Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Engineering
Department of Civil and Environmental Engineering

Jonas Lund

E-Scooter speed choice

A statistical analysis of route planning data

Master's thesis in Bygg- og Miljøteknikk
Supervisor: Thomas Jonsson
March 2021

NTNU
Kunnskap for en bedre verden

Jonas Lund

E-Scooter speed choice

A statistical analysis of route planning data

Master's thesis in Bygg- og Miljøteknikk
Supervisor: Thomas Jonsson
March 2021

Norwegian University of Science and Technology
Faculty of Engineering
Department of Civil and Environmental Engineering

NTNU
Norwegian University of
Science and Technology

Summary

This paper investigates the efficacy of using publicly available data about e-scooter locations with a high update frequency for the purposes of analyzing speed choice. Other uses for the same data are also explored. The data is gathered from the EnTur e-scooter API and linear regression, correlation analysis and univariate general models are applied to the data. The structure of the collected data complicates analysis of speed by making central variables impossible to properly apply and this resulted in weak correlations. However, the methodology shows some promise and there are other uses than analysis of speed for which the data might prove very useful.

Contents

Introduction

The increasing prevalence of e-scooter sharing services over the past few years has faced scrutiny in public discourse (Adressa, 2019). While controversial, the scooters have something to offer those with an inclination for big data and traffic analysis: their locations are constantly tracked by GPS, creating an opportunity to do large scale studies on individual transportation. This paper investigates the efficacy of using publicly available data from route planning service Entur in predicting speed of e-scooters based on factors of the environment in which they are operated. Additionally, some consideration is given to alternative application for the data. An understanding of what components of the environment impact speed could have implications for the design of various features of infrastructure and help orient traffic safety efforts toward relevant features of the built environment. Speed being a major contributor to both the probability of accidents and the severity of any accidents that do occur would make such knowledge highly valuable. Furthermore, there may be transferable lessons to, for instance, cycling. A mode of transport for which data about speed is not as readily available.

Methodology

Data for this analysis was gathered through the EnTur (2020) API for e-scooter locations using a Python script (Appendix 1). The script recorded the locations of all available e-scooters in the Trondheim region at the update interval, 15 seconds. There were in the analysis period three companies with e-scooters active in the region, however, only two of these, Tier and Voi, report data to EnTur. For this analysis, data from a total of 3 separate data collection periods were used.

- 15.10.2020 18:08 – 17.10.2020 12:11
- 22.10.2020 17:02 – 23.10.2020 00:53
- 26.11.2020 16:50 – 27.11.2020 19:12

All the data was collected in October and November of 2020 and therefore, seasonal variance, which may be significant, cannot be accounted for. Each observation contained the individual identifying number for the e-scooters, current remaining battery life as a percentage, name of the operator, and GPS coordinates. A timestamp was added to each observation using the current time when the data was collected.

Data about e-scooters were only reported by the API when they were available for use and this presented a problem for the analysis. Because of this, the useful information was the absence of data on an e-scooter, implying that it was in use and therefore unavailable. This also meant observations were much further apart than the expected 15 seconds and that the error of calculating the distance travelled from one observation to the next as a straight line was unacceptably high.

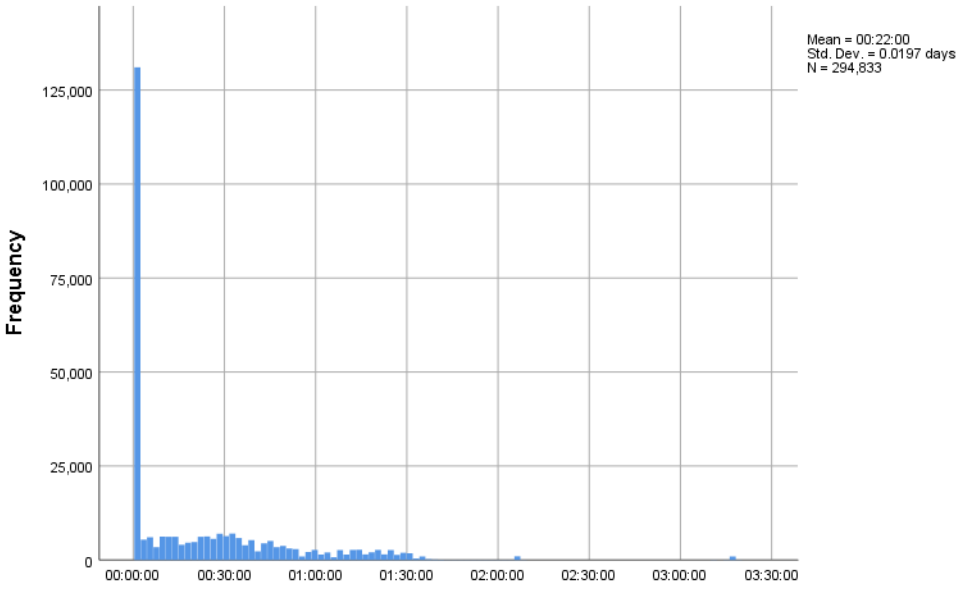


Figure 1 Frequency of trip lengths in terms of time in a sample of the original data

Therefore, it became necessary to perform a network analysis to calculate the shortest path for each route as an approximation of distance travelled. The task then was to identify the last observation prior to missing data, the origin of the trip, and the first observation after missing data, the destination.

A python script (Appendix 2) was run on the data to mark origins and destinations, and to pair them up with an ID number to show they were part of the same trip. There were two approaches to cleaning the data with this in mind, a trip being a change in location or a change in time from one point to the next. Time was chosen due to being easier to handle in Python and due to massive false positives due to GPS drift when using distance as the evaluator. This necessitated choosing a cut-off point greater than 15 seconds where a trip would be considered real. The cut-off was set at 2 minutes (Zou et al, 2020) and consequently all trips included in the analysis are at least 2 minutes long. Figure 1 shows a sample of the data prior to cleaning showing the need for a time cut-off.

Still, quite a few trips with lengths of zero, or close to zero, remained implying false positives from network disconnects, or possibly round trips. A length cut-off was therefore instituted at 200 meters to exclude the former for not being real trips and the latter for the actual length travelled, and therefore the speed, not being knowable. Finally, there were some remaining instances of trips where the remaining battery capacity after a trip was greater than prior to the trip. In these cases, it is likely that the e-scooter was run until it was out of power and was later charged and reappeared in the data, meaning the time measurement is entirely unreliable. These trips were also excluded.

In summary, the exclusion criteria were as follows:

- Trips shorter than 2 minutes
- Trips shorter than 200 meters
- Trips with a negative battery usage

The network analysis was done using a network dataset built in ArcMap with base data from Vegkart (Statens vegvesen, 2021), including non-restricted roads and paths for pedestrians and cyclists, as well as elevation data. The point data was allocated to the network by closest fit and the network analysis was run on the pairs of points accumulating the length of the trips, and the horizontal and vertical climb of each trip. Horizontal climb being the total distance travelled at any incline during the trip, and vertical climb being the total distance climbed vertically during those inclines. These climb data were then transformed into climb rate, horizontal climb divided by length, and climb gradient, vertical climb divided by horizontal climb, in SPSS.

Data about AADT, speed limits for cars, sidewalk and bike road locations, and land use designations were all collected initially, but were not implemented due to the revealed nature of the collected data. Implementation of these potentially relevant data was not done because it is both difficult to determine in what way to distribute the measures across a trip rather than assigning it to a point, and impossible to do so in a manner that does not introduce massive errors as the data necessitated an approximation of path and these factors depend on knowing the specific path to be of any relevance. For example, choosing a side street to the shortest path may have a negligible impact on total distance travelled, but could impact a variable like AADT by orders of magnitude.

Downpour and temperature data was collected from Norsk Klimaservicesenter (2020) and associated with the trips using Excel. The data was gathered from 3 station surrounding the city and then averaged across those 3 observations. The stations were selected for their even distribution around the city and representative elevation. While it is

unlikely for these data points to be significant for speed choice, they may well impact the number of trips taken at a given time and can be useful in contextualizing other data.

Several variables were calculated from the existing data. Speed as a function of the time spent on a trip and the shortest path from origin to destination in the network, battery use as a function of battery before and after a trip, climb rate as horizontal climb divided by length, and climb gradient as vertical climb divided by horizontal climb. Further, the operators, Tier and Voi, were given the number 0 and 1 respectably. The data was then put through statistical analysis in SPSS to explore relationships between the variables.

Results

Distribution of speed

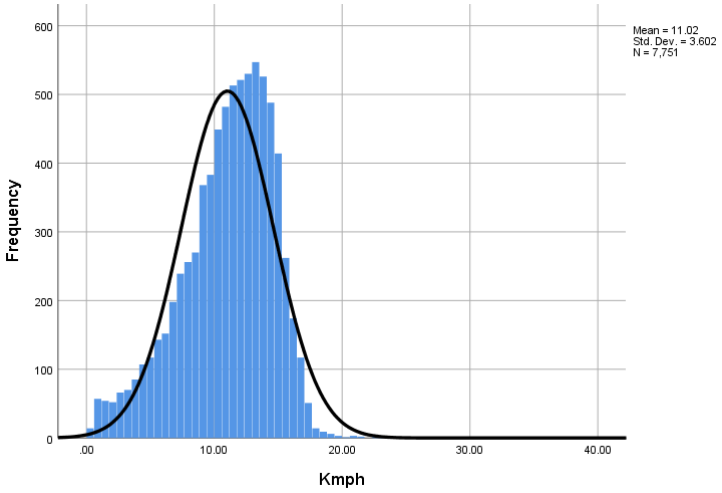


Figure 2 Distribution of speed and normal curve fit

The majority of trips have an average speed between 10 and 15 km/h as shown in figure 2. This is as expected, given that the scooters have a legal maximum speed of 20 km/h (Samferdselsdepartementet, 2018) and some level of impedance is likely. But there are also some rare instances of average speeds above 20 km/h. While that seems intuitively impossible, there are several possible explanations. E-scooters may go faster than 20 km/h downhill if the breaks kick in slow and there is also a chance that some riders chose paths that are shorter than what the network analysis calculated where there are links in the real-life network that, for whatever reason, does not show up in the network data the model is based on. As for distribution, speed appears to be a skewed normal distribution, tilting towards higher values, the skew likely caused by the limitations on speed.

Distribution of trips through the day

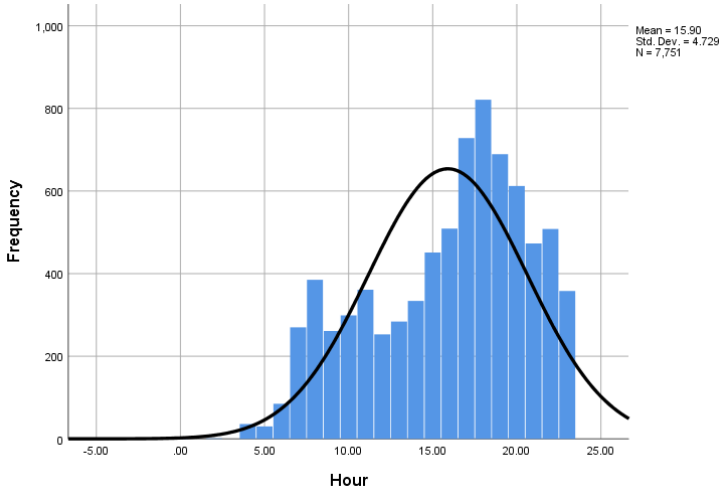


Figure 3 Distribution of trips throughout the day and normal curve fit

Per figure 3, most recorded trips occur in the afternoon and evening, but there is also a spike between 08:00 and 09:00. This lines up with rush hour times to some extent, but there are significantly more trips later in the day. The fit to the normal distribution is not great, meaning that resulting correlations must be treated warily. Also of note is the vanishingly small number of night-time observations. Lastly, the periods of data collection do not match up with full days. Reviewing the observation periods, the afternoon and evening are overrepresented. This is likely the cause of the significant skew of this distribution.

Distribution of travel times

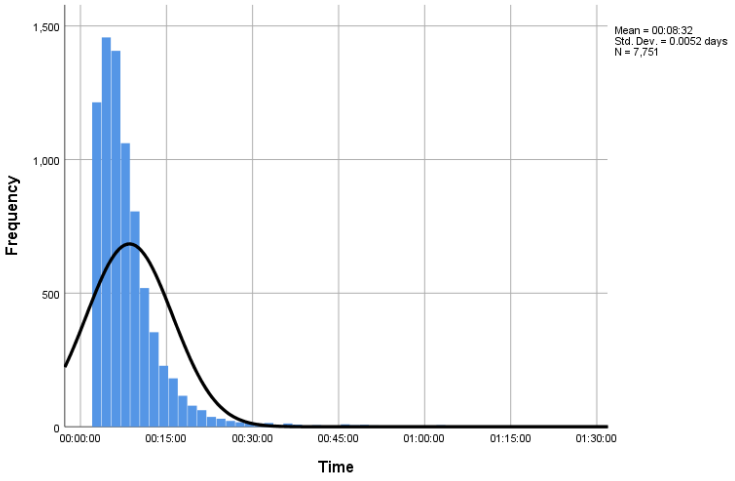


Figure 4 Distribution of travel times and normal curve fit

As mentioned earlier all trips shorter than 2 minutes are removed as part of the cleaning process for the data. This explains the steep drop off at the bottom of the range. Based on this graph there is clearly missing data at the lower end. As travel times approach 0, the frequency spikes considerably and this happens at roughly 2 minutes. It seems likely that some real trips were removed alongside the false trips. The distribution seems to be a left skewed normal distribution, but with the left-most part missing which causes a worse fit than what would be the case if the real and false trips at the lower end could be separated.

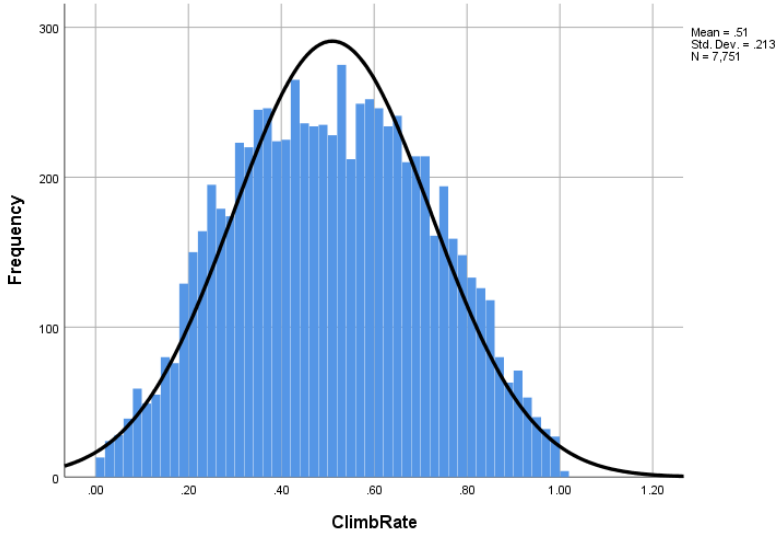


Figure 6 Distribution of climb rate and normal curve fit

Distribution of climb rate

The climb rate, calculated as horizontal climb divided by trip length, measures the proportion of a trip that is travelled at an incline. Climb rate is very close to a normal distribution. This is what one would expect to see if trips were truly random as every uphill is a downhill when travelling the other way. Horizontal climb is the number of meters travelled horizontally at an incline, and even minute inclines are counted.

Distribution of climb gradient

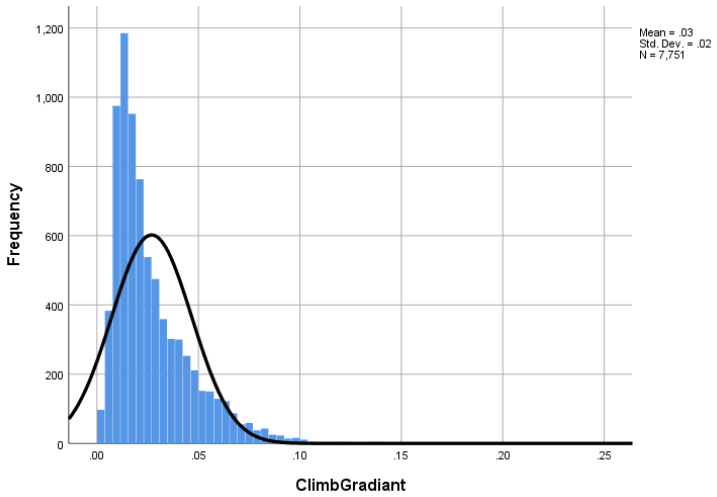


Figure 5 Distribution of climb gradient and normal curve fit

The climb gradient, calculated as vertical climb divided by horizontal climb, is a measure of average steepness of inclines. Vertical climb being the number of meters travelled upwards vertically during all inclines of a trip. Therefore, vertical climb can be greater than the difference in elevation between an origin and a destination. The distribution of climb gradients is clearly skewed towards lower values. This does not map well onto a normal distribution.

Scatter plots:

Before looking at correlations, scatter plots are drawn in SPSS to give an initial impression of the nature of the relationship between speed and other variables. Doing so will prove helpful when interpreting results of various analyses later (Schober, Boer and Schwarte, 2018).

Scatter plot of climb rate and speed

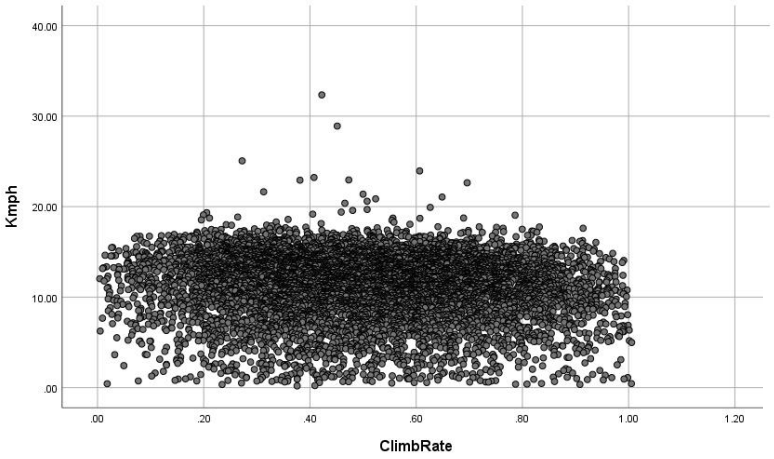


Figure 7 Scatter plot of climb rate and speed

There appears to be little to no correlation between climb rate and speed based on the scatter plot in figure 7. While there seems to be a faint downward trend in speed as climb rate increases, the majority of the data is in a roughly rectangular form with some outliers above it. Looking at those outliers in table 1 shows that they are all short trips in terms of time, compared to a mean of 8 minutes, and that they all have average or below climb gradients as per figure 5. The outliers have no obvious correlation with distance, climb rate and hour of the day and there are too few data points for an analysis to be revealing. Based on these observations it seems these outliers are instances where some circumstance on a portion of the trip lead to a higher speed than the speed limit and that because the trips are short, the impact of that circumstance was significant to the average speed. This may be the case if there were a significant downhill or a shortcut not present in the network analysis.

Distance	Time	Kmph	ClimbRate	ClimbGradiant	Hour
1482.53	0:02:45	32.35	.42	.02	13.00
1405.04	0:02:55	28.90	.45	.02	19.00
1364.01	0:03:16	25.05	.27	.02	20.00
1436.64	0:03:36	23.94	.61	.02	19.00
1367.33	0:03:32	23.22	.41	.03	9.00
1370.53	0:03:35	22.95	.47	.03	12.00
1241.48	0:03:15	22.92	.38	.03	18.00
1351.63	0:03:35	22.63	.70	.01	16.00
1880.93	0:05:13	21.63	.31	.03	16.00
1056.81	0:02:58	21.37	.50	.02	6.00
1924.84	0:05:29	21.06	.65	.02	8.00
1500.79	0:04:19	20.86	.52	.02	15.00
1127.29	0:03:17	20.60	.51	.02	16.00
1012.86	0:02:59	20.37	.47	.02	20.00

Table 1 Outliers in the climb rate and speed scatter plot

Scatter plot of climb gradient

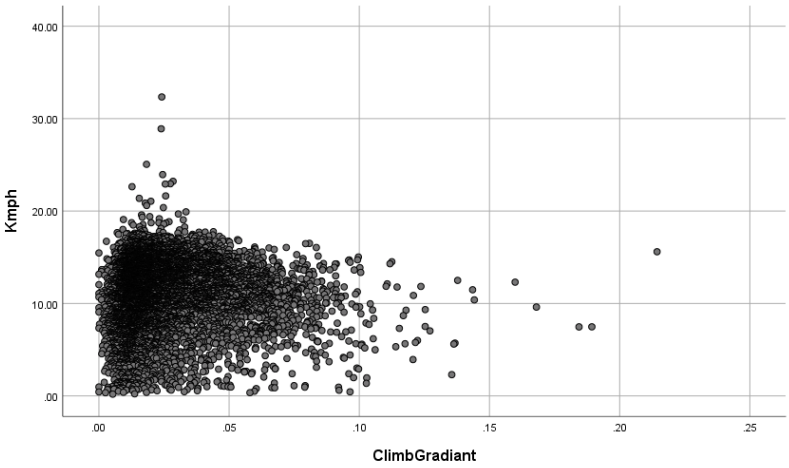


Figure 8 Scatter plot of climb gradient and speed

The scatter plot, figure 8, of climb gradient against speed shows a large chunk of points where both variables are low and outliers at the extremes of one or the other, but not both. The high-speed outliers are the same as in table 1, short trips with a low climb gradient. The high climb gradient outliers make sense intuitively and the initial impression from this plot is that there will be some correlation between climb gradient and speed. The relationship appears to be inverse when considering the whole but looking only at the gradient outliers there is a rising trend. Given the limited number of data points that are outliers, it is difficult to draw strong conclusion based on these observations.

Scatter plot of hour of the day and speed

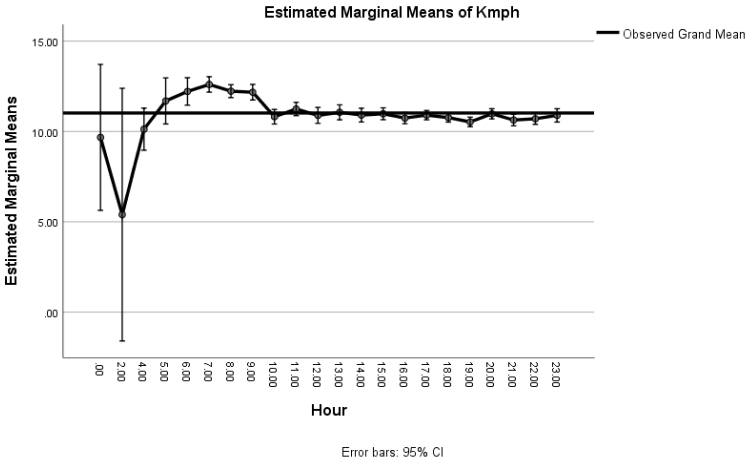


Figure 9 Bar plot of hour of the day and speed

While a scatter plot for hour of the day would be of interest, the variable does not carry minutes with it. Therefore, the scatter plot would be very difficult to read and glean anything useful from. The data is therefore instead put through a univariate general linear model in SPSS to produce a more easily readable bar plot.

This plot shows that speeds are higher early in the morning and go on to hit the mean at about 10. There are very few datapoints for the night-time hours and so that data is unreliable as seen in table 2. From 7 and on though, there is enough data to rely on the results. The increase in speed approximately coincides with morning rush hours and this would imply these are work trips. Given the layout of Trondheim, with a low lying downtown and more elevated residential areas, this may be explained as a prominence of downhill trips to work or school in this period. There is no equivalent drop in speed in the afternoon however, which may imply that other modes are preferred when making the return trip, having to travel up significant gradients. This complicates the interpretation of linear regression later as the hour of the day variable does not seem to have a linear relationship with speed. There appears to be two sets of

Scatter plot of battery use and speed

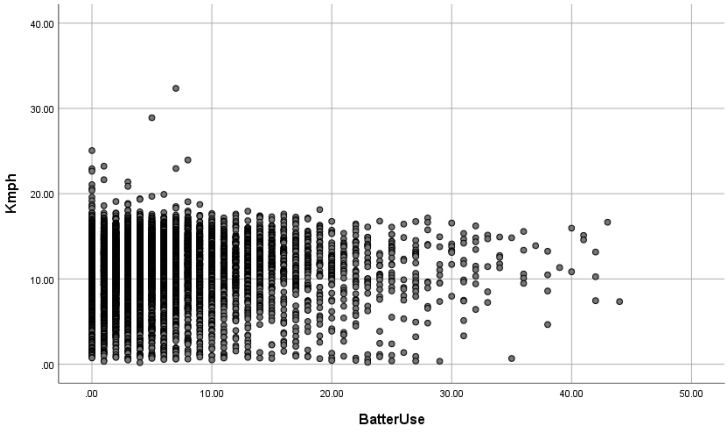


Figure 10 Scatter plot of battery use and speed

Figure 10 shows the scatter plot of battery use and speed. Most of the data crowds the low end of both variables, with a few speed outliers standing out and a gradual dissipation into outliers as battery use increases. A relationship between battery use and speed would speak to efficiency of the vehicle. There is no obvious correlation present in the plot, except that all speed outliers are at the low end of battery use. However, as per table 1, these outliers are short trips, and it would not be surprising for short trips to require little in terms of power.

Scatter plot of temperature and speed

The scatter plot of temperature and speed found in figure 11 reveals nothing obviously correlational between the variables. Speed outliers are evenly

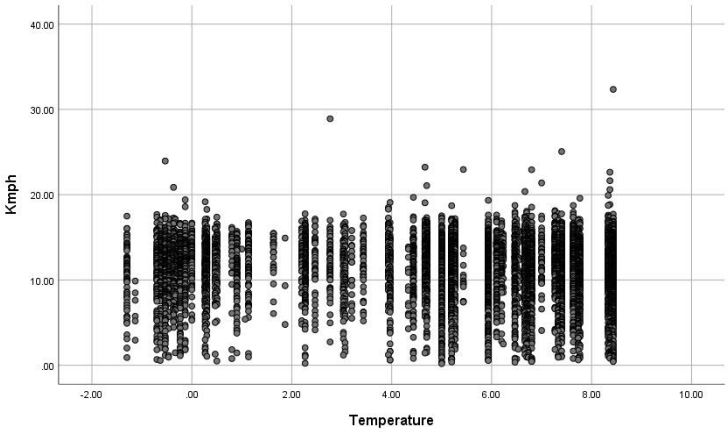


Figure 11 Scatter plot of temperature and speed

distributed and there is no clear trend in the speeds of trips as temperature increases.

Scatter plot of downpour and speed

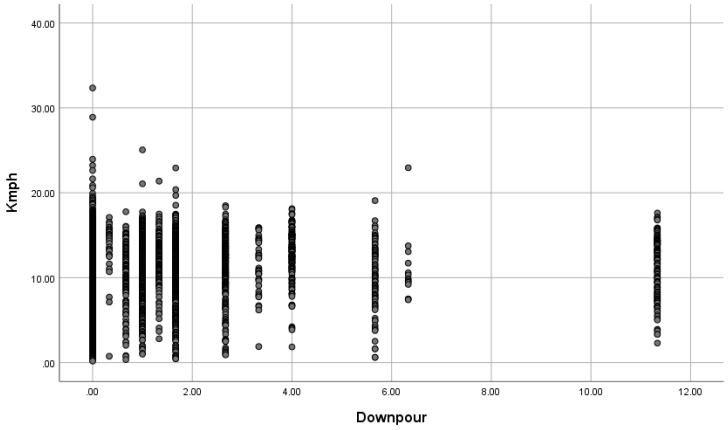


Figure 13 Scatter plot of downpour and speed

Figure 13 shows the scatter plot of downpour and speed. The plot is sparsely populated outside of 0 downpour and no obvious trend can be gleaned from it. That being said, this might imply a relationship between downpour and the number of trips taken. The limited number of trips during downpour could be a result of either limited downpour in the observation periods, or other modes of transport replacing e-scooters when there is downpour.

Scatter plot of operator and speed

The scatter plot of operator and speed, figure 12, contains little in ways of useful information. Operator 0 is Tier and 1 is Voi. The Voi data is more populated just above the 20 km/h mark, but there are more severe outliers on the Tier side. There is little here to indicate any significant correlation and no clear case to be made for any type of mathematical relationship.

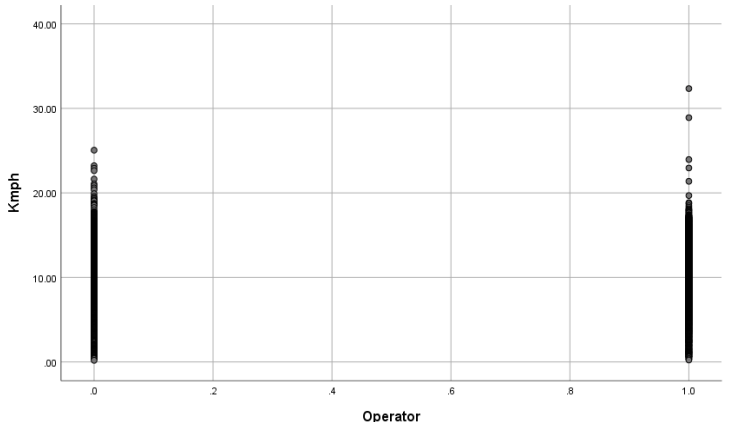


Figure 12 Scatter plot of operator and speed

Linear regression and correlations

Speed in the context of this analysis is average speed throughout trips. This has implications for how one ought to interpret correlation between the variables. In the real world, speed varies significantly throughout trips. A rider may have to stop at a traffic light, slow down to avoid hitting pedestrians, or decrease speed to safely make a sharp turn. Ideally, observations would be frequent enough to capture this variation and these effects would either be accounted for directly or represented by proxies to reveal the true impact of each effect. However, seeing as the nature of the source data dictates an averaging of

Correlations

		Kmph	ClimbGradiant	Temperature	Downpour	ClimbRate	Operator	Hour	BatterUse
Pearson Correlation	Kmph	1.000	-.012	-.052	.022	-.079	-.092	-.106	.017
	ClimbGradiant	-.012	1.000	-.005	-.004	.302	-.009	.026	.229
	Temperature	-.052	-.005	1.000	.139	-.035	.123	-.051	.015
	Downpour	.022	-.004	.139	1.000	-.032	.027	-.255	.020
	ClimbRate	-.079	.302	-.035	-.032	1.000	-.011	.056	.240
	Operator	-.092	-.009	.123	.027	-.011	1.000	.041	.476
	Hour	-.106	.026	-.051	-.255	.056	.041	1.000	.093
	BatterUse	.017	.229	.015	.020	.240	.476	.093	1.000

Table 2 Correlation from linear regression

speed across trips, there is an expectation that correlations to speed be reduced. Therefore, it is reasonable to consider lower thresholds for correlation. Obviously, this is a weakness compared to using data with higher resolution.

The data are put through a linear regression in SPSS which, among other things results in table 3 containing correlations. Note that there are no strong correlations with speed to any of the tested factors here. Hour, operator and climb rate stand out slightly as the strongest contenders for significant variables, but even here the correlation is very weak. And the variable hour is not necessarily reliable given the discussion related to figure 3. However, these correlations must be viewed in light of the earlier scatter plots and the averaging of speed across trips.

Hour of the day had the strongest Pearson correlation. But there are several reasons to be sceptical of the result. It might be better to simply stick to figure 9 which does not have as many complicating factors.

Operator was the second strongest Pearson Correlation, but still very weak. The scatter plot did not help contextualize this part of the data in any way, and although there may be a difference between the speed capabilities of the scooters of these two companies, it appears to be insignificant.

Climb gradient not having a significant impact on speed is a mind-boggling outcome. It is blatantly obvious that an e-scooter will slow down when faced with a sufficiently steep gradient. But e-scooters, having enough power to go faster than 20 km/h if there were not a hard limit on speed, would also have enough power to maintain top speed up moderate inclines. There is then a threshold at which a gradient becomes significant to speed.

The fact that these significant inclines fail

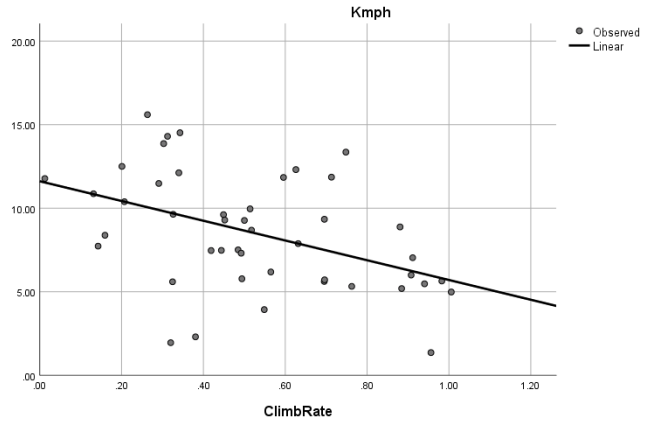


Figure 14 Curve estimation of climb rate and speed

to show up in the data needs an explanation. One reasonable hypothesis would be that there is a selection bias in the data stemming from riders being able to choose their mode of transport knowing the trip in advance, and that the slope of their planned trip might influence their choice. That is, when a rider knows their planned trip has an incline that would be significant to the speed of an e-scooter they may choose to ride the bus instead. Another hypothesis would be that the data is simply too averaged. While there are inclines that significantly impact speed as they are traversed, the impact is so small relative to the overall speed of the trip as to become insignificant when averaged out. If this is the case an analysis of data recorded throughout trips would be better suited for determining the true significance of gradients.

These hypotheses can to some extent be tested by reviewing the distributions of gradient and climb rate. Climb rate has a mean of 0.51 and is very close to a true normal distribution per figure 6. That would mean trips have just as much uphill as downhill travel. Ergo, there is no clear case to be made for riders avoiding uphills in general. To include in this the slope of these inclines, consider the distribution of gradient from figure 5. There is here a strong prevalence of lower gradients, the distribution being severely skewed to the

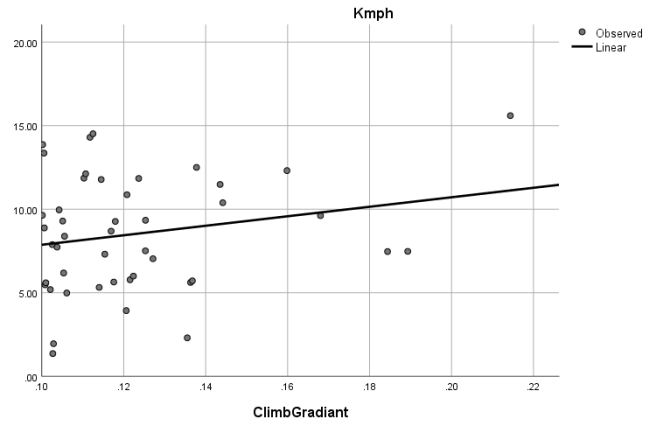


Figure 15 Curve estimation of climb gradient and speed with gradient outliers

left. This lines up decently well with both hypotheses, in either case one would expect to see few trips with steep gradients. However, there are some trips with steep gradients. Query whether there is correlation between speed and gradient when insignificant gradients are taken out of the dataset.

Limiting the dataset to only those trips that have a climb gradient above 10%, there appears to be a correlation between speed, climb rate and climb gradient as seen in figures 14 and 16. This can also be seen in table 4, containing the Pearson correlations of speed, climb rate, and climb gradient on that selection of data.

The negative correlation between speed and climb rate is unsurprising, but the positive relationship between speed and climb gradient is, again, puzzling. However, consider the initial scatter plot of climb gradient from figure 8. The scatter plot had two distinct formations of outliers, one set with speed outliers and one with gradient outliers and the scatter plot seemed to reveal a negative relationship between speed and gradient when

Correlations

		Kmph	ClimbRate	ClimbGradiant
Pearson Correlation	Kmph	1.000	-.443	.210
	ClimbRate	-.443	1.000	-.198
	ClimbGradiant	.210	-.198	1.000

Table 3 Correlation between speed, climb rate and climb gradient

considering the outliers in isolation. Using this information, data with gradients above 10 % or speeds above 20 km/h are selected. Doing another curve estimation in SPSS on this selection of data shows a very strong, and negative, relationship between speed and gradient as can be seen in table 3 and figure 15.

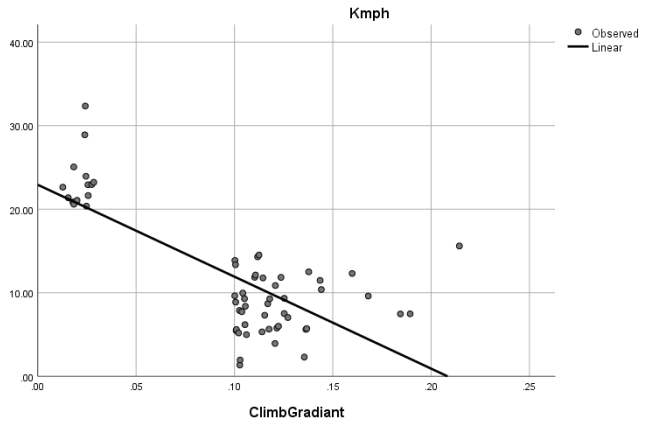


Figure 16 Curve estimation of climb gradient and speed with speed and gradient outliers

As a caveat, only 45 of the nearly 8000 recorded trips have a gradient higher than 10 % and only 59 have either a gradient over 10 % or an average speed over 20 kmph. The sample size is very limited, and while the correlation is strong, it depends to a large extent on the chosen selection of the data. While a decent argument can be made for selecting the data in this manner, the results must be viewed with some degree of scepticism as when dealing with this small a dataset it is very easy to force almost any result by selecting the ranges.

Validation of model presumptions

For the linear regression analysis to be valid a series of assumptions must be verified. Linearity of the relationships,

Correlations

		Kmph	ClimbRate	ClimbGradiant
Pearson Correlation	Kmph	1.000	-.268	-.739
	ClimbRate	-.268	1.000	-.016
	ClimbGradiant	-.739	-.016	1.000

Table 4 Correlations between speed, climb rate and climb gradient with speed and gradient outliers

normality of the variables, limited multicollinearity, and homoscedasticity (Alexopoulos, 2010).

The distributions from earlier, figures 2 through 6, show a decent degree of normalization across variables.

The scatter plots, figures 7 through 13, show either no apparent relationship or something similar to a linear relationship,

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Hour	ClimbRate	Temperature	BatterUse
1	1	1.958	1.000	.02	.02			
	2	.042	6.869	.98	.98			
2	1	2.848	1.000	.01	.01	.02		
	2	.115	4.979	.02	.23	.81		
	3	.037	8.809	.97	.77	.17		
3	1	3.608	1.000	.00	.01	.01	.02	
	2	.247	3.823	.00	.02	.10	.83	
	3	.112	5.675	.02	.28	.72	.04	
	4	.033	10.438	.97	.69	.17	.11	
4	1	4.195	1.000	.00	.00	.01	.01	.02
	2	.427	3.136	.00	.00	.00	.10	.86
	3	.234	4.231	.01	.04	.10	.74	.11
	4	.111	6.154	.02	.27	.72	.03	.02
	5	.033	11.262	.97	.69	.17	.11	.00

a. Dependent Variable: Kmph

Figure 17 Collinearity diagnostics

however faint. This is not a clear-cut case, but the analysis is at least partially validated on this test.

To test for multicollinearity, consider the condition indexes given by the regression analysis in SPSS shown in figure 17. The indexes are barely any higher than 11. This is well within acceptable limits implying that there are no significant issues with collinearity (Belsley, 1991).

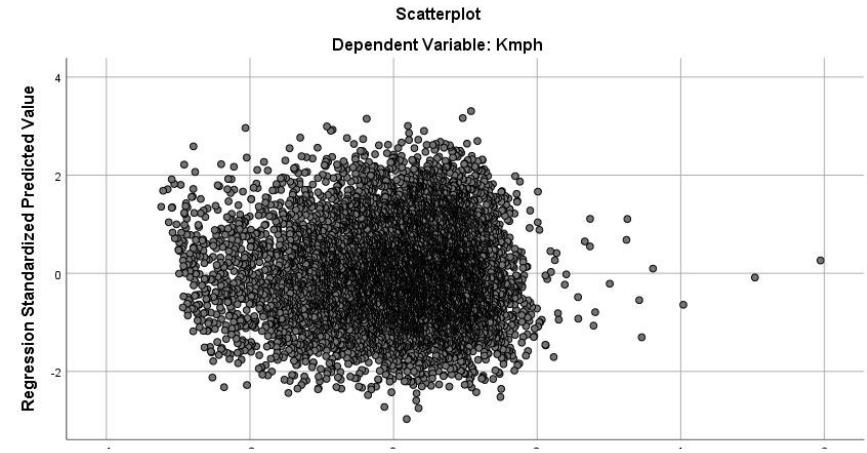


Figure 18 Homoscedasticity test

Finally, homoscedasticity can be tested for by plotting predicted values against residuals as seen in figure 18. The expectation is that the shape of the data will be close to a rectangular shape. This is largely the case.

Alternative use cases for the data

Data not being captured during e-scooter use lead to generation of trips and data at lower resolution than what would be best for a speed analysis. This does not, however, present as much of a challenge to an analysis of usage rates and their relationship to environmental variables.

As e-scooter riders are left exposed to the elements, it would be reasonable to assume there may be a correlation between the number of trips and weather conditions. Using the gather downpour and temperature data, a correlation analysis and a univariate general model is run in SPSS.

Trips and downpour

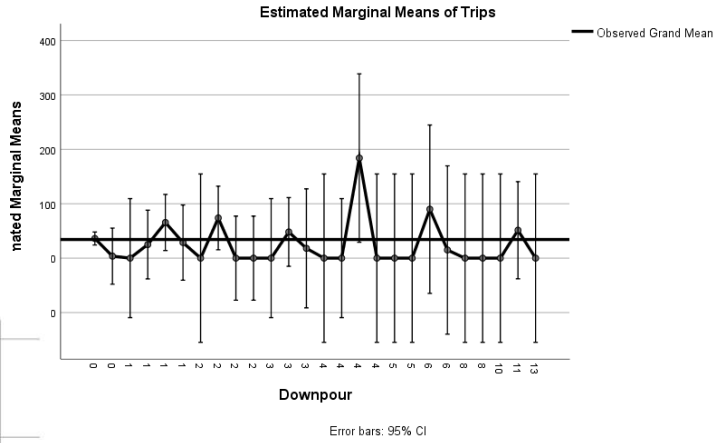


Figure 19 Bar plot of trips and downpour

Figure 19 shows how trip numbers vary with downpour. The most noteworthy piece of information in the plot is the fact that trip numbers have very low variance at no downpour and very high variance elsewhere. There

are not enough trips in downpour of any level to say anything with certainty about its relationship with trip numbers based on the bar plot. However, observe that most of the non-zero datapoints are below the mean. This implies a negative relationship, though not with much precision. A larger dataset, with more weather variance, would greatly benefit this part of the analysis.

Trips and temperature

As seen in figure 20, there is a pretty clear-cut case for temperature having a positive relationship with trip numbers. There are quite a few datapoints at the no trips line but considering figure 3 this is entirely unsurprising. There are times of the day when hardly anyone rides e-scooters, naturally this leads to some datapoints landing on the x-axis regardless of temperature.

Correlation

Given figures 19 and 20, there is reason to expect a positive and a negative relationship between temperature and downpour respectively, and the number of trips. Table 5 shows that there is in fact a

Correlations

		Trips	Temperature	Downpour
Trips	Pearson Correlation	1	.389**	-.106
	Sig. (2-tailed)		.000	.354
	N	78	78	78
Temperature	Pearson Correlation	.389**	1	.296**
	Sig. (2-tailed)	.000		.008
	N	78	78	78
Downpour	Pearson Correlation	-.106	.296**	1
	Sig. (2-tailed)	.354	.008	
	N	78	78	78

**. Correlation is significant at the 0.01 level (2-tailed).

Table 5 Correlation between number of trips and weather

strong positive correlation (Schober, Boer and Schwarte, 2018) between temperature and trips, while the relationship to downpour is far less certain, though negative if it is anything. This lines up well

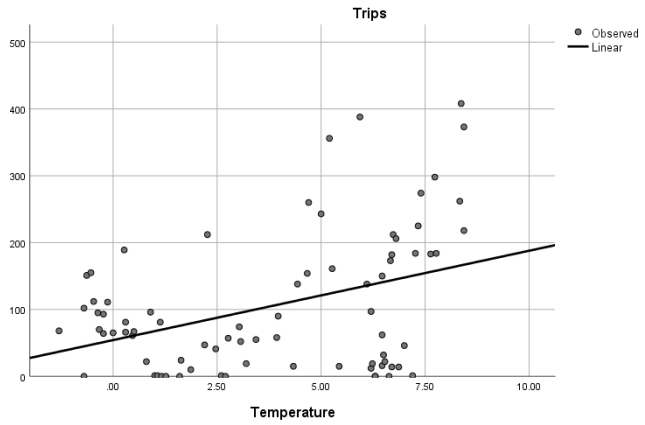


Figure 20 Curve estimation for temperature and trip numbers

with what seems intuitive, that bad weather leads to less usage of e-scooters.

General descriptions of e-scooter usage patterns

Finally, the data makes it possible to analyse the nature of e-scooter travel patterns. Origins and destinations are known and can be separated and mapped out. This has utility for city planners who need to understand where e-scooters are likely to take up space.

By plotting other variables than speed against each other it is possible to analyse how consumers use e-scooters. Mapping time of day against travel time for instance, reveals that trips between 10 and 11 are significantly longer than the rest of the day as seen in figure 21.

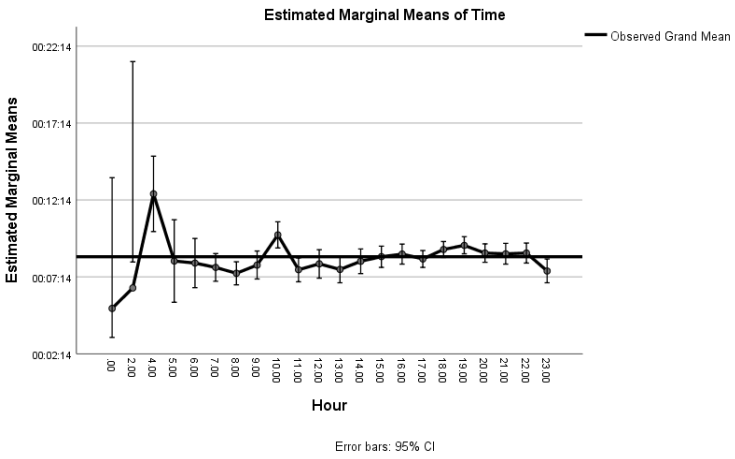


Figure 21 Bar plot of travel time and time of day

Conclusions

By using this analytical approach to route planning data, some correlation was discovered between speed and other variables. However, the correlations were weak and there were likely many relevant variables missing. It is the case then, that this particular attempt at using big data as a means for investigating environmental causes of speed choice on e-scooters, has failed. The low resolution of the data made it impossible to apply these, likely significant, features of the infrastructure environment, and the results suffered for it. This does not mean that the whole methodology must fall though. There are examples of similar studies, most of which had access to data during e-scooter use, that found far more significant results (Zou et al, 2020). The lesson then is to ensure observations can be made during trips if one intends to use the data for speed analysis or any other type of study that required high resolution data and strictly localized variables. This would likely require the cooperation of e-scooter companies.

There are some weaknesses inherent to the methodology as well. The methodology suffers from the lack of data about the riders. Gender, age and weight among other things, likely impact speed. These are unknowable factors when not making observations in the field.

There does, however, appear to be useful application of the data for the purposes of researching usage rates of e-scooters and their relationship with environmental variables. Here, the data is far more applicable as the exact path of the trips is largely irrelevant to the investigatory purpose. Data about how long trips on e-scooters typically are, where they tend to begin and end, at what times they happen, and in which weather are all interesting pieces of information useful to city planners. The data gathered for this analysis had little seasonal variance which somewhat limits the ability to accurate predict usage rates of e-scooters. Further research using a larger set of data collected in various seasons would be beneficial.

References

Adressa (2019) Kommunen krever at de nye el-sparkesyklene fjernes, 17.06.2019, available at: https://www.adressa.no/pluss/nyheter/2019/06/17/Kommunen-krever-at-de-nye-el-sparkesyklene-blir-fjernet-19274470.ece (29.03.2021)

Alexopoulos, E.C. (2010) Introduction to Multivariate Regression Analysis, *Hippokratia* 14: 23-28

Belsley, D. (1991). A Guide to Using Collinearity Diagnostics, *Computer Science in Economics and Management,* 4: 33-50

EnTur (2020) Find all Scooters near a coordinate. Available from: https://developer.entur.org/pages-mobility-docs-scooters (29.03.2021)

Norsk Klimaservicesenter (2020). Seklima, observasjoner og værstatistikk. Available from: https://seklima.met.no/observations/ (29.03.2021)

Schober, P., Boer, C. and Schwarte, L. (2018) Correlation Coefficients: Appropriate Use and Interpretation, *Anesthesia & Analgesia* 126: 1763-1768.
DOI: 10.1213/ANE.0000000000002864

Samferdselsdepartementet (2018) Forskrift om krav til sykkel. Available from: https://lovdata.no/dokument/SF/forskrift/1990-02-19-119 (29.03.2021)

Statens vegvesen (2021) Vegkart. Available from: https://vegkart.atlas.vegvesen.no/#kartlag:geodata/@268102,7030868,9/hvor:~(kommune~(~5001)) (29.03.2021)

Zou, Z., Younes, H., Erdogan, S. and Wu, J. (2020) Exploratory Analysis of Real-Time E-Scooter Trip Data in Washington, D.C. *Transportation Research Record*, 2674(8) 285-299.
DOI: 10.1177/0361198120919760

Appendices

Appendix 1 – Python, data collection

```
import requests
import datetime
import time

filename = 'dataset ' + str(datetime.datetime.now().strftime("%Y-%m-%d %H-%M-%S")) + '.csv'

while True:
    file = open(filename, 'a')

    resp = requests.get('https://api.entur.io/mobility/v1/scooters',
                        params={'lat': 63.430, 'lon': 10.394, 'range': 5000, 'max': 2000},
                        headers={'ET-Client-Name': 'ntnu - masteroppgave_jonas_lund'})

    if resp.status_code != 200:
        print('WARNING: unhandled http status code ' + str(resp.status_code))
        print(resp.text)
    else:
        curtime = str(datetime.datetime.now().strftime("%Y-%m-%d %H:%M:%S"))
        print(curtime + ' results: ' + str(len(resp.json())))

        for params in resp.json():
            out = [
                curtime,
                params['id'],
                params['operator'],
                str(params['lat']),
                str(params['lon'])
                ]
            if 'battery' in params:
                out.append(str(params['battery']))
            elif 'batteryLevel' in params:
                out.append(params['batteryLevel'])
            else:
                out.append('UNKNOWN')
                print(params)
            #print(','.join(out))
            file.write(','.join(out) + '\n')

        file.close()

    time.sleep(13)
```

Appendix 2 – Python, data cleaning

```
import pandas as pd
import datetime as dt

#Finds the difference in time between an observation
#and the next observation of the same scooter
def checkTime(df, i, j, l,fmt):
    if j >= l or df.iloc[i, 1] != df.iloc[j, 1]:
        return dt.timedelta(seconds=0)
    return abs(dt.datetime.strptime(df.iloc[i,0],fmt) - dt.datetime.strptime(df.iloc[j,0],fmt))

#Loads CSV-file to 'df'
df = pd.read_csv(r'E:\Master\APIProsjekt\dataset 2020-11-26 16-50-48.csv')
#Sort the CSV according to scooter ID and time of observation
df.sort_values(by=['ID', 'Time'], inplace=True) # sort by ID and then by timestamp
l = len(df.index)
print(l)
k = 0
i = 0
fmt = '%Y-%m-%d %H:%M:%S'
#Set the first row, 7th column to -1 in order for Excel to
#recognize the column as integer later on
df.iloc [i,6] = -1

while i in range(l):
    d = checkTime(df,i,i+1,l,fmt)
#If the difference in time between the current row and the next is between
#2 and 90 minutes, tag current row and next row as being part of trip number k
    if d.total_seconds() >= 120 and d.total_seconds() <= 5400:
        df.iloc [i,6] = k
        df.iloc [i+1,6] = k
        k = k+1
        i = i+2
    else:
#        df.iloc [i,6] = -1
        i = i+1
#Print progress every 100'000 rows
    if i % 100000 == 0:
        print(i)
#Write results to new CSV-file
df.to_csv(r'E:\Master\APIProsjekt\CSVer\TimeKeep\TimeKeep2 2020-11-26 16-50-48.csv', index=False)
print('DONE!')
```

Appendix 3 - Progress Report

Adapting to imperfect data was the primary challenge in working on this paper. A misconception about the manner in which data was gathered lead to a lot of time wasted and created a situation in which many of the most likely causes for speed choice, which this paper attempts to investigate, could not be implemented in the analysis.

First, I had to learn to program in Python. I chose Python because it is implemented in QGIS and ArcMap, the two GIS-programs I was considering using, and because it has a library for handling CSV-files, pandas, which was supposed to be relatively easy to learn and efficient.

The initial plan was to use a dataset containing observations about the GPS-locations of e-scooters at 15 second intervals obtained from Entur as a starting point for an analysis using a QGIS module built by Anita Graser (https://anitagraser.com/movement-data-in-gis/). The module takes observations along a set of routes as inputs and creates illustrations of aggregated travel, revealing travel patterns. However, the module has limited supporting information making implementation in QGIS somewhat difficult to accomplish. Additionally, it later turned out that the structure of the Entur data was not compatible with this module and so the idea would not have been workable either way.

Then I moved on to the idea of doing a multivariate analysis of speed choice on e-scooters. Still believing the data to be observations at 15 second intervals I planned to estimate speed and gradient at all points using a simple forward/backward script. The script calculated the difference in location between each datapoint, and the previous and following datapoint. With the approximated distance in hand and the time between observations and elevation known, calculating gradient and speed was trivial. While this script worked as intended, it was incredibly slow, taking days to compute each dataset, and also, when looking over the results, I found that e-scooters do not report their updated location to Entur while in use. This meant that the results from the script would not be useable. The simple straight-line calculation may have been a decent estimate when the difference in time when it was thought to be only 15 seconds, but when a vehicle is not observed for on average 20 minutes there is a need for a more thorough approach. This problem would likely have been discovered much earlier if not for the fact that the sheer size of the datasets prohibited use of Excel in any meaningful way. Excel could not hold more roughly 3 hours of data and attempts at adding formulas to all rows to quickly review the data lead to insufficient memory crashes.

Having discovered that the data available through Entur had a much lower resolution than anticipated, I had to reconsider my methodology. While the lack of data during use is irrelevant to the intended use of the data, route planning, it presents quite the challenge for any analysis of trips. The available data is essentially a negative of scooter movements where the useful data is identified by a distinct lack of data in a measure of time where, at the end of that time, the position of the scooter has changed significantly. To resolve this issue, I decided to map the trip onto a network model by using network analysis in GIS software. I first considered QGIS but could find no way for the network analysis tool to handle a series of orgins and destinations. I then moved on to ArcMap, using the Model Builder to create a script that could run through the CSV-files and map paths from origins to destinations. However, I could not find a way of solving it without using at least two iterators and Model Builder only allows for one. Finally, I found that the easiest solution would be to first number the pairs of origins and destinations using a Python script and then work on the Model Builder from there. Now, as it turns out, with

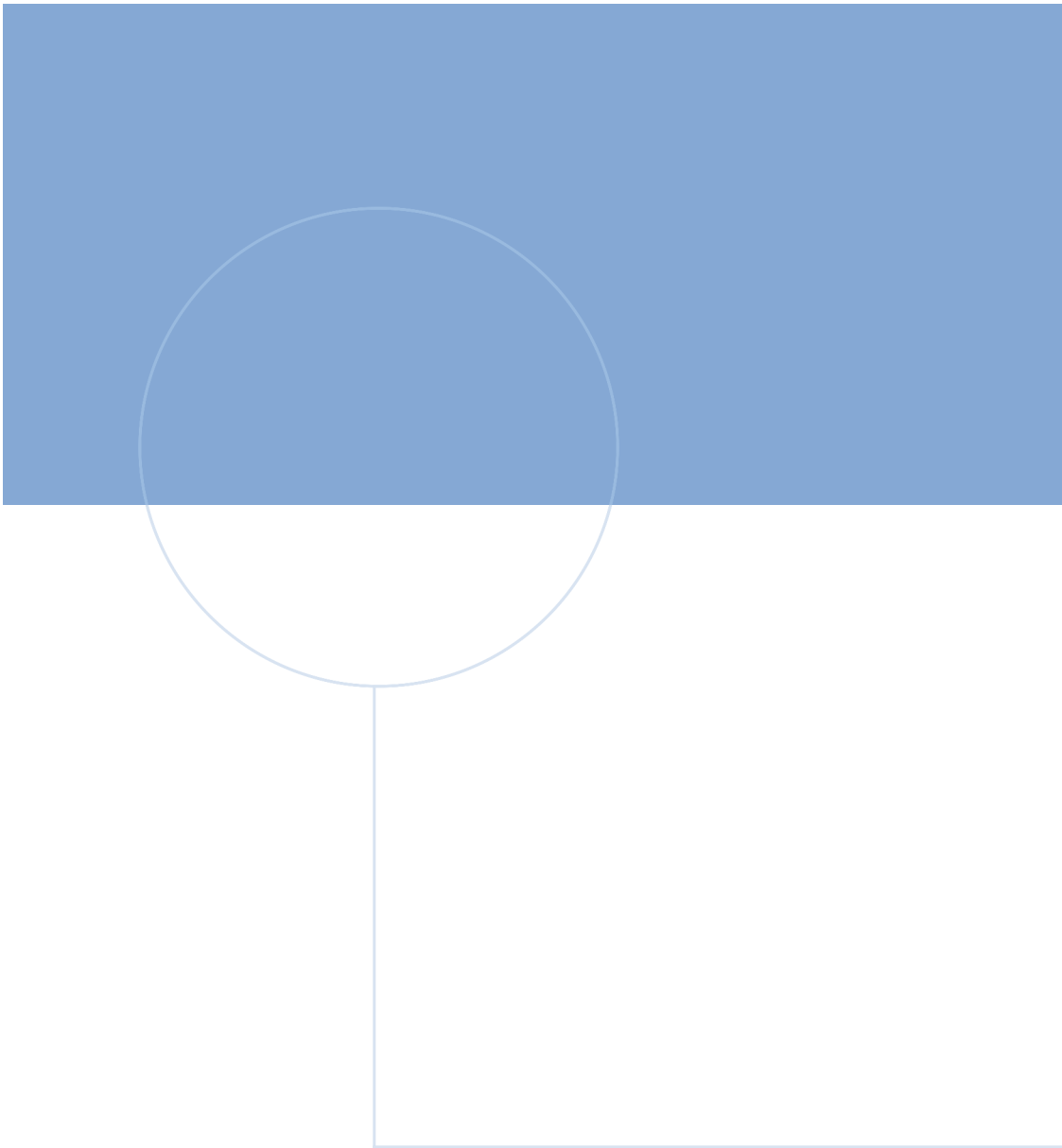
numbered pairs of points the standard network analysis tool in ArcMap could handle the task.

Seeing that the distance script from earlier was inefficient and armed with the knowledge that the data I was interested in was marked by a significant gap in time between observations, I decided to base the new calculation on time differences. I wrote a Python script that computed the time difference between observations of each e-scooter and then also tagged each pair of points with the same ID number. These points become our origin and destination points. Second, I had to build a network model of the available paths, which I did by using elevation data for Trondheim I downloaded for use in a previous GIS course and used road network data from Vegkart. To get a measure of speed, I had to find the distance travelled. There is no way to know exactly what path each traveller took, but assuming the travellers, incentivised by the cost of using an e-scooter, chose the shortest path from their origin to their destination, a network analysis of the paired points to estimate the distance of each trip would give a decent estimate. This distance, together with the known times when the scooters were at their origins and destinations gives an acceptable estimate of average speed.

At this point it occurred to me that this estimation limited the amount of data that could be reliably mapped onto the observations for use in the statistical analysis. While distance travelled would not be significantly different if the rider chose to, for instance, travel on a parallel road, this would not be true for all relevant datapoints. AADT, speed limit, whether there was a sidewalk or not, and other environmental factors specific to a street, cannot be reliably applied to data of this kind. I therefore was forced to limit the analysis to data that could be reasonably said to not be tainted by the method of route estimation.

Distance, as I have already discussed, is likely estimated fairly well by the GIS software. And with time between observations known, speed can be calculated. The network analysis run in GIS accumulated horizontal and vertical climb as well as total distance, and while these data points are not necessarily the exact same as the real data for the trip, the difference in height between origin and destination is known and it is unlikely that this would be off by much. In retrospect, it seems obvious that horizontal and vertical decline ought to have been accumulated as well. To get some more factors to work with, I also downloaded weather data, temperature and downpour specifically, in part to investigate any correlation with travel speed, which I deemed unlikely, but also to look for correlation between weather and the number of trips.

Having done the analysis it seems to me that this approach is generally good and useful for the purpose of speed analysis, but requires data gathering during e-scooter rides for the results to be reliable and interesting. However, the data as it is can be useful for lots of other purposes, such as investigating general travel patterns in a city more broadly, looking at when e-scooters are used, or in what manner or weather.



NTNU

Kunnskap for en bedre verden