Estimating cycling travel times accounting for individual capabilites

Hugo Georgenthum
Semester Paper presentation

ETHZ

June, 2024

Motivation

- **Increasing Popularity of Biking:**
  - Biking is becoming a preferred mode of transport.
- **Understand key Variables:**
  - Studies on the influence of personal features on biker's behaviour.
  - Gather variables to have the best representation on bike trips
- **Need for Accurate Trip Duration Estimates:**
  - Assists in optimizing route choices.

EBIS Project Overview

- **Project Duration:**
  - 2022 - 2024
- **Project Team:**
  - Joint project by ETH Zurich and University of Basel.
  - Funded by the Swiss Federal Office of Energy (SFOE).
- **Project Goals:**
  - Investigate the impact of e-biking on CO2 emissions in the transport sector.
  - Improve knowledge on transport behavior using GPS tracking and surveys.
- **Data:**
  - Survey on personal features and owned bikes (bike, pedelec, S-pedelec)
  - Tracking data representing trips

Identify Variables

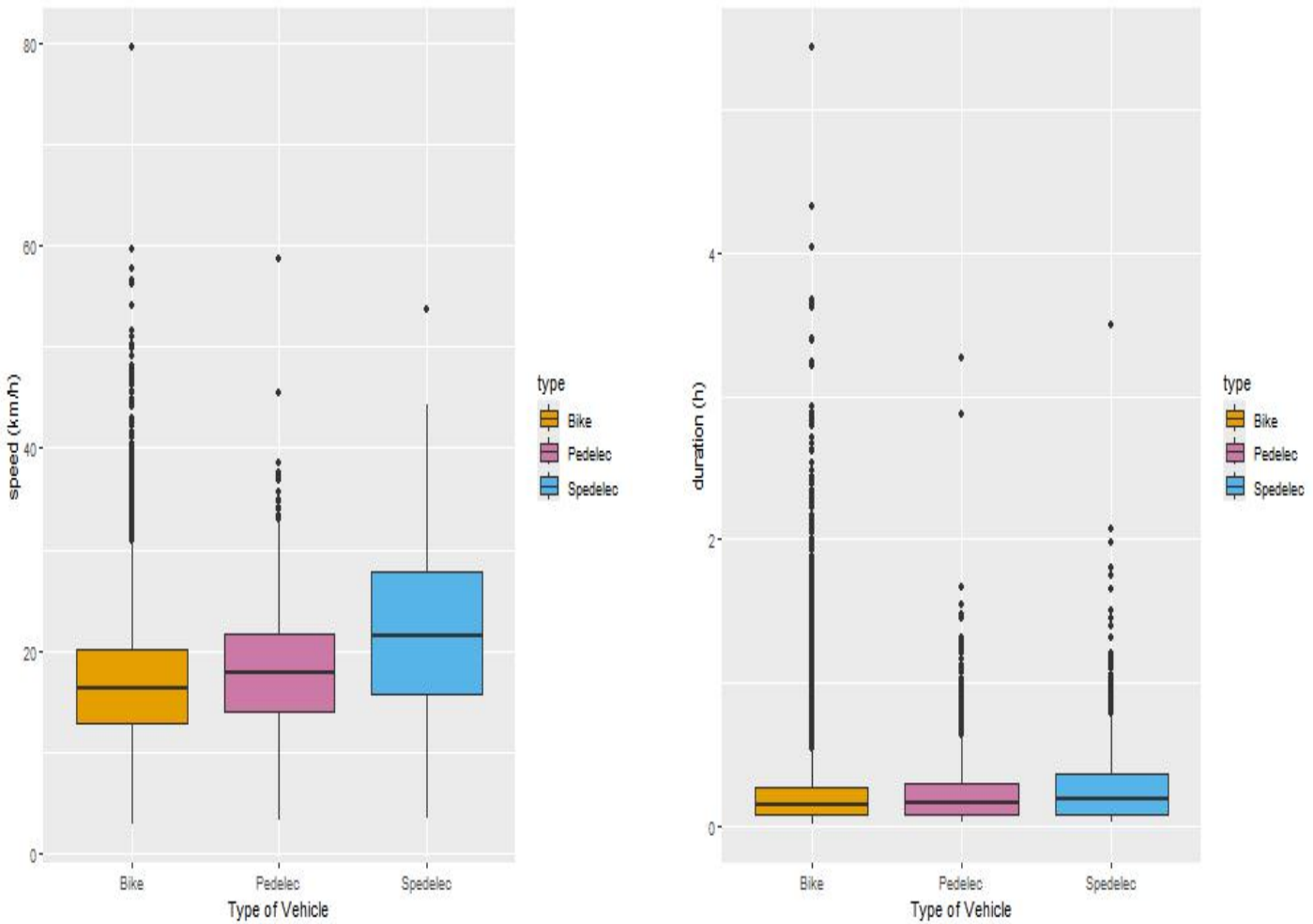

Figure: Speed (left) and Distance (right) distributions for each bike type

Identify Variables

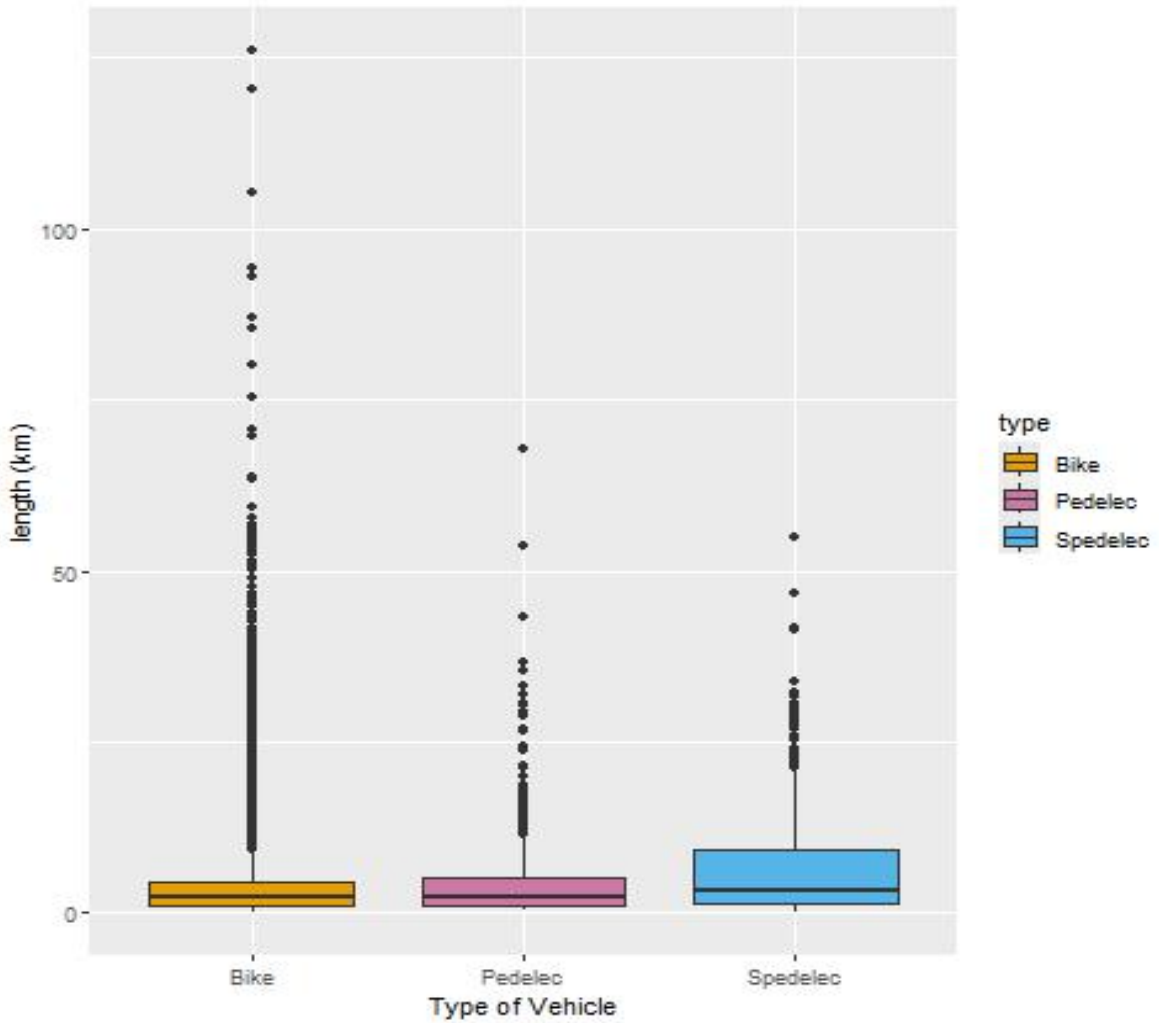

Figure: Length (km) distributions for the different bike types

Identify Variables

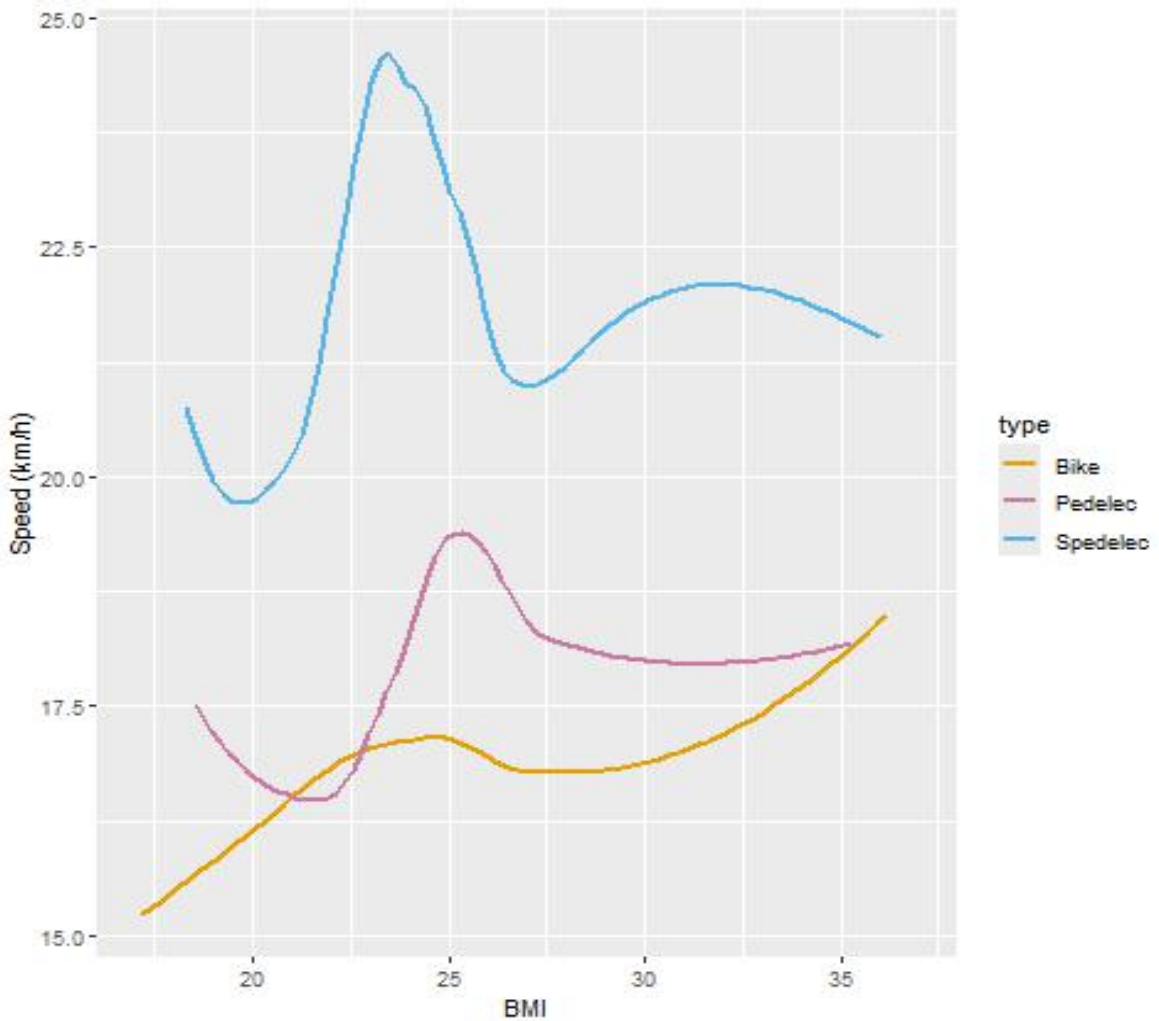

Figure: Smoothed representation of speed as a function of BMI

Identify Variables

- **Biker variables:**
  - Intense days: number of exercise days in a week
  - BMI (height, weight)
  - Power (gender, age, Vo2max)
  - Health status (Excellent, Very good, Good, Fair)
- **Road variables:**
  - Distance
  - Elevation (positive ascent, difference between lowest and highest points,...)
  - Steepness (maximum, minimum, average)
  - Highway type (17 different categories)
  - cycle lane %
  - one way road %
  - Intersections (crossings, traffic lights)

Identify Variables

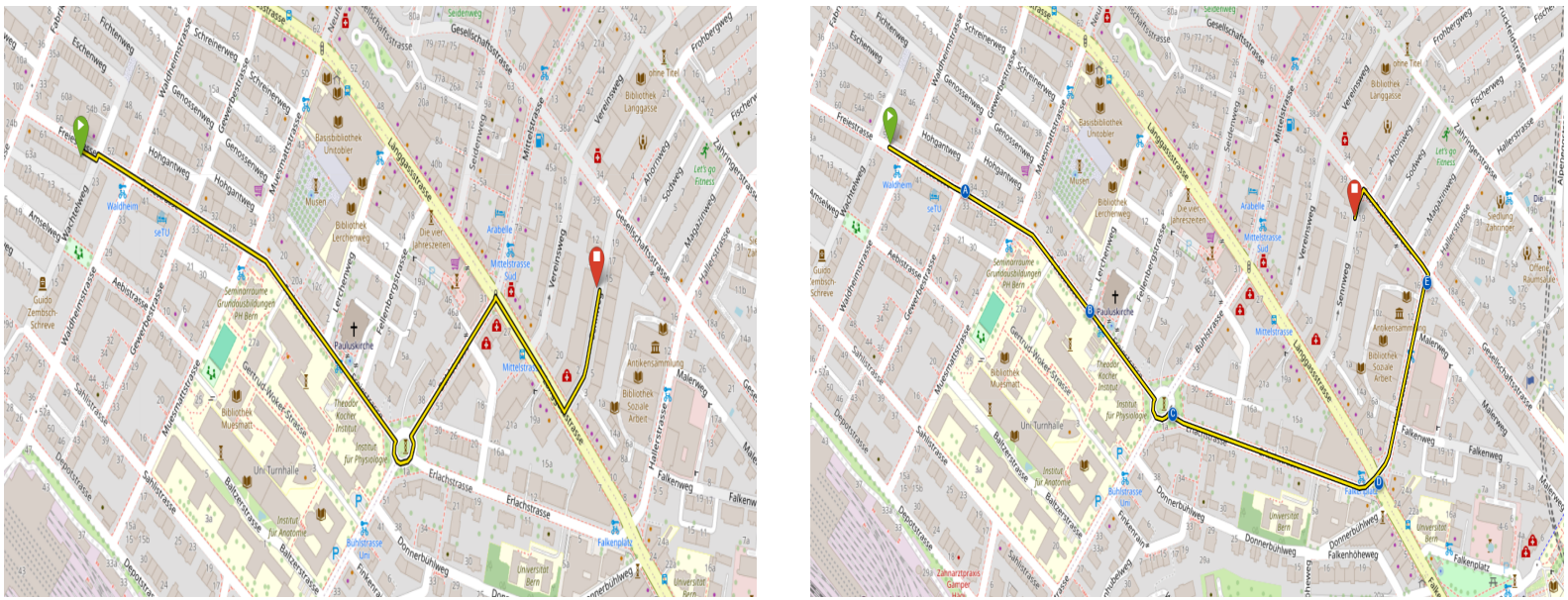

Figure: Route computed with brouterR, without (left) and with (right) intermediate points

Linear Models

- **Ordinary Least Squares:** Minimuize sum of squared distance between observed and predicted values

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \beta)^2 \tag{1}$$

- **Robust Linear Models:** Robust to assumptions violations

$$\beta^{(t)} = \arg\min_{\beta} \sum_{i=1}^{n} w_i^{(t)} (y_i - \mathbf{x}_i^\top \beta)^2 \tag{2}$$

- **Akaike Information Criterion (AIC):** Stepwise feature selection

$$\mathrm{AIC} = 2k - 2\ln(L) \tag{3}$$

Model

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

Linear Models

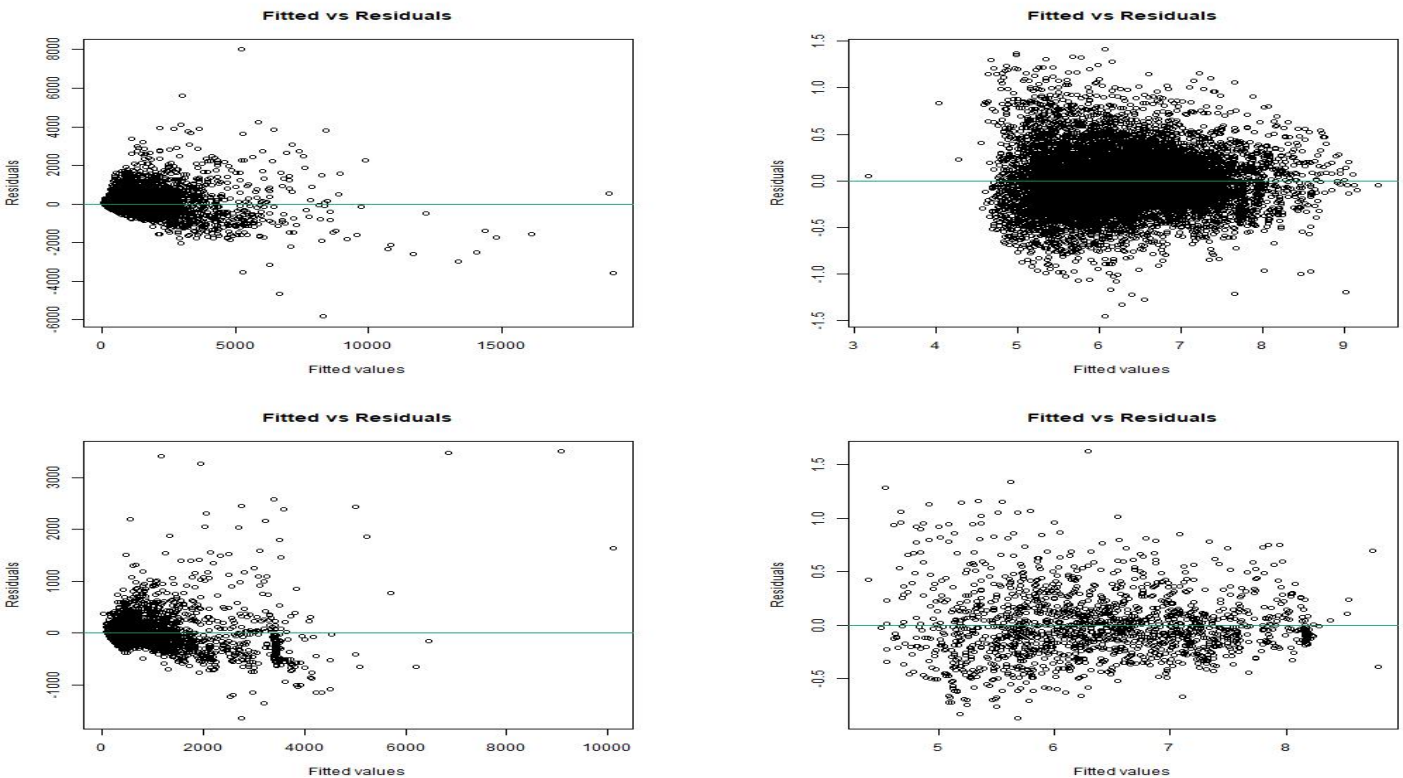

Figure: Fitted vs Residual plots using OLS for Bike and e-Bike

Linear Methods

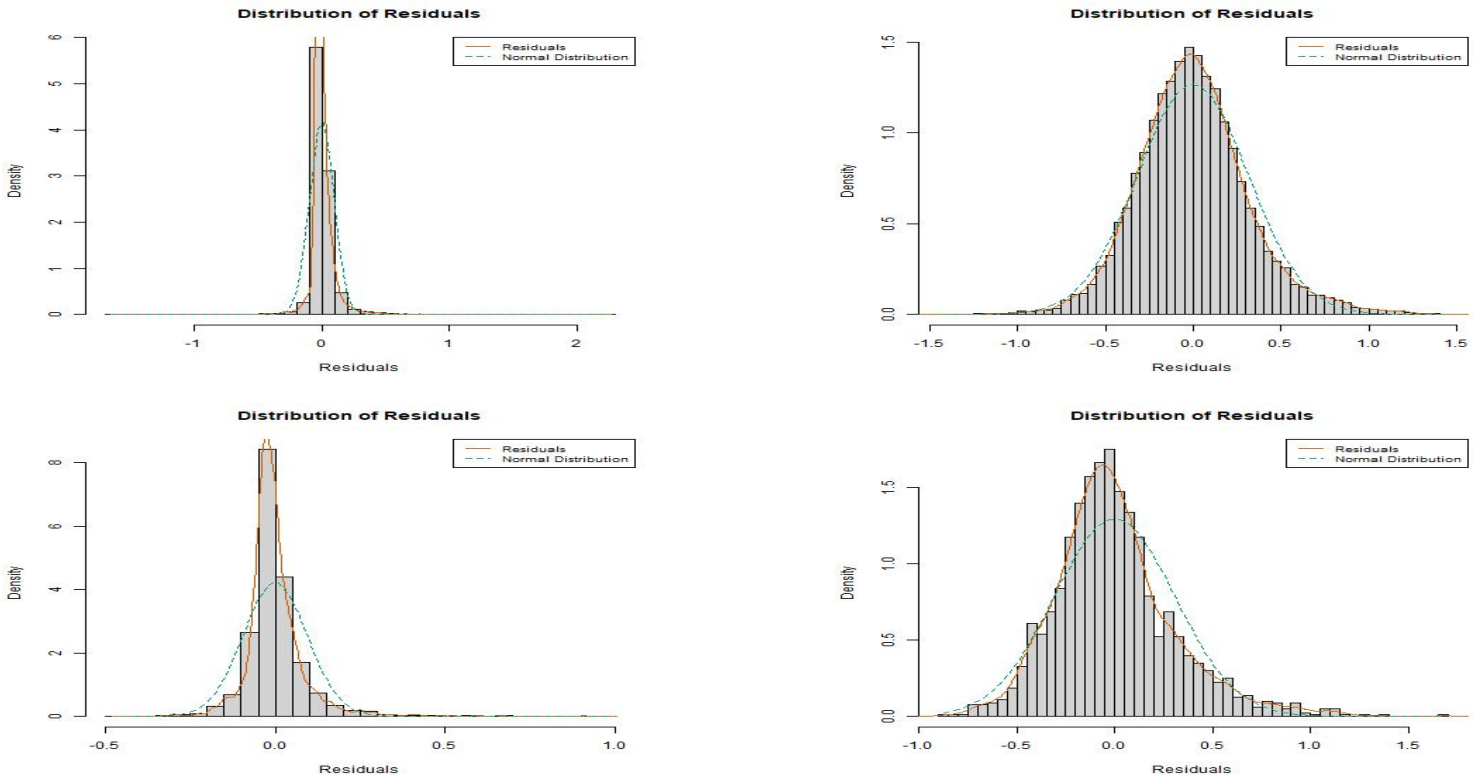

Figure: Normality of Residuals using OLS for Bike and e-Bike

Linear Methods

	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: Variance Inflation Factor (VIF)

Methods

- **Advantages:** No Assumptions, Capture more complex relationships
- **Random Forest:** multiple decision trees and combines their predictions

$$\hat{y}_i^{RF} = \frac{1}{T} \sum_{b=1}^{T} T_b(\mathbf{x}_i) \tag{4}$$

- **XGboost:** trade-off between loss and complexity

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{5}$$

Ordinary Least squares

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: Bike: Linear Model Summary

Identify Variables

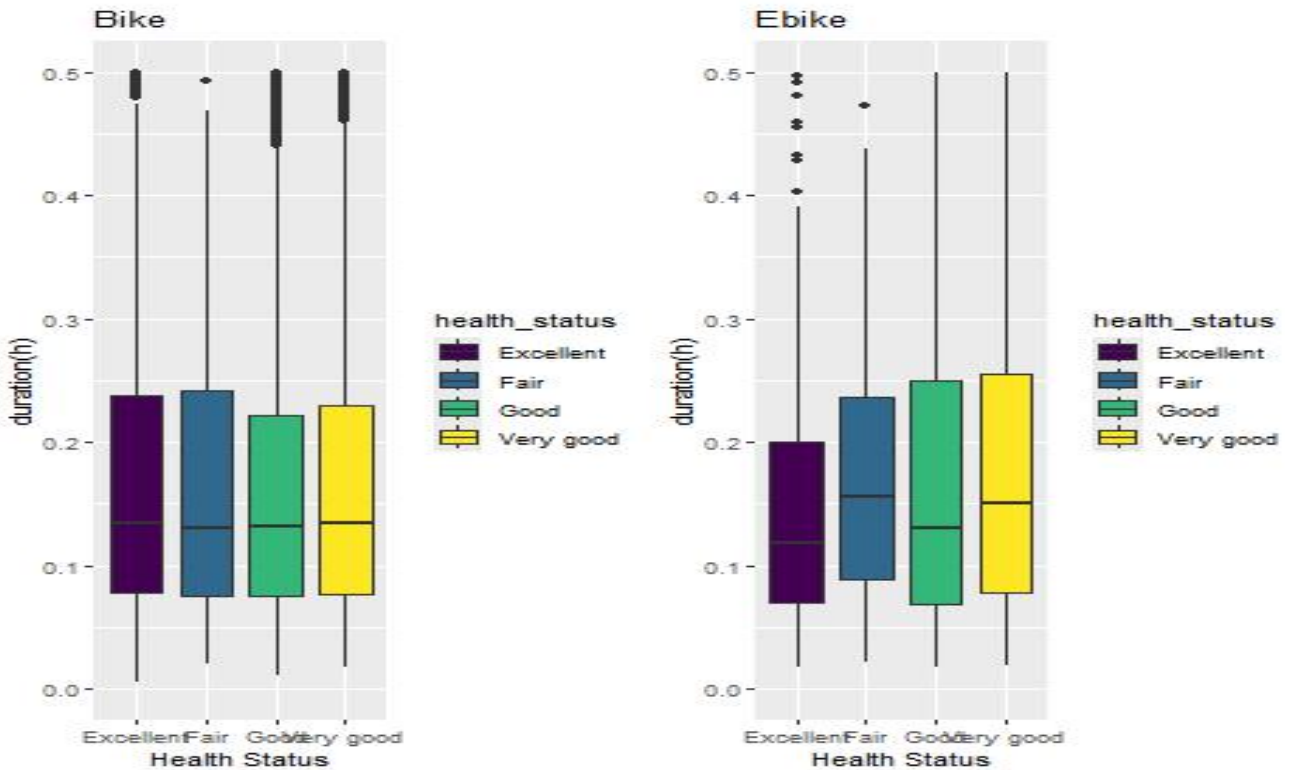

Figure: Duration Distribution (without outliers) for each health status

Status	Excellent	Fair	Good	Very good
Median	0.149	0.138	0.143	0.146

Ordinary Least squares

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: E-Bike: Linear Model Summary

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			
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: OLS ANOVA for Bike (up) and e-Bike (down)

Linear Methods

- **Robust Linear Models:**
  - Same performance in terms of Mean squared Error and Explained Variance

- **Akaike Criterion Information:**
  - 52 retained variables
  - $avg\_steepness * BMI$, $length^2$
  - $num\_traffic\_signals * pa\_intense\_days$, $cycleway\_percentage^{-1}$

Random Forest

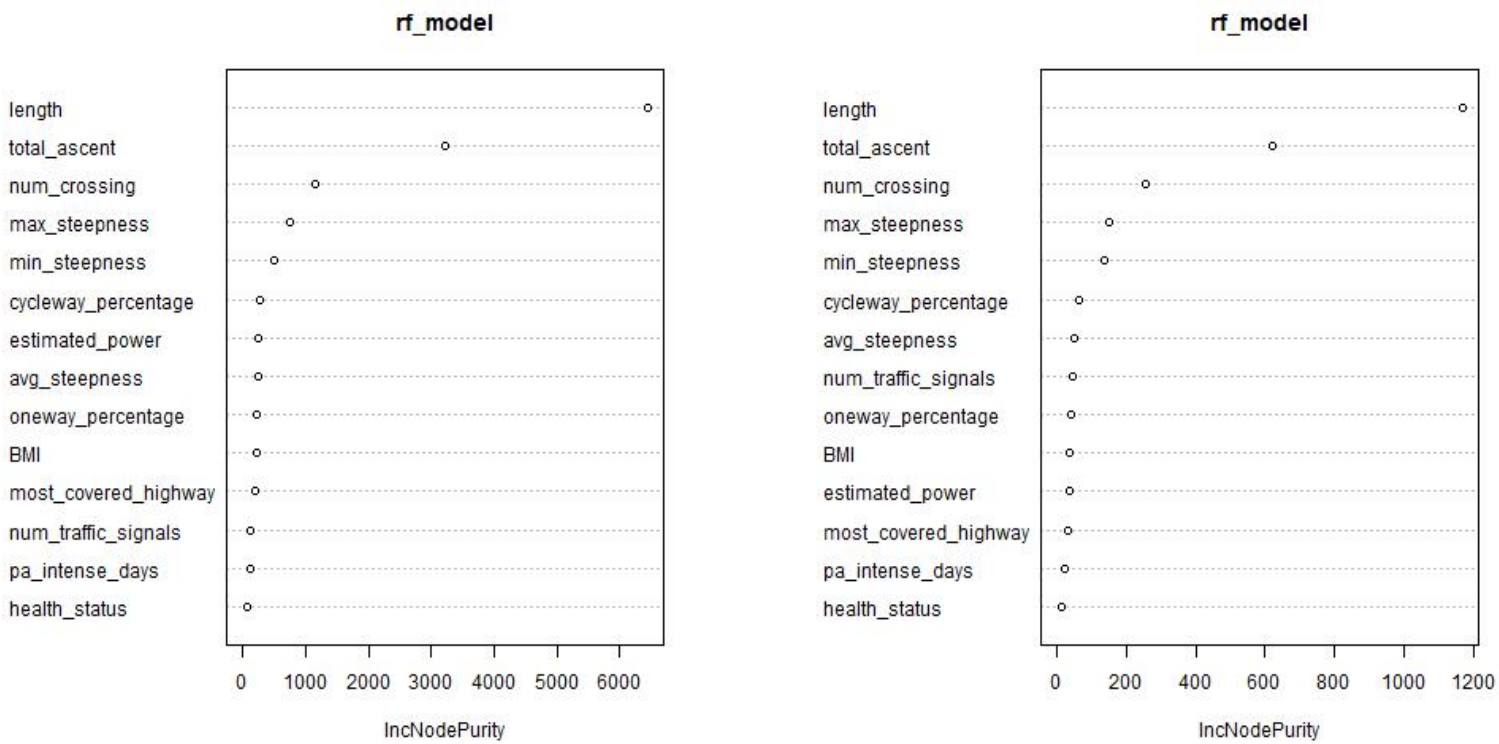

Figure: Importance of covariates from Random Forest decision making

XGboost

		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
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

Table: Importance of variables from XGboost

Model Evaluation Metrics

To evaluate the performance of a regression model, we use the following measures:

- **Residual Squared Error (RSE):**

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **R-squared (R^2):**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Comparison

	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

Table: Comparison of Models for Bike and e-Bike

- Random Forest shows better performance.
- e-Bikes have better results.

Conclusion

- Improved knowledge on key features influencing Bike trips.
- Promising results of the prediction models.
- **further work:** include more detailed route statistics.

References

"https://brouter.m11n.de/map=19/47.38606/8.55026/CyclOSM". In: ().

"https://ebis.ethz.ch/". In: ().

"https://wiki.openstreetmap.org/wiki/Key:highwayLink*roads*". In: ().

Meyer de Freitas, Lucas and Kay W. Axhausen. "The influence of individual physical capabilities for cycling adoption: Understanding its influence and mode-shift potentials". In: *Transportation Research Part A: Policy and Practice* 185 (2024), p. 104105. ISSN: 0965-8564. DOI: https://doi.org/10.1016/j.tra.2024.104105. URL: https://www.sciencedirect.com/science/article/pii/S0965856424001538.

Poliziani, Cristian et al. "Cyclist's waiting time estimation at intersections, a case study with GPS traces from Bologna". In: *Transportation Research Procedia* 62 (2022). 24th Euro Working Group on Transportation Meeting, pp. 325–332. ISSN: 2352-1465. DOI: https://doi.org/10.1016/j.trpro.2022.02.041. URL: https://www.sciencedirect.com/science/article/pii/S2352146522001685.