mathcal{G}; j \neq i} \frac{d_{ij}^{\text{Eucl}}}{d_{ij}},$$

where d_{ij} are weighted (i.e., actual) distances and d_{ij}^{Eucl} are the Euclidean distances between nodes i and j.

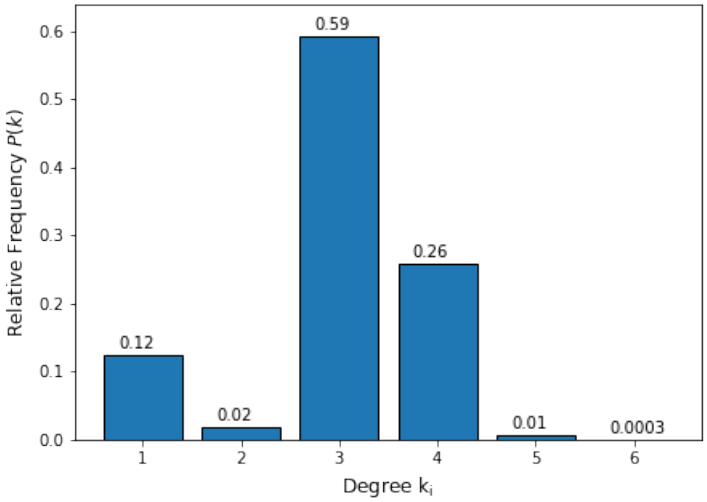
Straightness centrality is directly related to the *route factor* (or *circuity*) (D. Levinson & El-Geneidy, 2009).

4.2 Descriptive Analysis

Comprehensive descriptive statistics for the graph used in this thesis were already provided by Boeing (2020b), see Table 9. In this section, we will present several measures for road networks put forth in the literature, and discuss them in the context of the Berlin road network. Since many of these measures were only implemented for undirected networks in the past, we will also use the undirected version of the graph unless otherwise indicated.

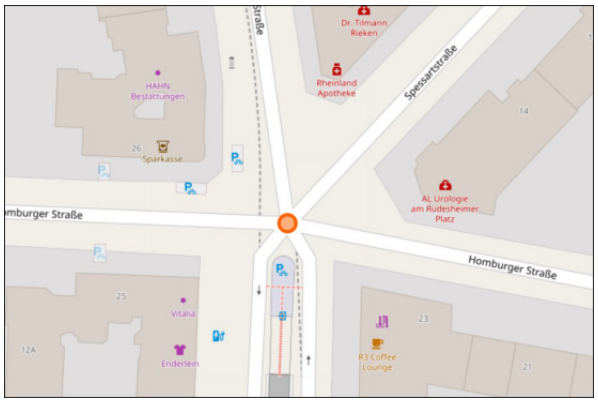
Looking at the average degree, we find $\langle k^{\text{in}} \rangle = 2.59$ for the directed version of the graph. When all directed edges are replaced by undirected ones, and redundant edges are removed, this number becomes $\langle k \rangle = 3.005$. Both numbers are roughly in line with Buhl et al. (2006)'s result of $\langle k \rangle \approx 2.43$ for 40 mostly European and North African non-planned settlements, who are using undirected network representations.

Figure 8: Degree distribution of Berlin's undirected road network



From the degree distribution (Figure 8), it is evident that the Berlin road network does not follow a "scale-free" distribution, which makes sense due to its spatial embedding. In accordance with both Lämmer et al. (2006) and Cardillo et al. (2006), we too observe that most nodes in the network have a degree less than 5, and that most nodes have either degree 3 or 4 (see Figure 9 for an example of a rare 6-way street intersection). However, we do not find that nodes with four neighbours are predominating, as Lämmer et al. (2006) suggest in their survey for 21 German cities, including Berlin. Taking into account that their network has a comparatively similar number of nodes but comprises of roughly 45,000 additional edges, one reason for this discrepancy might be that they enforce planarity on their networks (possible consequences are discussed in Boeing (2020d)). Another is that they

Figure 9: Example of a node with degree six (OSM ID: 26646276).



probably consider Berlin in its administrative boundaries[12]. The above degree distribution also underlines Cardillo et al.'s (2006) finding that self-organized cities (such as Berlin) tend to display $P(k = 3) > P(k = 4)$ while the inverse seems to be true for single-planned cities. Courtat, Gloaguen, and Douady's (2011) *organic ratio*

$$r_N = \frac{N(1) + N(3)}{\sum_{j \neq 2} N(j)}, \tag{8}$$

where $N(j)$ is the number of edges with degree j, is also used to indicate whether or not a city had been planned. If a city was planned, $N(4)$ would be very large due to the presence of many four-way intersections and r_N would consequently approach 0. On the other hand, $r_N = 1$ if there are no four-way intersections at all. Note that this assumes that $N(2) = 0$ and that there are no crossings with more than 4 adjacent roads which is usually the case. For the undirected Berlin UC network, $r_N = 0.73$, which supports the previous finding. The related fraction of tree-like structures, i.e., the fraction of dead-end roads amounts to 10.99%, which is almost as much as Lämmer et al. (2006) find for Berlin. Finally, we can also calculate the *compactness* of the graph, as mentioned in (Barthélemy, 2011, p. 9):

$$\Psi = 1 - \frac{4A}{(\ell_T - 2\sqrt{A})^2}, \tag{9}$$

where A is the total area of the city and ℓ_T the total length of the road network. This measure gives an indication of how much a city consists of roads. For the Berlin UC network, this measure is 0.99, but it must be noted that no comparison is available, and that this ratio is not necessarily equal to the fraction of space that is captured by roads.

[12]However, when using the OSM road network of Berlin's administrative boundaries, the discrepancy is still at about 30,000.

4.2.1 Centrality Measures

Following Lämmer et al. (2006), we take a look at the (weighted) betweenness centrality of Berlin's road network (Figure 10 (a)). Here, the city's ring-like structure is clearly visible, as well as the radial streets converging at the city centre. Looking at traffic volume, as measured by waypoints on street segments, Figure 10 (b) yields a similar picture. In both plots, thicker roads mean higher betweenness, or volume, respectively. Comparing the two images, they look quite alike with a few differences. First, (b) may exhibit thick lines and therefore high traffic volumes, that are not highlighted in (a). This is because in the graph representation, some roads at the graph's boundary are not at all "between", because fewer network roads are available to be connected in this restricted graph than in reality. A second difference is that the city centre seems much more cluttered when looking at traffic volume than when looking at betweenness. This might in fact be a consequence of bad traffic and congestion, as waypoints on a given road segment naturally increase when there is congestion.

Edge betweenness has been used as proxy for predicted traffic volume in the past (Lämmer et al., 2006), even though this has later been contested (Gao, Wang, Gao, & Liu, 2013; Kazerani & Winter, 2009). We test the goodness of this proxy by calculating the Pearson correlation coefficient. This gives us $r \approx 0.44, p < 0.001$, demonstrating that betweenness might be a reliable but imperfect measure of expected traffic volume or flow.

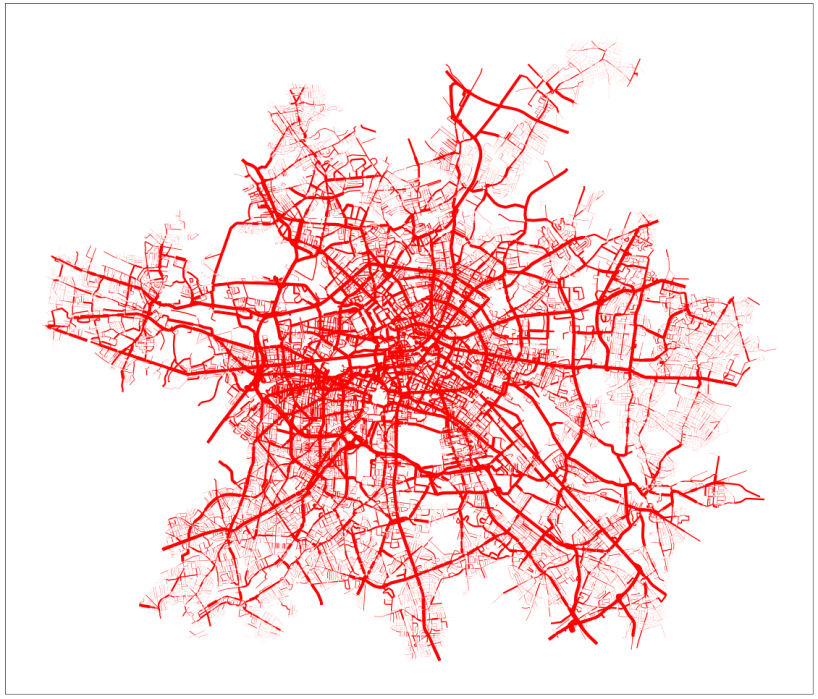
We also reproduce Crucitti et al.'s 2006 plots for weighted node betweenness and closeness centrality, as well as straightness centrality (Figure 11). Node betweenness centrality (Figure 11 (a)) reveals a similar picture as seen above, with the main corridors of the city more well-pronounced. Node betweenness reaches its highest values on the *Autobahn* (freeway) A100 south of the city centre. The *Bundesstraßen* (federal highways) 1, 2, 96, 96a, which are similarly important as primary movement channels, also exhibit high betweenness values. Closeness centrality (Figure 11 (b)) is highest in the city centre and decreases from there. This is due to the nature of the road network, the index itself, and because we are looking at a part of the whole road network with artificial bounds. There are some dark blue dots scattered across the map, and by manual inspection we discovered that they are junctions in parts of the city which are effectively gated from the public or nodes where there is a barrier such as a gate[13]. Straightness centrality yields a different picture (Figure 11 (c)). As it gives an estimate of the mean ratio of direct (Euclidean) distances and actual road network

[13]Examples for this are some ways in the *Kleingartenanlage Fabiansche Erben* (a garden plot), and the industrial area *Das Neue Gartenfeld*, both of which belong to the little cluster of dark blue points in the north west.

Figure 10: Betweenness centrality and traffic in Berlin

(a)

(b)



Notes: Betweenness centrality (a) and traffic (b); thicker lines indicate higher betweenness or volume, respectively. To allow for an easy comparison, the line thickness is binned into quantiles, so that the highest quantile has the thickest lines. For a "raw" version of the plot, see Figure 23 in the appendix.

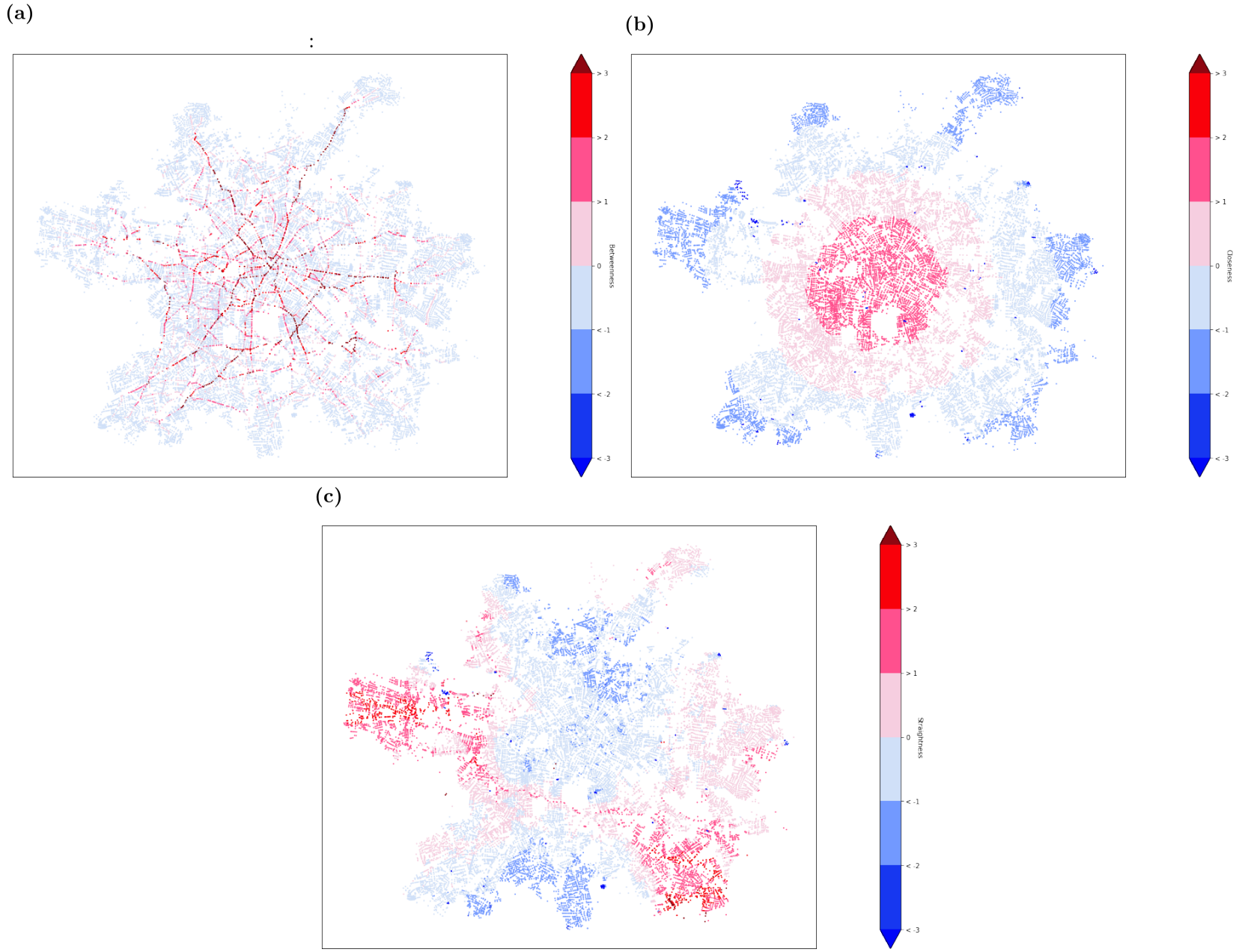
distances, values are low where the network is inefficient, i.e., where the network distance is far higher than the Euclidean distance. Equivalently, one can say that the higher the straightness at a given node, the more easy it is to reach from all other nodes in the network. We observe that the whole city centre exhibits rather low straightness, confirming that travel from outer regions to the center requires complex routing. On the other hand, straightness is high in the boroughs Spandau and Treptow-Köpenick, as well as along the Autobahn A100 and A113, which makes sense, as they are two of the main connecting corridors. Parts of Tempelhof-Schöneberg and Steglitz-Zehlendorf in the south and Pankow in the north make up the other side of the spectrum, making them the boroughs in Berlin where destinations are reachable least efficiently considering the entire Berlin UC road network.

Lastly, we consider isochrones, as previously done by Lämmer et al. (2006). Isochrones are lines on which every point is reachable within the same amount of time. For this, we use the administrative definition of the Berlin road network, since some connectors are missing in the UC representation. For the isochrones, we assume that the maximum possible speed can be and is driven. When there is no maximum speed information available, we impute the mean trip speed of 32.73 km/h (Table 8). We calculate for each road segment the time it needs to pass it by dividing its length by its maximum speed. Figure 12 displays isochrone polygons. Given that the isochrones expand spherically from the starting point in the center, it seems that Berlin's road network is equally efficient in all directions from the city centre. However, this is only the case under excellent traffic conditions. We note the difference to Dresden's non-spherical isochrones from Lämmer et al. (2006), which we reproduced in Figure 21 in the appendix.

4.2.2 Scope Constraints and Limitations

Beyond the measures discussed here, other descriptive statistics have been proposed in the literature, which may be worth examining. Many of these are outside the scope of this work, as code implementations are inapplicable to our data our entirely unavailable. Examples of these are ringness, webness, circuitness, and treeness (Xie & Levinson, 2007), which all make a statement about predefined connection patterns of arterial roads in cities. Moreover, we did not elaborate on cost and efficiency of the network, which is also related to transport performance (Latora & Marchiori, 2003). It would also be interesting to compare the performance and results of different graph-theoretic community detection methods of city road networks in general and of the Berlin road network in particular. Finally, since we only consider the

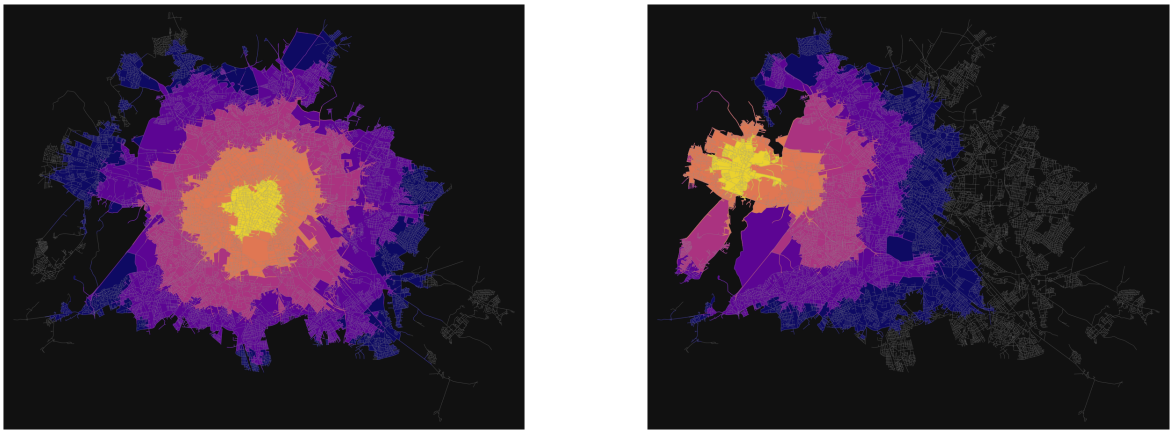
Figure 11: Node betweenness, closeness, and straightness centrality in Berlin



*Notes:*These maps depict road network graphs detailing (a) betweenness, (b) closeness, and (c) straightness centrality of road network junctions. Values are binned into a colour scale ranging from -3 (dark blue) to +3 (dark red) standard deviations from the mean.

Figure 12: Isochrones in Berlin

(a) **(b)**



Notes: This map depicts the road segments and polygonised isochrones of Berlin. Point of reference is a point in Berlin's city center (a) or a point in Spandau in the north west (b) Brighter areas are reachable faster, and the brackets are 5, 10, 15, 20, and 25 minutes under ideal circumstances (driving with allowed maximum speed and no traffic lights). For instance, someone starting in the Berlin city centre could reach the Zehlendorf intersection (*Kreuz Zehlendorf*, end of long straight line segment in the south-west) within 20 minutes (dark purple), given she rides with maximum permissible speed on every road on her way. A driver from Spandau needs at least 20-25 minutes to the city centre.

primal representation of Berlin's road network, we are not able to comment on properties of the network's dual representation.

Concerning mobility networks, our data does not allow for in-depth analyses of motifs (Schneider et al., 2013) and mobility energy considerations (Kölbl & Helbing, 2003), because the lack of unique driver IDs complicates trip bundling at the individual level. Appropriate spatio-temporal and stochastic trip chaining is a complex problem and was also considered out of scope for this thesis. However, we encourage researchers to expand on these topics, if suitable data is available, and build on the existing code base provided with this thesis.

A final limitation of past research in this field, as well as of this work, is that primarily the undirected road network is analysed. Here, it is done for better comparison with past work, but in the future, analysing the directed network may yield more accuracy and more appropriate statistics, e.g., when calculating centrality measures.

5 Case Study: COVID-19 and Traffic

5.1 Background

In this part, we use the network representation that we have so far described theoretically and put it to use for a practical application. The rapid spread of the Coronavirus Disease (COVID-19), the first global pandemic in the 21st century, has led to a scenario without precedence. Most countries of the world imposed "lockdowns", closing all non-essential businesses for a certain amount of time. Substantial restrictions in travelling were enforced worldwide, and, in many countries including Germany, employers were asked to offer home office when and where possible. The economic consequences of the pandemic are yet to be gauged. Naturally, with a significant share of people working from home and many cultural and social activities unavailable, road traffic was heavily reduced (see European Data Portal (2020) and Section 5.3.2). Road traffic reductions result(ed) in reduced congestion and an overall reduced number of accidents, too (Retallack & Ostendorf, 2020; Statistisches Bundesamt, 2021). In the following, we want to estimate the economic savings resulting from the reduction in traffic accidents in 2020 due to less traffic on Berlin's streets. Using the GPS traffic data as mapped on the network provided by OSM, we (1) identify road segments that were *critical* in 2017 by employing a method from He, Yan, Liu, and Ma (2016), and (2) validate this identification method by checking whether it captures road segments with many accidents, polluted road segments, or those with high betweenness centrality implying a high propensity for congestion (cp. Lämmer et al. (2006)). On identified critical road segments, (3) we compare different predictive models for the number of accidents on a given road segment, (4) estimate the number of traffic-relevant accidents in 2020, where the traffic volume was much lower in comparison to previous years with higher traffic volume, and, finally (5), estimate the economic savings associated with this reduction.

5.2 Methodology

5.2.1 Identifying Critical Road Segments

Multiple ways have been presented in the literature to identify congested or otherwise critical roads in a road network. A common approach is the full network scan (Jenelius, Petersen, & Mattsson, 2006; Taylor, Sekhar, & D'Este, 2006), by which each link is iteratively removed from the graph and the removal's effects on network performance are measured. Critical links are then identified by comparing the effect of removal across all possible links (Chen, Lam,

Sumalee, Li, & Li, 2012). Since this can become computationally burdensome, especially in larger networks, methods that work similarly but consider subnetworks instead of the whole network (Erath, Birdsall, Axhausen, & Hajdin, 2009) or partial networks (Yang, Liu, Li, & He, 2016) have been introduced. The idea is to reduce the search space by considering only those areas that are most probably impacted by the node failure in question. More recent advancements of the full-scan method take origin-destination traffic flows into account (F. Li, Jia, Luo, Li, & Yang, 2020).

A related strand of literature focuses on the identification and effects of bottlenecks (see Hale et al. (2016) for a comprehensive review). While many researchers study congestion propagation by using data from known congested locations (Nguyen, Liu, & Chen, 2017; Tao, Xi, & Li, 2016; Y. Wang, Cao, Li, & Gu, 2016; Xu, Yue, & Li, 2013), C. Li, Yue, Mao, and Xu (2020) identify bottlenecks by taking both the cost of congestion at the potential bottleneck as well as the "contagion" cost of congestion propagating from the original bottleneck into account. Tian et al. (2021) consider cascading failures as well and introduce a criticality index for road links that is the product of road link vulnerability (probability of failure) and road link importance (a function of the impact of failure on subsequent road links). Solé-Ribalta, Gómez, and Arenas (2016) use an analytical approach to identify congestion hotspots and introduce the Microscopic Congestion Model, which can be solved using real traffic flow data and simulated data. Finally, He et al. (2016) define intuitive Road Segment Congestion and Network Congestion indices.[14].

Unfortunately, few to none of the methods described above are readily implemented or otherwise available in standard software. Therefore, we adopt He et al.'s (2016) method and implement it ourselves. In the following, we first describe the *Road Segment Congestion Index* which lies at the heart of their method. Second, we discuss its implementation using the waypoint data at hand. Third, we evaluate the accuracy of the method by comparing the obtained indices with empirically observed data, such as typical congestion hotspots in Berlin and air quality data from a public network of air quality measuring stations. Fourth, we discuss feasible models for predicting the number of road segments on Berlin's streets.

[14]Notably, edge betweenness is seldom used directly as a measure for proneness to congestion. According to F. Li et al. (2020), this is because it is not always accurately representing traffic flow and density. For example, high-betweenness edges can have low traffic flow, thus the impact of corresponding node failures might be less than expected.

5.2.2 The Road Segment Congestion Index

In He et al. (2016), the authors build on an existing definition of the *Speed Performance Index*, an index used to assess the traffic state level on a given road segment at a given time. The index reflects the ratio between average speed and maximum permissible speed:

$$R_{i,t}^{v} = \frac{v_{i,t}}{V_i^{\max}} \cdot 100, \tag{10}$$

where $R_{i,t}^{v}$ is the Speed Performance Index on road segment i at time t, $v_{i,t}$ is the average travel speed on road segment i at time t, and $V_i^{\max}$ is the maximum permissible speed on the road segment[15]. Based on this index, He et al. (2016) categorise road segments into four traffic state levels using three threshold values: *heavy congestion* ($R_{i,t}^{v} \in [0, 25]$), *mild congestion* ($R_{i,t}^{v} \in (25, 50]$), *smooth* ($R_{i,t}^{v} \in (50, 75]$), and *very smooth* ($R_{i,t}^{v} \in (75, 100]$). With these categories, the fraction of time that the road is not congested, R_i^{NC}. can be calculated:

$$R_i^{NC} = \frac{t_i^{NC}}{T_i}, \tag{11}$$

where t_i^{NC} is the duration during which road segment i is not congested (i.e., $R_{i,t}^{v} \leq 50$, so that it falls into either the *heavy congestion* or *mild congestion* brackets) and T_i is the length of the observation period. Finally, the *Road Segment Congestion Index, RSI*, or R_i is calculated as follows:

$$R_i = \frac{\bar{R}_i^{v}}{100} \cdot R_i^{NC}, \tag{12}$$

where $\bar{R}_i^{v}$ is the average Speed Performance Index taken over the time period of observation. Note that a low R_i stands for bad traffic conditions.

For the calculation of the indices, we use only waypoints from the two "rush hour" time windows also used by Koch et al. (2021) (06:00 - 09:59 and 14:00-19:59) during weekdays, as these are the main traffic hours (*Hauptverkehrszeiten*) (Senatsverwaltung für Umwelt, Verkehr und Klimaschutz, 2019). We calculate $R_{i,t}^{v}$ for each road segment and 15-minute time period in these time windows. Within road segments, we only keep data from time intervals which have at least 20 observations. That is, if there are less than 20 waypoints on a given road segment at a given 15 minute time period, this time interval will be dropped for this road segment. The rationale is that the Speed Performance Index can become very noisy when there are only few observations. 20 was chosen as a threshold value because it

[15] While it is possible for a road segment to have different maximum speeds depending on the time of day, we only consider a fixed road segment speed.

corresponds to the median value of all observation counts at a given time period and road segment[16].

We retrieve permissible road speed data from OSM using `OSMNx` for every road segment. Some road segments consist of sub-segments with differing road speeds. In these cases, we take the road speed that is valid for the longest distance within the segment[17]. In some situations, this might lead to an $R^v_{i,t} > 1$, in which case this value is replaced by 1. For 4.95% of the 47,293 road segments considered, there was no road speed data available from OSM. We drop these segments and do not consider them for critical road segment identification as we cannot calculate the corresponding RSI. Visual inspection showed that the road segments without road speed information belong to side roads or rural areas too insignificant for OSM users to collect information. To check this assumption, 25 road segments without RSI were manually looked up on OSM[18]. Table 10 in the appendix and the related Figures 20 (a)-20 (f) indicate that mostly residential roads and living streets are not covered, supporting this conjecture.

Before calculating R^{NC}_i, we employ another filtering step at the time period level. Across all road segments, we observe a median time period coverage of 85% of the 40 15-minute-periods on a given day. Hence, for more than half of all road segments, we do not observe any traffic in at least one 15-minute period in the specified rush hour time window *during the whole year*. We argue that these road segments are probably not heavily affected by large traffic volumes and congestion, so we drop all road segments for which we do not have coverage of at least 34 (85% of 40) of the 40 15-minute time periods. This amounts to a number of 21,900 road segments, or 46.31% of all road segments.

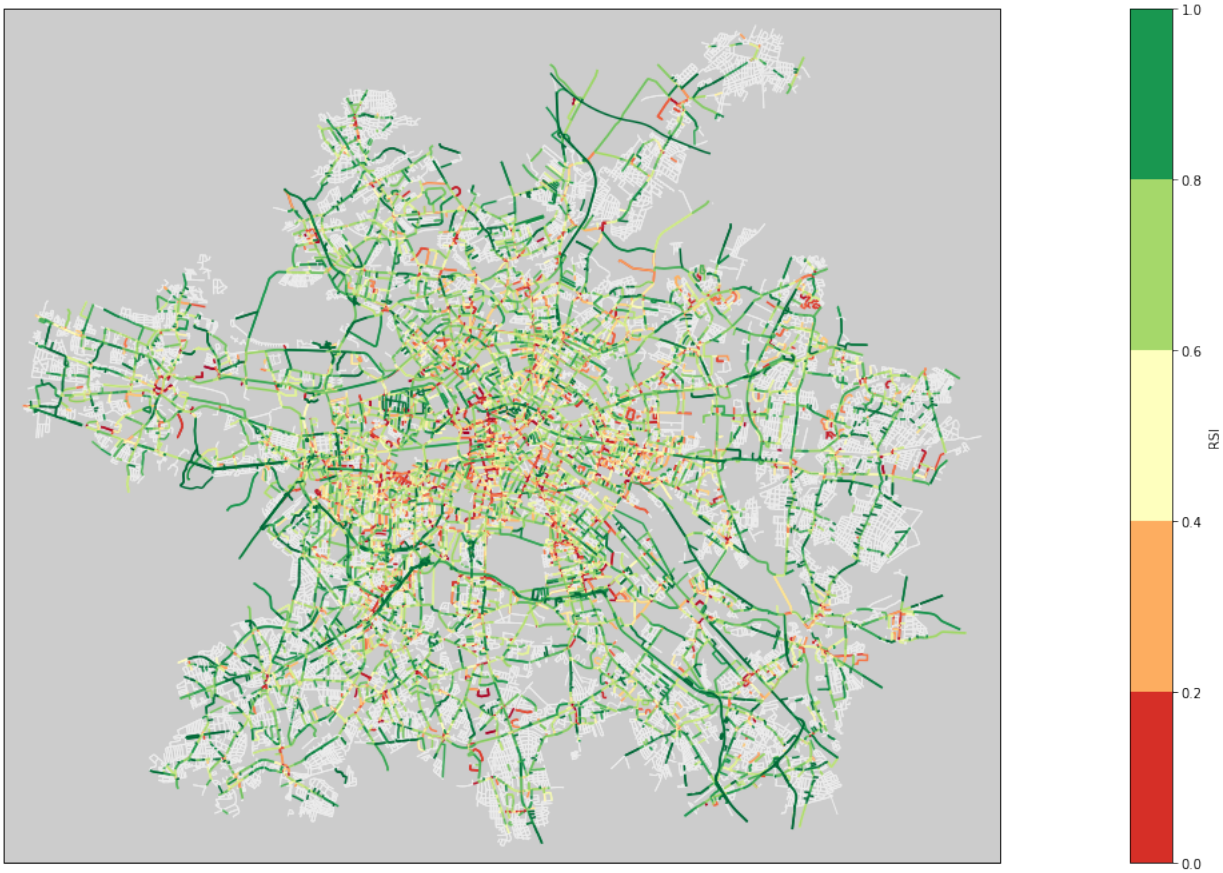
Finally, the RSI can be computed for 23,053 (48.75%) of all road segments. Figure 13 indicates that we can calculate the RSI for all major roads and traffic arterials as well as the majority of the inner city, while the truncation of the sample affects suburban areas and side streets (indicated in grey).

[16] An arbitrary choice, higher values likely lead to more precision but at the cost of losing even more of possible information for the analysis.

[17] For example, if there is a 100m road segment with permissible road speeds 30 km/h and 50 km/h, and on 60m of this road segment, 50 km/h is the valid permissible maximum speed, we assume it to be valid for the whole 100m road segment.

[18] The OSM IDs can be checked by visiting `https://www.openstreetmap.org/way/26545704` and replacing the id in the url.

Figure 13: RSI of road segments in Berlin



Notes: This map depicts the road segments and colour codes them according to their RSI as described in Section 5.2.2. Low values (red) denote bad traffic and possibly high congestion. Green indicates good traffic conditions. For grey road segments, the RSI was not calculated.

5.2.3 RSI Validation

The RSI is a measure to identify critical road segments where much traffic happens and which are likely to be congested (or, at least, where traffic goes slow when compared to the maximum speed limit). But are these segments really *critical*? We test the RSI's relation to two important real-life consequences of traffic, namely, the number of accidents and, the average pollution. We also test its association to edge betweenness centrality. We define a road segment i with $R_i \leq 0.5$ as critical. These are road segments which are either congested ($R^v_{i,t} \leq 50$) during most of the 15-minute-periods, or streets affected by rather slow average road speeds as compared to the maximum permissible speed, or both.

To get the accident counts on a given road segment, we retrieve the number of accidents for each road segment as follows: first, since we only consider the traffic during the main hours on working days within the Berlin UC area, we filter the accident data accordingly. After this step, 792 accidents remain. Second, we account for the fact that reported accidents are always mapped onto the nearest junction and for the influence of accidents in the vicinity of busy road segments by creating a 100m buffer around each road segment (cp. Figure 14). Then, for each road segment, we count how many accidents fall into the resulting road segment polygon[19]. A visual inspection indicates that the RSI identifies almost all of the places in Berlin, where accidents happen, and many where they do not (Figure 15). We also estimated Spearman's Rank correlation coefficient of the number of accidents on road segment i and RSI_i, which results in a test statistic of $r \approx -0.16$, indicating a statistically significant negative correlation ($p < .001$ in a one-sided test), as expected: road segments with bad traffic conditions see more accidents.

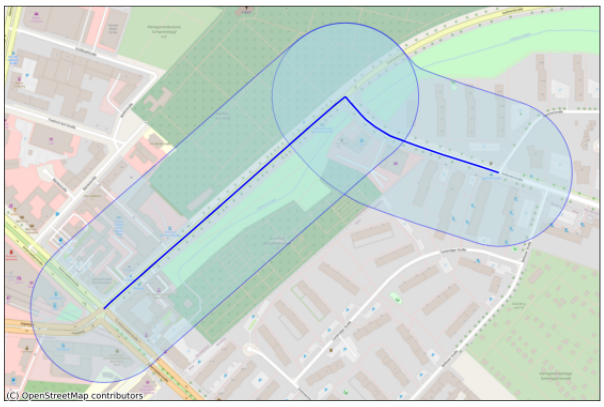
Furthermore, we investigate the relation of the RSI and the degree of pollution. To do so, we download sensor data including measured PM2.5 and PM10 concentration (in $\mu g/m^3$) from different locations in Berlin for every day in 2017. This data is available from the public open-source project sensor.community[20] and its respective archive API[21]. Users at sensor.community can buy and install their own sensors capturing pollution, temperature, humidity, noise, and other environmental factors and make it accessible for everyone by sending the data to a public server. Sensor.community features data from over 13,000 active sensors worldwide, and operates in 69 mostly European countries. Our sample of the data

[19] Note, that this artificially inflates the total number of accidents, because some accidents are counted twice or more, so that the inflated number of accidents amounts to 4,047. We account for this inflation at a later stage.

[20] `https://sensor.community`

[21] `https://archive.sensor.community/csv_per_month/`

Figure 14: Spatial Buffering of road segments



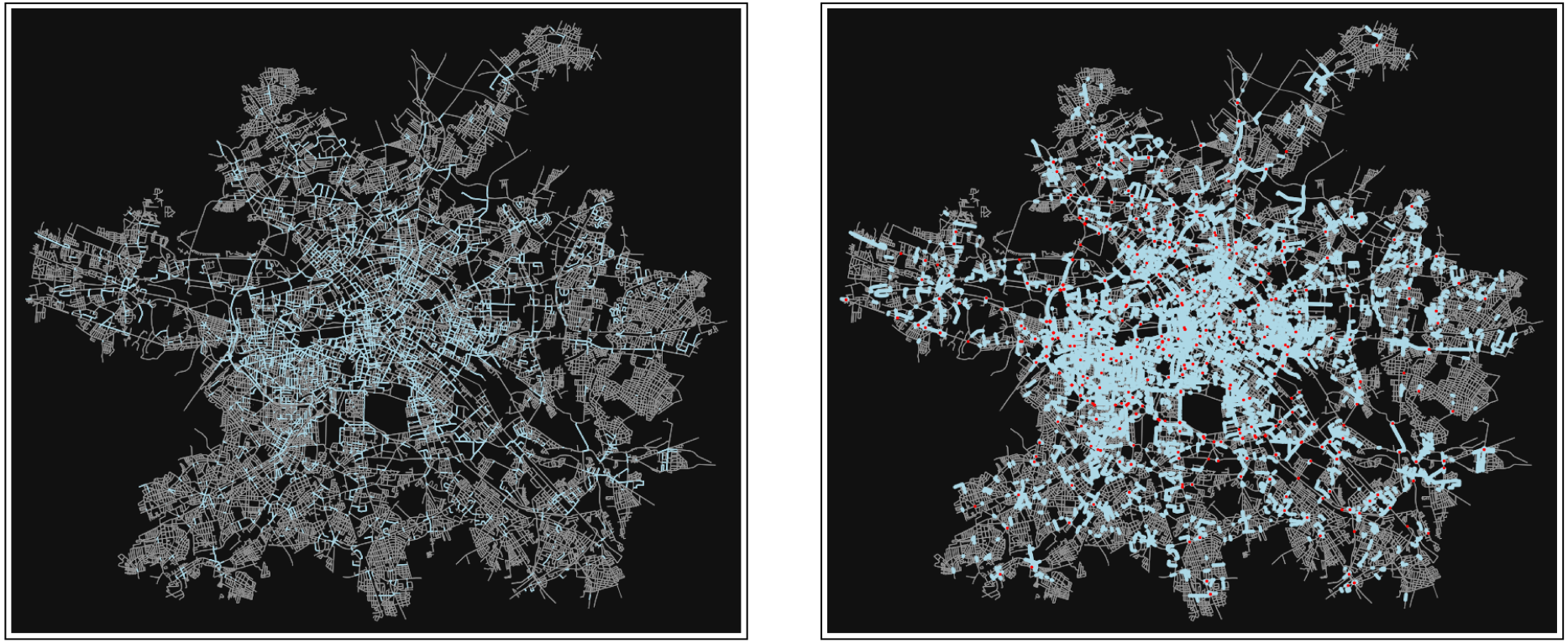
Notes: Two road segments (blue lines) and a 100m buffer in both directions (transparent light blue)

comprises of daily measurements at around 8 AM taken from all sensors of the model SDS011 (as indicated by the API).

Considering measurements from the Berlin UC area only leaves us with data from 136 sensor stations. Then, for each station, we aggregate the data by taking the mean (and, equivalently, the sum) over the PM2.5 and PM10 measurements, so that we have a yearly mean, or total pollution exposure. We then lay a circular buffer of 100m around every sensor station and subsequently intersect these buffers with the line geometries of our road segments. The intuition here is that the buffer captures the pollution from the nearby road segments, i.e., the traffic of a road segment falling into the buffer area of a station contributes to the measured pollution. After the intersection, 647 road segments remain for correlation, meaning that, on average, one station's buffer overlaps with 4 to 5 road segments. Finally, we calculate the Pearson correlation between RSI and PM2.5 and PM10 values, respectively. Both correlation coefficients are very small and far from statistical significance ($r_{\text{RSI, PM2.5}} \approx 0.04,\ p \approx 0.30;\ r_{\text{RSI, PM10}} \approx 0.01,\ p \approx 0.84$). Thus, the RSI is not indicative in terms of air quality. After all, weather conditions, such as wind, air pressure, temperature, and humidity play an important role for the concentration of particulate matter (UCAR Center for Science Education, 2021), and yearly aggregated may not provide sufficient statistical variation for the correct identification of such effects.

Finally, we also calculate the Pearson correlation coefficient between the RSI and edge betweenness and we find a practically insignificant positive correlation of $r \approx 0.02, p < .10$. This is interesting, because it suggests that road betweenness is not at all helpful in predicting goodness of traffic flow, which in turn can mean that highly between roads in the graph-theoretic sense do not have to be congested if planning is done well.

Figure 15: Distribution of critical road segments and accidents



Notes: Both panels depict road segments with $RSI_i \leq 0.5$. The panel on the left shows such segments (light blue) and all other road segments in grey, while the right panel depicts their 100m buffers and the 792 accidents used in the analysis (red).

In conclusion, road segments with a lower RSI have a higher risk of traffic accidents, but it does not seem to be a good direct predictor in terms of pollution and air quality, at least not at the yearly level.

5.2.4 Estimating Accident Frequency

Our goal is to obtain a predictive model for the number of accidents on a given *critical* ($RSI_i \leq 0.5$) road segment in a year. Many factors contribute to the event of a traffic accident, including weather conditions (Andreescu & Frost, 1998; Eisenberg, 2004), time of day and road design aspects (Ivan, Wang, & Bernardo, 2000), vehicle characteristics (Lie, Tingvall, Krafft, & Kullgren, 2006), driver behaviour (de Winter & Dodou, 2010; Gicquel et al., 2017), and traffic conditions (Retallack & Ostendorf, 2019). As we are predicting traffic accidents on a yearly basis, individual driver behaviour is not helpful, and meaningful vehicle characteristic averages are impossible to obtain. The problem with fine-grained predictor variables such as temperature and precipitation is that for them to be useful, we would need to model accidents on a very high resolution, as these variables only make sense for a sensible time interval, such as hours. The yearly number of accidents in Berlin that we consider is very small, so we end up with an excessive number of zero outcomes whenever we aggregate over too small time windows. Moreover, our unit of observation are road segments, and the available data is limited as to the number of possible predictors. For example, we have no information about the curvature of an edge segment. For part of the data, information about the number of lanes and the type of paving is available, but restricting the data further might induce other issues, such as lower predictive performance due to high estimator variance. Hence, we aim to predict the number of traffic accidents using the traffic volume (total number of vehicles on a given road segment in the defined rush hour time window) per year and the maximum permissible speed on the road segment.

5.2.5 Model Choice

Even on this very coarse level, and even though we consider only the busiest roads in terms of RSI, there are still many road segments left with zero outcomes, i.e., road segments where no accidents happen. Of the 8,017 road segments in question, 76.13% do not experience any traffic-relevant accident in the year 2017. As common linear models are not well equipped to handle data of this type, and because the outcome in question (number of accidents) is discrete, two types of generalised linear models were considered: Zero-Inflated Poisson

Regression and Zero-Inflated Negative Binomial Regression. Both models are mixture models consisting of a binary classification model, such as a logit or probit model, and a Poisson, or, respectively, Negative Binomial model. The binary response part models the probability of a non-zero outcome, the second part models the counts, adjusted for the excessive zero counts (Hilbe, 2011). Theory suggests that excessive zeros are generated by an independent separate process (UCLA: Statistical Consulting Group, n. d.a).

Zero-Inflated Poisson Regression Let Y_i denote the random variable counting the number of accidents to happen on road segment i and y_i denote a realisation of Y_i. It is assumed that the generation of the excess zeros and the generation of the accident counts follow separate processes. In the Zero-Inflated Poisson (ZIP) Model Lambert (1992),

$$Y_i \sim \begin{cases} 0, & \text{w.p. } \pi_i \\ Poisson(\lambda_i), & \text{w.p. } 1-\pi_i. \end{cases} \tag{13}$$

The probability of zero accidents happening on road segment i, π_i can be estimated by a binary response model of the form

$$\pi_i = G(\mathbf{z}_i'\boldsymbol{\gamma}), \tag{14}$$

where z_i are predictors, e.g., attributes of road segment i, and γ is an estimand of G. The most common case for G is to be the logistic link function

$$G(\mathbf{z}_i, \boldsymbol{\gamma}) = \frac{\exp(\mathbf{z}_i^T\boldsymbol{\gamma})}{1+\exp(\mathbf{z}_i^T\boldsymbol{\gamma})}, \tag{15}$$

but other models, e.g., probit models, are also possible. The standard Poisson model for a count outcome can be written as follows:

$$P(y_i|\lambda_i) = \frac{\lambda_i^{y_i}}{y_i!}\exp(-\lambda_i), \tag{16}$$

with

$$\lambda_i := \mathbb{E}[Y|x_i] = \exp(\mathbf{x}_i^T\boldsymbol{\beta}), \tag{17}$$

the mean of the predicted Poisson distribution. Hence, assuming Y follows a zero-inflated

Poisson probability process, the probability distribution of Y can be written

$$P(Y = y_i|\mathbf{x}_i, \mathbf{z}_i) = \begin{cases} \pi_i + (1 - \pi_i)\exp(-\lambda_i), & y_i = 0 \\ (1 - \pi_i)\frac{\lambda_i^{y_i}}{y_i!}\exp(-\lambda_i), & y_i > 0 \end{cases} \tag{18}$$

with conditional expectation

$$\mathbb{E}[y_i|\mathbf{x}_i, \mathbf{z}_i] = 0 \cdot \pi_i + \lambda_i \cdot (1 - \pi_i) = \lambda_i(1 - \pi_i) \tag{19}$$

Note, that $\mathbf{z}_i$ and $\mathbf{x}_i$ need not necessarily be the same. The solutions for the parameters are obtained by Maximum Likelihood estimation, i.e., by finding the optimum of

$$\begin{aligned} \mathcal{L} = &\sum_{i:y_i=0} \log\left(\exp(\mathbf{z}_i^T\boldsymbol{\gamma}) + \exp(-\lambda_i)\right) + \sum_{i:y_i>0} y_i \log(\lambda_i) - \log(y_i) - \lambda_i \\ &- \sum_{i=1}^{n} \log(1 + \exp(\mathbf{z}_i^T\boldsymbol{\gamma})). \end{aligned} \tag{20}$$

See Appendix A.1 for the derivation of the log-likelihood. This does not yield a closed-form solution, which is why numerical procedures have to be used. Nanjundan and Naika (2012) discuss the solution methods in further detail.

Zero-Inflated Negative Binomial Regression The Zero-Inflated Negative Binomial (ZINB) Model is a generalisation to the ZIP model in that it allows for overdispersion, meaning that the variance is not restricted to be equal to the mean. Much like in the ZIP model, we assume

$$Y_i \sim \begin{cases} 0, & \text{w.p. } \pi_i \\ NB(\lambda_i), & \text{w.p. } 1 - \pi_i. \end{cases} \tag{21}$$

And again, π_i is estimated by some binary response model. One parametrisation of the negative binomial distribution is given by

$$h(y_i) := P(Y = y_i|\lambda_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)}(1 - r_i)^{\alpha^{-1}} r_i^{y_i}, \tag{22}$$

where α is a dispersion parameter and $r_i := \frac{\lambda_i}{\lambda_i + \alpha^{-1}} = \frac{\lambda_i \alpha}{1 + \lambda_i \alpha}$. Thus,

$$h(y_i) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(\frac{1}{1 + \lambda_i \alpha}\right)^{\alpha^{-1}} \left(\frac{\lambda_i \alpha}{1 + \lambda_i \alpha}\right)^{y_i} \tag{23}$$

λ_i is the same as in the previous section. The probability distribution of Y under this model is

$$P(Y = y_i | \mathbf{x}_i, \mathbf{z}_i) = \begin{cases} \pi_i + (1 - \pi_i) \left(\frac{1}{1+\lambda_i \alpha}\right)^{\alpha^{-1}}, & y_i = 0 \\ (1 - \pi_i) h(y_i), & y_i > 0 \end{cases} \tag{24}$$

with conditional expectation

$$\mathbb{E}[y_i | \mathbf{x}_i, \mathbf{z}_i] = 0 \cdot \pi_i + \lambda_i \cdot (1 - \pi_i) = \lambda_i (1 - \pi_i), \tag{25}$$

as in the ZIP model (Berk & MacDonald, 2008). Again, regression coefficients and α are estimated by maximum likelihood (see Appendix A.2 for the log-likelihood), using numerical techniques. Details are found in Hilbe (2011) and Berk and MacDonald (2008).

5.3 Empirical Results

5.3.1 Model Estimation and Testing

For both types of models, we estimate the parameters with log traffic volume and maximum permissible speed as explanatory variables. We choose a logistic regression for the zero-inflation part of the model. Before the estimation, we removed road segments with more than 20 accidents in one year, because they are clear outliers with leverage, and not conducive to the quality of the prediction model (cp. Figure 22 in the appendix) and we focus on predictive aspects as foundation four our simulation outcomes, rather than on causal inference. Results for this estimation are shown in Tables 6 and 7 in the Appendix. We summarise the most important results here.

First, we note that in both cases, all coefficient estimates are statistically significant. The main parameter values for log traffic volume (`log_flow_n`)[22] and maximum permissible speed are positive, indicating that more accidents happen when traffic volume or maximum road speed are higher. Both intuition and literature suggest that this is the correct relationship, as congestion (related to traffic volume during the peak hours) is often found to have a linear positive association with traffic accidents (Retallack & Ostendorf, 2019). However, some

[22]Technically, we deal with yearly volume, which is a variant of flow, as flow is defined as the number of passing vehicles per unit of time, here: year.

studies also find a U-shaped relationship with highest occurrence of accidents in low and high levels of congestion (Retallack & Ostendorf, 2019). Likewise, the parameter values for the two independent variables for the logistic regression are negative in both models. Since the outcome "no accident happens on road segment i" is coded as 1, this does make sense as well, since it indicates that the probability of an accident not happening on a road segment declines with increasing traffic or maximum permissible speed, ceteris paribus.

The interpretation of the parameters is the same for both models, but not straightforward. It is $\hat{\boldsymbol{\beta}}_j = \log(\hat{y}_{x_j=x_j+1}) - \log(\hat{y}_{x_j=x}) = \log\left(\frac{\hat{y}_{x_j=x_j+1}}{\hat{y}_{x_j=x}}\right)$, i.e., they are the natural logarithm of the ratio of expected counts if j increases by a unit (UCLA: Statistical Consulting Group, n. d.b). Fortunately, our goal is prediction rather than inference, and so the coefficient itself does not necessarily need to have an economic interpretation. For the ZINB, the dispersion parameter α was estimated to be $\alpha \approx 2.03 \iff \alpha^{-1} \approx 0.49$, thus the model detected some overdispersion in the data and accounted for that[23]. Finally, we note that the log-likelihood of the ZINB is greater than that of the ZIP.

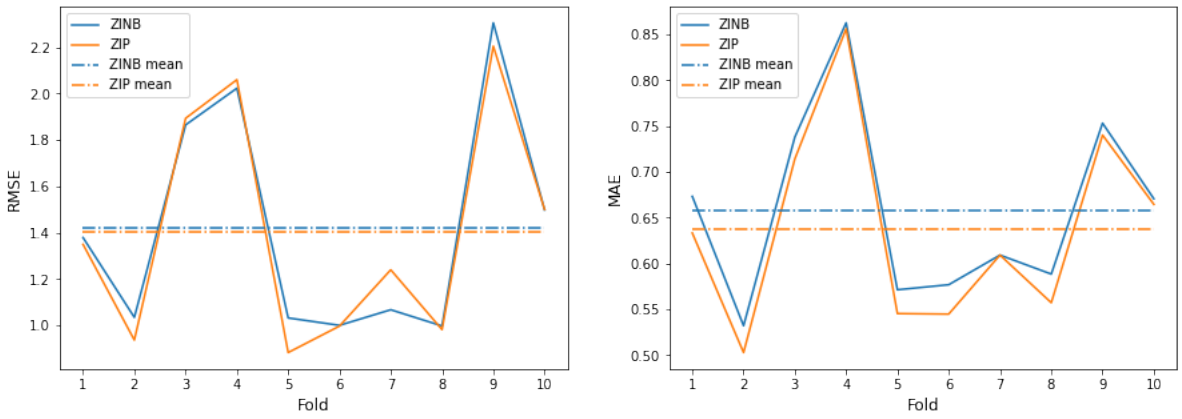
To select the model with the best predictive performance, we evaluate both models on the full dataset using 10-fold cross-validation (Hastie, Tibshirani, & Friedman, 2009, Chapter 7) with root mean squared error (RMSE) and mean absolute error (MAE) as evaluation measures[24]. Both models perform similarly under both evaluation criteria, with the ZIP model performing slightly better in both cases (Figure 16, $RMSE_{\text{ZIP}} \approx 1.40 \leq 1.42 \approx RMSE_{\text{ZINB}}$; $MAE_{\text{ZIP}} \approx 0.64 \leq 0.66 \approx MAE_{\text{ZINB}}$. We therefore choose the ZIP model as our preferred predictive model.

We perform some further model inspection. Figure 17 displays the probability mass function and the cumulative distribution function of the counts in the data and as predicted by our model. The ZIP model predicts zero counts slightly more often as there are in the data, and compensates by predicting less one counts. Two and three counts are predicted slightly more often than there are in the data, but otherwise the two distributions align very well. More concisely, the model predicts 57.7% of the counts exactly correctly, overshoots the count by one in 25.9% of the cases, and underestimates the count by one in 9.2% of the cases. In 92.9% of the cases, the prediction is therefore correct within a tolerance area of size one. This is exemplified in Figure 18.

[23]Note, that $\mathbb{V}[Pois(\lambda_i)] = \lambda_i$ and $\mathbb{V}[NB(\lambda_i, \alpha)] = \lambda_i(1 + \alpha\lambda_i)$ in this parametrisation.

[24]For true outcome vector $\mathbf{y}$ and prediction vector $\hat{\mathbf{y}}$, $RMSE := \sqrt{\frac{1}{n}\left|\left|\hat{\mathbf{y}} - \mathbf{y}\right|\right|_2^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ and $MAE := \frac{1}{n}\left|\left|\hat{\mathbf{y}} - \mathbf{y}\right|\right|_1 = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)|$.

Figure 16: Model testing - 10-fold cross-validation (RMSE / MAE)



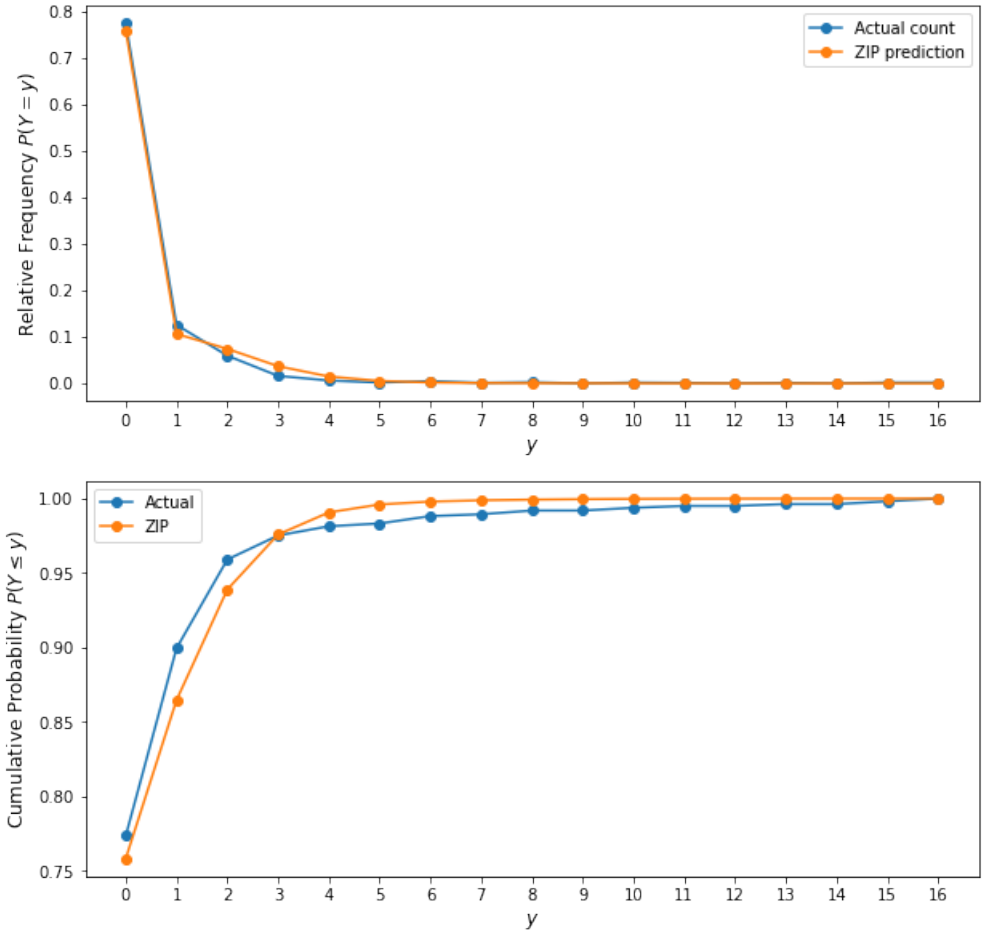
Notes: Root mean squared error (left) and Mean absolute error (right) of the ZIP and ZINB models on each training fold. Dashed lines represent overall means.

5.3.2 Traffic Accidents in 2020

Having decided on the predictive model to use, we now want to estimate the number of accidents that were avoided in 2020 due to a reduction in traffic volume. To assess the actual reduction of traffic volume, we downloaded traffic detector data from the years 2017 and 2020 from the traffic information centre Berlin[25]. This dataset contains data from 529 unique traffic detectors at different locations in Berlin. For each traffic detector, hourly vehicle counts and mean speeds are available. We filter this data to only include measurements from the specified rush hour time window on weekdays. We further restrict the sample to only include measurements from April to December. Even though the first lockdown in Germany came into effect on March 16$^{\text{th}}$, we cannot take these two more weeks into consideration, because the data for 2017 is only available from April onward. We also remove any measurements which may be inaccurate (as indicated by a "quality" variable in the data). Observations with less than perfect quality amount to 1.92% of the filtered data. Then, we aggregate the data on the detector level by taking the mean. This is done to prevent missing measurements in either the 2017 or 2020 dataset to distort the final estimate. Finally, we take the sum of the aggregated means for each of the years and divide the sum of 2020 by the sum of 2017 to arrive at an estimate of how much less traffic there was in 2020. The estimate is 87.29%, meaning that over the course of post-pandemic Berlin in 2020, total traffic went down by about 13% compared with the year 2017. As a sanity check, we made the same comparison with the year 2019 and it produced a similar estimate of 88.47%. In Figure 19, daily vehicle

[25] `https://api.viz.berlin.de/daten/verkehrsdetektion`

Figure 17: Model fit - density and distribution functions



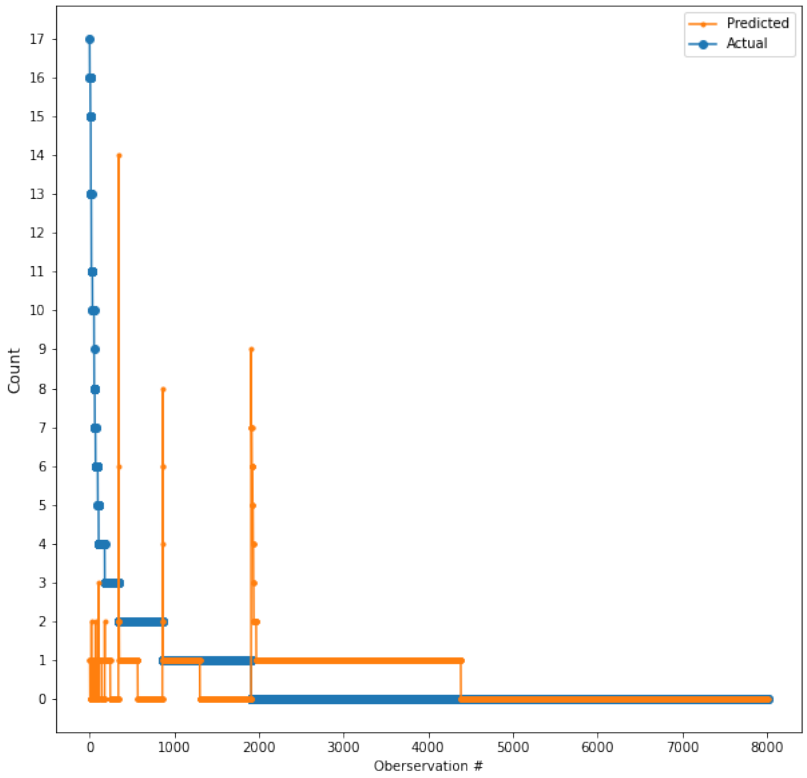
Notes: Probability mass function (top) and cumulative distribution function (bottom) of the data and ZIP model predictions.

counts are plotted against the yearly mean vehicle count against the mean (at every detector) of the year 2017. The stark drop at the time of the first lockdown is clearly visible, the second one beginning on December 16_{th} is even more pronounced[26].

As a next step, we draw a 87.29% sample of our trip data. As the data is assumed to be representative, we abstain from stratified sampling in order to keep it that way. Using this sample, we count the new traffic on Berlin's roads as described in Section 5.2.2. We then select those 8,011 road segments that were found to be *critical* ($RSI_i \leq 0.5$) in the overall 2017 data. These are fed into the ZIP model from Section 5.3.1. Our model predicts a total of 3,217 accidents, thus a reduction by 616 if traffic is reduced by approximately 13%. This constitutes an unproportional reduction in traffic accidents by 16%. When we correct for the accident inflation induced by our variable specification, this reduces to a predicted reduction

[26]Similar figures for single streets were available at `https://viz.berlin.de/2020/12/verkehrsstaerke/` and can still be found using internet archives.

Figure 18: Comparison of predicted and actual vehicle counts



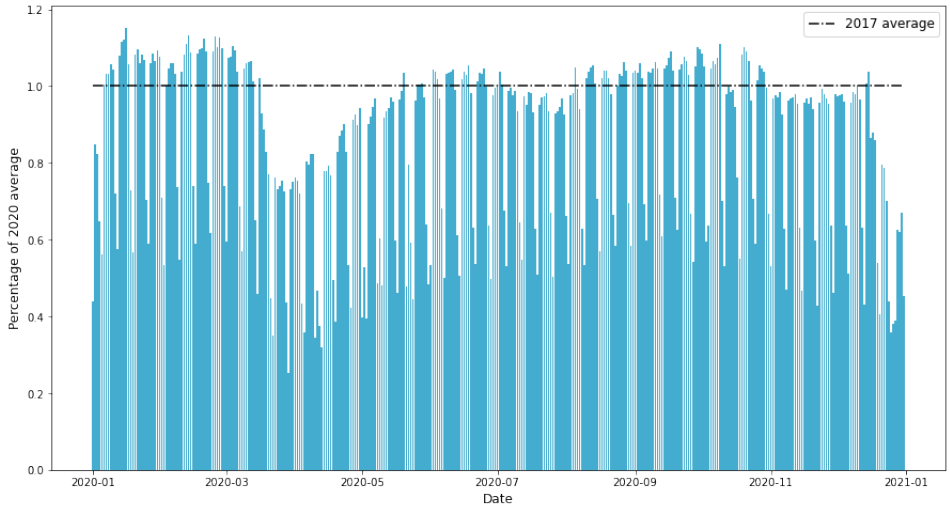
Notes: Predicted accident counts and actual accident counts, sorted by count of actually observed vehicles in descending order. X axis displays observations in the full dataset.

of 127 traffic-relevant accidents[27].

Since our accident data does not include information about injuries, fatalities, or material damage, we have to perform a rough estimation in order to evaluate the monetary impact associated with the reduction in accidents. The German Federal Statistical Office provides traffic accident statistics for every year and state (*Bundesland*), with a subdivision of accidents into different types (Statistisches Bundesamt, 2018) and the Federal Road Research Institute calculates the average cost per traffic accident of each type for every year. Table 3 combines this information and reports the total number and percentage of accidents observed on German roads by category, as well as their estimated economic cost (Bundesanstalt für Straßenwesen, 2021). In general, accidents are subdivided into accidents *with injury to per-*

[27] 3,217 accidents are more than the original 792 accidents in our data, but recall that for the original model estimation, accident data was artificially inflated because accidents were counted for several road segment buffers. Inflated accidents for the original data (without dropped outliers) amount to 3,833, i.e., the original accident data was inflated by the factor 4.84. Assuming that in 2020, accidents would have scaled in the same way, we divide the difference of inflated accidents (616) by this factor and obtain $127.27 \approx 127$ traffic-relevant accidents in Berlin that happened less in 2020 when compared to 2017. This kind of extrapolation probably comes with systematic noise, because less overall traffic might lead to even less accidents exactly there where many normally busy roads (with high RSI) overlap. However, this implies that the inflation factor will be smaller for areas of the network with less traffic. Therefore our estimate obtained will likely be conservative.

Figure 19: Traffic volume as counted by traffic detectors in Berlin



Notes: Only traffic between 06:00 - 08:59 and 14:00 - 19:59 is considered. Blue bars represent traffic in 2020 for each day, the black dashed line is the mean of 2020 (= 100%) during weekdays. For each detector, the daily traffic volume in 2020 was computed and then compared to the 2017 mean on that detector. Finally, percentages were averaged over all detectors.

sons and *damage-only*-accidents. In both cases, the upper part of Table 3 only displays the cost associated with the material damage. In the 14,493 accidents that caused (any) bodily injury, 17,420 people were injured. This means that in every accident that induced bodily injury, on average 1.2 persons were hurt. The lower part of Table 3 reports the cost for each and all cost associated with the injuries for different types of personal damage. In total, the 143,442 accidents that happened in Berlin entailed a total economic cost of 1,457,369,659 €. This sum corresponds to 4.26% of the total cost caused by traffic accidents in Germany in the year 2017. This percentage is also approximately equal to Berlin's share of the total population in 2017.

Assuming that the distribution into categories is the same for 2020, we can now calculate the economic savings. Drawing from Tang and Ahlfeldt (2020), we express the predicted economic savings on the critical road segments as

$$B = \sum_{k=1}^{K} (Y_k^{2020} - \hat{Y}_k^{2017}) \times M_k^{2017}, \tag{26}$$

where B is the amount of predicted economic savings, Y_k^{2017} is the number of accidents of each type k in Berlin in the year 2017, as specified above, $\hat{Y}_{k,2020}$ is the prediction for this type of accident given the simulated traffic data of 2020, and M_k^{2017} denotes the average monetized cost from an additional accident of type k. Instead of providing average values for

Table 3: Accidents and associated costs in Berlin and Germany in the year 2017

Accidents		**Count**	**Share**	**avg. cost / accident**	**Total Cost**
with injury to persons		14,493	10.10%	17,391* €	252,047,763 €
damage-only	in the strict sense†	1,233	0.86%	22,839 €	28,160,487 €
	other accidents under the influence of intoxicants	806	0.56%	6,177 €	4,978,662 €
	others	126,910	88.47%	6,177 €	783,923,070 €
Subtotal		**143,442**			**1,069,109,982 €**
Accident victims		**Count**	**Share**	**avg. cost / person**	**Total Cost**
killed		36	0.21%	1,150,234 €	41,408,424 €
seriously injured		2,316	13.30%	116,335 €	269,431,860 €
slightly injured		15,068	86.50%	5,138 €	77,419,384 €
Subtotal		**17,420**			**388,259,384 €**
German Total Population		82,521,653	Total cost Germany		34,230,000,000 €
Berlin Total Population		3,574,830	Total cost Berlin		1,457,369,650 €
Share		4.33%	Share		4.26%

Notes: Accidents in 2017 and associated incurred economic cost. Accident counts and economic costs per accident / person taken from Statistisches Bundesamt (2018) and Bundesanstalt für Straßenwesen (2021). In the 14,493 accidents with damage to persons, a total of 17,420 (factor: 1.20) people were injured or killed. Berlin has a share of total incurred economic cost by accidents that is approximately equal to its share of population. Numbers may not sum up to totals because of rounding errors.
*: Costs in the upper panel are only associated with the material damage, as costs related to physical damage to persons are outlined in the bottom part.
†: Accidents with only material damage in the strict sense is a definition by Statistisches Bundesamt (2018) and is used for accidents whose cause is a misdemeanour or a criminal act relating to road traffic, and after which a vehicle has to be towed away from the accident site.

M_k^{2017} directly, we instead calculate the projected number of accidents of every type for the total of 127 accidents that were avoided in 2020 as predicted by our model in Table 4. The accuracy of these calculations depends on three factors: (1) the accuracy of our predictive model, (2) the accuracy of the "inflation factor" outlined above, and (3) whether or not the distribution of accident types is the same in 2020 as in 2017. We arrive at an estimate of 1,251,525 € in total savings. Considering that these are only the economic savings associated with avoided traffic-relevant accidents on critical road segments in the Berlin UC area during the rush hour time windows, this figure is considerable.

But this is just the estimate for a very limited sample of accidents, namely the traffic-relevant accidents during the rush hour on weekdays. If we extrapolate this number to the total number of avoided traffic-relevant accidents in Berlin, we arrive at $\frac{127.17}{791} \times 3,224 \approx 518$

Table 4: Estimated traffic-relevant accidents and associated savings in Berlin and Germany for the year 2020 during the rush hour time windows defined in Section 5.2.2

Accidents		**Count**	**avg. cost / accident**	**Total Savings**
with injury to persons		$127 \times 10.10\% \approx 13$	17,391 €	226,083 €
damage-only	in the strict sense	$123 \times 0.86\% \approx 1$	22,839	22,839 €
	other accidents under the influence of intoxicants	$127 \times 0.56\% \approx 1$	6,177 €	6,177 €
	others	$127 \times 88.47\% \approx 112$	6,177 €	691,824 €
Subtotal		**127**		**946,923 €**
Accident victims		**Count**	**avg. cost / person**	**Total Savings**
killed		$16 \times 0.21\% \approx 0$	1,150,234 €	0 €
seriously injured		$16 \times 13.30\% \approx 2$	116,335 €	232,670 €
slightly injured		$16 \times 86.50\% \approx 14$	5,138 €	71,932 €
Subtotal		$\mathbf{13 \times 1.2 \approx 16}$		**304,602 €**
Total Savings				**1,251,525 €**

Notes: 127 accidents were predicted to happen less on critical road segments in the Berlin UC area during rush hours on weekdays. Distributional probabilities (percentages) were taken from 2017 (Table 3). In the 13 accidents with injury to persons, 16 people were hurt overall, but no one died. The factor 1.2 is also taken from the reference Table 3. Numbers may not sum up to totals because of rounding errors.

avoided accidents[28], making up for savings of 5,051,417 € (cp. Table 11 in the appendix). Finally, extrapolating the number to the total number of accidents avoided in Berlin amounts to $\frac{127.17}{791} \times 143,442 \approx 23,061$ accidents, saving 234,502,891 € and, importantly, six lives. The estimates produced here turn out to be relatively precise when evaluated against the true accident numbers, which are published in Statistisches Bundesamt (2021). In fact, the number of accidents in Berlin is 123,928, so that our estimate of $(1 - \frac{127.17}{791}) \times 143,442 \approx 120,381$ total traffic accidents is only off by 3,547, or 2.86%. Using the same cost figures for each accident type as in 2017, the total actual cost of traffic accidents in Berlin 2017 amounts to 1,281,917,077 €, our estimate is 1,222,866,759 €, therefore we underestimate the cost by 59,050,318 €, again only 4.61%. A good part of this deviance is explained by an extraordinarily high number of 50 killed accident victims in 2020, while our model only predicted 30 for 2020. Interestingly, too, the factor of injured people in each accident with bodily injury is almost as high as in 2017 (1.17 vs. 1.20). These findings are summarised in Table 5, Table 12 displays actual accident numbers and associated costs for 2020.

[28]Recall that 791 is the number of accidents in the Berlin UC area during rush hours and 3,224 is the total number of accidents in Berlin as per the INRIX accident data (Section 3.2).

Table 5: Extrapolated accident costs of accidents in 2020 in Berlin, estimated savings, and estimation deviations

Accidents		**Count**	**Deviation**	**avg. cost / accident**	**Total Cost**	**Deviation**	**Deviation in %**
with injury to persons		$120,381 \times 10.10\% \approx 12,163$	-698	17,391 €	211,526,733 €	-12,138,918 €	-5.43
damage-only	in the strict sense	$120,381 \times 0.86\% \approx 1,035$	155	22,839 €	23,638,365 €	3,540,045 €	17.61
	other accidents under the influence of intoxicants	$120,381 \times 0.56\% \approx 676$	-2	6,177 €	4,175,652 €	-12,354 €	-0.29
	others	$120,381 \times 88.47\% \approx 106,507$	-3,002	6,177 €	657,893,739 €	-18,543,354 €	-2.74
Subtotal		**120,381**	**-3,547**		**897,234,489 €**	**-27,154,581 €**	**-2.94**
Accident victims		**Count**	**Deviation**	**avg. cost / person**	**Total Cost**	**Deviation**	**Deviation in %**
killed		$14,619 \times 0.21\% \approx 30$	-20	1,150,234 €	34,507,020 €	-23,004,680 €	-40.00
seriously injured		$14,619 \times 13.30\% \approx 1,944$	-63	116,335 €	226,155,240 €	-7,329,105 €	-3.14
slightly injured		$14,619 \times 86.50\% \approx 12,645$	-304	5,138 €	64,970,010 €	-1,561,952 €	-2.35
Subtotal		$12,163 \times 1.2 \approx 14,619$	**-387**		**325,632,270 €**	**-31,895,737 €**	**-8.92**
Total Cost					**1,222,866,759 €**	**-59,050,318 €**	**-4.61**
Difference to 2017 Total (Estimated savings)					**234,502,891 €**		

Notes: Estimated accidents for Berlin 2020 and deviations from actual accident numbers. Negative numbers mean that more accidents happened / higher costs incurred in actuality (cp. Table 12). By this extrapolation, 120,381 accidents are estimated to happen in Berlin 2020, so that it underestimates the real number by 3,547. Moreover, a total cost of 1,222,866,759 €is estimated to be caused by accidents in 2020, 234,502,891 €less than the real figure for 2017 (cp. Table 3). Numbers may not sum up to totals because of rounding errors.

6 Conclusion

In this thesis, we analysed the Berlin traffic network using a network representation with data from OpenStreetMap. We found that Berlin's OSM UC network representation is similar in properties to previously used representations, and we pointed out special characteristics pertaining to Berlin's traffic network. Using high-resolution real-world traffic data, we also implemented a method to identify potentially congested road segments in the streets of Berlin, and used these critical segments to build a model predicting accidents on busy roads. Finally, using this model, we calculated the estimated economic savings of less busy roads during the COVID-19 pandemic in 2020 and arrived at total savings of approximately 1.251M € resulting only from avoided accidents on *critical* roads during the rush hour on weekdays. Extrapolating the data onto total accidents in Berlin yielded an astoundingly precise estimate of 120,381 accidents, with the real number for 2020 being 123,928. Estimated savings for 2020 amount to approximately 230M € while real savings were only about 175M € mostly due to a different distribution of accident types in 2020. In any case, our model predicts the true number of accidents within a margin of 2.86% and based on this, our calculations come as close as 4.61% to the truth. Both savings figures are economically significant.

A second contribution of our work is that we exemplified how approaches from computational network science can aid in economic analyses. Computational network science is well-equipped to work with big data such as ours, and the approach taken in our case study serves as evidence that even with relatively simple economic models, very precise estimations can be performed if the right data and method are at hand. We also note that the approach taken here can be used for any street network and is not limited to Berlin. Furthermore, any reduction in traffic can be modelled with our approach, given suitable data, and the COVID-19 case study is just an example.

Our results also indirectly hint at policy implications: less automotive traffic on the streets not only reduces pollution but also accidents. Getting people off the streets, for example by enhancing public transportation offerings and incentivising the usage of alternative transportation methods and carpooling, saves money and lives.

As we only looked at one very specific, real traffic reduction, we encourage future researchers to examine the effects of differing levels of traffic reduction to examine possible effects of scale, and study the association of traffic accidents and traffic volume across volume levels. It would also be interesting to dig deeper into the spatial and temporal distribution of accidents to find areas or times where a traffic reduction would make sense most. Moreover,

to validate the findings of this thesis, examples from other German and international cities would be insightful.

Considering city road and mobility networks, further research could focus on comprehensive reviews and graph-theoretic comparisons of Urban Centre road networks as defined and provided by Boeing (in press). Given the many technological developments of the past few years (OSM, `OSMNx`), this has become a feasible endeavour. Using similar data as ours future research in computational network science and economics could also focus on topic such as "traffic justice", e.g., by examining who the main causers of traffic on critical road segments are, and what other interactions critical road segments have with their environment (e.g. concerning economic activity or environmental aspects).

Finally, we note that this work has limitations, too. First, in the network analysis of the Berlin road network, we primarily focussed on its undirected representation. We did this to have direct comparisons with prior work in the field, but as direction is an important aspect in real-life traffic, omitting this information might influence results, for example, when looking at centrality measures. Second, in the case study, we were constrained by the data to produce a model that accommodates more predictors actually associated with traffic accidents, such as weather conditions. Although the model already performs well with only two explanatory variables, the estimates produced might be even more precise with additional relevant information. Moreover, the accident counting method allows for some imprecision and leads to an inflated total number of accidents, a factor that we have to account for, and which might induce additional noise. Finally, we base our extrapolations on data stemming from only traffic-relevant accidents during the rush hour, off-hour accidents or non-traffic-relevant accidents might have other characteristics, inducing more systematic noise into our calculations. Nevertheless, when compared with the real figures, our estimates are still very precise.

References

Abdulla, B., & Birgisson, B. (2021). Characterization of vulnerability of road networks to random and nonrandom disruptions using network percolation approach. *Journal of Computing in Civil Engineering*, *35*(1), 04020054. doi: 10.1061/(asce)cp.1943-5487.0000938

Albert, R., Albert, I., & Nakarado, G. L. (2004). Structural vulnerability of the north american power grid. *Physical Review E*, *69*, 025103. Retrieved from `https://link.aps.org/doi/10.1103/PhysRevE.69.025103` doi: 10.1103/PhysRevE.69.025103

Alderson, D., Li, L., Willinger, W., & Doyle, J. (2005). Understanding internet topology: principles, models, and validation. *IEEE/ACM Transactions on Networking*, *13*(6), 1205-1218. doi: 10.1109/TNET.2005.861250

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, *8*(6), 450–461. doi: 10.1038/nrg2102

Amézquita-López, J., Valdés-Atencio, J., & Angulo-García, D. (2021). Understanding traffic congestion via network analysis, agent modeling, and the trajectory of urban expansion: A coastal city case. *Infrastructures*, *6*(6), 85. doi: 10.3390/infrastructures6060085

Andreescu, M., & Frost, D. (1998). Weather and traffic accidents in montreal, canada. *Climate Research*, *9*, 225–230. Retrieved from `https://doi.org/10.3354/cr009225` doi: 10.3354/cr009225

Barabási, A.-L., & Albert, R. (1999, October). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. Retrieved from `https://doi.org/10.1126/science.286.5439.509` doi: 10.1126/science.286.5439.509

Barabási, A.-L., & Pósfai, M. (2016). *Network science*. Cambridge: Cambridge University Press. Retrieved from `http://barabasi.com/networksciencebook/`

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., ... Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports*, *734*, 1-74. Retrieved from `https://www.sciencedirect.com/science/article/pii/S037015731830022X` doi: https://doi.org/10.1016/j.physrep.2018.01.001

Barrington-Leigh, C., & Millard-Ball, A. (2017, 08). The world's user-generated road map is more than 80complete. *PLOS ONE*, *12*(8), 1-20. Retrieved from `https://doi.org/10.1371/journal.pone.0180698` doi: 10.1371/journal.pone.0180698

Barthélemy, M. (2011, February). Spatial networks. *Physics Reports*, *499*(1-3), 1–101. Retrieved from `https://doi.org/10.1016/j.physrep.2010.11.002` doi: 10.1016/

j.physrep.2010.11.002

Barthelemy, M. (2018). *Morphogenesis of spatial networks*. Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-319-20565-6 doi: 10.1007/978-3-319-20565-6

Bazzani, A., Giorgini, B., Rambaldi, S., Gallotti, R., & Giovannini, L. (2010). Statistical laws in urban mobility from microscopic gps data in the area of florence. *Journal of Statistical Mechanics: Theory and Experiment*, *2010*, 05001.

Berk, R., & MacDonald, J. M. (2008, April). Overdispersion and poisson regression. *Journal of Quantitative Criminology*, *24*(3), 269–284. Retrieved from https://doi.org/10.1007/s10940-008-9048-4 doi: 10.1007/s10940-008-9048-4

Boeing, G. (2017). Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, *65*, 126-139. Retrieved from https://www.sciencedirect.com/science/article/pii/S0198971516303970 doi: 10.1016/j.compenvurbsys.2017.05.004

Boeing, G. (2018, August). A multi-scale analysis of 27, 000 urban street networks: Every US city, town, urbanized area, and zillow neighborhood. *Environment and Planning B: Urban Analytics and City Science*, *47*(4), 590–608. Retrieved from https://doi.org/10.1177/2399808318784595 doi: 10.1177/2399808318784595

Boeing, G. (2019a). The morphology and circuity of walkable and drivable street networks. In *The mathematics of urban morphology* (pp. 271–287). Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-030-12381-9_12 doi: 10.1007/978-3-030-12381-9_12

Boeing, G. (2019b, August). Urban spatial order: street network orientation, configuration, and entropy. *Applied Network Science*, *4*(1). Retrieved from https://doi.org/10.1007/s41109-019-0189-1 doi: 10.1007/s41109-019-0189-1

Boeing, G. (2020a). *Global Urban Street Networks GraphML.* Harvard Dataverse. Retrieved from https://doi.org/10.7910/DVN/KA5HJ3 doi: 10.7910/DVN/KA5HJ3

Boeing, G. (2020b). *Global Urban Street Networks Indicators.* Harvard Dataverse. Retrieved from https://doi.org/10.7910/DVN/ZTFPTB doi: 10.7910/DVN/ZTFPTB

Boeing, G. (2020c). Off the grid... and back again? the recent evolution of american street network planning and design. *SSRN Electronic Journal*. Retrieved from https://doi.org/10.2139/ssrn.3665445 doi: 10.2139/ssrn.3665445

Boeing, G. (2020d). Planarity and street network representation in urban form analy-

sis. *Environment and Planning B: Urban Analytics and City Science*, *47*(5), 855-869. Retrieved from https://doi.org/10.1177/2399808318802941 doi: 10.1177/2399808318802941

Boeing, G. (2021, February). Spatial information and the legibility of urban form: Big data in urban morphology. *International Journal of Information Management*, *56*, 102013. Retrieved from https://doi.org/10.1016/j.ijinfomgt.2019.09.009 doi: 10.1016/j.ijinfomgt.2019.09.009

Boeing, G. (in press). Street network models and indicators for every urban area in the world. *Geographical Analysis*, *n/a*(n/a). Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/gean.12281 doi: 10.1111/gean.12281

Buhl, J., Gautrais, J., Reeves, N., Solé, R. V., Valverde, S., Kuntz, P., & Theraulaz, G. (2006). Topological patterns in street networks of self-organized urban settlements. *Eur. Phys. J. B*, *49*(4), 513-522. Retrieved from https://doi.org/10.1140/epjb/e2006-00085-1 doi: 10.1140/epjb/e2006-00085-1

Bullmore, E., & Sporns, O. (2009, March). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3), 186–198. doi: 10.1038/nrn2575

Bundesanstalt für Straßenwesen. (2021, May). *Volkswirtschaftliche kosten von straßenerkehrsunfällen in deutschland* (Tech. Rep.). Bergisch Gladbach. Retrieved from https://www.bast.de/BASt_2017/DE/Statistik/Unfaelle/volkswirtschaftliche_kosten.pdf?__blob=publicationFile&v=15 on September 21st, 2021.

Cardillo, A., Scellato, S., Latora, V., & Porta, S. (2006, Jun). Structural properties of planar graphs of urban street patterns. *Phys. Rev. E*, *73*, 066107. Retrieved from https://link.aps.org/doi/10.1103/PhysRevE.73.066107 doi: 10.1103/PhysRevE.73.066107

Chen, B. Y., Lam, W. H., Sumalee, A., Li, Q., & Li, Z.-C. (2012, March). Vulnerability analysis for large-scale and congested road networks with demand uncertainty. *Transportation Research Part A: Policy and Practice*, *46*(3), 501–516. Retrieved from https://doi.org/10.1016/j.tra.2011.11.018 doi: 10.1016/j.tra.2011.11.018

Courtat, T., Gloaguen, C., & Douady, S. (2011, Mar). Mathematics and morphogenesis of cities: A geometrical approach. *Phys. Rev. E*, *83*, 036106. Retrieved from https://link.aps.org/doi/10.1103/PhysRevE.83.036106 doi: 10.1103/PhysRevE

.83.036106

Crucitti, P., Latora, V., & Porta, S. (2006, March). Centrality measures in spatial networks of urban streets. *Physical Review E*, *73*(3). Retrieved from https://doi.org/10.1103/physreve.73.036125 doi: 10.1103/physreve.73.036125

de Winter, J., & Dodou, D. (2010, December). The driver behaviour questionnaire as a predictor of accidents: A meta-analysis. *Journal of Safety Research*, *41*(6), 463–470. Retrieved from https://doi.org/10.1016/j.jsr.2010.10.007 doi: 10.1016/j.jsr.2010.10.007

Dumedah, G., & Garsonu, E. K. (2021, January). Characterising the structural pattern of urban road networks in ghana using geometric and topological measures. *Geo: Geography and Environment*, *8*(1). Retrieved from https://doi.org/10.1002/geo2.95 doi: 10.1002/geo2.95

Eisenberg, D. (2004, July). The mixed effects of precipitation on traffic crashes. *Accident Analysis & Prevention*, *36*(4), 637–647. Retrieved from https://doi.org/10.1016/s0001-4575(03)00085-x doi: 10.1016/s0001-4575(03)00085-x

Erath, A., Birdsall, J., Axhausen, K. W., & Hajdin, R. (2009, January). Vulnerability assessment methodology for swiss road network. *Transportation Research Record: Journal of the Transportation Research Board*, *2137*(1), 118–126. Retrieved from https://doi.org/10.3141/2137-13 doi: 10.3141/2137-13

EROS Earth Resources Observation and Science Center. (2017). *Shuttle radar topography mission (srtm) void filled.* U.S. Geological Survey. Retrieved from https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-void doi: 10.5066/F7F76B1X

European Data Portal. (2020). *The COVID-19 related traffic reduction and decreased air pollution in Europe.* Retrieved from https://data.europa.eu/en/impact-studies/covid-19/covid-19-related-traffic-reduction-and-decreased-air-pollution-europe on August 20th, 2021.

Florczyk, A. J., Melchiorri, M., Corbane, C., Schiavina, M., Maffenini, M., Pesaresi, M., ... Zanchetta, L. (2019). *Description of the GHS Urban Centre Database 2015: Public release 2019 : Version 1.0.* LU: Publications Office of the European Union.

Gao, S., Wang, Y., Gao, Y., & Liu, Y. (2013). Understanding urban traffic-flow characteristics: A rethinking of betweenness centrality. *Environment and Planning B: Planning*

and Design, *40*(1), 135–153. Retrieved from https://doi.org/10.1068/b38141 doi: 10.1068/b38141

Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying online social networks. *Journal of Computer-Mediated Communication*, *3*(1). doi: 10.1111/j.1083-6101.1997.tb00062.x

Gicquel, L., Ordonneau, P., Blot, E., Toillon, C., Ingrand, P., & Romo, L. (2017). Description of various factors contributing to traffic accidents in youth and measures proposed to alleviate recurrence. *Frontiers in Psychiatry*, *8*. Retrieved from https://doi.org/10.3389/fpsyt.2017.00094 doi: 10.3389/fpsyt.2017.00094

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th python in science conference* (p. 11 - 15). Pasadena, CA USA.

Haggett, P., & Chorley, R. J. (1969). *Network analysis in geography.* London: Edward Arnold.

Haklay, M. (2010). How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682-703. Retrieved from https://doi.org/10.1068/b35097 doi: 10.1068/b35097

Hale, D., Jagannathan, R., Xyntarakis, M., Su, P., Jiang, X., Ma, J., ... Krause, C. (2016). *Traffic Bottlenecks: Identification and Solutions.*

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction* (2nd ed.). Springer.

He, F., Yan, X., Liu, Y., & Ma, L. (2016). A traffic congestion assessment method for urban road networks based on speed performance index. *Procedia Engineering*, *137*, 425–433. Retrieved from https://doi.org/10.1016/j.proeng.2016.01.277 doi: 10.1016/j.proeng.2016.01.277

Hilbe, J. (2011). *Negative binomial regression.* Cambridge University Press.

Ivan, J. N., Wang, C., & Bernardo, N. R. (2000). Explaining two-lane highway crash rates using land use and hourly exposure. *Accident Analysis & Prevention*, *32*(6), 787–795. Retrieved from https://doi.org/10.1016/s0001-4575(99)00132-3 doi: 10.1016/s0001-4575(99)00132-3

Jenelius, E., Petersen, T., & Mattsson, L.-G. (2006). Importance and exposure in road network vulnerability analysis. *Transportation Research Part A: Policy and Practice*,

40(7), 537-560. Retrieved from https://www.sciencedirect.com/science/article/pii/S096585640500162X doi: 10.1016/j.tra.2005.11.003

Jiang, B. (2007). A topological pattern of urban street networks: Universality and peculiarity. *Physica A: Statistical Mechanics and its Applications*, *384*(2), 647–655. Retrieved from https://doi.org/10.1016/j.physa.2007.05.064 doi: 10.1016/j.physa.2007.05.064

Jiang, B., & Claramunt, C. (2004, February). Topological analysis of urban street networks. *Environment and Planning B: Planning and Design*, *31*(1), 151–162. Retrieved from https://doi.org/10.1068/b306 doi: 10.1068/b306

Kalapala, V., Sanwalani, V., Clauset, A., & Moore, C. (2006, February). Scale invariance in road networks. *Physical Review E*, *73*(2). Retrieved from https://doi.org/10.1103/physreve.73.026130 doi: 10.1103/physreve.73.026130

Kansky, K. J. (1965). *Structure of Transportation Networks: Relationships Between Network Network Geometry and Regional Characteristics, by K.J. Kansky.* (Unpublished doctoral dissertation). University of Chicago, 1963, Chicago.

Kazerani, A., & Winter, S. (2009). Can betweenness centrality explain traffic flow. In *Proceedings of 12th AGILE International Conference on Geographic Information Science, Hanover, Germany.*

Koch, N., Ritter, N., Rohlf, A., & Thies, B. (2021, Jul). *Machine Learning from Big GPS Data about the Costs of Congestion* (Working Paper). Berlin: Mercator Research Institute on Global Commons and Climate Change.

Kölbl, R., & Helbing, D. (2003, May). Energy laws in human travel behaviour. *New Journal of Physics*, *5*, 48–48. Retrieved from https://doi.org/10.1088/1367-2630/5/1/348 doi: 10.1088/1367-2630/5/1/348

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, *34*(1), 1–14. Retrieved from http://www.jstor.org/stable/1269547

Lämmer, S., Gehlsen, B., & Helbing, D. (2006). Scaling laws in the spatial structure of urban road networks. *Physica A: Statistical Mechanics and its Applications*, *363*(1), 89–95. doi: 10.1016/j.physa.2006.01.051

Latora, V., & Marchiori, M. (2003, March). Economic small-world behavior in weighted networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, *32*(2), 249–263. Retrieved from https://doi.org/10.1140/epjb/e2003-00095-5 doi: 10.1140/epjb/e2003-00095-5

Levinson, D., & El-Geneidy, A. (2009, November). The minimum circuity frontier and the journey to work. *Regional Science and Urban Economics*, *39*(6), 732–738. Retrieved from https://doi.org/10.1016/j.regsciurbeco.2009.07.003 doi: 10.1016/j.regsciurbeco.2009.07.003

Levinson, D., & Wu, Y. (2005, March). The rational locator reexamined: Are travel times still stable? *Transportation*, *32*(2), 187–202. Retrieved from https://doi.org/10.1007/s11116-004-5507-4 doi: 10.1007/s11116-004-5507-4

Levinson, D. M., & Kumar, A. (1994, September). The rational locator: Why travel times have remained stable. *Journal of the American Planning Association*, *60*(3), 319–332. Retrieved from https://doi.org/10.1080/01944369408975590 doi: 10.1080/01944369408975590

Li, C., Yue, W., Mao, G., & Xu, Z. (2020). Congestion propagation based bottleneck identification in urban road networks. *IEEE Transactions on Vehicular Technology*, *69*, 4827-4841.

Li, F., Jia, H., Luo, Q., Li, Y., & Yang, L. (2020, April). Identification of critical links in a large-scale road network considering the traffic flow betweenness index. *PLOS ONE*, *15*(4), e0227474. Retrieved from https://doi.org/10.1371/journal.pone.0227474 doi: 10.1371/journal.pone.0227474

Liang, X., Zhao, J., Dong, L., & Xu, K. (2013, October). Unraveling the origin of exponential law in intra-urban human mobility. *Scientific Reports*, *3*(1), 2983. doi: 10.1038/srep02983

Lie, A., Tingvall, C., Krafft, M., & Kullgren, A. (2006, March). The effectiveness of electronic stability control (ESC) in reducing real life crashes and injuries. *Traffic Injury Prevention*, *7*(1), 38–43. Retrieved from https://doi.org/10.1080/15389580500346838 doi: 10.1080/15389580500346838

Ludwig, I., Voss, A., & Krause-Traudes, M. (2011). A comparison of the street networks of navteq and osm in germany. In S. Geertman, W. Reinhardt, & F. Toppen (Eds.), *Advancing geoinformation science for a changing world* (pp. 65–84). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-642-19789-5_4 doi: 10.1007/978-3-642-19789-5_4

Marchetti, C. (1994, September). Anthropological invariants in travel behavior. *Technological Forecasting and Social Change*, *47*(1), 75–88. Retrieved from https://doi.org/10.1016/0040-1625(94)90041-8 doi: 10.1016/0040-1625(94)90041-8

Mason, O., & Verwoerd, M. (2007, March). Graph theory and networks in Biology. *IET systems biology*, *1*(2), 89–119. doi: 10.1049/iet-syb:20060038

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th acm sigcomm conference on internet measurement* (p. 29–42). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1298306.1298311 doi: 10.1145/1298306.1298311

Münch, F. V., Thies, B., Puschmann, C., & Bruns, A. (2021). Walking through twitter: Sampling a language-based follow network of influential twitter accounts. *Social Media + Society*, *7*(1), 2056305120984475. Retrieved from https://doi.org/10.1177/2056305120984475 doi: 10.1177/2056305120984475

Nanjundan, G., & Naika, T. R. (2012). Asymptotic comparison of method of moments estimators and maximum likelihood estimators of parameters in zero-inflated poisson model. *Applied Mathematics*, *03*(06), 610–616. Retrieved from https://doi.org/10.4236/am.2012.36095 doi: 10.4236/am.2012.36095

NASA/METI/AIST/Japan Spacesystems & U.S./Japan ASTER Science Team. (2009). *ASTER Global Digital Elevation Model.* NASA EOSDIS Land Processes DAAC. Retrieved from https://lpdaac.usgs.gov/products/astgtmv002/ doi: 10.5067/ASTER/ASTGTM.002

Neis, P., Zielstra, D., & Zipf, A. (2012). The street network evolution of crowdsourced maps: Openstreetmap in germany 2007–2011. *Future Internet*, *4*(1), 1–21. Retrieved from https://www.mdpi.com/1999-5903/4/1/1 doi: 10.3390/fi4010001

Nguyen, H., Liu, W., & Chen, F. (2017, June). Discovering congestion propagation patterns in spatio-temporal traffic data. *IEEE Transactions on Big Data*, *3*(2), 169–180. Retrieved from https://doi.org/10.1109/tbdata.2016.2587669 doi: 10.1109/tbdata.2016.2587669

OpenStreetMap. (2021). *OpenStreetMap stats.* Retrieved from https://www.openstreetmap.org/stats/data_stats.html on July 7th, 2021.

Porta, S., Latora, V., Wang, F., Rueda, S., Strano, E., Scellato, S., ... Latora, L. (2012). Street centrality and the location of economic activities in barcelona. *Urban Studies*, *49*(7), 1471–1488. Retrieved from http://www.jstor.org/stable/26150937

Porta, S., Strano, E., Iacoviello, V., Messora, R., Latora, V., Cardillo, A., ... Scellato, S. (2009, January). Street centrality and densities of retail and services in bologna, italy.

Environment and Planning B: Planning and Design, *36*(3), 450–465. Retrieved from https://doi.org/10.1068/b34098 doi: 10.1068/b34098

Retallack, A. E., & Ostendorf, B. (2019, September). Current understanding of the effects of congestion on traffic accidents. *International Journal of Environmental Research and Public Health*, *16*(18), 3400. Retrieved from https://doi.org/10.3390/ijerph16183400 doi: 10.3390/ijerph16183400

Retallack, A. E., & Ostendorf, B. (2020, February). Relationship between traffic volume and accident frequency at intersections. *International Journal of Environmental Research and Public Health*, *17*(4), 1393. Retrieved from https://doi.org/10.3390/ijerph17041393 doi: 10.3390/ijerph17041393

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., & González, M. C. (2013, July). Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, *10*(84), 20130246. Retrieved from https://doi.org/10.1098/rsif.2013.0246 doi: 10.1098/rsif.2013.0246

Scott, J. (1988). Social network analysis. *Sociology*, *22*(1), 109-127. doi: 10.1177/0038038588022001007

Seifert, D. (2021). *OpenStreetMap streetlist evaluation on regio-osm.de.* Retrieved from https://regio-osm.de/listofstreets/index.html on July 7th, 2021.

Senatsverwaltung für Umwelt, Verkehr und Klimaschutz. (2019, Feb). *Nahverkehrsplan berlin 2019-2023* (Tech. Rep.). Berlin. Retrieved from https://digital.zlb.de/viewer/resolver?urn=urn:nbn:de:kobv:109-1-15384699 on September 27th, 2021.

Shang, W.-L., Chen, J., Bi, H., Sui, Y., Chen, Y., & Yu, H. (2021, March). Impacts of COVID-19 pandemic on user behaviors and environmental benefits of bike sharing: A big-data analysis. *Applied Energy*, *285*, 116429. Retrieved from https://doi.org/10.1016/j.apenergy.2020.116429 doi: 10.1016/j.apenergy.2020.116429

Solé-Ribalta, A., Gómez, S., & Arenas, A. (2016, October). A model to identify urban traffic congestion hotspots in complex networks. *Royal Society Open Science*, *3*(10), 160098. Retrieved from https://doi.org/10.1098/rsos.160098 doi: 10.1098/rsos.160098

Statistisches Bundesamt. (2018, Aug). *Verkehr: Verkehrsunfälle 2017* (Tech. Rep.). Retrieved from https://www.statistischebibliothek.de/mir/receive/DEHeft_mods_00083585 on September 21st, 2021.

Statistisches Bundesamt. (2021, Jul). *Verkehr: Verkehrsunfälle 2020* (Tech. Rep.). Retrieved from https://www.statistischebibliothek.de/mir/receive/DEHeft_mods

_00135650 on September 21st, 2021.

Strano, E., Cardillo, A., Iacoviello, V., Latora, V., Messora, R., Porta, S., & Scellato, S. (2007). Street centrality vs. commerce and service locations in cities: a kernel density correlation case study in bologna, italy. *arXiv: Physics and Society*.

Taaffe, E. J., Morrill, R. L., & Gould, P. R. (1963, October). Transport expansion in underdeveloped countries: A comparative analysis. *Geographical Review*, *53*(4), 503. Retrieved from https://doi.org/10.2307/212383 doi: 10.2307/212383

Tang, C. K., & Ahlfeldt, G. (2020, Dec). *Do speed cameras save lives?* (Working Paper). Retrieved from https://www.dropbox.com/s/rupfe4rxqziog9r/Speed_Camera_Dec2020.pdf?dl=0 on September 21st, 2021.

Tao, R., Xi, Y., & Li, D. (2016, July). Simulation analysis on urban traffic congestion propagation based on complex network. In *2016 IEEE international conference on service operations and logistics, and informatics (SOLI).* IEEE. Retrieved from https://doi.org/10.1109/soli.2016.7551690 doi: 10.1109/soli.2016.7551690

Taylor, M. A. P., Sekhar, S. V. C., & D'Este, G. M. (2006, August). Application of accessibility based methods for vulnerability analysis of strategic road networks. *Networks and Spatial Economics*, *6*(3-4), 267–291. Retrieved from https://doi.org/10.1007/s11067-006-9284-9 doi: 10.1007/s11067-006-9284-9

Thomas, K., Benjamin, R.-K., Fernando, P., Brian, G., Matthias, B., Jonathan, F., ... et al. (2016). Jupyter notebooks - a publishing format for reproducible computational workflows. *Stand Alone*, *0*(Positioning and Power in Academic Publishing: Players, Agents and Agendas), 87–90. Retrieved from https://doi.org/10.3233/978-1-61499-649-1-87 doi: 10.3233/978-1-61499-649-1-87

Tian, Y., Liu, X., Li, Z., Tang, S., Shang, C., & Wei, L. (2021, April). Identification of critical links in urban road network considering cascading failures. *Mathematical Problems in Engineering*, *2021*, 1–11. Retrieved from https://doi.org/10.1155/2021/9994347 doi: 10.1155/2021/9994347

Travers, J., & Milgram, S. (1969, December). An experimental study of the small world problem. *Sociometry*, *32*(4), 425. Retrieved from https://doi.org/10.2307/2786545 doi: 10.2307/2786545

UCAR Center for Science Education. (2021). *How Weather Affects Air Quality.* Retrieved from https://scied.ucar.edu/learning-zone/air-quality/how-weather-affects-air-quality on September 3rd, 2021.

UCLA: Statistical Consulting Group. (n. d.a). *Poisson regression.* Retrieved from https://stats.idre.ucla.edu/r/dae/zinb/ on September 5th, 2021.

UCLA: Statistical Consulting Group. (n. d.b). *Zero-inflated negative binomial regression.* Retrieved from https://stats.idre.ucla.edu/stata/output/poisson-regression/ on September 6th, 2021.

Vanajakshi, L., & Rilett, L. R. (2004, January). Loop detector data diagnostics based on conservation-of-vehicles principle. *Transportation Research Record: Journal of the Transportation Research Board*, *1870*(1), 162–169. Retrieved from https://doi.org/10.3141/1870-21 doi: 10.3141/1870-21

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual.* Scotts Valley, CA: CreateSpace.

Wang, P., Hunter, T., Bayen, A. M., Schechtner, K., & González, M. C. (2012, December). Understanding road usage patterns in urban areas. *Scientific Reports*, *2*(1). Retrieved from https://doi.org/10.1038/srep01001 doi: 10.1038/srep01001

Wang, Y., Cao, J., Li, W., & Gu, T. (2016, May). Mining traffic congestion correlation between road segments on GPS trajectories. In *2016 IEEE international conference on smart computing (SMARTCOMP).* IEEE. Retrieved from https://doi.org/10.1109/smartcomp.2016.7501704 doi: 10.1109/smartcomp.2016.7501704

Watts, D. J., & Strogatz, S. H. (1998, June). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. doi: 10.1038/30918

Xie, F., & Levinson, D. (2007, July). Measuring the structure of road networks. *Geographical Analysis*, *39*(3), 336–356. Retrieved from https://doi.org/10.1111/j.1538-4632.2007.00707.x doi: 10.1111/j.1538-4632.2007.00707.x

Xu, L., Yue, Y., & Li, Q. (2013, November). Identifying urban traffic congestion pattern from historical floating car data. *Procedia - Social and Behavioral Sciences*, *96*, 2084–2095. Retrieved from https://doi.org/10.1016/j.sbspro.2013.08.235 doi: 10.1016/j.sbspro.2013.08.235

Yang, X., Liu, L., Li, Y., & He, R. (2016, June). Identifying critical links in urban traffic networks: a partial network scan algorithm. *Kybernetes*, *45*(6), 915–930. Retrieved from https://doi.org/10.1108/k-05-2015-0144 doi: 10.1108/k-05-2015-0144

Zegura, E., Calvert, K., & Donahoo, M. (1997). A quantitative comparison of graph-based models for internet topology. *IEEE/ACM Transactions on Networking*, *5*(6), 770-783. doi: 10.1109/90.650138

A Calculations

A.1 Zero-Inflated Poisson Regression Log-Likelihood

With $\frac{\pi_i}{1-\pi_i} = \exp(\mathbf{z}_i^T \boldsymbol{\gamma})$, and $1 - \pi_i = \frac{1}{1+\exp(\mathbf{z}_i^T \boldsymbol{\gamma})}$,

$$
\begin{aligned}
&\log\left(\prod_i P(y_i|\mathbf{x}_i, \mathbf{z}_i)\right) \\
&= \log\left(\prod_{i:y_i=0} P(y_i|\mathbf{x}_i, \mathbf{z}_i)\right) + \log\left(\prod_{i:y_i>0} P(y_i|\mathbf{x}_i, \mathbf{z}_i)\right) \\
&= \sum_{i:y_i=0} \log\left(\pi_i + (1-\pi_i)\exp(-\lambda_i)\right) + \sum_{i:y_i>0} \log(1-\pi_i) + y_i \log(\lambda_i) - \log(y_i) - \lambda_i \\
&= \sum_{i:y_i=0} \log\left((\pi_i + (1-\pi_i)\exp(-\lambda_i))\frac{(1-\pi_i)}{(1-\pi_i)}\right) \\
&\qquad + \sum_{i:y_i>0} \log(1-\pi_i) + y_i \log(\lambda_i) - \log(y_i) - \lambda_i \\
&= \sum_{i:y_i=0} \log\left(\left(\frac{\pi_i}{1-\pi_i} + \exp(-\lambda_i)\right)(1-\pi_i)\right) \\
&\qquad + \sum_{i:y_i>0} \log(1-\pi_i) + y_i \log(\lambda_i) - \log(y_i) - \lambda_i \\
&= \sum_{i:y_i=0} \log\left((\exp(\mathbf{z}_i^T \boldsymbol{\gamma}) + \exp(-\lambda_i)\right) + (1-\pi_i) \qquad (27) \\
&\qquad + \sum_{i:y_i>0} \log(1-\pi_i) + y_i \log(\lambda_i) - \log(y_i) - \lambda_i \\
&= \sum_{i:y_i=0} \log\left((\exp(\mathbf{z}_i^T \boldsymbol{\gamma}) + \exp(-\lambda_i)\right) + \log(1-\pi_i) \\
&\qquad + \sum_{i:y_i>0} \log(1-\pi_i) + y_i \log(\lambda_i) - \log(y_i) - \lambda_i \\
&= \sum_{i:y_i=0} \log\left((\exp(\mathbf{z}_i^T \boldsymbol{\gamma}) + \exp(-\lambda_i)\right) + \sum_{i:y_i>0} y_i \log(\lambda_i) - \log(y_i) - \lambda_i + \sum_{i}^{n} \log(1-\pi_i) \\
&= \sum_{i:y_i=0} \log\left((\exp(\mathbf{z}_i^T \boldsymbol{\gamma}) + \exp(-\lambda_i)\right) \\
&\qquad + \sum_{i:y_i>0} y_i \log(\lambda_i) - \log(y_i) - \lambda_i \\
&\qquad - \sum_{i=1}^{n} \log(1 + \exp(\mathbf{z}_i^T \boldsymbol{\gamma})).
\end{aligned}
$$

A.2 Zero-Inflated Negative Binomial Regression Log-Likelihood

With $\frac{\pi_i}{1-\pi_i} = \exp(\mathbf{z}_i^T\boldsymbol{\gamma})$, and $1-\pi_i = \frac{1}{1+\exp(\mathbf{z}_i^T\boldsymbol{\gamma})}$,

$$
\begin{aligned}
&\log\left(\prod_i P(y_i|\mathbf{x}_i,\mathbf{z}_i)\right) \\
&= \log\left(\prod_{i:y_i=0} P(y_i|\mathbf{x}_i,\mathbf{z}_i)\right) + \log\left(\prod_{i:y_i>0} P(y_i|\mathbf{x}_i,\mathbf{z}_i)\right) \\
&= \sum_{i:y_i=0} \log\left(\pi_i + (1-\pi_i)\left(\frac{1}{1+\lambda_i\alpha}\right)^{\alpha^{-1}}\right) + \sum_{i:y_i>0}\log(1-\pi_i) + \log(h(y_i)) \\
&= \sum_{i:y_i=0} \log\left(\left(\pi_i + (1-\pi_i)\left(\frac{1}{1+\lambda_i\alpha}\right)^{\alpha^{-1}}\right)\frac{1-\pi_i}{1-\pi_i}\right) \\
&\qquad + \sum_{i:y_i>0} -\log(1+\exp(\mathbf{z}_i^T\boldsymbol{\gamma})) + \log\left(\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i+1)}\right) \\
&\qquad + \log\left(\left(\frac{1}{1+\alpha\lambda_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right)^{y_i}\right) \\
&= \sum_{i:y_i=0} \log\left(\left(\frac{\pi_i}{1-\pi_i} + (1+\lambda_i\alpha)^{-\alpha^{-1}}\right)(1-\pi_i)\right) \\
&\qquad + \sum_{i:y_i>0} -\log(1+\exp(\mathbf{z}_i^T\boldsymbol{\gamma})) + \log\left(\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})}\frac{1}{\Gamma(y_i+1)}\right) \\
&\qquad + \log\left(\left(\frac{1}{1+\alpha\lambda_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right)^{y_i}\right) \\
&\overset{(*)}{=} \sum_{i:y_i=0} \log\left(\exp(\mathbf{z}_i^T\boldsymbol{\gamma}) + (1+\lambda_i\alpha)^{-\alpha^{-1}}\right) - \log\left(1+\exp(\mathbf{z}_i^T\boldsymbol{\gamma})\right) \\
&\qquad + \sum_{i:y_i>0}\sum_{j=0}^{y_i-1}\log(j+\alpha^{-1}) - \log(1+\exp(\mathbf{z}_i^T\boldsymbol{\gamma})) - \log(y!) \\
&\qquad - \alpha^{-1}\log(1+\alpha\lambda_i) + y_i(\log(\alpha\lambda_i) - \log((1+\alpha\lambda_i)) \\
&= \sum_{i:y_i=0} \log\left(\exp(\mathbf{z}_i^T\boldsymbol{\gamma}) + (1+\lambda_i\alpha)^{-\alpha^{-1}}\right) \\
&\qquad + \sum_{i:y_i>0} -\log(y!) - (y_i+\alpha^{-1})\log(1+\alpha\lambda_i) + y_i(\log(\alpha)+\log(\lambda_i)) \\
&\qquad + \sum_{i:y_i>0}\sum_{j=0}^{y_i-1}\log(j+\alpha^{-1}) \\
&\qquad - \sum_{i=1}^{n}\log\left(1+\exp(\mathbf{z}_i^T\boldsymbol{\gamma})\right).
\end{aligned}
\tag{28}
$$

At $(*)$, we used

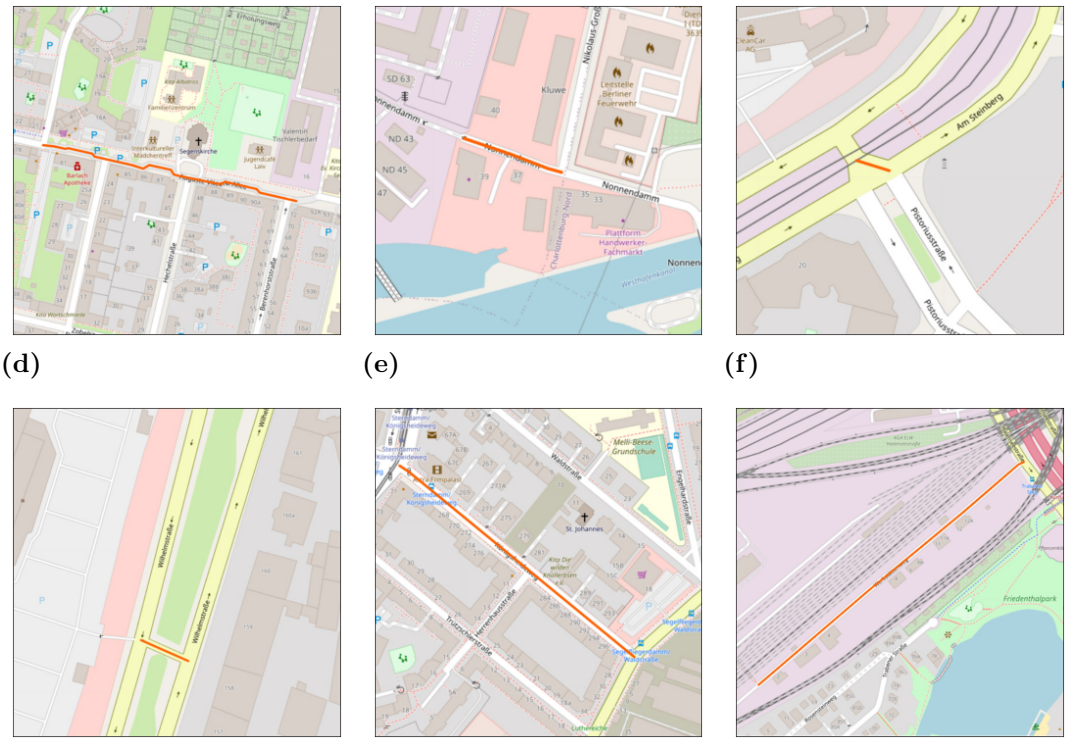
$$\ln\left(\frac{\Gamma(x_i+\alpha^{-1})}{\Gamma(\alpha^{-1})}\right)=\sum_{j=0}^{x_i-1}\ln(j+\alpha^{-1}), \tag{29}$$

which is easily shown by applying the property $\Gamma(x+1)=x\Gamma(x)$ repeatedly.

B Figures

Figure 20: Examples of main road segments without maximum speed information

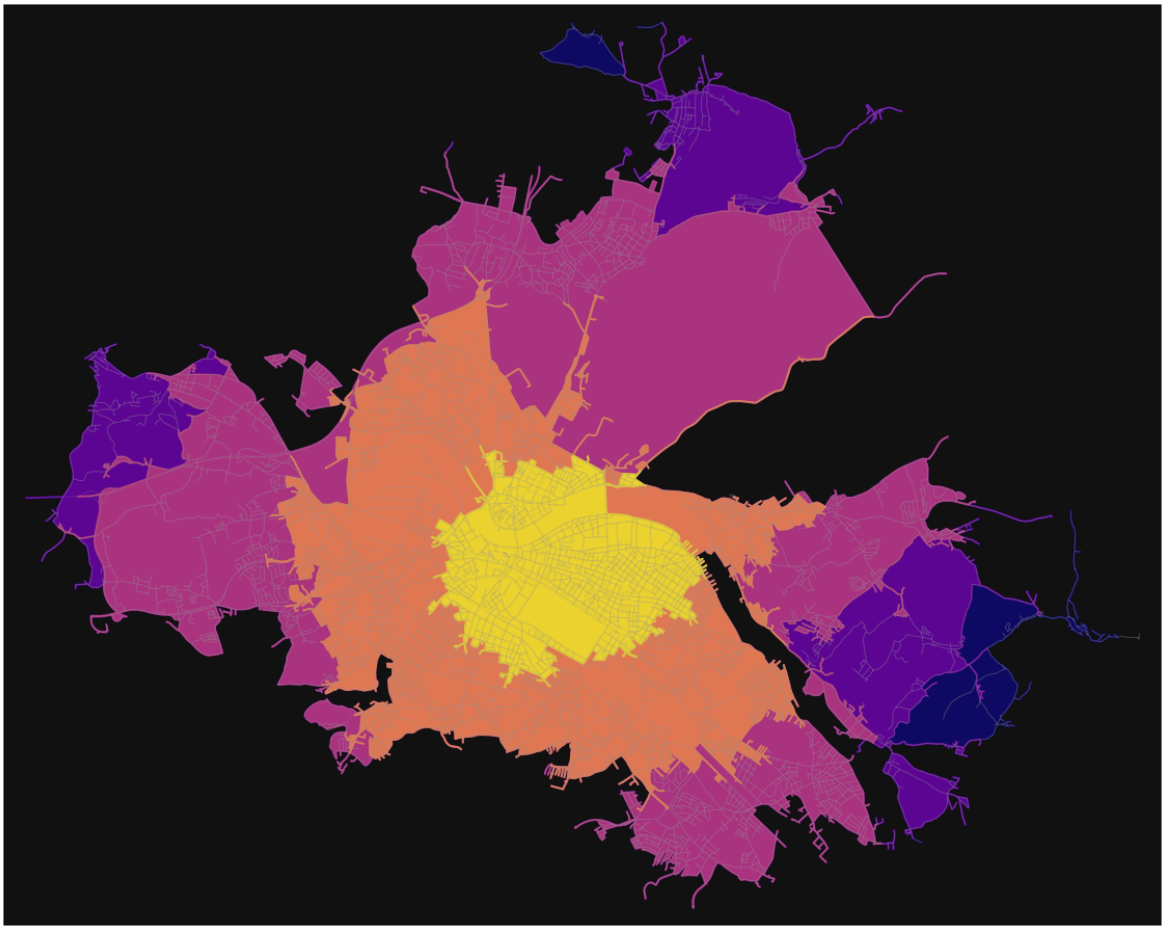
(a) **(b)** **(c)**



(d) **(e)** **(f)**

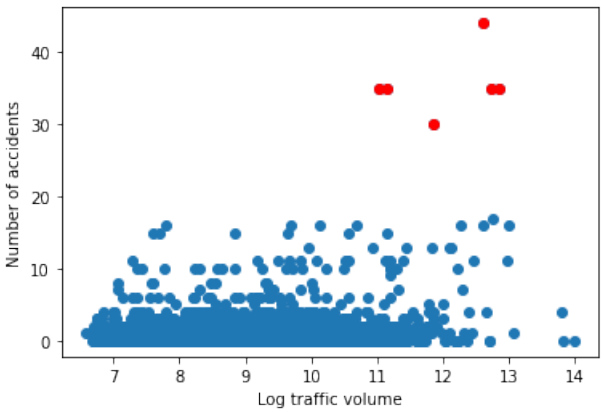
Notes: These panels depict road segments for which no maximum permissible road speed information is available on OSM. They are classified as (a) "living street", (b) "residential", (c) "secondary", (d) "secondary_link", (e) "tertiary", (f) "unclassified" and demonstate why these types of segments can reasonably be expected to often lack max speed information.

Figure 21: Isochrones in Dresden



Notes: This map depicts the road segments and polygonised isochrones of Dresden. Point of reference is a point in Dresden's city center. Other than in Berlin (Figure 12), isochrones do not extend radially, as there seem to be faster roads towards the north than towards the east of the city. Again, brighter areas are reachable faster from a point of the city centre, and the brackets are 5, 10, 15, 20, an 25 minutes.

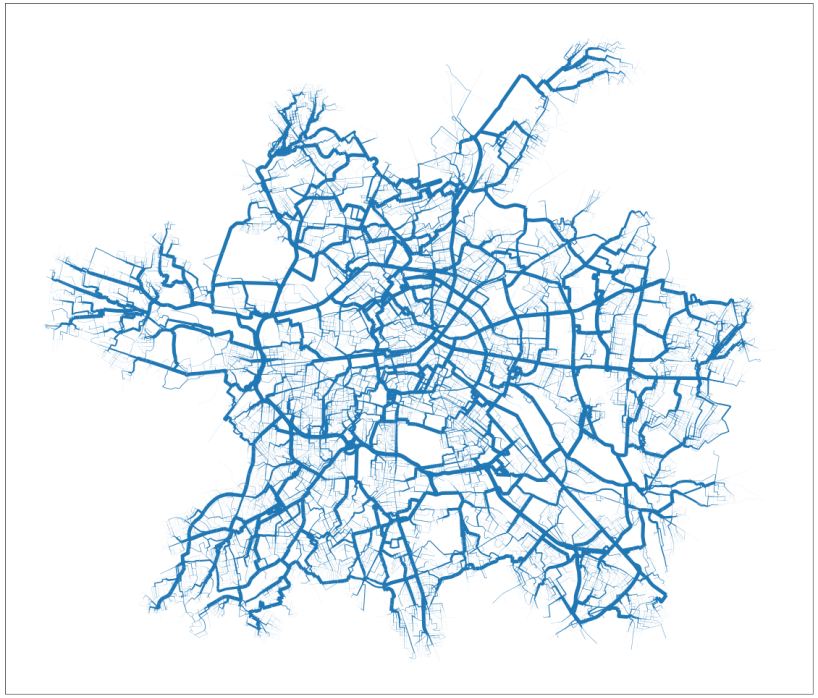
Figure 22: Log traffic volume on road segments vs. number of accidents



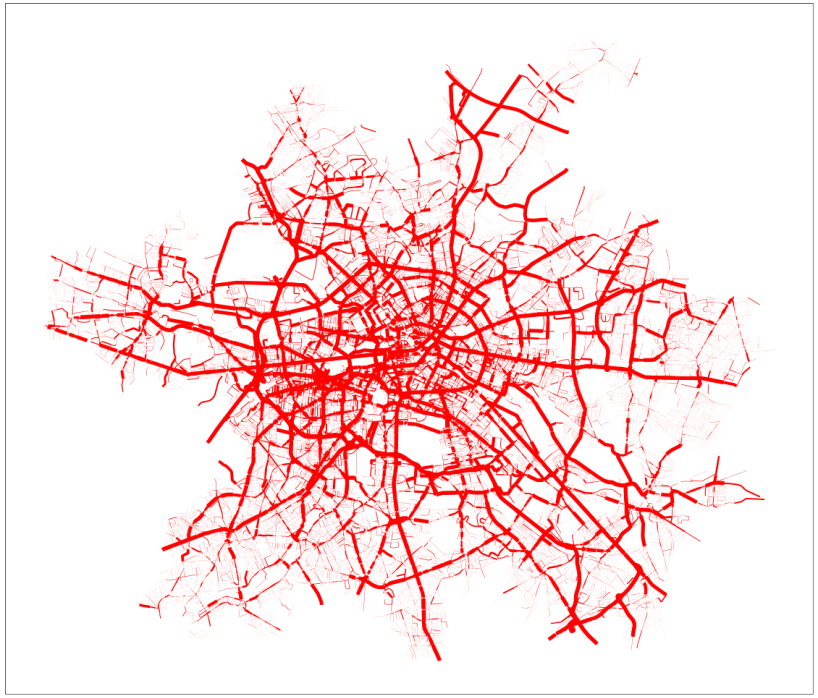
Notes: Red observations are considered outliers and have thus been removed.

Figure 23: Betweenness centrality and traffic in Berlin

(a)



(b)



Notes: Betweenness centrality (a) and traffic (b); thicker lines indicate higher betweenness or volume, respectively.

C Tables

Table 6: Zero-Inflated Poisson Regression Results

Dep. Variable:	n_accidents	**No. Observations:**	8011
Model:	ZeroInflatedPoisson	**Df Residuals:**	8008
Method:	MLE	**Df Model:**	2
converged:	True	**Log-Likelihood:**	-7041.5

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Logistic Regression						
const	5.7469	0.304	18.903	0.000	5.151	6.343
log_flow_n	-0.3608	0.030	-11.932	0.000	-0.420	-0.302
maxSpeed	-0.0396	0.005	-7.875	0.000	-0.050	-0.030
Poisson Regression						
const	-2.7266	0.153	-17.791	0.000	-3.027	-2.426
log_flow_n	0.1567	0.017	9.220	0.000	0.123	0.190
maxSpeed	0.0347	0.003	12.223	0.000	0.029	0.040

Notes: The dependent variable is the number of accidents, the independent variables are log traffic volume (`log_flow_n`) and maximum permissible road speed.

Table 7: Zero-Inflated Negative Binomial Regression Results

Dep. Variable:	n_accidents	**No. Observations:**	8011
Model:	ZeroInflatedNegativeBinomial	**Df Residuals:**	8008
Method:	MLE	**Df Model:**	2
converged:	True	**Log-Likelihood:**	-6594.2

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Logistic Regression						
const	9.7783	1.028	9.513	0.000	7.764	11.793
log_flow_n	-0.7011	0.107	-6.550	0.000	-0.911	-0.491
maxSpeed	-0.1054	0.014	-7.579	0.000	-0.133	-0.078
Neg. Bin. Regression						
const	-3.5612	0.334	-10.650	0.000	-4.217	-2.906
log_flow_n	0.2520	0.031	8.168	0.000	0.192	0.313
maxSpeed	0.0183	0.005	3.656	0.000	0.009	0.028
alpha	2.0305	0.178	11.402	0.000	1.681	2.380

Notes: The dependent variable is the number of accidents, the independent variables are log traffic volume (`log_flow_n`) and maximum permissible road speed. `alpha` is the dispersion parameter α, as indicated in 5.2.5.

Table 8: Waypoint and trip statistics in three different areas

	n_{points}	n_{trips}	avg. speed (km/h)	max. speed (km/h)	avg. distance (km)	max. distance (km)	avg. trip duration (min.)	max trip duration (min.)	earliest date	latest date
Berlin UC	850,587,925	32,106,737	32.44	200.00	21.93	1998.95	26.97	1209.12	2017-01-01	2018-01-01
Berlin adm.	894,920,983	32,606,521	32.73	200.00	22.20	1998.95	27.00	1209.12	2017-01-01	2018-01-01
Brandenburg & Berlin	1,357,551,853	34,145,757	34.22	200.00	25.43	1998.95	28.86	1209.12	2017-01-01	2018-01-01

Notes: Waypoint and trip statistics for Trips in and through the Berlin Urban Centre (UC), Berlin's administrative boundaries, and the combined Berlin and Brandenburg region. Waypoints are counted exactly within the area, while trips are counted if the trip either ends or originates in the respective area. For example, a trip starting and ending in Brandenburg but passing through Berlin would not be counted in the Berlin UC or administrative areas and a trips that starts outside of Brandenburg and end outside the Berlin Brandenburg region is not counted for the Berlin Brandenburg region. The high values for average trip duration and speed even within Berlin stem from the fact that a trip ending in Berlin may well start outside Berlin, traveling far distances at a high speed. If the sample is constrained to trips starting and ending in Berlin, both numbers are lower.

Table 9: Attributes of Berlin's UC network

Variable Name	Description	Value
country	Primary country name	germany
core_city	Urban area core city name	berlin
uc_id	Urban area's unique identifier in UCD	2851
cc_avg_dir	Avg clustering coefficient (unweighted, directed)	0.037
cc_avg_undir	Avg clustering coefficient (unweighted, undirected)	0.052
cc_wt_avg_dir	Avg clustering coefficient (weighted, directed)	0.001
cc_wt_avg_undir	Avg clustering coefficient (weighted, undirected)	0.001
circuity	Ratio of street lengths to straight-line distances	1.039
elev_iqr	Interquartile range of node elevations, meters	12
elev_mean	Mean node elevation, meters	44.709
elev_median	Median node elevation, meters	44
elev_range	Range of node elevations, meters	66
elev_std	Standard deviation of node elevations, meters	8.240
grade_mean	Mean absolute street grade (incline)	0.018
grade_median	Median absolute street grade (incline)	0.009
intersect_count	Count of physical street intersections	24,908
intersect_count_clean	Count of physical street intersections (after merging nodes within 10 meters geometrically)	21,083
intersect_count_clean_topo	Count of physical street intersections (after merging nodes within 10 meters topologically)	21,189
k_avg	Avg node degree (undirected)	3.008
length_mean	Mean street segment length (undirected edges), meters	142.339
length_median	Median street segment length (undirected edges), meters	113.449
length_total	Total street length (undirected edges), meters	6,066,366.461
node_count	Count of nodes	28,339
orientation_entropy	Entropy of street bearings	3.571
orientation_order	Orientation order of street bearings	0.011
pagerank_max	Maximum PageRank value of any node	0.000
prop_4way	Proportion of nodes that represent 4-way street intersections	0.263
prop_3way	Proportion of nodes that represent 3-way street intersections	0.601
prop_deadend	Proportion of nodes that represent dead-ends	0.121
self_loop_proportion	Proportion of edges that are self-loops	0.002
straightness	The inverse of circuity	0.963
street_segment_count	Count of street segments (undirected edges)	42,619
resident_pop	Total resident population, 2015 (UCD)	3,271,872
area	Area within boundary polygon, km2 (UCD)	686
built_up_area	Built-up surface area in 2015, km2 (UCD)	433.529

Notes: Selected variables, variable descriptions, and their values from Berlin's UC network (taken from Boeing (2020b)). Numerical descriptive statistics are rounded to the third decimal.

Table 10: Twenty-five road segments without available maximum speed information from OSM

OSM ID	Street name and Post code	OSM highway tag
4515896	Neufertstraße, 14509	living street
4978512	Lynarstraße, 13585	residential
5116453	Winckelmannstraße, 12487	residential
6229931	Bornitzstraße. 10365	residential
8792775	Werkstättenweg, 14055	unclassified
11405545	Schwarzheider Straße, 12627	residential
12305962	Auguste-Viktoria-Allee, 13403	living street
15920080	Klausenerplatz, 14059	living street
23453218	Helmstraße, 10827	living street
25368898	Königsheideweg, 12487	tertiary
26406925	Hansastraße, 13088	residential
27156341	Lichterfelder Allee, 14513	living street
29112676	Am Steinberg, 13086	secondary
44202827	Schillerstraße, 10625	living street
89583721	Chemnitzer Straße, 12621	living street
90094622	Helga-Haase-Straße, 13053	living street
98723438	Wilhelmstraße, 13595	secondary link
160130964	Nonnendamm, 13627	residential
167410411	Straße 49, 13089	residential
195386065	Platanenstraße, 13156	residential
415858612	Joachim-Böhmer-Straße, 13053	living street
454285423	Große Hamburger Straße, 10115	living street
793150994	Crellestraße, 10827	living street
821773653	Große-Leege-Straße, 13055	residential
868550032	Riedemannweg, 13607	unclassified

Notes: Most of the road segments are tagged residential roads or living streets. The few secondary streets are road segments on bigger streets for which assigning a speed limit is nonsensical, as outlined in Figures 20 (a)-20 (f). Tag descriptions are available at `https://wiki.openstreetmap.org/wiki/Category:Tag_descriptions_for_key_%22highway%22`

Table 11: Estimated traffic-relevant accidents and associated savings in Berlin and Germany for the year 2020

Accidents		**Count**	**avg. cost / accident**	**Total Savings**
with injury to persons		$518 \times 10.10\% \approx 52$	17,391 €	904,332 €
damage-only	in the strict sense other accidents	$518 \times 0.86\% \approx 4$	22,839 €	91,356 €
	under the influence of intoxicants	$518 \times 0.56\% \approx 3$	6,177 €	18,531 €
	others	$518 \times 88.47\% \approx 458$	6,177 €	2,829,066 €
Subtotal		**519**		**3,843,285 €**
Accident victims		**Count**	**avg. cost / person**	**Total Savings**
killed		$63 \times 0.21\% \approx 0$	1,150,234 €	0 €
seriously injured		$63 \times 13.30\% \approx 8$	116,335 €	930,680 €
slightly injured		$63 \times 86.50\% \approx 54$	5,138 €	277,452 €
Subtotal		$52 \times 1.2 \approx 63$		**1,208,132 €**
Total Savings				**5,051,417 €**

Notes: Estimated traffic-relevant accidents. For 2020, 518 accidents are predicted to be avoided. Savings are calculated using the cost estimates (based on 2017 figures from Bundesanstalt für Straßenwesen (2021)) for each accident type. Numbers may not sum up to totals because of rounding errors.

Table 12: Accidents and associated costs in Berlin in the year 2020

Accidents		**Count**	**Share**	**avg. cost / accident**	**Total Cost**
with injury to persons		12,861	10.38%	17,391 €	223,665,651 €
damage-only	in the strict sense other accidents	880	0.71%	22,839 €	20,098,320 €
	under the influence of intoxicants	678	0.55%	6,177 €	4,188,006 €
	others	109,509	88.37%	6,177 €	676,437,093 €
SUbtotal		**123,928**			**924,389,070 €**
Accident victims		**Count**	**Share**	**avg. cost / person**	**Total Cost**
killed		50	0.33%	1,150,234 €	57,511,700 €
seriously injured		2,007	13.37%	116,335 €	233,484,345 €
slightly injured		12,949	86,29%	5,138 €	66,531,962 €
Subtotal		**15,006**			**357,528,007 €**
Total Cost					**1,281,917,077 €**
Difference to 2017 Total (Real savings)					**175.452.573 €**

Notes: Accidents in 2020 and associated incurred economic cost. Accident counts and economic costs per accident / person taken from Statistisches Bundesamt (2021) and Bundesanstalt für Straßenwesen (2021). Note that the costs per accident and per person are taken from the year 2017 for better comparison. In the 12,861 accidents with injury to persons, 1.17 as many, i.e. 15,006, people were hurt. Numbers may not sum up to totals because of rounding errors.

Declaration of Authorship

I hereby confirm that I have authored this Master's thesis independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Berlin, September 30, 2021

Ben Thies