Estimating cycling travel times accounting for individual capabilites

Hugo Georgenthum

May 2024

Abstract

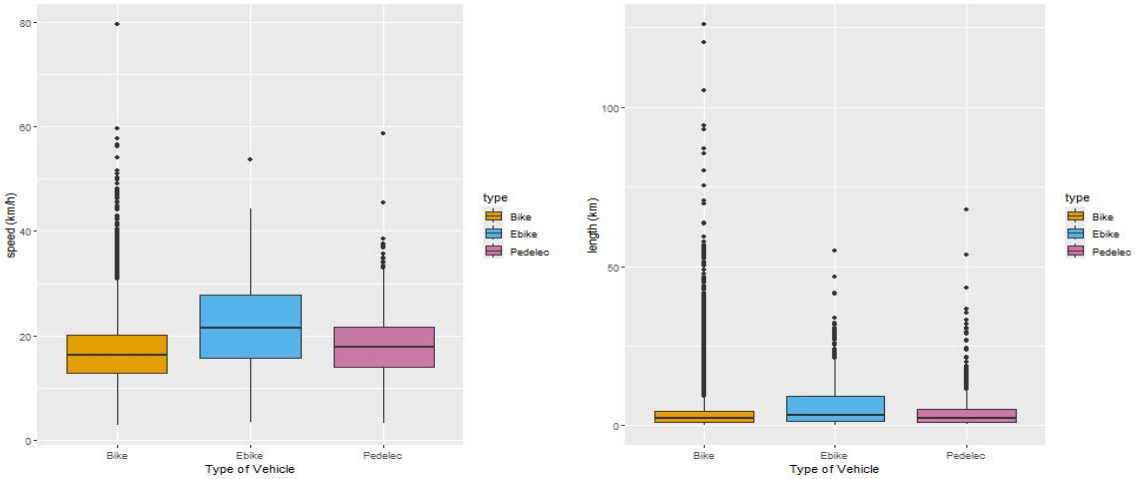
Predicting the duration of bicycle and electric bicycle trips is essential for urban transportation planning and promoting sustainable mobility. This study develops predictive models to analyze the influence of rider characteristics, trip specifics, and road conditions on trip duration. By utilizing statistical methods and machine learning techniques, we aim to improve the accuracy of these predictions, supporting more efficient route planning and infrastructure design, ultimately contributing to sustainable urban development.

1 Introduction

In the realm of travel duration prediction, existing research has explored various factors influencing trip duration, including infrastructure characteristics, individual attributes, and environmental conditions. For instance, recent studies have investigated cyclists' waiting times at intersections, highlighting their significance in overall trip duration estimation. Papers such as Poliziani et al. (1) have developed algorithms to estimate waiting times using GPS traces, allowing for a more nuanced understanding of how intersection types and cyclist attributes affect travel duration. Additionally, research on cyclists' physical capabilities, has shown how factors such as age, frequency of cycling, and gender can impact trips duration (2). By leveraging insights from these studies, future research in travel duration prediction can integrate variables such as estimated power of the cyclist and intersection characteristics to enhance the accuracy of duration forecasts, thus providing valuable tools for transportation planning and decision-making processes.

Bikes with electric assistance have emerged in bikers daily life, and can have a significant impact on bikers behaviour. A normal bike is purely human-powered, while a pedelec provides pedal-assist up to 25 km/h. An S-Pedelec offers more powerful pedal-assist, boosting speeds up to 45 km/h. To illustrate the distinct behaviors of the different bike types, one can examine the distribution of trip speeds. As shown in Figure 1, the median time of a trip with a normal bike (16.33 km/h) is lower than the one with pedelecs (17.89 km/h) and S-pedelecs (21.57 km/h). The length distribution for electric assisted bikes is more concentrated with fewer long-distance trips compared to bikes, which exhibit more outliers. Additionally, the personal features of the biker and route are expected to have different impacts on the speed. In Figure 2, while pedelecs and S-pedelecs exhibit similar behaviors of speed as a function of body-mass-index (BMI), Bike trips have a unique relationship between these covariates. These differences indicate a variation in trip purpose depending on the bike type and highlight the necessity for different models to accurately predict trip duration for the different bike types.

This study aims to address several key questions: What are the important features for duration prediction? How do these features influence trip duration differently for bikes and e-bikes? Are non-parametric methods more effective than linear models in predicting trip duration? By addressing these questions and objectives, this study seeks to improve the understanding of trip duration dynamics in bikes and electric bikes trips and contribute to the development of better duration prediction models.



(a) Speed (b) Length

Figure 1: Speed and length distributions for each bike type

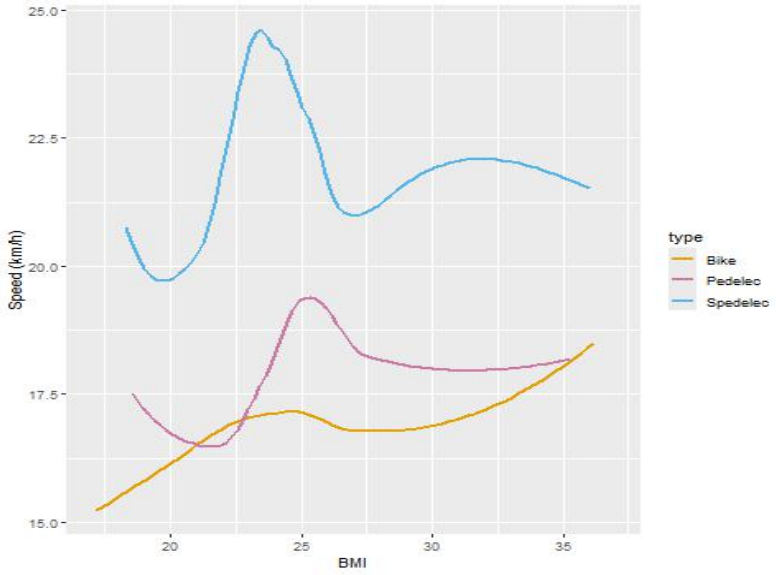


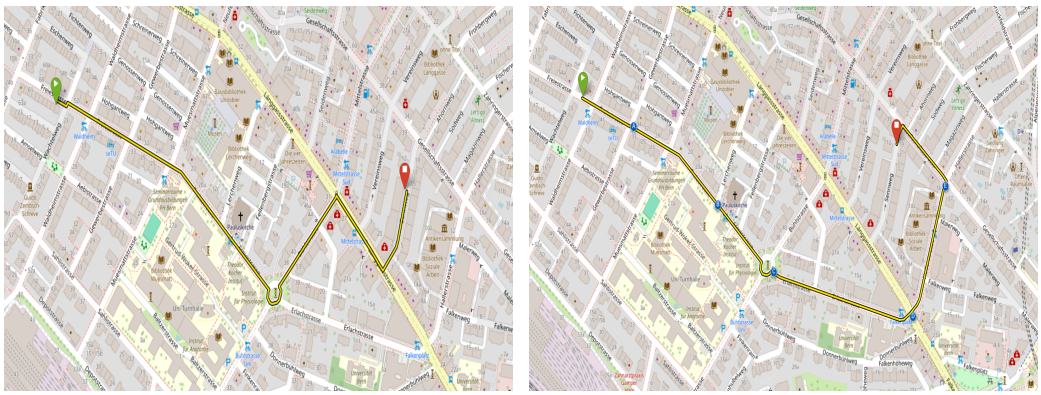
Figure 2: Smoothed representation of speed as a function of BMI

2 Methods

2.1 Dataset Creation

The dataset utilized in our analysis is derived from the EBIS project (3), a robust survey initiative conducted in Switzerland aimed at comprehensively documenting individuals' transportation behaviors. This dataset comprises two distinct parts: the survey data, encompassing personal details of the participants, and the tracking data, which offers a description of the various trips undertaken, thereby providing valuable insights into travel patterns and preferences. The survey provides a diverse array of personal details concerning users and their bicycles. Meanwhile, the tracking data offers a comprehensive collection of trips, comprising duration, starting and ending locations, spatial trajectory, and unique user identifiers for each journey.

The geometry allows the creation of key features duration prediction. The WKB hexadecimal character representation of the geometry is converted into WGS 84 coordinates, which creates a succession of latitude/longitude points. These points form a succession of segments forming end to end the true route. The number of intermediate points is controlled since every segment is computed by calling the brouterR (4), which causes high running time. By keeping only 1 intermediate point over 10, a running time of approximately 10 hours for a computer with 32Go of RAM and i7 processor is needed. The image below shows the importance of using the actual geometry by comparing the path calculated by brouterR using only the starting and ending points with the actual path taken using intermediate points. A higher number of points in the route results in a more precise computation of the route and thus more accurate data for the models.



(a) Without intermediate points (b) With intermediate points

Figure 3: Route computed with brouterR

After simulating the routes, additional variables are incorporated into the dataset, such as maximum, minimum, and average steepness, total positive elevation, and the number of intersections. These variables enhance the representation of route characteristics that potentially influence duration.

Considering the cyclist's potential power is vital, as it depends on factors like age, gender, physical activity level, height, and weight. To estimate this power, a function is used, calculating the VO2 max—a key indicator of aerobic fitness—based on these variables. Then, it derives the power output, considering the VO2 max, BMI, and activity level. This approach ensures a more accurate estimation of the cyclist's potential power during biking activities.

The dataset is categorized into three types of bicycles: standard bikes, Pedelecs , and S-Pedelecs, being used in 2 different models for predictions: One for regular bikes and one for electric assisted bikes.

2.2 Data analysis

To ensure the quality and reliability of the data, a thorough cleaning process was conducted. This involved filtering out rows with implausible or nonsensical values, such as excessively high BMI measurements or average speeds exceeding 200 km/h. By eliminating these outliers and erroneous entries, the integrity of the dataset was improved, ensuring that subsequent analyses and modeling efforts were based on accurate and meaningful information.

A study of different features on the trip duration also allows a better representation of the important features for predictions and get a first selection of the covariates. In figure 4, a higher variance in the influence of positive ascent on the travel time for bike than e-bikes is observed. This plot suggests that electric assisted bike compensate external factors, in this case elevation, with the motor power to stabilize the duration of the trip.

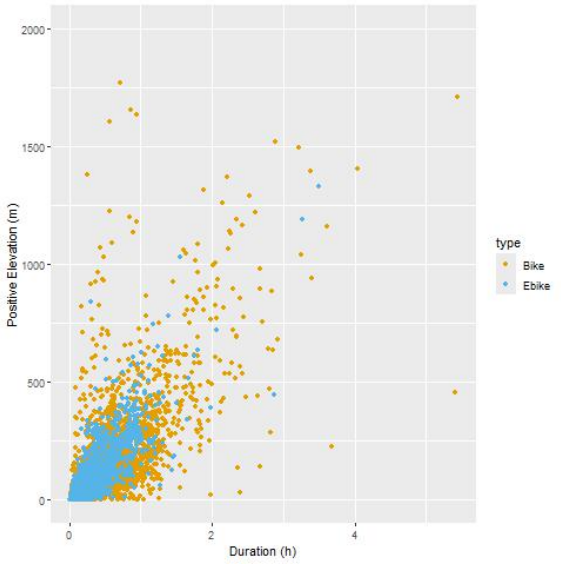


Figure 4: Positive Elevation of the trip against the duration

Considering the road type is crucial for predicting bike trip duration because different road types have varying characteristics, such as traffic density, surface quality, and presence of obstacles, which can significantly impact cycling speed and safety. Understanding these variations helps in creating more accurate and reliable models for trip duration prediction. Below is a table for the different Highway types provided by Open Street Map (Highway in a British English meaning). The description comes from the Wiki (5) is shown in Table 1.

2.3 Linear models

Log transformations were applied to the dataset to normalize skewed distributions and stabilize variances, enhancing the robustness and interpretability of the data for statistical analyses (see Annex). The formula used to model the duration is the following:

Highway	Description
Residentary	Roads which serve as an access to housing, without function of connecting settlements. Often lined with housing.
Cycle way	For designated cycle ways.
Tertiary	Often link small towns and villages
Track	Roads for mostly agricultural or forestry uses.
Secondary	Often link towns
Path	A non-specific path.
Primary	Often link large towns.
Unclassified	Minor roads of a lower classification than tertiary, but which serve a purpose other than access to properties.
Living Street	For living streets, which are residential streets where pedestrians have legal priority over cars, speeds are kept very low and this is can use for narrow roads that usually using for motorcycle roads.
Foot way	For designated footpaths; i.e., mainly/exclusively for pedestrians. This includes walking tracks and gravel paths.
Service	For access roads to, or within an industrial estate, camp site, business park, car park, alleys, etc.
Pedestrian	For roads used mainly/exclusively for pedestrians in shopping and some residential areas
Steps	For flights of steps (stairs) on foot ways.
Corridor	For a hallway inside of a building.
Secondary Link	The link roads (sliproads/ramps) leading to/from a secondary road from/to a secondary road or lower class highway.
Trunk	The most important roads in a country's system that aren't motorways.
Construction	For roads under construction.

Table 1: Highway Tags Descriptions

$$
\begin{aligned}
\text{duration} = \beta_0 &+ \beta_1 \cdot \text{length} + \beta_2 \cdot \text{max\_steepness} + \beta_3 \cdot \text{min\_steepness} + \beta_4 \cdot \text{avg\_steepness} \\
&+ \beta_5 \cdot \text{total\_ascent} + \beta_6 \cdot \text{num\_traffic\_signals} + \beta_7 \cdot \text{num\_crossing} \\
&+ \beta_8 \cdot \text{cycleway\_percentage} + \beta_9 \cdot \text{oneway\_percentage} + \beta_{10} \cdot \text{pa\_intense\_days} \\
&+ \mathbf{1}_{\text{Fair}} \cdot \beta_{11} \cdot \text{health\_status\_Fair} + \mathbf{1}_{\text{Good}} \cdot \beta_{12} \cdot \text{health\_status\_Good} \\
&+ \mathbf{1}_{\text{Very good}} \cdot \beta_{13} \cdot \text{health\_status\_Very\_good} + \beta_{14} \cdot \text{BMI} \\
&+ \mathbf{1}_{\text{cycleway}} \cdot \beta_{15} \cdot \text{most\_covered\_highway\_cycleway} + ... \\
&+ \mathbf{1}_{\text{construction}} \cdot \beta_{29} \cdot \text{most\_covered\_highway\_construction} + \epsilon
\end{aligned}
$$

With $\mathbf{1}_{\text{condition}}$ being the indicator function defined as:

$$
\mathbf{1}_{\text{condition}}(\text{trip}) = \begin{cases} 1 & \text{if condition holds in the trip} \\ 0 & \text{otherwise} \end{cases}
$$

These covariates have been selected to avoid multicollinearity. For example, key features like height and weight are omitted due to the high correlation with BMI (BMI $= \frac{\text{weight}}{\text{height\_m}^2}$), and gender, age, number of intense days by week are used to compute estimated power.

Ordinary least squares (OLS) is constrained to linear relationships between the response and the coefficients, i.e. the model is linear with respect to his coefficients. To address this limitation, non-linear terms such as the inverse of variables, squares, and interactions between pairs of variables were added to the model. This approach helps the model to capture more complex relationships. The Akaike Information Criterion (AIC) was utilized to compare models and select

the best-fitting one. The Model with the lowest AIC value were preferred, indicating a better balance between model fit and complexity:

$$\text{AIC} = 2k - 2\ln(L) \tag{1}$$

with k being the number of kept covariates, and $ln(L)$ the logarithm of the likelihood of the model.

2.4 Non parametric methods

Given the potential limitations of linear models, non-parametric methods are also explored. These do not assume a specific functional form for the relationship between predictors and the response variable. The explored methods are:

• Random Forest is an ensemble learning method that constructs multiple decision trees during training and merges their results to improve accuracy and control over-fitting. Each tree is built using a random subset of the data and a random subset of features. The final output is computed as being the average output of all the trees T_b. In this study, we have $T = 500$ trees.

$$\hat{y}_i^{RF} = \frac{1}{T}\sum_{b=1}^{T} T_b(\mathbf{x}_i) \tag{2}$$

This method is good for capturing complex patterns in the data, as it reduces over-fitting by averaging multiple trees. It also provides a measure of feature importance, helping in the interpretation of the model.

• XGBoost (Extreme Gradient Boosting) builds decision trees sequentially, where each tree attempts to correct the errors of the previous ones. The aim of this method if to optimize the objective function $\mathcal{L}$, composed of a loss term between predicted of real values, and a regularization term:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

Where $\Omega(f_k)$ represents the regularization term of tree f_k that penalizes the complexity of the model (norm of the leaf weights). It also provides insights into feature importance, which helps in understanding the influence of different features.

3 Results

3.1 Linear Models

The linear model results (table 2) indicates significant predictors for normal bike trip duration. An analysis of the estimates, and in particular their sign offers a representation of the effect of each covariate on the response variable. One can observe that length and average steepness of the trip affect the response variable positively, meaning that higher distance and steeper trips a lead to longer trips. Other variables such as *total ascent*, and *num traffic signals* have the same effect since they tend to slow the biker. On the other hand, *min steepness*, *cycleway percentage*, *oneway percentage* and *estimated power* have a negative estimate, meaning that a higher value of the variable leads to shorter trips. In general, we observe that a biker with poorer health status makes shorter trips ($\beta = -0.06$ for Fair against $\beta = 0.04$ for Very good). As seen in Figure 5, healthier biker tends to do longer trips, while it is not the case for e-bikers. In a three-year study, it is possible that a positive feedback loop exists between health and bike trip duration. Initially, healthier participants may take longer bike trips, which could further improve their health. This improvement might then enable them to ride even longer distances, creating a reinforcing cycle of health benefits and increased bike trip duration. Also, the percentage of the trip on a one-way road and on a cycle lane promotes shorter travel times. Surprisingly, *num crossing* has a negative estimate $\beta = -0.001$. One could also observe the estimates for the different highway types of each trip. Compared to residential highway, chosen as reference, cycle way, secondary and corridor have a negative estimate. Alternatively, Steps, Path or Pedestrian have positive estimate.

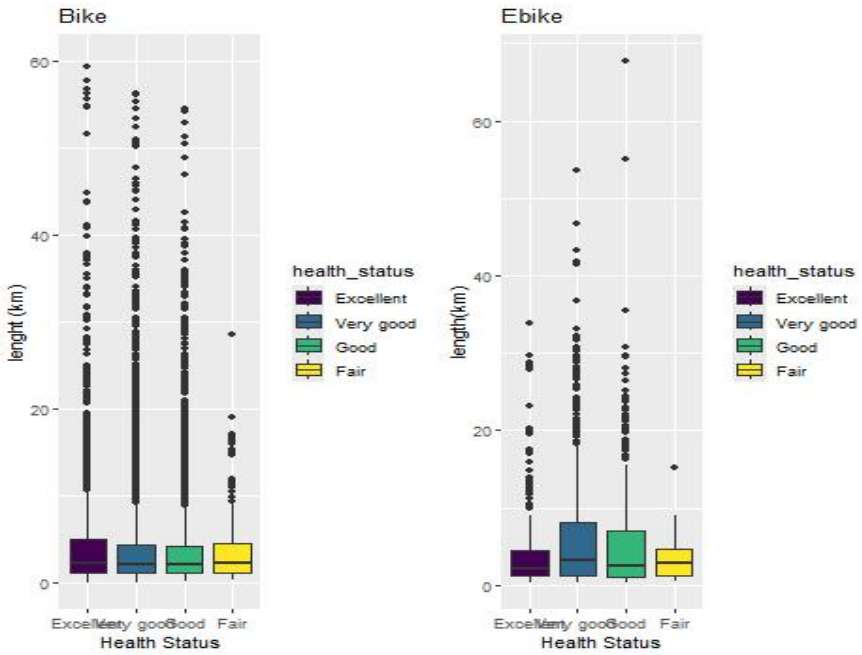


Figure 5: Length distribution for each health status and bike type

The model explains a high proportion of the variance in trip duration ($R^2 = 0.87$), suggesting that the included predictors provide substantial insight into factors influencing normal bike trip duration, as well as Residual Standard Error of 0.31.

Variable	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(**Intercept**)	2.0164	0.1072	18.81	0.0000	***
Length	0.7206	0.0048	150.76	0.0000	***
Maximum steepness	0.0054	0.0027	2.01	0.0448	*
Minimum steepness	-0.0027	0.0006	-4.12	0.0000	***
Average steepness	0.0510	0.0054	9.52	0.0000	***
Total ascent	0.0793	0.0039	20.18	0.0000	***
Number traffic signals	0.0699	0.0052	13.52	0.0000	***
Number crossing	-0.0010	0.0031	-0.32	0.7487	
Cycle way percentage	-0.0022	0.0002	-9.70	0.0000	***
One way percentage	0.0002	0.0001	1.30	0.1939	
Most covered highway: Residential (reference)					
Most covered highway: Cycle way	-0.0206	0.0135	-1.53	0.1271	
Most covered highway: Tertiary	-0.0638	0.0077	-8.29	0.0000	***
Most covered highway: Track	0.1599	0.0120	13.31	0.0000	***
Most covered highway: Secondary	-0.0744	0.0076	-9.81	0.0000	***
Most covered highway: Path	0.0404	0.0124	3.27	0.0011	**
Most covered highway: Primary	-0.0883	0.0097	-9.08	0.0000	***
Most covered highway: Unclassified	-0.0093	0.0121	-0.77	0.4419	
Most covered highway: Living street	0.0566	0.0221	2.57	0.0102	*
Most covered highway: Foot way	-0.0082	0.0095	-0.87	0.3868	
Most covered highway: Service	0.0551	0.0192	2.86	0.0042	**
Most covered highway: Pedestrian	0.1263	0.0272	4.65	0.0000	***
Most covered highway: Steps	0.1465	0.1785	0.82	0.4118	
Most covered highway: Corridor	-0.8288	0.3100	-2.67	0.0075	**
Most covered highway: Secondary link	-0.5940	0.3092	-1.92	0.0547	*
Most covered highway: Construction	-0.1344	0.2187	-0.61	0.5388	
Pa intense days	0.0111	0.0015	7.46	0.0000	***
Health status: Excellent (reference)					
health status: Fair	-0.0622	0.0187	-3.32	0.0009	***
health status: Good	0.0160	0.0074	2.15	0.0312	*
health status: Very good	0.0442	0.0065	6.82	0.0000	***
BMI	0.0145	0.0139	1.04	0.2972	
Estimated power	-0.2906	0.0155	-18.77	0.0000	***

Table 2: Bike: Linear Model Summary

The results using the Electric Bike model (table 3) can be compared to the normal Bike. Not surprisingly, we observe that the effect of health on the trip duration is less significant for e-Bikes, since the potential performance differences are compensated by the electric assistance. This is confirmed by *Estimated power* estimate which has less effect on e-Bikes than on Bikes (-0.14 against -0.29 respectively). Also, the maximum steepness estimate becomes negative.

The results for the e-bike duration prediction model show a higher R^2 value of 0.89 compared to the normal bike model, indicating that it explains a slightly larger proportion of the variance in trip duration. Additionally, the residual standard error is similar at 0.31. Overall, these metrics indicate that the e-bike model performs slightly better than the model for normal bike duration prediction.

Variable	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(**Intercept**)	1.9039	0.2727	6.98	0.0000	***
Length	0.7126	0.0124	57.25	0.0000	***
Maximum steepness	-0.0019	0.0069	-0.28	0.7817	
Minimum steepness	-0.0048	0.0037	-1.33	0.1846	
Average steepness	0.0109	0.0131	0.83	0.4058	
Total ascent	0.0712	0.0105	6.80	0.0000	***
Number traffic signals	0.0882	0.0121	7.31	0.0000	***
Number crossing	-0.0172	0.0081	-2.12	0.0345	*
Cycle way percentage	-0.0007	0.0007	-1.09	0.2739	
One way percentage	0.0017	0.0003	5.33	0.0000	***
Most covered highway: Residential (reference)					
Most covered highway: Track	0.2737	0.0321	8.52	0.0000	***
Most covered highway: Tertiary	-0.0508	0.0200	-2.54	0.0111	*
Most covered highway: Foot way	-0.0117	0.0246	-0.48	0.6335	
Most covered highway: Primary	-0.0695	0.0236	-2.94	0.0033	**
Most covered highway: Cycle way	-0.1434	0.0282	-5.08	0.0000	***
Most covered highway: Secondary	-0.1297	0.0177	-7.32	0.0000	***
Most covered highway: Service	0.1173	0.0542	2.17	0.0304	*
Most covered highway: Unclassified	0.1115	0.0312	3.58	0.0004	***
Most covered highway: Pedestrian	0.2524	0.0938	2.69	0.0072	**
Most covered highway: Living_street	-0.0368	0.0655	-0.56	0.5744	
Most covered highway: Path	-0.0765	0.0261	-2.93	0.0034	**
Most covered highway: Trunk	-0.0965	0.3105	-0.31	0.7559	
Pa intense days	-0.0047	0.0046	-1.02	0.3065	
Health status: Excellent (reference)					
Health status: Fair	-0.0708	0.0424	-1.67	0.0950	.
Health status: Good	-0.0279	0.0223	-1.25	0.2115	
Health status: Very good	-0.0451	0.0213	-2.12	0.0342	*
BMI	-0.1941	0.0439	-4.42	0.0000	***
Estimated power	-0.1471	0.0395	-3.73	0.0002	***

Table 3: E-Bike: Linear Model Summary

The Analysis of variance (ANOVA) analysis in table 4 and 5 analysis that length and total ascent are the most significant predictors for both bike and e-bike models. However, the main difference between the models is that the number of intense days is more important for predicting bike trip duration, whereas BMI is more crucial for e-bike trip duration. Additionally, minimum steepness is not a significant factor in the e-bike model.

Variable	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Significance
length	1	11725.56	11725.56	122800.42	0.0000	***
Maximum steepness	1	76.38	76.38	799.95	0.0000	***
Minimum steepness	1	1.64	1.64	17.21	0.0000	***
Average steepness	1	40.46	40.46	423.79	0.0000	***
Total ascent	1	48.17	48.17	504.47	0.0000	***
Number traffic signals	1	5.97	5.97	62.52	0.0000	***
Number crossing	1	12.61	12.61	132.04	0.0000	***
Cycle way percentage	1	23.68	23.68	248.03	0.0000	***
One way percentage	1	0.00	0.00	0.03	0.8607	
Most covered highway	15	43.21	2.88	30.17	0.0000	***
Pa intense days	1	3.12	3.12	32.73	0.0000	***
Health status	3	9.15	3.05	31.96	0.0000	***
BMI	1	0.07	0.07	0.72	0.3946	
Estimated power	1	33.65	33.65	352.38	0.0000	***
Residuals	18882	1802.94	0.10			

Table 4: Bike - OLS Anova

Variable	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Significance
length	1	2350.42	2350.42	24591.24	0.0000	***
Maximum steepness	1	9.28	9.28	97.08	0.0000	***
Minimum steepness	1	0.45	0.45	4.66	0.0309	*
Average steepness	1	2.46	2.46	25.77	0.0000	***
Total ascent	1	8.74	8.74	91.39	0.0000	***
Number traffic signals	1	2.20	2.20	22.98	0.0000	***
Number crossing	1	3.46	3.46	36.25	0.0000	***
Cycle way percentage	1	0.83	0.83	8.65	0.0033	***
One way percentage	1	3.00	3.00	31.36	0.0000	
Most covered highway	12	18.97	1.58	16.54	0.0000	***
Pa intense days	1	0.23	0.23	2.37	0.1235	
Health status	3	0.52	0.17	1.82	0.1408	
BMI	1	2.22	2.22	23.21	0.0000	***
Estimated power	1	1.33	1.33	13.90	0.0002	***
Residuals	2978	284.64	0.10			

Table 5: e-Bike - OLS ANOVA

The assumptions of linear models were found to be inadequate, notably due to the presence of heteroskedasticity (see Annex). Indeed, Non constant variance test rejects the null hypothesis of a constant variance in the residuals with a $p - value < 10^{-16}$, such as the normality of residuals rejected by the Anderson-Darling test. In response, robust linear models (RLM) were examined as an alternative. Unlike ordinary least squares (OLS) models, RLMs do not consider these assumptions, theoretically offering improved performance. However, in practice, the results yielded by RLMs were remarkably similar to those obtained by traditional OLS models. This suggests that while RLMs provide a robust framework against heteroskedasticity, their practical benefits over OLS may be minimal in this specific context.

The plots of the robust linear models show residual patterns and fitted values that are very similar to those obtained from the ordinary least squares models. This similarity suggests that the robust models, despite addressing heteroskedasticity, do not significantly deviate from the results of the OLS models in this case.

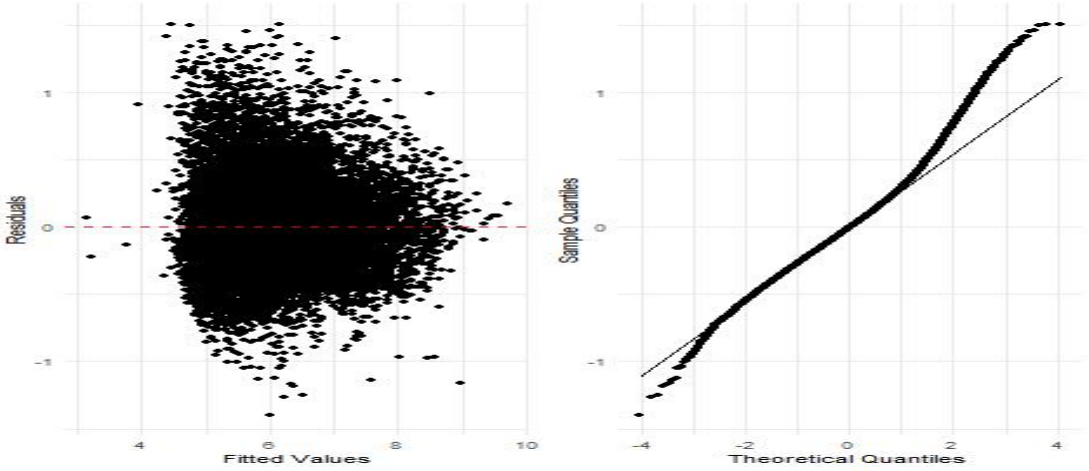


Figure 6: Robust Linear Model for Bike

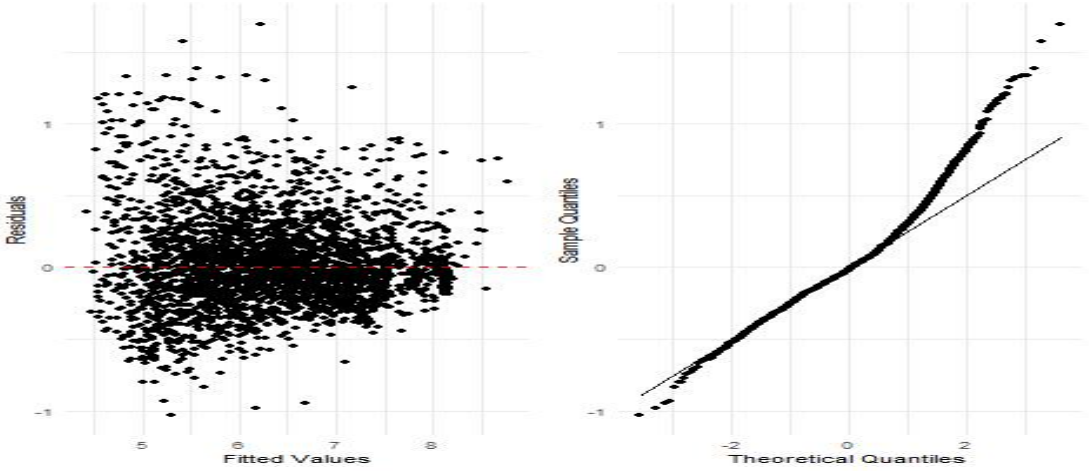


Figure 7: Robust Linear Model for e-Bike

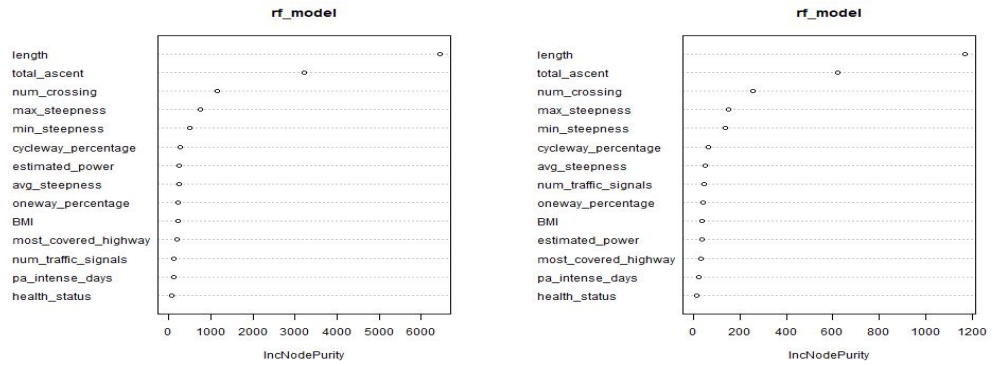
The results of the AIC model selection incorporating non-linear terms, squares, inverses, and first-degree interactions between covariates, also show very similar results as the OLS models. The AIC selection process retained a total of 52 significant covarites, and the increased complexity from

adding these covariates reduces the model's interpretation.
Adding interactions manually is also an interesting strategy (for example length*BMI, total_ascent*avg_steepness,...) but these additional covariates have a too high correlation values with original covariates.

3.2 Non parametric methods

The Figure below displays the importance of covariates in the random forest algorithm giving some explanation. IncNodePurity is a measure of feature importance in Random Forest models, indicating how much a feature reduces impurity when used for splitting nodes. Higher IncNodePurity values suggest the feature contributes more significantly to making accurate splits in the decision trees.

One can observe that the covariates *length*, *total_ascent* and *num_crossing* are the most important features for both models. However, the lower general IncNodePurity values of electric Bikes models translates a duration more regulated by the motor, leading to less variability and lower impurity reductions.



(a) Bike (b) e-Bike

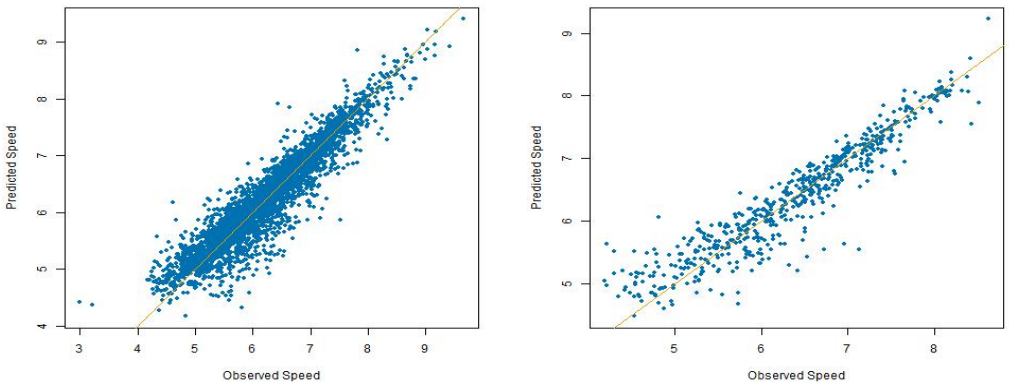
Figure 8: Importance of covariates from Random Forest decision making

In XGBoost, three key measures are used to assess feature importance: Gain, Cover, and Frequency. Gain measures the improvement in accuracy brought by a feature to the branches it splits; higher values indicate more significant features. Cover represents the proportion of observations affected by the feature's splits, showing how widely a feature is used in the model. Frequency counts how often a feature is used for splitting, indicating its prevalence in the decision trees.

To predict trip duration, the feature length is notably more important for bikes (Gain of 0.90) compared to e-bikes (Gain of 0.87). Beyond this difference, the results for other features are relatively similar between the two models, indicating that while trip length plays a more critical role in predicting bike trip duration, other features contribute in comparable ways for both bike and e-bike trip predictions (table 6).

	Feature	Gain	Cover	Frequency
Bike:				
1	Length	0.90	0.18	0.19
2	Total ascent	0.03	0.06	0.07
3	Estimated power	0.01	0.16	0.11
4	BMI	0.01	0.16	0.10
5	Average steepness	0.01	0.10	0.07
6	Maximum steepness	0.01	0.06	0.09
7	Minimum steepness	0.01	0.05	0.09
8	Most covered highway	0.00	0.04	0.04
9	One way percentage	0.00	0.07	0.07
10	Number crossing	0.00	0.04	0.05
11	Cycle way percentage	0.00	0.05	0.04
12	Pa intense days	0.00	0.01	0.03
13	Number traffic signals	0.00	0.01	0.02
14	Health status	0.00	0.01	0.01
e - Bike:				
1	Length	0.87	0.18	0.23
2	Total ascent	0.04	0.07	0.08
3	BMI	0.01	0.10	0.08
4	Minimum steepness	0.01	0.09	0.10
5	Estimated power	0.01	0.11	0.07
6	Maximum steepness	0.01	0.07	0.09
7	Average steepness	0.01	0.12	0.07
8	Number crossing	0.01	0.04	0.06
9	One way percentage	0.01	0.10	0.08
10	Most covered highway	0.01	0.03	0.04
11	Cycle way percentage	0.00	0.06	0.05
12	Pa intense days	0.00	0.02	0.02
13	Number traffic signals	0.00	0.02	0.02
14	Health status	0.00	0.01	0.02

Table 6: Importance of variables from XGboost



(a) Bike (b) e-Bike

Figure 9: Predicted versus Observed values using XGboost

3.3 Comparison

In all models examined, both length and total ascent consistently emerge as the most influential predictors of trip duration. Conversely, one-way percentage generally shows minimal impact across models, suggesting it is not a significant determinant of trip duration. Interestingly, the number of crossings exhibits notable performance, especially evident in linear regression and random forest models, where it consistently shows robust predictive capability.

Explained variance (R^2) quantifies the proportion of variance in the dependent variable that is explained by the model, offering insight into its predictive power. On the other hand, residual standard error (RSE) represents the average distance between observed and predicted values, providing a measure of prediction accuracy. Together, these metrics offer a balanced assessment of both predictive performance and model fit. The table below summarizes the results for the different methods presented comparing these measures.

Table 7: Comparison of Models for Bike and e-Bike

	Bike		e-Bike	
	RSE	R^2(%)	RSE	R^2(%)
OLS	0.31	86.6	0.32	88.7
OLS-AIC	0.31	87.0	0.31	89.6
RLM	0.31	88.6	0.32	88.6
RandomForest	0.30	88.0	0.30	89.6
XGBoost	0.30	87.3	0.33	88.7

In general, E-bike models show the best performance due to the motor assistance, which provides consistent speeds and minimizes the impact of rider effort and fitness levels. This assistance compensate the potential external factors such as weather and elevation, as well as the biker characteristics. As a result, models can capture the patterns more effectively, leading to better prediction accuracy for e-bike trips compared to regular bikes. Random forest outperforms linear methods by capturing complex relationships between predictors and the response variable more effectively.

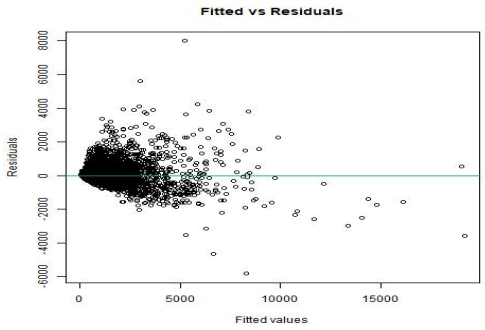
3.4 Conclusion

Promising results were obtained in predicting trip duration, with slightly better accuracy for electric bikes compared to regular bikes. However, there is potential for further improvement. Introducing more variables into the models, particularly those that account for different types of crossings would also be beneficial, in particular to address heteroskdedasticity. As shown in paper (1), there are significant differences in duration based on the type of turn, such as right turns versus left turns. Incorporating these nuances could significantly refine the predictive power of the models.

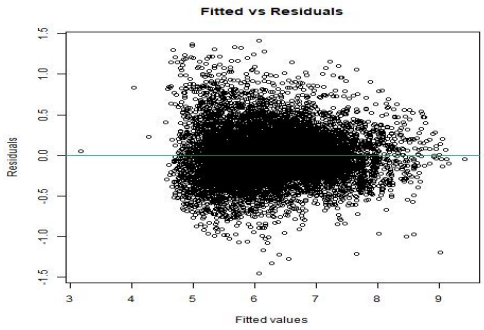
4 Annex

4.1 Linear Models Assumptions

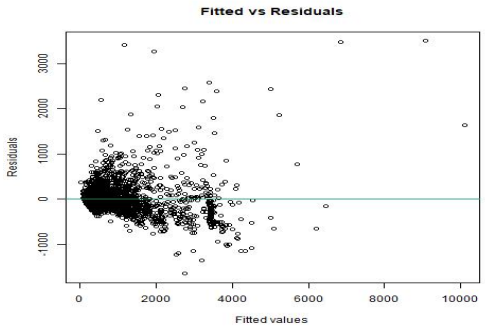
Figure 10 and Figure 11 display the tests for heteroskedasticity and normality of the residuals, respectively. The results show an improvement for the transformed data, indicating that the log-transformation better meets the assumptions of the linear model. However, heteroskedasticity can still be observed after the transformation. The same improvement can be observed for the normality of the residuals, but still not enough to meet linear model assumptions. Table 8 represent the Generalized Variance Inflation Factor (GVIF), which measures the extent of multicollinearity in a regression model, with higher values indicating greater multicollinearity. A maximum threshold of 3 for the value $GVIF^(1/(2*Df))$ was chosen as tolerable.



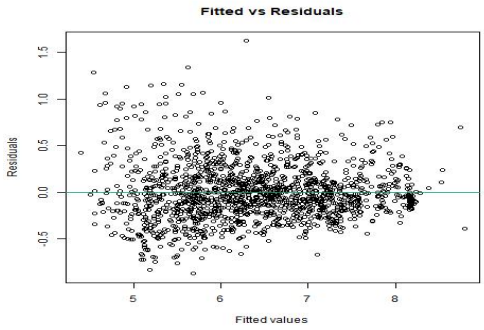
(a) Original plot (Bike)



(b) Plot after log-transformation (Bike)

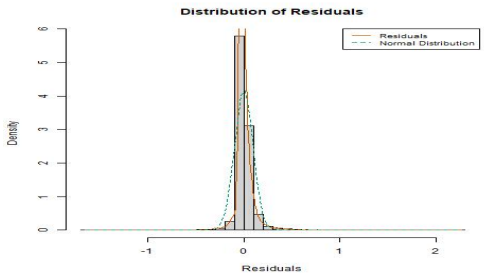


(c) Original plot (e-Bike)

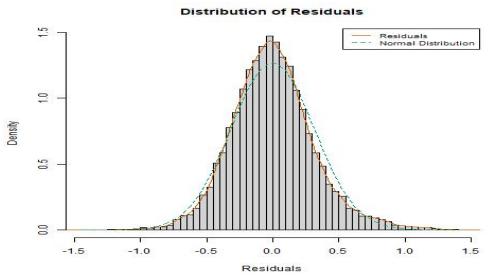


(d) Plot after log-transformation (e-Bike)

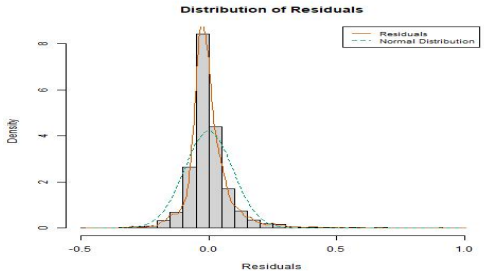
Figure 10: Fitted vs Residual plots using OLS for Bike and e-Bike



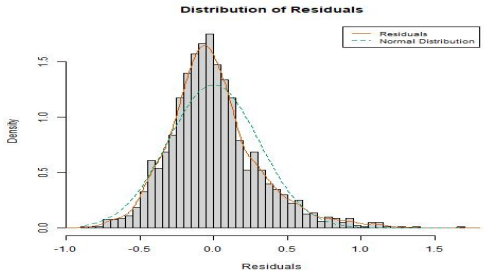
(a) Original plot (Bike)



(b) Plot after log-transformation (Bike)



(c) Original plot (e-Bike)



(d) Plot after log-transformation (e-Bike)

Figure 11: Normality of residuals using OLS for Bike and e-Bike

	GVIF	Df	GVIF^(1/(2*Df))
Bike			
length	4.12	1.00	2.03
Maximum steepness	2.71	1.00	1.65
Minimum steepness	1.09	1.00	1.04
Average steepness	1.51	1.00	1.23
Total ascent	6.15	1.00	2.48
Number traffic signals	1.77	1.00	1.33
Number crossing	3.10	1.00	1.76
Cycle way percentage	1.24	1.00	1.11
One way percentage	1.31	1.00	1.14
Most covered highway	1.78	15.00	1.02
Pa intense days	1.12	1.00	1.06
Health status	1.20	3.00	1.03
BMI	1.12	1.00	1.06
Estimated power	1.08	1.00	1.04
E-Bike			
length	6.12	1.00	2.47
Maximum steepness	2.95	1.00	1.72
Minimum steepness	1.26	1.00	1.12
Average steepness	1.95	1.00	1.40
Total ascent	8.33	1.00	2.89
Number traffic signals	1.95	1.00	1.40
Number crossing	3.66	1.00	1.91
Cycle way percentage	1.24	1.00	1.11
One way percentage	1.30	1.00	1.14
Most covered highway	2.16	12.00	1.03
Pa intense days	1.19	1.00	1.09
Health status	1.37	3.00	1.05
BMI	1.26	1.00	1.12
Estimated power	1.15	1.00	1.07

Table 8: Variance Inflation Factor (VIF)

4.2 Code

This function is structured to estimate power output using empirical formulas and physiological constants, adjusted by demographic and physical activity factors. It serves as a simplified model to estimate power output in a health and fitness context, potentially useful for fitness assessments or training planning.

```
estimate_power <-function(age, gender, sport, height, weight){
  FreiPhysActWeek <- ifelse(sport %in% c(2,3,4,5), 180,
                     ifelse(sport %in% c(1), 90, 75))
  pass <- ifelse(sport %in% c(5,6,7), 8,
           ifelse(sport %in% c(2,3,4), 6,
           ifelse(sport %in% c(1), 4, 2)))
  SexFormula <- ifelse(gender=="Male",1,ifelse(gender=="Female",0,NA))
  BMI <- 10*height/weight
  VO2max <- 57.402-0.372*age+8.596*SexFormula+1.396*pass-0.683*BMI
  O <- (VO2max*weight)/10
  b <- BMI*50/25
  l <- ifelse(BMI<19 | BMI>30, 0.48,
        ifelse((BMI<=30 & BMI>25) & FreiPhysActWeek<=75, 0.55,
          ifelse((BMI<=25 & BMI>19) & FreiPhysActWeek<=75,0.55,
            ifelse((BMI<=30 & BMI>19) & FreiPhysActWeek<180 & FreiPhysActWeek>75, 0.6,
              ifelse((BMI<=30 & BMI>25) & FreiPhysActWeek>=180, 0.6,
                ifelse((BMI<=25 & BMI>=19) & FreiPhysActWeek>=180,0.7,NA))))))
  a <- (O*l)
  W <- a-b
  return(W)
}
```

Listing 1: Power Estimation function in R

References

[1] C. Poliziani, F. Rupi, J. Schweizer, M. Saracco, and D. Capuano, "Cyclist's waiting time estimation at intersections, a case study with GPS traces from Bologna," *Transportation Research Procedia*, vol. 62, pp. 325–332, 2022. 24th Euro Working Group on Transportation Meeting.

[2] L. Meyer de Freitas and K. W. Axhausen, "The influence of individual physical capabilities for cycling adoption: Understanding its influence and mode-shift potentials," *Transportation Research Part A: Policy and Practice*, vol. 185, p. 104105, 2024.

[3] "https://ebis.ethz.ch/,"

[4] "https://brouter.m11n.de/map=19/47.38606/8.55026/CyclOSM,"

[5] "https://wiki.openstreetmap.org/wiki/Key:highwayLink$_r oads$,"

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every written paper or thesis authored during the course of studies. In consultation with the supervisor, one of the following three options must be selected:

I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used no generative artificial intelligence technologies[1].

I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used and cited generative artificial intelligence technologies[2].

I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used generative artificial intelligence technologies[3]. In consultation with the supervisor, I did not cite them.

Title of paper or thesis:

Estimating cycling travel times accounting for individual capabilites

Authored by:
If the work was compiled in a group, the names of all authors are required.

Last name(s):	**First name(s):**
Georgenthum	Hugo

With my signature I confirm the following:

- I have adhered to the rules set out in the Citation Guide.
- I have documented all methods, data and processes truthfully and fully.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for originality.

Place, date	**Signature(s)**
Zürich, 20/06/24	

If the work was compiled in a group, the names of all authors are required. Through their signatures they vouch jointly for the entire content of the written work.

[1] E.g. ChatGPT, DALL E 2, Google Bard
[2] E.g. ChatGPT, DALL E 2, Google Bard
[3] E.g. ChatGPT, DALL E 2, Google Bard