Untapping the potential of crowdsourcing app data to measure urban bicycle flow

LUNDS UNIVERSITET
Lunds Tekniska Högskola

LTH Ingenjörshögskolan vid Campus Helsingborg
<Institution / avdelning>

Examensarbete:
Sebastian Palomino Achulli

LTH Ingenjörshögskolan vid Campus Helsingborg
Lunds universitet
Box 882
251 08 Helsingborg

LTH School of Engineering
Lund University
Box 882
SE-251 08 Helsingborg
Sweden

Tryckt i Sverige
Media-Tryck
Biblioteksdirektionen
Lunds universitet
Lund 2024

Förord

Examensarbetet är det avslutande momentet i högskoleingenjörsutbildningen i byggteknik med inriktning inom väg- och trafikteknik vid Lunds Tekniska Högskola. Jag vill tacka båda mina handledare Carmelo D'agostino och Zhankun Chen som har hjälpt mig genom hela processen. Dessutom vill jag tacka Andreas Persson för att ha väglett mig genom examensarbetet och för att ha tagit rollen som examinator.

Sebastian Palomino
2024-05-04

Sammanfattning

De nuvarande metoderna som tillämpas av olika företag och forskare när det gäller om att räkna cykelvolym i olika tätorter har förbättrats över tiden. Dock har det aldrig varit felfri, vilket har lett till fortsatta studier och ännu mer utrymme att förbättra de befintliga metoderna. Därför är syftet med rapporten att ytterligare analysera det som redan har gjorts och hur man kan överträffa nuvarande förväntningar.

En djupdykning i olika cyklisters beteende gjordes genom att analysera antal fel som inträffade vid insamling av data, samt orsaken till att dessa felaktigheter uppstår i första hand. Flera analyser gjordes i hela Lund stad, med hjälp av kommunen och applikationen TravelVu, både i hjärtat av staden och även i utkanten.

I hopp om att ta reda på varför dessa extremvärden inträffar i första hand gjordes en jämförelse genom att göra en litteraturgenomgång på flera vetenskapliga artiklar. Olika metoder användes av var och en, vissa var mer lika än andra men de gav var och en tydlig inblick i hur de bedrev sin forskning.

De flesta problemen är relaterade till kraftigt blandade trafikmiljöer som uppstår, speciellt under rusningstid. Hindrandet av trafik samt andra faktorer orsakar dessa brister under datainsamling, men som tidigare nämnt finns det fortfarande mer att undersöka och förbättra på vad vi redan vet.

Abstract

The current methods employed by different companies and researchers when it comes to measuring bike volume at different urban environments have been improving over time. However, it has never been completely flawless, leading to future studies and even more room to improve the existing methods. Which is why, the purpose of this report is to further analyze what has already been done and how to exceed current expectations.

A deep dive into different cyclists' behavior was conducted by analyzing the number of errors that occurred while gathering data, as well as the reason as to why these inaccuracies arise in the first place. Multiple analyzes were done throughout the town of Lund, with the help of the municipality and the app provided by TravelVu, both in the heart of the city and even on the outskirts. The miscalculation of data across these different urban locations differs from place to place but there's also similarities between them.

In hopes of finding out why these outliers happen in the first place a comparison was made by doing a literature review on multiple scientific papers. Different methods were employed by each one, some were more similar than others but they each provided a clear insight into how they conducted their research.

Most of the problems are related to heavily mixed traffic environments that occur, especially during peak hours. The obstruction of traffic as well as other factors cause these deficiencies during data collection, but as previously mentioned, there's still more to research and improve on what we already know.

Innehållsförteckning

1 Introduction

1.1 Background

The city known as Lund is renowned for having a high amount of bicycle trips daily, aiming towards becoming climate neutral by the year of 2030 (Lunds kommun 2024). Among the key characteristics of Lund in terms of its bicycle culture is the infrastructure, Lund has an extensive network of dedicated bike paths, as well as multiple bike parking facilities. This has led to a significant proportion of the residents to prefer biking as their primary mode of transportation, additionally the large student population contributes to the high rate of bicycle usage. The city continuously encourages its residents to, among other things, reduce carbon emissions and think in a more environmentally sustainable way.

Modern scientific research has been observant to the importance of bikes in general and how impactful it has been to the society, mostly related to the number of bicycle trips. In some European countries bike usage is relatively high compared to other types of transportation, for some people it's their daily routine when they for example take the bike to their school, work, home, etc. In Sweden, there are few studies that address the analysis of bike trips using crowdsourced data, one of those few examples is the scientific paper written by Sonja Forward, *"Hållbart resande – möjligheter och hinder" (2014*). Even fewer studies are done on urban areas, especially in the context of smaller cities such as Lund even though it is considered to be a city with a pronounced bike culture. Therefore, to address this perceived research gap, this study strives to understand the reasoning behind possible inaccuracies and to further investigate outliers in the specified models.

Crowd sourced data is generated or contributed to by a significant share of the total population that collaborates on a large scale, usually through collaborative efforts or online platforms such as apps. This is a more efficient way of gathering data compared to the traditional way of data collection since the latter usually relies on smaller groups/organizations to collect information. Instead, it accumulates data on a wider area, for example through mass participation where everyone provides a piece of information. In the current study they were voluntarily contributing since they provided data during their free time without financially benefitting from it (Blohm, I, Leimeister, JM, and Zogaj, S 2018). In the methodology presented in the thesis, the crowd sourced data can be used to analyze certain bicycle behaviors, such as where cyclists usually cycle, the estimated volume in specific areas, possible peak times and other relevant indicators.

There are multiple different ways of collecting crowd sourced data, some of them involve using GPS by downloading an application or manual counting, but more specific methods include using pneumatic road sensors, or also known as pneumatic tube counters which is a general method still used by various organizations. The procedure can vary a lot, but the most common method employed is by using mobile devices to track the cyclist's location.

1.2 Purpose and objective

In short, this paper should give an answer to how the quality and quantity of crowdsourced data affect the reliability of predictions for bicycle volume in different urban environments. This is done by tempting to gain a better understanding of how bike volumes are miscounted as well as what different methods are deployed to ensure a more accurate and detailed result. Just like it's written by Ulf Paulsson in his book *"Examensarbeten - Att skriva uppdragsbaserade uppsatser och rapporter"* (2020) this paper uses previous knowledge that is accessible by anyone and has been scientifically approved, but it also aims to further improve current research within the transportation world and provide valuable insight. It ought to contribute with valuable information for instance whilst modeling traffic safety measures and for future research purposes, since traffic volume is the factor that is most correlated with traffic accidents, meaning that obtaining an accurate bike volume will help producing better accident prediction models. This is extremely important since most European countries aim towards accomplishing vision zero, which means to avoid fatalities or injuries that involve road accidents (Cyclomedia, 2023).

Different perspectives from different places, such as United States, Canada, Germany and Austria, provide a broader insight into how people analyze and research cyclist's behavior. Multiple literatures sources were reviewed to better comprehend how cyclist volume is counted across different locations as well as to what the cause for error could be.

2 Literature review

2.1 Tube counters

In the report *"Validation of Bicycle Counts from Pneumatic Tube Counters in Mixed Traffic Flows"* (Brosnan, M., et al. 2015) it is written that, like the approach applied in the current study, that there were two different providers. However instead of using GPS and manual counting, like TravelVu and Lunds Municipality, pneumatic tube counters and standard tube counters were installed at different locations in the city of Minneapolis.

2.1.1 Work procedure

Tube counters, also known as pneumatic tube counters work by utilizing air pressure changes to detect the passage of vehicles, including bicycles. Firstly, a tube is stretched across the road surface which is filled with air and then sealed. This makes the passing vehicle compress the air inside creating a pressure wave that travels along the tube. Secondly, a sensor device is connected at the end of each tube, detecting the pressure changes caused by the vehicle movement, in this case bike movements. The sensor thus registers the pressure change. Lastly, the collected information is transmitted to a counting device or something similar which records the passage of vehicles. Collected data can include speed, number axles, direction of travel and so on. The recorded data is analyzed to provide different traffic patterns, vehicle counts and other relevant information.

The MetroCount counters, provided by Minnesota DOT, were specifically acquired to assess their effectiveness in counting bicycles. Concurrently, Timemark counters, a routine tool for Hennepin County, are regularly employed for comprehensive traffic volume assessments as part of statewide vehicular monitoring initiatives overseen by the Minnesota DOT. Installation of these counters was conducted by the Hennepin County engineer, who holds the primary responsibility for the county's traffic monitoring program, along with research assistants from the University of Minnesota.

Notably, MetroCount advice a tube spacing of 1 meter for classification counts, while Timemark recommends a roughly 3 meters separation between tubes. Both systems operate on a similar principle: passing vehicles or bicycles generate air pulses recoded as "axle hits." Subsequently, computer algorithms, during post-processing, classify the vehicle type or identify bicycles based on the time intervals between axle hits, considering the distance between tubes. Given the distinct interpretations of axle strikes, these systems enable the concurrent collection of bicycle and vehicular counts.

At the Portland Avenue site, two MetroCount pneumatic tube counters were installed in September of 2013. The tubes, spaced 1 meter apart, extended from the curb across one travel lane and one bike lane, which is almost 3,5 meters. Meanwhile, at the University Avenue site, a more comprehensive approach involved the deployment of four pneumatic tube counters in parallel formation in June 2014. Standard tubes provided by the manufacturers were used regularly across all of the installations.

In addition, Brosnan (2015) writes that the systematic refinement of bicycle classification methodologies, guided by manufacturer recommendations and innovative validation techniques, is a pivotal aspect of contemporary transportation research. The Hennepin County engineer strategically adjusted the default settings of Timemark counters to enhance their effectiveness in concurrently classifying bicycles and motorized vehicles.

The tailored adjustments included the following modifications:

- Registration of events with weaker air pulses to account for bicycles' smaller mass and narrower ties, which displace less air than motorized vehicles.
- Registration of events for slower-moving vehicles, aligning with the characteristic lower speeds of bicycles compared to motorized counterparts.
- Registration of events associated with shorter axle spacing, acknowledging bicycles' inherently shorter axle spacing relative to motorized vehicles.

Additionally, Hennepin County implemented a complementary validation strategy by installing video cameras at each monitoring location. Student employees from the Minnesota Traffic Observatory meticulously reviewed the captured footage, noting the precise time each bicycle passed. The incorporation of video validation served as a robust quality assurance measure, contributing with to the general reliability of the automated bicycle counts.

The study's diverse configurations, surrounding sites such as Portland Avenue and University Avenue with varying lane configurations, required a more precise criteria for classification. The MetroCount system, offering a range of algorithms for data classification, was deployed with three distinct approaches:

- The usage of the MetroCount algorithm ARXCycle, aligning with Australian standards and incorporating bicycle counts in its output.

- Customized classification through postprocessing of motorcycle counts derived from the MetroCount ARX algorithm. In this approach, researchers extracted events with axle bases less than 4 feet and speeds below 25mph.
- Adoption of the BOCO algorithm developed by Boulder County, Colorado. This algorithm was specifically designed to identify bicycles classified as motorized vehicles within standard algorithms.

2.1.2 Reliability of tube counters

Following adjustments to support bicycle counting, the researchers utilized the standard Timemark classification output for bicycle counts, configurated to provide data in 15-minute bins.

To validate the accuracy of the automated counts, the study followed established procedures from prior assessments of automated counters. Manual counts derived from video footage were regarded as the "standard" or best estimate of actual bicycle numbers. The assessment of relative accuracy involved the calculation of percentage differences, with the following formula:

$$d_i = \frac{(a_i - v_i)}{v_i}$$

- d_i represents the percentage difference for site *i*,
- a_i represents the automated count for site i,
- v_i represents the video-confirmed count for site *i*.

To comprehend the reasons for inconsistencies between manual video counts and automated tube counts, as written by Brosnan (2015), researchers carefully matched time stamps from both sources. This matching process involved identifying matches, false positives (bikes classified by the automated counter but not on video), and false negatives (bicycles missed by the automated counter but recorded on video). The alignment of time stamps employed manual inspection for the Portland Avenue site, while an algorithm programmed in Excel facilitated the process at the University Avenue site.

To quantify the absolute error rate relative to the actual number of bicycles, the following equation was employed:

$$e_i = \frac{(f^+ + f^-)}{v_i}$$

- e_i represents the absolute error rate for site i,
- f^+ represents false positives,
- f^- represents false negatives.

Brosnan (2015) also explains that an agreement rate for each configuration and classification scheme was determined as the percentage of instances where time stamps from the video and counter were in agreement. The absolute error rates were exclusively calculated for MetroCount devices, as Timemark counts were reported in 15-minute bins and lacked individual event inspection. To account for systematic counter errors, calibration equations were estimated by regressing hourly manual counts on hourly automated counts using ordinary least squares regression, with a forced origin through zero. The application of calibration equations involved adjusting hourly counts from the counters, and the resulting estimates were plotted against actual hourly volumes.

The acquired result was that the comprehensive analysis of bicycle volumes across diverse monitoring sites and days provides a refined understanding of the performance of pneumatic tube counts and the associated challenges in accurately estimating bicycle traffic. The range of bicycle volumes, spanning from nine to 30 bicycles per hour across different sites sets the stage for evaluating the accuracy of pneumatic tube counts.

The examination of estimates derived from pneumatic tube counts revealed a general trend of undercounts, with variations across sites, configurations, and classification algorithms. The observed percentage errors underscore the complexity of accurately capturing bicycle volumes, presenting a range from an undercount of 57% at the University Avenue site (with one bike lane, three car lanes, and a Timemark device) to a minor overcount of approximately 6% at the Portland Avenue site (featuring one bike lane, one car lane, a MetroCount device, and the BOCO classification algorithm).

Lastly, a couple of key findings, according to Brosnan (2015), from this analysis include site-specific variations, meaning that percentage error was notably lower at the Portland Avenue site, especially in scenarios with lower overall bicycle and traffic volumes. This finding suggests a potential correlation between the absolute volume of both bicycles and vehicles and the accuracy of counts. Additionally, percentage errors were higher at the University Avenue site, particularly in configurations with one bicycle lane and three car lanes. The algorithmic performance also varied, the percentage error for the three classification algorithms used with the MetroCount devices were generally similar, with the BOCO algorithm demonstrating the lowest percentage error across all sites and configurations. There were also device discrepancies, percentage errors for counts from Timemark devices were consistently higher than those from MetroCount devices, regardless of the classification algorithm used for MetroCount data. Further on, the agreement or match rates between video and counter time stamps varied widely, ranging from slightly less than 50% to approximately 90%. This variability underscores the complexity of achieving consistent alignment between automated counts and manually validated counts from video footage. Lastly, absolute error rates were notably higher than percentage error rates across all classifications schemes, indicating the presence of a substantial number of both false negatives and false positives. False negatives, indicating missed bicycles, were more frequent than false positives in most cases.

2.2 Bicycle counting on a yearly scale

The author of *"A spatial modeling approach to estimating bike share traffic volume from GPS data"* (Brown, M. J., et al. 2022) explained that the city of Hamilton in Ontario, Canada, provided a unique context for the exploration of cycling dynamics, particularly with the establishment of Hamilton Bike Share (HBS) in 2015. With a population of 536,917 in 2016, Hamilton represents a major urban center where cycling initiatives have gained prominence. HBS, operational year-round, distinguishes itself as one of the few Bike Share Systems (BSSs) in North America that perseveres through cold and snowy winters. As of April 2020, HBS had 132 hubs and approximately 825 GPS-equipped bikes in operation.

2.2.1 Networks specifically for cycling

This study offers a detailed examination of cycling activity throughout the entirety of the year of 2018 with the help of GPS data. Brown (2022) discloses that with a dataset originally comprising of 347 079 unique GPS trip trajectories, the research aims to capture cycling behavior across all seasons. The application of the GIS-based map matching algorithm developed by Dalumpines and Scott (2011) and the GIS-based Episode Reconstruction Toolkit (GERT; Dalumpines and Scott, 2018) becomes influential in converting raw GPS trajectories into meaningful routes aligned with the existing cycling network.

A fundamental aspect of the research presented involves the creation of a comprehensive cycling network, leveraging both road and trail data to ensure the accurate representation of cycling routes. The study draws on an open dataset from Hamilton's Open Data Portal, enriched with bikeway information and trail features from McMaster Library's Maps, Data, & GIS Centre and the City of Hamilton. The integration of CanMap® Content Suite from DMTI Spatial further enhances the network's capability.

Brown (2022) moreover describes that manual digitization efforts, concentrated in high-traffic areas such as McMaster University, were undertaken to create "unofficial" pathway features. These features, informed by satellite imagery and raw GPS tracks, are designed for the routes commonly utilized by HBS users. The combination of diverse data sources, including the 2016 Canadian Census and the City of Hamilton's parcel data, supplements the network with crucial population and employment information. This supplementary data plays a vital role in testing various accessibility variables.

The transformation of the network into a network dataset, supported by ArcGIS® facilitates comprehensive analysis. Importantly, the dataset is configured to allow mutual travel along network links, recognizing that cyclist do not strictly stick to traffic rules as observed for automobiles. This subtle approach to network construction ensures the accurate capture of cycling trips, lining up with the unique dynamics of cycling in an urban environment.

On top of that, Scott et al. (2021) clarifies that HSB users had preferences, that indicate a general inclination towards routes with cycling infrastructure. Building on this, Lu et al. (2018) uncovers the tendency of HSB users to deviate from the shortest path, opting for routes featuring designated bike lanes, moderate traffic bike routes, or separated bike paths. Major roads lacking bikeway classifications are strategically chosen as the reference level, considering their high automobile traffic and comparatively reduced attractiveness to cyclists.

2.2.2 Accuracy of GPS trajectories

During the processing phase, roughly 14% of invalid GPS trajectories (48,693) were identified and filtered out by GERT. An additional 3% of trajectories (11,799) were lost during the map-matching process, attributed to network topology or GPS errors. Following this careful processing, the total number of map-matched routes available for analysis amounted to 286,587.

To derive the annual bike share traffic (ABST) volume for network links, the route features generated by the map-matching process were intersected with the cycling network, resulting in the creation of new route-link features. These features were then merged into a unified dataset and summed for each network link. ABST, characterized by a positively inclined distribution with a mean of 873 and a median of 33.

Brown (2022) further delves into the substantial impact of accessibility on daily ridership at HSB hubs, which were calculated using the method in Scott and Ciuro (2019). Accessibility metrics, including population, employment, and hub-trip distance accessibility, are comprehensively derived, drawing on the 2016 Canadian census data and the unique trip patterns across the city. Taking the three measures into consideration, the following formula was employed, inspired by Hansen (1959):

$$A_i = \sum_j O_j f(C_{ij})$$

- A_i represents the accessibility link of i,
- O_j represents the number of employees, residents, or the discovered trips in the vicinity of hub j,
- C_{ij} represents the cost of traveling linking i and j,
- $f(C_{ij})$ represents an impedance function.

The methodology by Scott and Horner (2008) is exercised to derive a negative exponential distance decay impedance function, calibrated with the decay parameter β determined through unique hub-to-hub trip distances. This distance-decay model enhances the spatial dimension of the analysis, capturing varying influences of trip distances on accessibility and the following formula was utilized:

$$I_k = \alpha \exp(-\beta t_k)$$

- I_k represents the number of trips for category k,
- k represents the distance category,
- β has an estimated value of 0.000628,
- t_k represents the trip distance for category k in 100 m increments.

Model estimates and summary statistics are carefully examined before and after implementing a spatial filter, revealing the hub-trip distance accessibility variable as the most influential predictor of ABST. The research identifies multiple correlation issues between population and employment accessibility measures, leading to the preference for hub-trip distance accessibility. Distance to the nearest hub and bus stop emerges as significant explanatory variables, while network distance to McMaster University and the Central Business District (CBD) proves less effective.

Following the application of the spatial filter, certain bikeway classifications lose significance, suggesting their initial importance may have been inflated by spatial autocorrelation. The spatial filter successfully reduced residual autocorrelation, enhancing the model's explanatory power, as evidenced by the adjusted R^2 of 0.89. The significance of all explanatory variables at the 0.001 level or better underscores the robustness of the spatially filtered model.

Furthermore, Brown (2022) clarifies that model interpretation involves expressing coefficients as a percent change in ABST for a one-unit change in the independent variable. The investigation makes use of the natural logarithmic transformation of ABST, and model predictions are validated through repeated k-fold cross-validation. The Root Mean Squared Error (RMSE) of 0.97 indicates the model's predictive value, particularly for planning decisions related to cycling infrastructure upgrades and new bike share hubs.

A couple of key findings are, for example, that for every unit increase in hub-trip distance accessibility, ABST increases by 2.63%. Secondly, as network distances to the closest hub increase, ABST decreases by 59.75% per kilometer. This finding suggests that bike share traffic is concentrated in the vicinity of bike share hubs, aligning with HBS policies penalizing bikes not returned to hubs. Thirdly, network distances to the closest bus stop show a 30.72% decrease in ABST per kilometer, highlighting the importance of the directness of routes. Fourthly, major roads with a bike lane experience a 155.23% increase in ABST compared to major roads with no bikeway classification, on the other hand minor roads with a bike lane see a substantial 201.22% increase in ABST compared to the reference level. Going even deeper, Brown (2022) sheds light on minor roads without a bikeway classification witness a decline of 68.78% in ABST compared to the reference, likewise on the pathways in McMaster University's campus and minor trails in public parks display negative coefficients, indicating decreases of 77.71% and 87%, respectively. Lastly, paved multi-use pathways exhibit a 42.76% increase in ABST compared to the reference.

2.3 Bicyclists' behavior at signalized intersections

As written in the scientific paper *"Traffic flow at signalized intersections with large volumes of bicycle traffic"* (Grigoropoulos, G., et al. 2022) the author analyzes motor vehicle and bicycle traffic flow at signalized intersections in Germany, emphasizing the data collection methodology and the subsequent analysis of bicyclists' operational behavior.

2.3.1 Methodology and simulation

Grigoropoulos (2022) discloses that before the data collection they took certain criteria into consideration, such as excluding high gradients and other environmental factors that could potentially influence bicyclists' behavior. This led to them considering 32 possible intersections at six different cities around Germany, but eventually they decided on only collecting data on 8 intersections to analyze traffic flow with varying cycling infrastructure.

The course of action revolved around gathering information through video, that was strategically used at the different junctions, aiming for the highest possible point of view as well as an optimal viewing angle. Two systems were utilized: a mobile system with cameras mounted on a 12 m high mast on the Urban Traffic Research Car (UTRaCar), used in Berlin, and a fixed system with a camera on a mast or high building, used in Munich and Freiburg. The data collection was carried out over a day at each intersection, the periods ranged from 1 to 2 hours during rush hour traffic. This approach provided with a focused examination of bicyclists' behavior during peak times.

Additionally, Grigoropoulos (2022) addressed trajectories of bicyclists automatically extracted from the video data. A couple of key parameters such as acceleration and average speed were then derived from these trajectories to provide insights into operational behavior. The data classification was done as following: Trajectories were classified according to vehicle type, distinguishing between cyclists, motor vehicles and pedestrians. This shed light on bicyclists' interactions and habits at signalized intersections. However, to ensure the accuracy of automated tracking and classification procedures, the trajectory data also underwent manual verification, successively enhancing the reliability of the accumulated data for subsequent investigation. However, pedestrian trajectories did not exhibit any significant interaction with bicyclists and this caused the pedestrian trajectories to be excluded from further analysis, instead mainly focusing on bicyclists and motor vehicle interactions.

The study utilised PTV Vissim, a comprehensive software for multi-modal microscopic traffic flow simulation. The software incorporates dedicated behavior models for bicyclists, intensifying realistic lateral movements based on empirical observations. For instance, bicyclists riding straight across the intersection typically remain within the width of the bicycle lane, influencing motor vehicle movements. Furthermore, the simulation models account for different bicyclist maneuvers, such as turning left with a direct maneuver. Bicyclists moving into the left lane upstream, passing waiting motor vehicles, and queuing in front of vehicles in the bike box are accurately represented in the simulations.

Simulation is used to derive data that is usually not noticeable, for example, Angenendt et al. (2005) provides insights into the lateral positioning of bicyclists concerning various bicycle infrastructure, for instance bike boxes, bicycle lanes and paths. These findings serve as a foundation for simulation models, ensuring realistic lateral movements in PTV Vissim. Results on bicyclist speed distribution, desired speed, and acceleration behavior from multiple studies (Figliozzi et al., 2013; Parkin and Rotheram, 2010; Taylor, 1993; Twaddle and Grigoropoulos, 2016) are integrated to model operational behavior in the simulation. The combination of empirical and literature-based data enhances the accuracy of the simulation, and the cumulative speed distribution function of bicyclists generated using the trajectory dataset from empirical studies serves as a validation mechanism for the simulation models. Further enhancing the accuracy of the simulated bicyclist behavior, aligning with past research.

2.3.2 Simulation results and bicycle behavior

Delving deeper into the results, Grigoropoulos (2022) makes it clear that video data were collected during the summer months to capture peak bicycle traffic, providing a comprehensive understanding of bicyclist behavior under favorable conditions. Depending on intersection conditions, video cameras were mounted on the UTRaCar, adjacent buildings, or masts to ensure optimal visibility of the intersection approaches.

Initial analysis involved manually viewing video segments to gain insights into queuing behavior, left-turn maneuvers and adaptations based on infrastructures as well as signal states surrounding the junction. To ensure collected data integrity, more attention was given to avoid irregular or recurring obstructions to bicycle traffic flow caused by other road users. Among the results, it was noted that bicyclists approaching the intersection using bike boxes tended to carry out direct left turns, emphasizing the importance of bike boxes in facilitating this maneuver and reducing delays. In addition, across three intersections, bicyclists had a tendency to use dedicated bicycle infrastructure if available, adapting their movement in order to make their intended actions easier. On the other hand, if dedicated infrastructure was absent, bicyclists were more likely to use motor vehicle or pedestrian infrastructure for left turns, adjusting their behavior based on traffic signal state.

Further on, Grigoropoulos (2022) makes it clear that amidst the key findings of his research, acceleration from stop is an essential parameter for simulating driving behavior, facilitating in the calibration of traffic simulation models. A reasonable speed distribution is also a critical input for accurately simulating bicyclist performance actions. Queue density, which refers to the number of bicyclists per square meter in a queue in front of a stop line on a cycling facility, provides valuable insights for infrastructure dimensioning and traffic simulation models. Average discharge time, indicating the time it takes for each bicyclist to exit a queue at a stop line, is likely to be more efficient on wider cycling facilities, although it isn't fully certain and thus further research was recommended. Lastly, occupancy time, which is the duration of the conflict area between right-turning motor vehicles and bicyclists riding straight across the intersection that is occupied by bicyclists, increased stepwise with the number of bicyclists, highlighting the impact of group size on conflict area occupation.

Simulation models are designed to study traffic performance parameters at intersection approaches with different types of bicycle infrastructure. Simulation studies allow the investigation of increasing bicycle traffic volumes that may not be observable in empirical studies. The study distinguished between calibration of input parameters and calibration based on simulation output. Acceleration functions and desired speed distributions in PTV Vissim were calibrated based on empirical study results. Empirically determined queue density, average discharge time and occupancy time were used to calibrate lengthwise and sideways behavior models. Vehicular traffic was fine-tuned to reproduce discharge times and capacities in accordance with HBS (Hessian Bicycle Standard). Observed effect of bicycle traffic on the average discharge time of the first vehicle turning right was used to validate the simulation models.

The comparison between empirical and simulation was approximately low in value, for example when comparing queue density on bicycle lanes, the relative error of the means is small (1.5%), but when comparing mean waiting times of motor vehicles turning right for each bicyclist crossing the intersection, the relative error is large (11%). Grigoropoulos (2022) describes that the results depended on relationships identified in observed data, extending beyond the range of observations. Strong matches of results to real-world applications were approached with caution, intensifying the role of simulation tools when empirical data was insufficient or impractical to collect.

Four simulation scenarios were analyzed, each featuring modified bicycle volume, vehicular volume, signal cycle length and green time ratio, with the presentation of diverse performance metrics. In the first scenario, it was noted that the capacity of right-turning vehicular traffic decreases with increasing bicycle traffic volume, though this decrease was less noticeable at very high bicycle traffic volumes. Capacity was primarily influenced by the actual green time ratio, but cycle length also played a role, especially at high bicycle traffic volumes and low green time ratios. Different approach infrastructures, such as bike boxes and bicycle lanes, impacted right-turning vehicle capacity. Bike boxes made it easier to direct left turns for bicycles but slightly reduce the average capacity of right-turning vehicles compared to bicycle lanes. During the second scenario, capacity of left-turning vehicular traffic was strongly influenced by the actual green time ratio, with the effect of cycle length becoming more significant with higher volumes of oncoming motor and bicycle traffic. Capacity thresholds were identified based on oncoming motor and bicycle traffic volumes, indicating a reduction in capacity. Longer cycle lengths made capacity reduction worse during phase changes. Throughout the third scenario, vehicular capacity crossing the intersection was relatively independent of bicycle traffic volume. Bicyclists queuing on the right side of the bike box did not significantly impact motor vehicle movement at the start of the green phase. Finally, in the fourth scenario, vehicular capacity decreased as the number of cyclists turning left increased. The presence of a bike box, where left-turning bicyclists queue in front of motor vehicles, led to capacity reductions, especially with smaller time gaps.

The point of these different scenarios, according to Grigoropoulos (2022), is that it highlights the intricate interactions between vehicular and bicycle traffic at intersections across various settings. It identifies key factors influencing vehicular capacity, such as bicycle volume, signal control parameters and intersection infrastructure. These findings provide valuable recognition for traffic planning and signal design, emphasizing the need for adjusting approaches to accommodate diverse traffic compositions and enhance overall intersection efficiency. In addition, the findings serve to shed light on the importance of balancing the needs of different road users for optimal traffic flow.

2.4 Voluntary mass bicycling

Throughout the year of 2015 there were multiple of cycling trips recorded in Vienna within a study described by Schnötzlinger, P., et al, in his scientific paper *"Volunteered mass cycling self-tracking data – grade of representation and aptitude for planning"* (2022). In this paper, the authors describe a varied approach on collecting data, such as incorporating data preprocessing, GIS technology and statistical analysis tools amongst other things.

2.4.1 Different approaches to acquire data

The primary source of data in this study is the 'Bike Citizens' app, designed to benefit urban cycling with features such as route navigation and journey logs. Users consent to the anonymized usage of their trips, which are manually collected by the user. The dataset includes two primary components - 'trackpoints' (point data) and 'tracks' (polyline data). Trackpoints provide detailed information about the cyclist's positions, while tracks encapsulate the entire journeys, comprising multiple trackpoints.

The first and last 100 meters of each trip were cut short to anonymize data and prevent the identification of specific individuals. The analytical tools included GIS tools, ArcMap 10.4.1 and QGIS 2.8. Their main usage was for spatial operations and analyses. Database and spatial extensions consisted of PostgreSQL 9.5 and PostGIS 2.3 which were both used for efficient data handling, storage and spatial examinations. Python 3.6 was utilized for data manipulation, calculations and statistical visualizations.

The dataset was improved with several key variables to enhance data cleaning and ease advanced analyses.

- dist_BC: Distance traveled according to Bike Citizens app.
- dist_GIS: Distance traveled measured with GIS.
- dist_GIS200: GIS-measured distance with an additional 200 meters for anonymization.
- dur_BC: Duration of the trip according to Bike Citizens.
- dur_GIS: Duration of the trip measured with GIS.
- dur_GIS200: GIS-measured duration with an additional time for the missing 200 meters.
- v_BC: Calculated speed using dist_BC and dur_BC.
- v_GIS: Calculated speed using dist_GIS and dur_GIS.

Since data cleaning is a crucial step in the analysis of cycling trajectory data, a three-staged filter was employed to remove faulty or unusable data while preserving valuable records for subsequent analysis. The initial step involved deleting all tracks located outside the municipal boundary, the focus here was on tracks with at least one segment within the boundary, ensuring relevance for later analyses of traffic intensity and velocity. Identifying and deleting duplicate tracks was essential to assemble the dataset.

The second stage aimed to effectively remove outliers by filtering tracks based on attribute values such as duration, distance and velocity. The study defined thresholds for duration, distance and velocity based on previous research findings. For instance, the lower threshold for duration followed the approach of Froehlich and Krumm (2008), setting it at 30 seconds. Upper thresholds were determined through a combination of literature review and data from the Austrian-wide mobility survey. Literature sources such as Stopher et al. (2005), Segadilha and Sanches (2014), and El-Geneidy et al. (2007) informed the determination of the maximum average speed, set at 36 km/h in the current study. The upper limit for distance was determined using surveys like the bicycle traffic survey in Vienna (2010), with a set limit of 50 km. Duration limits were defined based on practical considerations and survey data.

Schnötzlinger (2022) describes that the last step involved adding a supplementary filter to remove trips taking longer than 150 minutes. This filtering process, encompassing spatial and attribute-based criteria, resulted in a refined dataset ready for in-depth trajectory analysis.

Removing tracks outside the city boundaries is a common practice to focus the analysis on relevant geographical areas. This step ensured that the dataset is confined to the study area of Vienna, resulting in a reduced but geographically relevant dataset. Identifying and removing duplicate tracks is essential for maintaining dataset integrity, duplicate tracks, often caused by recording errors or system glitches, can skew subsequent analyses. Tracks lacking geographical dimension, represented by zero length, were excluded, which in its turn ensured that only tracks with meaningful spatial information were retained for analysis. The study determined specific thresholds for duration, distance and speed based on literature findings and practical considerations, which aided in filtering out tracks with attributes outside the defined ranges, contributing to the overall data quality.

Positional accuracy is crucial in trajectory analysis, especially in urban environments where satellite visibility may be obstructed. The map-matching algorithm was introduced to assign recorded tracks to the digital route network, enhancing the accuracy of position estimation. While not achieving 100% accuracy, map-matching significantly improves the reliability of the dataset (Quddus et al., 2007). Challenges in accurately determining the real position, especially in cycling data, arise due to the narrow width and high mobility of bikes. Existing literature, such as Jagadeesh et al. (2004), highlights the difficulties in precisely determining the real position of cyclists.

To analyze mean speed along chosen segments within Vienna’s cycling network, a careful selection of segments was conducted. Focus was placed on inner districts, evenly distributing segments across different facility types. Segment lengths were standardized to 100 meters to balance interference reduction and speed calculation accuracy. Prior to mean speed calculation, a preliminary data cleaning was performed, filtering partial tracks based on distance and speed criteria and changing tracks based on direction of movement which ensured that only relevant and accurate data contributed to the mean speed analysis.

2.4.2 Outcome of the acquired data

The research implemented an Origin-Destination (OD) matrix to visualize the share of cycling traffic based on recorded tracks within Vienna's municipal districts. The matrix provides insights into the connection and traffic patterns between different districts. Notably, the inner districts exhibited higher connections and significant traffic flows could be observed even between distant districts. Results from the analysis revealed strong traffic flows between districts, indicating the unified nature of these areas.

Schnötzlinger (2022) further explains that to validate the reliability of GPS-based data, discoveries from the investigation compares traffic volumes derived from automatic counting stations with GPS-based values. The comparison involved 12 cyclist counting stations in Vienna and the traffic volume was calculated based on the number of selected tracks per station and month, considering different types of days (working days, weekends and holidays). It revealed a clear correlation between GPS-based values and counts from automatic stations. High correlation coefficients (0.886-0.935) suggest a strong association, with the strongest correlation observed on working days. Despite the generally representative nature of GPS data, the examination acknowledges lower Direct Traffic Volume (DTV) values in GPS data compared to counts. While the GPS data lack precision in detailing the spatial distribution of cycling traffic, results draw attention to the overall illustration of GPS-based values. The analysis acknowledged variations in correlation coefficients based on the traffic volumes and density of the data point cloud, highlighting the impact of variations on correlation strength.

The research applied linear equations to calculate a parameter describing the grade of representation of recorded cycling data from the mobile app compared to ground counts. This parameter varied spatially across counting stations, offering insights into the characteristics of the app data in different locations. The calculated parameters revealed variations in the grade of representation across counting stations, indicating that the spatial distribution of the app-recorded pathways was not entirely representative. Observations acknowledged the influence of spatial conditions, such as residential density and the availability of alternative cycling facilities, on the need for a route planner.

Temporal component of data representation was explored by calculating monthly coefficients, providing a better understanding into how well the app data represented ground counts throughout the year. Seasonal variation in data representation was also identified, with lower representation during the first quarter (January to March) and a sharp increase in April. The values declined in June, reaching a second low point, followed by a continuous rise, peaking in December. This unusual course was attributed to the initial availability of the app at cost, followed by a sharp increase in usage after being offered free of charge through a municipal cooperation.

When comparing BC (Bike Citizen) calculations and survey data, only one station had a BC-calculated average distance travel lower than the surveyed one, while at other stations, BC values were consistently higher. Discrepancies between survey and BC data were considered, with the possibility of longer distances being overestimated and shorter distances underestimated due to personal evaluations in survey responses. Environmental design and spatial structure were suggested as potential factors influencing individual awareness of time and distance.

Schnötzlinger (2022) notes that the average trip length provides limited insights. Counting stations showed substantial variations in trip lengths, influenced by spatial and environmental factors. Some observations exhibited a higher number of short-distance trips.

Furthermore, cycling infrastructure was categorized based on organizational form (interaction with other transport modes) and facility type (structural design). The study considered on-street mixed, on-street marked, off-street and traffic-calmed infrastructure types. Altitude differences was also introduced along cycling segments as a significant external factor influencing speed. The digital terrain model of the City of Vienna was used to calculate altitude differences for selected segments. Correlations were performed to estimate the impact of altitude differences on speed. While altitude differences were considered, the need to explore variations in altitude outside the segment was also addressed. The influence of intersections and traffic light phases on speed is recognized as a potential factor not directly addressed in the analysis.

Among the results, it is noted that there was a negative correlation between altitude differences and speed. With each additional meter in altitude or each additional percentage point of slope, speed decreased by 0.8 or 1.0 km/h. The low correlation suggests that factors beyond altitude differences contributed to variations in speed.

To analyze the impact of cycling infrastructure on speed, segments with the same infrastructure type and a similar number of tracks in both directions were selected, which ensured a focused examination of speed differences along specific infrastructure types. A negative correlation (-0.479) was identified between climbing percentage and average speed. Thus, average speed decreases by 1.9 km/h with each additional percentage in climbing. The observed scattering of values was attributed to the diversity of cycling infrastructure types. High average speeds were found along multi-purpose lanes, dedicated cycle routes and in mixed traffic. Pedestrian areas displayed the lowest mean speed. When aggregating by organizational form, on street-facilities showed high average speeds, while traffic-calmed areas had the lowest. The results indicate that, according to Schnötzlinger (2022), cycling infrastructure heavily influences cycling speed. The impact of altitude differences as well as additional factors such as intersections and traffic lights impact the variations in speed.

2.5 Calculating ridership with crowdsourced data

In the scientific paper written by Jestico, B., et al, *"Mapping ridership using crowdsourced cycling data"* (2016) the author mentions that in the city of Victoria, BC, Canada, a high cycling rate is noticeable, making it an intriguing location for studying cycling patterns. Among the things explained in the report, it gives an overview of the cycling infrastructure, data collection methods and integration of manual counts and crowdsourced Strava data in Victoria.

2.5.1 Data gathering in Victoria

In 2013, 18 locations in Victoria underwent manual cyclist counts as part of the regional bike count program. Counts occurred during different months, data revealed hourly, peak and daily counts at intersections, major roadways, residential streets and multi-use trails. Strava, a popular fitness app, provided a crowdsourced cycling dataset for 2013. The dataset included a road network shapefile with information on the number of Strava users cycling on specific roadways. Strava data featured high spatial and temporal coverage, offering counts comparable to manual counts.

A comparison was made between manual and Strava data which showed variations in cyclist counts at specific locations and time periods. The study utilized a portion of the Strava data for direct comparisons, while the larger dataset was used to create prediction maps for cyclist volumes at unknown locations. Compiling crowdsourced data into hourly intervals and aligning it with days of manual counts made this comparison easier.

Using PostgreSQL, the study summarized crowdsourced counts for each road segment, enabling an evaluation of the relationship between manual counts and crowdsourced counts. R^2 values from simple linear regression indicated the strength of this relationship, with increasing accuracy for larger time windows. The study then explored the usefulness of predicting cycling volumes in Victoria using a Generalized Linear Model (GLM). Explanatory variables, including time of year, were included based on their significance in previous studies. Crowdsourced cyclist volume data from Strava, exhibiting nearly continuous coverage, was a key explanatory variable. Predictions were made at a daily level, covering AM and PM peak traffic periods for each season.

The GLM aimed to predict cycling volumes for all unsampled road segments in Victoria. Non-significant explanatory variables were removed, and collinearity was addressed using Variance Inflation Factors (VIF). The model included a cross-validation step, dividing data into 90% training and 10% testing subsets, repeated 100 times to assess prediction accuracy. Model error was evaluated through cross-validation, comparing predicted cycling volumes to observed volumes in the 10% testing subset. Percent differences between predicted and observed volumes were calculated. Classification accuracy was also assessed, categorizing volumes into low, medium and high classes. Following this, Jestico (2016) explained that maps were created using the prediction model and selected classification levels for all road and trail segments in Victoria. Maps were generated for each count season, offering a visual representation of cycling volume variations throughout the year.

2.5.2 Results from crowdsourced data

The GLM identified five significant explanatory variables associated with cycling volumes: crowdsourced data volumes, segment slope, posted speed limit, time of year and the presence of on-street parking facilities. Crowdsourced data volumes positively correlated with manual count volumes, while an increase in segment slope, higher posted speed limits and the presence of on-street parking were associated with decreased cyclist volumes. The different seasons played a crucial role, with May, July and October showing increased volumes compared to January. To illustrate, an increase of one unit in crowdsourced cyclists was linked with an estimated rise of 51 cyclists at a given location, whereas a 1% increase in slope corresponded to a decrease in72 cyclists.

Cross-validation using a random 90% and 10% subset revealed an overall average model error of 38%. Over half of the predictions (55%) had errors of less than 30%. An evaluation of five different categorical scenarios for predicted cycling volumes was done, categorizing them into low, medium and high classes. Scenario 3, with low volumes (0-199), medium volumes (200-999) and high volumes (1000+), demonstrated the highest predictive accuracy across all categories, with accuracies of 76%, 77% and 85% respectively. Prediction maps were generated for each season, classifying cycling volumes into low, medium and high categories based on the Scenario 3 breakdown. May and July exhibited overall higher volumes of cyclists on all roadways compared to January and October. Roadways with high volumes in the winter and fall generally remained high throughout the year.

3 Method

3.1 Data

The data on bike movements was provided by the Lund Municipality from a previous project, "*Hitta dolda cykelpotentialer – ny förståelse av trafik genom att kombinera IoT med traditionella data*" (2020), and the method used to perform this task was by counting cyclists in certain urban areas across Lund at specific roadway links. The following places were chosen to conduct bicycle counts:

- Nilstorp
- Vipeholm
- Norra Fäladen
- Gunnesbo
- Klostergården
- Lund central

A total of 180 intersections were considered when performing these counts, this equals to 459 individual stations/points and at each of these stations, four 15 minutes sessions of counting cyclists were accomplished spread out over working days, during the year of 2021. It varied between 7:30 am to 5:10 pm and the data received by the municipality was in both pdf and Excel format where the studied sections had counts distributed in different columns.

3.2 Procedure

The method used to predict bicycle volume was done through an app called TravelVu. It was installed on participants mobile phones where it would track their movements in real time with the assistance of GPS and afterwards the volunteers provided Lund Municipality with different sorts of data after using the application.

GIS data from the TravelVu App was issued in a GIS Shapefile that consisted of the trip sections. Trip properties included (but was not limited to) travel time, measured in seconds, travel distance in meters, and trip segment ID. All the data collection was conducted during the year of 2021, more precisely between September 8th and October 17th, and the number of participating phones was 220 in total.

We identified the different intersections through Google Maps and visited each site to check on the environment, more specifically for data points that stood out the most. Aspects that were considered were for example how the different lanes were dimensioned, the amount of bike lanes and other relevant factors that could potentially affect the different cycling behaviors.

3.3 Linear model

A linear model is a mathematical representation of a linear relationship between the output and input variables. It is usually demonstrated as a linear combination of the input variables, each multiplied by a corresponding coefficient/weight, and intercept term, and a normally distributed error term.

An equation for a simple linear model with a single input variable is usually written as:

$$y = mx + b + \varepsilon \qquad (1)$$

- b is the y-intercept, it indicates the value of y when x is zero,
- m is the inclination of the line, represents the effect of the input variable on the output,
- x is the input, which is an independent variable,
- y is the output which is a dependent variable,
- ε is the normally distributed error term.

3.4 Outliers in linear model

An outlier in linear models is a data point/observation that tends to stand out from the rest of the dataset, potentially affecting the accuracy of conclusions which is later drawn from the data. The reasons could differ from point to point, their presence generally indicate errors in data collection, unusual phenomena or measurement variability. The most notable key point when it comes to outliers is the deviation from the norm, especially the data points that significantly deviate from the typical pattern or distribution of the rest of the data. It could either be unusually high or low values in comparison to the rest of the observations. Probable causes of this are for example measurement errors which result from errors during data collection. Natural variation is also a cause that occurs during rare or unique cases. This in turn influences measures of central tendency, which is sensitive to extreme values, and outliers distorts the value making it less representative of the typical observation in the dataset.

3.5 Influential data points

A threshold is a point or a level at which something occurs or shifts, in regression analysis it refers to a specific value of an independent variable at which there is a partial change in the relationship between the dependent variable. It represents a critical point and is associated with non-linear relationships, typically in linear models the relationship between variables is presumed to be constant across all values of the independent variables, but when a threshold is present the relationship could follow different rules in different ranges. In general, thresholds hold a strong influence in researching purposes since it is a valuable concept in understanding relationships that exhibit non-linear behavior which helps identify more nuanced patterns in the data. (Fong, Y., et al. 2017)

Cook's distance is used to recognize data points that might adversely affect the regression model, they are considered influential data points. Its primarily use is to measure how much of a difference there'll be in the model once a certain point is deleted, a large value typically indicates that it has a strong influence over the rest of the values. But it doesn't necessarily mean that it must be deleted, it identifies the influential data points that stand out the most from the rest since there could be some other reasons as to why the value is either too low or too high. (Zach, 2019)

The formula used to measure Cook's distance was as following:

$$D_i = \left(\frac{{r_i}^2}{p*MSE}\right) * \left(\frac{h_{ii}}{(1-h_{ii})^2}\right) \tag{2}$$

- h_{ii} is the ith leverage value,
- MSE is the mean squared error,
- p is the number of coefficients in the regression model,
- r_i is the ith residual.

Leverage refers to the influence a data point has on the estimation of the regression coefficients, high leverage strongly influences the slope and intercept of the regression line. It can indicate influential observations, but it doesn't necessarily imply outliers. (Chris, M 2016)

In the model applied in this study, X represents the bicycle count data from the TravelVu app, the aggregated counts that was obtained through the app installed on the participant's phones, and Y represents the count data from the municipality, which was acquired through the traditional method.

The following formula was used to calculate the leverage:

$$h_{ii} = \frac{1}{n} + \frac{1}{n-1}\left(\frac{x_i - \bar{x}}{sx}\right)^2 \qquad \textbf{(3)}$$

- n is the number of observational points,
- x_i is the count data from the municipality,
- $\bar{x}$ is the mean value of the join count,
- sx is the standard deviation.

4 Result

The following result is from the collected data done in 2021 throughout the months of September and October, with a total of 120 observations. All the data was analyzed with the help of Excel as well as Cook's distance which measures how the values/observations in the model change depending on which data point is deleted.

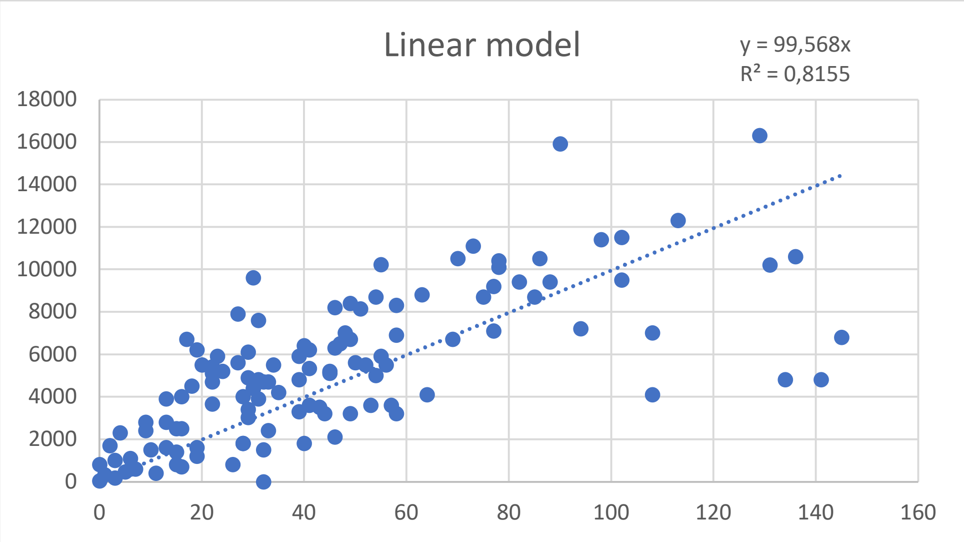

Figure 1: Linear model representing all the observations

Figure 1 shows the linear model that represents the spatial relationship between TravelVu data, the y-column (Join_Count) and manual pedestrian count, the x-column (AggrCount). The R^2 measures the amount of residuals explained by the model, the higher value the better results, a lower value would mean that more covariates should be included. The existence of outliers usually decreases R^2 values since there's a high probability for overlooked heterogeneity.

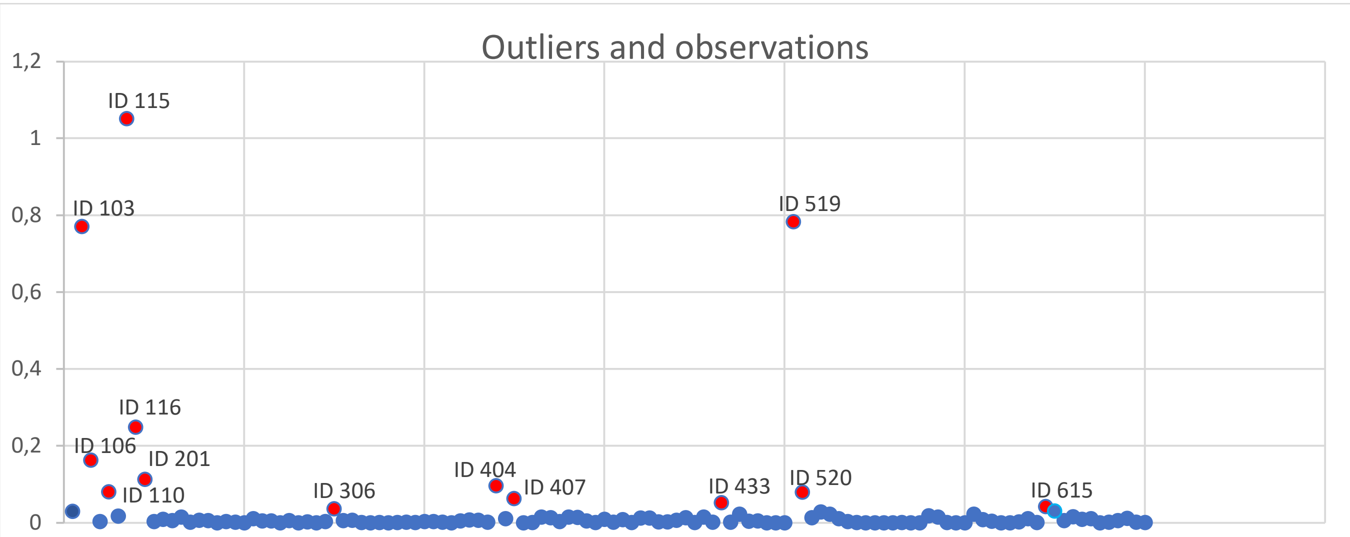

Figure 2: All the observations as well as the outliers, indicated by the red dots

In total there were 13 outliers, indicated by red dots and their corresponding ID number in Figure 2, according to the results acquired when using Cook's distance. This would mean that, their distance strays heavily from the linear model, while the blue dots in the figure display the rest of the observations that didn't stand out as much.

A common guideline is that any point with a Cook's distance greater than 4 divided by the total number of data points (n) is considered an outlier (Zach, 2019). This simplified formula was used to determine which of the results were an outlier:

$$Q = \frac{4}{n} \qquad \textbf{(4)}$$

When inserting the values on the formula (4), the answer it gave was approximately 0.033, meaning that if Cook's distance was higher than that, the observation should be deemed an outlier.

- Q is the limit of acceptable Cook's distance,
- n is the number of observations.

5 Analysis

5.1 Outliers nearby the center of the city

A closer examination of the outliers presented in the previous section is provided here. From this examination, it's evident that certain roads like Stortorget, Lilla Fiskaregatan and Kyrkogatan (ID 119 (103+116+115)) are major routes used by various types of vehicles. Even though the intersection between Kyrkogatan, Lilla Fiskaregatan and Stortorget doesn't meet the criteria of being an outlier, it stands out from the rest and could most likely have been considered one if the data wasn't collected during the pandemic. Furthermore, the notable deviations in Figure 1 could be attributed to the fact that these roads have been around for quite some time without much operation and maintenance. Another factor to consider is the diverse mix of daily traffic, including e-bikes, scooters, motorcycles and pedestrians, passing through these intersections. Given that the study is specifically focused on bicycle volume, there might be some slight discrepancies in the calculations when other types of vehicles and even pedestrians are in the mix.

Figure 3: Intersection between Stortorget with Kyrkogatan and Lilla Fiskaregatan, extracted from Google Maps[1]

[1] Google Maps - https://maps.google.com/

The central location of these roads in the heart of the city means that a wide range of people, from professionals heading to work to students going to school, pass through them regularly. This urban context might introduce subtle errors in the data collection process, whether through the app interface or manual counting, potentially leading to missed or duplicated observations of the vehicles passing through.

By utilizing Google Maps' standard traffic tool, a noticeable trend emerged, indicating consistently elevated traffic levels at the intersection, no matter which time or day it was. This observation reinforced the intersection's sustained crowded nature, setting it apart from others in Lund. The persistently high traffic volume, a characteristic feature of this junction, may have contributed to miscalculations or errors during the data collection phase. Notably, the intersection's design entails reductions in vehicle speed, potentially serving as a contributing factor to the inconsistencies encountered.

While the traffic intensity doesn't compare to that of the central hub of Lund, a comparative analysis against other intersections revealed a significant disproportion in daily vehicular flow. Google Maps indicated higher speeds on alternate routes, suggesting lower traffic density and a lack for lower speed. This inconsistency in traffic dynamics prompts an exploration of potential influences on the accuracy of data collection, particularly as it relates to the unique traffic patterns influenced by the intersection's design. Further examination is warranted to make out the extent to which these nuanced traffic conditions may have impacted the precision of recorded data, shedding light on potential complexities that contribute to the observed deviations in the dataset.

Moreover, a noteworthy outlier is the intersection between Tunavägen, Warholms väg and Ole Römers väg (ID 520), see figure 4. This junction experiences a notable volume of activity during peak hours that extends into morning and evening hours, due to the substantial influx of pedestrians and cyclists since both the university and student dorms are very close by. While the traffic flow at this intersection is comparatively modest in contrast to previously discussed outlier, its distinctive attribute resides in the proximity of numerous parking facilities. Notably, the nearness of a grocery store and an elementary school introduces a variable that plausibly may contribute to inaccuracies during the data collection phase. An error scenario could be caused by vehicles utilizing nearby parking spaces, potentially avoiding detection or failing to be recognized within the intersection, particularly if they do not resume motion promptly.

Figure 4: Intersection between Warholms väg, Tunavägen and Ole Römers väg, extracted from Google Maps[2]

Basically, static traffic could be a reason for the extreme values in some of these outliers while dynamic traffic tends to give a more precise result as well as a clearer picture on how different people behave when it comes to cycling, driving or walking.

5.2 Outliers on the outskirts of the city

Another outlier that was examined in-detail is the intersection of Getingevägen and Scheelevägen (ID 407), which presents an interesting case, highlighting that data collection challenges are not confined to city centers. In contrast to the previously discussed intersection, this particular junction has a more high-capacity character primarily for vehicular traffic, with fewer provisions for pedestrians or cyclists. Nevertheless, the presence of a roundabout introduces a unique dynamic that could have influenced the accuracy of bicycle counts. The intricacies of the roundabout, especially the interaction with cars maneuvering around it, may have introduced variability during both application-driven data collection and the municipality's counting process.

[2] Google Maps - https://maps.google.com/

Figure 5: Intersection between Getingevägen and Scheelevägen, extracted from Google Maps[3]

The design of this intersection made it necessary to perform a peculiar examination of the potential impact of the roundabout's design on the reliability of the recorded bicycle volumes. Exploring these specific features provided insights into how varying road configurations, even those outside city centers, contribute to the complexities of data collection and the subsequent interpretation of outliers in the dataset. Additional investigation is warranted to understand the interplay between road design elements and data accuracy in capturing the true nature of bicycle traffic at this intersection.

Delving further into the intricacies of outliers, it becomes evident that their origins extend beyond mere issues of infrastructure or data collection methodologies. The localization of the intersection itself emerges as a substantial factor influencing the observed variations. Specifically, the road next to the intersection of Spelmansvägen, Tunavägen, and Sångarevägen is none other than E22, renowned for its consistently high annual traffic volume as reported by Trafikverket. This distinctive feature introduces a potential layer of complexity that might impact both the municipality's data collection efforts and the precision of TravelVu's application.

[3] Google Maps - https://maps.google.com/

TRAFIKVERKET

Avsnitt: 2240152 Län: M Vägnummer: 22

Årsmedeldygnstrafik

Avsnitt	Fr o m	Till	Mätkod	Mätår	Mätriktning	ÅDT(OS) Samtliga fordon	ÅDT(OS) Tunga fordon	ÅDT(OS) Axelpar
2240152	1994-01-01	1998-01-01	2	1993	1	10985±(8%)	707±(14%)	11606±(8%)
2240152	1997-01-01	1998-01-01	2	1993	2	11014±(8%)	834±(14%)	11775±(8%)
2240152	1998-01-01	2002-01-01	2	1998	1	12278±(6%)	914±(8%)	13116±(6%)
2240152	1998-01-01	2002-01-01	2	1998	2	12139±(6%)	956±(8%)	13095±(6%)
2240152	2002-01-01	2006-01-01	2	2002	1	14490±(6%)	1158±(8%)	15521±(6%)
2240152	2002-01-01	2006-01-01	2	2002	2	14554±(6%)	1277±(8%)	15727±(6%)
2240152	2006-01-01	2011-01-01	2	2006	1	16487±(6%)	1463±(7%)	17760±(6%)
2240152	2006-01-01	2011-01-01	2	2006	2	16664±(6%)	1542±(7%)	18037±(6%)
2240152	2011-01-01	2015-01-01	2	2011	1	15542±(8%)	1604±(7%)	16753±(8%)
2240152	2011-01-01	2015-01-01	2	2011	2	16258±(12%)	1602±(9%)	17546±(12%)
2240152	2015-01-01	2019-01-01	2	2015	1	19190±(5%)	1829±(6%)	20458±(5%)
2240152	2015-01-01	2019-01-01	2	2015	2	19363±(5%)	1942±(6%)	20763±(5%)
2240152	2019-01-01	9999-12-31	3	2019	1	21040	2155	22435
2240152	2019-01-01	9999-12-31	3	2019	2	21015	2195	22555

Figure 6: Annual trafflic flow for roadway E22, extracted from Trafikverket[4]

The busy nature of E22, distinguished by a continuous flood of vehicles, may influence on the accuracy of data collected in the vicinity of the intersection. The immense traffic flow on E22, a prominent highway, may accidentally introduce variations in the measurements conducted by the municipality or TravelVu's application. It is possible that the substantial traffic flow, or perhaps the swiftness with which vehicles navigate this roadway, could introduce a level of distortion in the recorded data.

Moreover, the observed patterns might also be attributed to a phenomenon previously mentioned—the deceleration of vehicles passing through the intersection. This potential deceleration, similar to the slowdowns encountered in city centers, finds resonance in Google Maps' depiction of typical traffic in the area. The suggestion of a generally slower traffic pace, occasionally marked by traffic jams or disturbances, underscores the likelihood of vehicular deceleration at this intersection.

[4] Trafikverket - https://vtf.trafikverket.se/SeTrafikinformation

Figure 7: Intersection between Spelmansvägen, Tuvavägen and Sångarevägen, extracted from Google Maps[5]

Fundamentally, the intersection of Spelmansvägen, Tunavägen, and Sångarevägen (ID 519) represents a complex combination of factors. This includes the notable high traffic volume along E22, potential disruptions in the flow of traffic, and the natural features that distinctly define a road intersection.

This complex perspective triggers a deeper question into the unique dynamics of this specific location, aiming to explain the intricate factors that play a role in the observed deviations within the data patterns. Further investigation is crucial to comprehensively understand the confluence of these factors and their impact on data accuracy.

[5] Google Maps - https://maps.google.com/

6 Discussion

The methodology presented by MetroCount and Timemark shows how careful they are about picking the right equipment and installation protocols to ensure the precision of bicycle and vehicular traffic monitoring. By deploying MetroCount and Timemark systems, a powerful framework is established for investigating traffic dynamics at specific locations, contributing valuable insights to the broader field of transportation research.

Furthermore, not only do these techniques align with the manufacturer's guidelines but also demonstrates the unique characteristics of bicycle traffic. This method enhances the reliability and accuracy of bicycle counts but also contributes significantly to the advancement of transportation research, as well as advancing the precision of automated traffic monitoring systems.

As written in the paper *"A spatial modeling approach to estimating bike share traffic volume from GPS data"* (Brown, M. J., et al. 2022) they also used a similar approach to the one presented in this thesis when collecting data, GPS and manual counting. The difference however is that the deployment of GPS data was not through an application like TravelVu but instead on specific (shared) bikes. This in its turn could have been a major factor as to why there were a significant difference in the results. In addition, the difference in data collection period length may have had an influence on the results. Usually, a longer period tends to give more accurate results.

In addition, the paper delves into the complexity of Hamilton's cycling landscape, emphasizing the integration of technology, open data, and manual digitization efforts to construct a detailed network dataset. The comprehensive nature of the represented dataset not only made it easier for accurate analysis but also acknowledged the distinctive attributes of cycling as a mode of transportation within the city. This in turn made the study significant in contributing to the broader discourse on urban cycling, accessibility, and the operational dynamics of bike share systems.

Another paper that stood out was *"Traffic flow at signalized intersections with large volumes of bicycle traffic"* (Grigoropoulos, G., et al. 2022) since it highlighted the integration of empirical findings and existing literature into PTV Vissim simulation models for bicyclist behavior at intersections. By combining real-world observations with simulation capabilities, the study established a solid foundation for comprehending and forecasting interactions among bicyclists in various situations, ultimately contributing to improved intersection design and safety measures. The utilization of simulation tools like PTV Vissim emerged as a valuable approach for gaining insights into complex traffic dynamics involving bicyclists.

Further on, the study sheds light on how different infrastructures influence left-turn maneuvers, with a particular emphasis on the role of bike boxes and dedicated bicycle lanes. Results highlighted the adaptability of bicyclists based on available infrastructure and signal states, offering insights for intersection design and safety improvements. The video-based approach provided a detailed understanding of real-world bicyclist behavior, playing a part in valuable knowledge to the field of transportation studies.

One more scientific paper that stood out was *"Mapping ridership using crowdsourced cycling data"* (2016) written by Jestico, B., et al, where the author explained the importance of the integration of crowdsourced and manual count data, coupled with predictive modeling. This provided a comprehensive approach to understanding and predicting cycling volumes, the methodology and findings contributed with valuable insights for urban planners, offering a potential tool for assessing cycling infrastructure and informing decisions related to sustainable transportation.

6.1 Method discussion

During the data collection, some of the data was either not calculated correctly, corrupted or not calculated at all which explains why the aggregated counts in some of the intersections are zero. There is one exception in a certain intersection which is in Tornavägen; where, according to the excel files provided by Lund Municipality, there is in fact a measured bicycle volume even though it says on the analysis file that the aggregated count is zero.

When comparing different methods, there's always a common theme between every method when collecting data; the human factor is one of the reasons as to why there are errors during the process. It is inevitable that we make mistakes during the data collection, but apart from that it is also the tools used to collect this sort of data, no matter what kind of method is utilized, whether it is pneumatic tubes, GPS, video, or manual counting: A percentage of error will always be visible. The amount of percentage error varies from method to method but one thing that is certain is that there is always room for improvement. Probably, in the near future after enough research and testing there will be a completely accurate system/technique to acquire data without inaccuracies.

The advantages of counting bike volume with the help of applications by gathering crowdsourced data are many, for starters it is widely used in every country since it’s one of the most reliable methods that exists. It’s also a cheaper and more efficient way since it doesn’t require a lot of manpower and installation, it’s enough to engage volunteers that want to download an application on their mobile devices and let the app do its work by tracking their movement with GPS. Most of the times it shows an accurate result, as previously shown in Figure 1 and 2, where only 13 out of the 120 observations were deemed an outlier, meaning that roughly 90% of the observations were correctly calculated. In comparison to other methods where people must manually count the volume or when having to install different equipment on different locations which take even more time and resources. The biggest disadvantages include technical errors which are not caused by the cyclist but rather by the application itself, e g a bug or a glitch that occurs with the app. It is also not fully developed and totally accurate since technology keeps developing over time, therefore it could provide some miscalculations. Overall, the advantages overweight the disadvantages and it’s more likely to be expanded on and further improved in the near future while the other methods such as using road tubes or manual/video counting, are less likely to not be continued.

Continuing, obstruction of some sort is a major source of why there are so many errors during the process of acquiring data as mentioned in *“Validation of Bicycle Counts from Pneumatic Tube Counters in Mixed Traffic Flows”* (2015). This is a valid reason since everything surrounding the area that’s being monitored could in some way disturb the equipment, for example pedestrians interfering with the installation potentially damaging it, or errors caused by natural causes, non-recognizable vehicles and so on.

7 Conclusion

There are multiple factors as to why the results of crowdsourced data aren't always accurate, but one of the main reasons for the outliers in the data points/observations is because of mixed traffic. It has become evident that different kind of motor vehicles affect the precision during data gathering. Particularly when you're trying to solely count bicycle volume, all types of traffic, even pedestrians, can have an impact when collecting data as everyone behaves differently, either when driving or walking. It becomes even worse during peak hours when the traffic is flooded with all kinds of automobiles, stationary and long queues/pauses during rush hour, which may impact the accuracy.

Another common connection between the outliers is that most of them occur in the center of the city, the most probable cause for this is because of how heavily trafficked it is. Pedestrians, cyclists, scooters and so on all have some sort of disturbance effect when gathering information. Not only that but also intersections near high-speed roads such as motorways also have multiple outliers, the reason for that may be related to how fast the vehicles drive which could be a basis as to why the method deployed didn't work as intended.

8 References

Trivector Traffic 2021. Cykeltrafikmängder - Lund 2021. Preliminärversion.

FORWARD, S (2014). ¨Hållbart resande – möjligheter och hinder¨
https://vti.diva-portal.org/smash/record.jsf?pid=diva2%3A706243&dswid=-5672

Blohm, I, Leimeister, JM, and Zogaj, S (2018). ¨How to Manage Crowdsourcing Platforms Effectively¨
https://web.archive.org/web/20180720145920/https://www.alexandria.unisg.ch/252464/1/BlohmEtAl_2018_HowToManageCrowdsourcingIntermediaries.pdf

Zach (2019). ¨How to Identify Influential Data Points Using Cook's Distance¨
https://www.statology.org/how-to-identify-influential-data-points-using-cooks-distance/

Fong, Y., et al. (2017). "chngpt: threshold regression model estimation and inference."
https://doi.org/10.1186/s12859-017-1863-x

Chris, M (2016). ¨Influence in Regression¨
https://www.lithoguru.com/scientist/statistics/Lecture22.pdf

BROSNAN, M., PETESCH, M., PIEPER, J., SCHUMACHER, S. & LINDSEY, G. (2015). Validation of Bicycle Counts from Pneumatic Tube Counters in Mixed Traffic Flows. *Transportation Research Record,* 2527**,** 80-89.
https://doi.org/10.3141/2527-09

BROWN, M. J., SCOTT, D. M. & PÁEZ, A. (2022). A spatial modeling approach to estimating bike share traffic volume from GPS data. *Sustainable Cities and Society,* 76**,** 103401.
https://www.sciencedirect.com/science/article/pii/S2210670721006740

GRIGOROPOULOS, G., LEONHARDT, A., KATHS, H., JUNGHANS, M., BAIER, M. M. & BUSCH, F. (2022). Traffic flow at signalized intersections with large volumes of bicycle traffic. *Transportation Research Part A: Policy and Practice,* 155**,** 464-483.
https://doi.org/10.1016/j.tra.2021.11.021

SCHNÖTZLINGER, P., BREZINA, T. & EMBERGER, G. (2022). Volunteered mass cycling self-tracking data – grade of representation and aptitude for planning. *Transportmetrica A: Transport Science,* 18, 1470-1495. https://doi.org/10.1080/23249935.2021.1948929

JESTICO, B., NELSON, T. & WINTERS, M. (2016). Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography,* 52**,** 90-97. https://doi.org/10.1016/j.jtrangeo.2016.03.006

DALUMPINES, R., SCOTT, D. M., (2011). GIS-based Map-matching: Development and Demonstration of a Postprocessing Map-Matching Algorithm for Transportation Research https://link.springer.com/chapter/10.1007/978-3-642-19789-5_6

DALUMPINES, R., SCOTT, D. M., (2018). GIS-based episode reconstruction toolkit (GERT): A transferable, modular, and scalable framework for automated extraction of activity episodes from GPS data https://www.sciencedirect.com/science/article/abs/pii/S2214367X17300406

SCOTT, D. M., LU, W., & BROWN, M. J. (2021) Route choice of bike share users: Leveraging GPS data to derive choice sets https://www.sciencedirect.com/science/article/abs/pii/S0966692320309807

Lu, M., Hsu, S. C., Chen, P. C., & Lee, W. Y. (2018). Improving the sustainability of integrated transportation system with bike-sharing: A spatial agent-based approach. *Sustainable Cities and Society, 41*, 44–51 https://www.sciencedirect.com/science/article/abs/pii/S2210670718304700

Scott, D. M., & Ciuro, C. (2019). What factors influence bike share ridership? An investigation of Hamilton, Ontario's bike share hubs. *Travel Behaviour and Society, 16*, 50–58 https://www.sciencedirect.com/science/article/abs/pii/S2214367X18300188

Hansen, W. G. (1959). How accessibility shapes land use. *Journal of the American Institute of Planners, 25*, 73–76 https://www.tandfonline.com/doi/abs/10.1080/01944365908978307

Scott, D. M., & Horner, M. (2008). Examining the role of urban form in shaping people's accessibility to opportunities: An exploratory spatial data analysis. *Journal of Transport and Land Use, 1*, 89–119 http://refhub.elsevier.com/S2210-6707(21)00674-0/sbref0056

Deutsches Zentrum für Luft and Raumfahrt, (2020). UTRaCar and MoMoCar https://www.dlr.de/de/ts/desktopdefault.aspx/tabid-1237/5441_

Angenendt, W., Blase, A., Kl¨ockner, D., (2005). Verbesserung der Radverkehrsführung an Knoten. Baumann, F., 2016. Berner Ampeln stehen auf grün https://www.velojournal.ch/aktuell/nachrichten/detail/berner-ampeln-stehen-auf-gr%c3%bcn

Figliozzi, M., Wheeler, N., Monsere, C., (2013). A methodology to estimate bicyclists' acceleration and speed distributions at signalized intersections. Transp. Res. Rec. J. Transp. Res. Board 2387, 66–75. https://journals.sagepub.com/doi/10.3141/2387-08

Froehlich, Jon, and John Krumm. (2008). "Route Prediction from Trip Observations." Paper presented at the SAE World Congress & Exhibition, April 14.

Stopher, Peter, Q. Jiang, and Camden Fitzgerald. (2005). Processing GPS Data from Travel Surveys." Paper presented at the 28th Australasian Transport Research Forum, ATRF 05, January 1.

El-Geneidy, Ahmed, Kevin J. Krizek, and Michael Iacono. (2007). "Predicting Bicycle Travel Speeds Along Different Facilities Using GPS data: A Proof of Concept Model." Paper presented at the 86th Annual Meeting of the Transportation Research Board, Washington DC.

Segadilha, Ana Beatriz Pereira, and Suely da Penha Sanches. (2014). "Analysis of Bicycle Commuter Routes Using GPSs and GIS." *Procedia - Social and Behavioral Sciences* 162: 198–207. https://www.sciencedirect.com/science/article/pii/S1877042814063010?via%3Dihub

Quddus, Mohammed A., Washington Y. Ochieng, and Robert B. Noland. (2007). "Current Map-matching Algorithms for Transport Applications: State-of-the art and Future Research Directions." *Transportation Research Part C: Emerging Technologies* 15 (5): 312–328. https://www.sciencedirect.com/science/article/abs/pii/S0968090X07000265?via%3Dihub

Jagadeesh, G. R., T. Srikanthan, and X. D. Zhang. 2004. "A map Matching Method for GPS Based Realtime Vehicle Location." *Journal of Navigation* 57 (3): 429–440.

https://www.cambridge.org/core/journals/journal-of-navigation/article/abs/map-matching-method-for-gps-based-realtime-vehicle-location/E2DFAE61CABA4C37CF6AF79E3CFB89F8

KOMMUN, L. 2020. *Hitta dolda cykelpotentialer - ny förståelse av trafik genom att kombinera IoT med traditionella data* https://www.vinnova.se/p/hitta-dolda-cykelpotentialer---ny-forstaelse-av-trafik-genom-att-kombinera-iot-med-traditionella-data/

PAULSSON, U. 2020. *Examensarbeten - Att skriva uppdragsbaserade uppsatser och rapporter.*

KOMMUN, L. 2024. *Ett klimatneutral Lund* https://lund.se/kommun-och-politik/sa-arbetar-vi-med/ett-klimatneutralt-lund

CYCLOMEDIA. 2023. *Aiming for zero* https://www.cyclomedia.com/en/resources/blog/aiming-zero