Case Studies on Transport Policy 21 (2025) 101508

Contents lists available at ScienceDirect

Case Studies on Transport Policy

journal homepage: www.elsevier.com/locate/cstp

‘Mind the Gap’—The impact of discrepancies between Google Maps API and reported travel data in the Global South

Faza Fawzan Bastarianto [a,b,*], Thomas O. Hancock [a], Anugrah Ilahi [c], Ed Manley [d], Charisma Farheen Choudhury [a]

[a] Choice Modelling Centre, Institute for Transport Studies, University of Leeds, Leeds, United Kingdom
[b] Department of Civil and Environmental Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia
[c] Emirates Center for Mobility Research, United Arab Emirates University, Abu Dhabi, United Arab Emirates
[d] School of Geography, University of Leeds, Leeds, United Kingdom

ARTICLE INFO

Keywords:
Google Maps API
Revealed preference
Mode choice
VTTS
Travel survey

ABSTRACT

Over the past decade, online navigation services have been adopted increasingly as a source of ‘ground truth’ in estimating choice alternatives during travel behaviour. These services, including Google Maps, Bing Map, and Waze, which are designed to provide real time traffic information and navigation guidance to the users, are believed to offer comprehensive and precise information regarding travel attributes. Nevertheless, discrepancies between the travel attributes collected from those services and the travel data that is reported by the travellers may introduce a systematic bias into travel behaviour analysis and modelling. This paper attempts to explore this challenge by investigating the discrepancy between the reported travel times and costs and the corresponding values derived from the Google Maps API. The comparison is conducted in the context of a developing country, through the use of travel diary survey data from Greater Jakarta, where there is a greater variety of transport modes and individuals may have varying capacities to gauge travel attributes due to the unpredictability of traffic conditions. Results show that even minor adjustments to which observations are included and which specific attribute treatments are used can completely change values of travel time savings (VTTS) estimates. Further, the characteristics of the observations excluded in the process of pre-processing are investigated to provide insight into preventing loss of data in future mobility surveys. Recommendations to address both of these issues are discussed along with policy implications.

1. Introduction

Over the last decade, the use of online navigation planning tools within mobility and geographic analysis has grown considerably (Auld et al., 2016; Qi et al., 2019; Wagner et al., 2020). These functions, provided by companies including Google, Bing, and Esri, provide users with travel times, costs, and distances for given origins, destinations, and departure times. Most services operate without providing public information on the specific nature of their data sources and algorithm, however, broadly these services produce a set of near-optimal or cost-minimal paths under the constraints given. The accessibility of these services, available through APIs (Application Programming Interfaces), mean they are a popular option for rapidly producing cost estimates in a range of contexts.

Models of travel behaviour are an important instance where we require knowledge of attributes of the alternatives to predict their mobility decisions. These models are typically built using data obtained through revealed preference (RP) and/or stated preference (SP) surveys. SP surveys focus on the stated behaviour of individuals in a hypothetical setting to determine potential sensitivity to each attribute in given scenarios. In contrast, RP surveys focus on the choices that individuals have made in the real world. The strength of this kind of survey is that it reveals the actual decisions made by individuals in a real world context and considers situational and individual constraints (dell’Olio et al., 2018).

Whilst RP data reduces hypothetical bias, a key challenge in using such data for travel behaviour modelling is inferring the attributes of the unchosen alternatives. In the past, studies using RP data addressed this issue by imputing mean attribute values for each unchosen alternative (Train, 1986; Train and Winston, 2007). Later, researchers relied on a variety of methods to collect data on the travel attributes of unchosen alternatives for RP surveys. In mode choice analysis, for

* Corresponding author at: Choice Modelling Centre, Institute for Transport Studies, University of Leeds, Leeds, United Kingdom.
E-mail address: faza.fawzan@ugm.ac.id (F.F. Bastarianto).

https://doi.org/10.1016/j.cstp.2025.101508
Received 29 October 2024; Received in revised form 29 March 2025; Accepted 29 May 2025
Available online 13 June 2025

instance, the travel mode attributes, including travel time and travel cost for unchosen alternatives, are often obtained using theoretical and empirical approaches involving a number of assumptions for estimating the travel time and cost of each transport mode (Enam and Choudhury, 2011; Fox et al., 2014; Bastarianto et al., 2019).

Over the last decade, researchers have used navigation map providers services to complete travel attributes in datasets of RP surveys. Malokin et al. (2019) utilised Google Maps Application Programming Interface (API) to estimate travel time and travel cost according to the fastest routes as suggested by Google Maps, which requires a number of assumptions for each mode of transport. Auld et al. (2016) applied Google Maps API to analyse the behaviour of travellers in Illinois. Fu et al. (2023) employed various navigation map providers, including Google Maps API, Bing Maps API, Esri Routing Web Service, and OpenStreetMap NetworkX to generate time estimation of 10,000 origin–destination pairs in the USA.

It has been claimed that publicly available traffic information from navigation map providers such as Google Maps and Bing Maps will provide comprehensive and more accurate details for all traffic participants, thus enabling data-driven decisions (Wagner et al., 2020). However, to date there has been a lack of focus on the validation of travel attributes obtained via the Google Maps API. Discrepancies between travel attributes generated from Google Maps API and those reported by travellers may introduce systematic biases in travel behaviour analysis and modelling. For instance, systematic overestimation or underestimation of certain travel attributes for public transport leads to biases that can have significant implications for policy decisions and infrastructure planning. Policy measures such as congestion pricing and public transport subsidies rely on accurate travel behaviour insights (Guzman and Oviedo, 2018; Wu et al., 2017). If travel attributes are inaccurately represented, policies may be ineffective or even counterproductive. Moreover, transport infrastructure planning depends on reliable estimates of travel demand, time, and cost (Zhang and Cheng, 2023). Biases in these estimates could lead to inefficient allocation of resources for transport infrastructure investments, such as road expansions or investment in public transport networks.

In particular, there have been few comparisons of travel attributes of known trips obtained from Google Maps API compared to other sources of travel data, including GPS-tracked and recorded travel data. Wu (2018) found that travel times from Google Maps API are systematically higher than those from Uber Movement, which is based on GPS timestamps. Additionally, Wagner et al. (2020) validated Google Maps API travel time estimates by comparing them to GPS-tracked data from real test drives, finding that the deviation was less than 6%, indicating a reasonably high level of accuracy between the two. Similarly, one study in London found disparities showing systematic variations based on various modes and purposes of travel Hillel et al. (2016). The systematic characteristics of the discrepancies indicate that modelled trip duration from navigation map providers still require significant adjustment in order to closely represent observed travel patterns from the London Transport Demand Survey (LTDS).

The context of developing countries raises further challenges, such as unreliable public transport services (Muñoz et al., 2020), which can affect the accuracy of API-generated travel times. Moreover, the wider range of transport mode options (Ilahi et al., 2021), including both formal and informal transport modes, may not be accurately captured by Google Maps API. Additionally, traffic conditions are often more uncertain (Martinez and Masron, 2020), which might lead to greater discrepancies between Google Maps API and reported travel data. Mobile device penetration is also often varied, which reduces the likelihood of route navigation guidance (e.g. Google Maps) being available to travellers. The high variability of travel time may make it difficult for people to predict traffic conditions. All these factors make the applicability of the Google Maps API in the context of developing countries questionable (Zannat et al., 2021) and warrant the need for a detailed investigation.

In addition, when analysing discrepancies between Google Maps API and reported travel data, it is also essential to consider the potential sources of error generated in the Google Maps API itself (Google Developers, 2023c). These include the prediction of travel time in traffic based on historical data, which may not always capture current traffic conditions accurately. Specifically, Google Maps assumes average congestion levels based on historical data, which might not reflect sudden changes in traffic flow, particularly in areas with high variability such as Jakarta. Jakarta experiences severe traffic congestion that varies depending on the time of day, along with a wide range of vehicles and unpredictable weather conditions that can cause delays and change travel times unexpectedly. Moreover, the coverage of transport modes in the Google Map API is primarily limited to certain formal transport options and may not adequately consider for informal transport modes, such as bikes, motorcycles and minibuses, which are prevalent in the context of developing countries. This lack of representation of informal transport further contributes to discrepancies in travel time and cost estimates.

Despite the growing use of travel attributes retrieved from Google Maps API, existing research has primarily focused on developed countries (Hillel et al., 2016; Malokin et al., 2019). However, in the context of developing countries, several critical gaps remain unaddressed. First, there is a lack of studies systematically comparing API-derived mode-specific attributes (such as travel time and cost) with recorded values from the RP dataset. Prior research has additionally primarily focused on the reliability of API-based travel times, often neglecting travel costs. Given the importance of both travel attributes in understanding travel behaviour and formulating transport policies, a study that investigates both attributes is needed. Second, researchers typically remove irrelevant observations from the dataset. It is unclear, however, how a researcher should best implement a procedure for excluding observations where there are large discrepancies, nor is there an understanding of the characteristics of such excluded data, which could help prevent data loss in future surveys. Third, it remains unclear how these discrepancies influence key transport modelling outputs, such as the Value of Travel Time Savings (VTTS), which plays a crucial role in economic appraisal and policy making. This paper aims to address the research gap by investigating the following research questions:

1. How does the discrepancy between mode-specific attributes (travel time and cost in particular) from the RP dataset and Google Maps API data vary in the context of the Greater Jakarta?
2. Should observations with large discrepancies between the RP and Google Maps API travel data be excluded, and what are the key characteristics of the excluded observations that indicate significant discrepancies?
3. What is the relative impact of pre-processing travel data and attribute treatment on model outputs, specifically Value of Travel Time Savings?

The remainder of this paper is organised as follows. Section 2 introduces the study scope and data used in this study. Section 3 outlines the methods for retrieving travel attributes provided by navigation map providers, pre-processing the data, and validation of the results. Section 4 presents the results and discussion. Lastly, Section 5 concludes the study and presents suggestions for future work.

2. Case study and data

2.1. The city and transportation system

The Greater Jakarta area is known as a megacity, having an extensive population of over 30 million inhabitants. This population size places it as the second largest metropolitan area worldwide after Tokyo (Martinez and Masron, 2020). Commute-related travel accounts for the largest proportion of daily trips in this megacity. According to the Greater Jakarta Commuter Survey in 2019 (Indonesia, 2019),

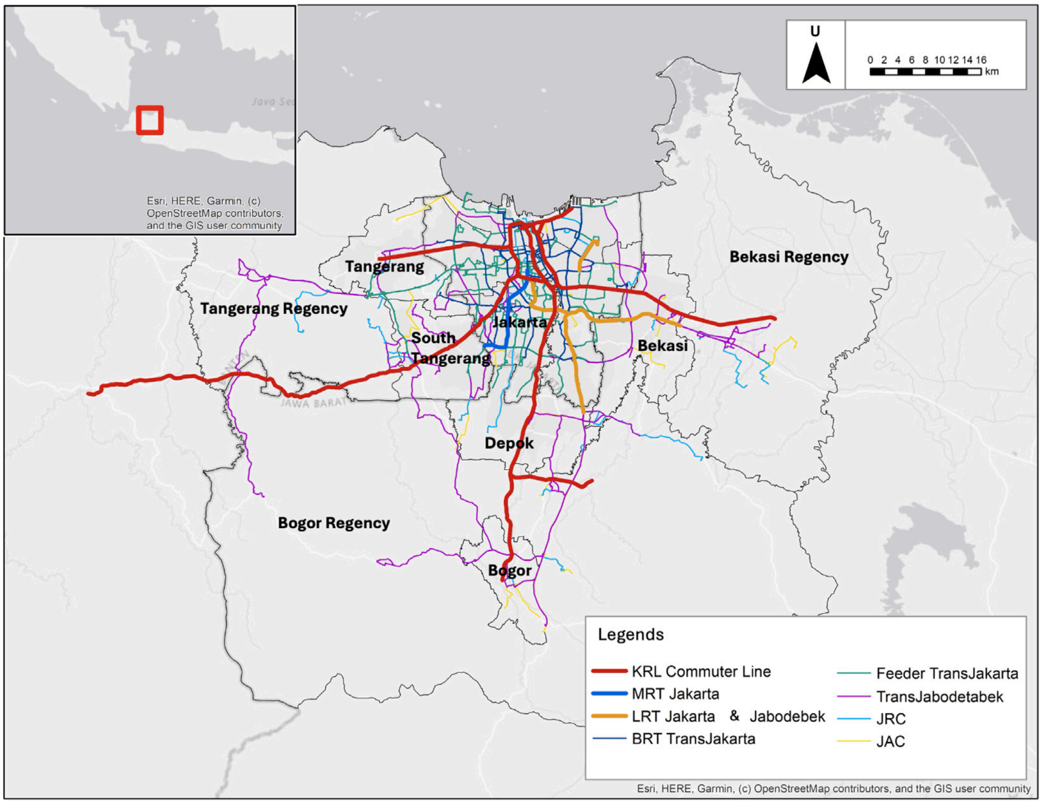


Fig. 1. Map of the mass transit network in Greater Jakarta.
Source: BPTJ (2023).

approximately 11% of the total population, or 3.2 million individuals, commute to Jakarta from outer districts for a variety of work-related activities in the capital. Fig. 1 shows the mass transit network in Greater Jakarta, including rail-based public transport (MRT, LRT, Commuter rail, etc.) and road-based transport (BRT Transjakarta and non-BRT).

The Jakarta Transport Agency has continually improved the ridership of public transport in the region by expanding the catchment services area of the TransJakarta Bus Rapid Transit (BRT) system, serving satellite areas and incorporating a network of busways. In addition, this region is served by new forms of public transport, including Mass Rapid Transit (MRT) and Light Rapid Transit (LRT), which connect the suburbs to the capital.

To better understand the demand for different travel modes in Jakarta, stated preference (Belgiawan et al., 2019) and revealed preference (Benita, 2023) data can be used to estimate preferences. However, the travel behaviour in Greater Jakarta differs substantially from developed countries in terms of socioeconomic structure (e.g., income, household size, family structure, age, gender roles), culture of work (e.g., working hours), advancements in technology (e.g., reliable internet access), and transportation landscape (e.g., public transport services, vehicle ownership, para-transit). These make Greater Jakarta an ideal location for investigating the discrepancies between the travel attributes obtained from surveys and the Google Maps API.

2.2. Mobility jakarta data

This study utilised a travel diary survey data namely Mobility Jakarta, described in detail in Ilahi et al. (2020). This data covers socio-demographic, stated preference (SP) and revealed preference (RP) data collected in Greater Jakarta. The survey on Mobility Jakarta was conducted between April and May of 2019. This study utilised the RP dataset of Mobility Jakarta data, primarily focusing on work-related travel purposes. The dataset consists of 2617 individuals and a total of 8770 observations of trips.

3. Methodology

3.1. Research design

This study employed a quantitative research approach, which allows for a systematic analysis of the discrepancies between the Google Maps API and reported travel data. The Google Maps Distance Matrix API service was used to extract travel data related to each origin–destination pair from Google Maps during different times of the day. The collected information included distance and travel time by different modes. Since in the context of Jakarta, travel data from Google Maps is limited only to cars, public transport, and walking, deriving travel times for the other modes (on-demand motorcycle, taxi motorcycle, bike, and motorcycle) required further data augmentation. The augmented data was subsequently cleaned to remove any trips that did not meet the study criteria, focusing only on first-leg work-related trips to the office and trips where both the origin and destination were within the Greater Jakarta area.

An analysis of the discrepancies between the Google Maps API and reported travel data was performed by examining the scatter plot of these datasets and the interquartile range (IQR) diagram of the discrepancies, which subsequently serve as the starting point for a threshold-based cutoff when filtering out observations for which the discrepancies are too large to be considered reliable.

This study utilised a multiple regression model to enable the identification of the variables that influence the exclusion as a result of the discrepancy between the two travel data. We then apply multinomial logit models to different combinations of datasets (with different data exclusion criteria) and different attribute treatments (use of reported or Google Maps API travel times/costs) to test the relative impact of these criteria and treatments on model outputs, specifically the estimated value of travel time savings (VTTS). Furthermore, a flowchart showing the methodological steps above can be seen in Fig. 2 below.

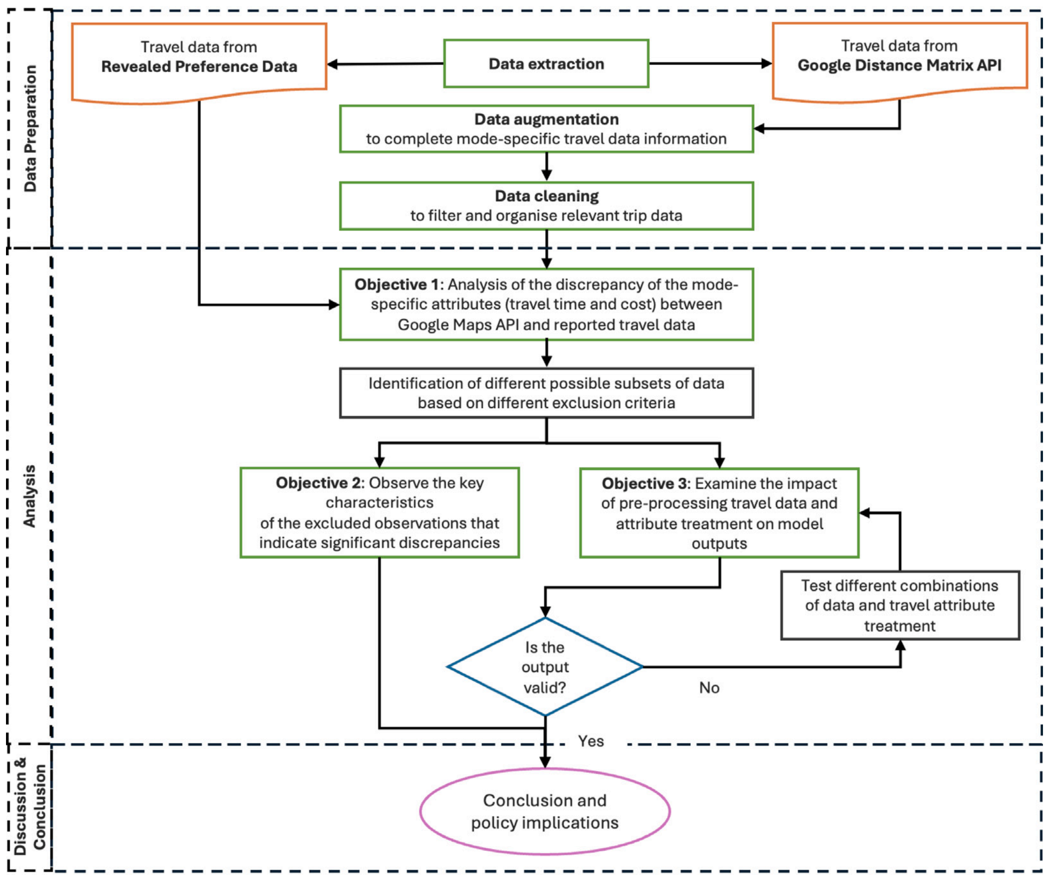


Fig. 2. Research method flowchart.

3.2. Google maps API

Google Maps API is a service that Google Inc. made available to the public in 2005,[1] allowing users to use the technology in a variety of applications. In this study, an effort is made to specifically use Google Maps Distance Matrix API (Google Developers, 2023c) to gather information of travel distance and time for a matrix of the full set of origins and destinations observed in the cleaned Mobility Jakarta RP dataset. This enables a comprehensive comparison of trip characteristics generated by the Google API and those reported by the respondent.

The Google Maps Distance Matrix API offers identical features to the Google Maps graphical end-user interface in a web browser, providing the distance and time required for a specific transport mode from a predefined origin to a predefined destination. The API allows the automated querying of multiple routing or travel requests, thus enabling the acquisition of a large number of travel distance and time data for a particular set of origin and destination matrices. In particular, the Distance Matrix API, unlike the Google Maps Directions API (Google Developers, 2023a), allows for the simultaneous structure of multiple origins and destinations and returns results for all possible combinations of travel time and distance.

The input for origin and destination matrix for this study has been converted from address format to latitude and longitude coordinates format using Geocoding API (Google Developers, 2023b). In addition, the optimistic traffic model was retrieved in this study. The procedures for utilising the aforementioned and other parameters can be discovered in Google Developers (2023c).

[1] https://cloud.google.com/blog/products/maps-platform/whats-next-google-maps-platform

3.3. Data augmentation

There are twelve modes of transport recorded from Mobility Jakarta Data: car, motorcycle, on demand transport (ODT) car, ODT motorcycle, car taxi, motorcycle taxi, minibus, conventional bus, bus rapid transit (BRT), commuter rail, bicycle, and walk. However, Google Distance Matrix API cannot calculate the travel duration and distance for some modes. The API, for instance, does not provide travel time and distance information for motorcycles and bicycles in some countries, including Indonesia. Only the travel time and distance for the chosen car-based, bus-based, rail, and walking modes can be retrieved via Google Distance Matrix API requests.

Thus, data augmentation was implemented to complete the travel time and travel cost for other modes for which Google Maps API services are not available. This meant that some assumptions were instead needed to calculate those travel time and cost. Table 1 below displays the functions used to determine the travel time and cost for different modes in this study.

The assumptions included, for instance, that the APICarDistanceNoTolls value above is the car's travel distance when the route avoided tolls. This value was used as an approach to determine the travel time for modes that cannot make use of tolls, such as bicycle, taxi motorcycle, and ODT motorcycle. The travel time is then calculated by dividing APICarDistanceNoTolls by the speed that corresponds to each mode. In addition, another assumption is that a motorbike travels three kilometres per hour faster than a car (Walton and Buchanan, 2012). The speed of the bicycle was assumed to be 15.3 kilometres per hour (Schleinitz et al., 2017). The calculation for BRT and Train were based on regulation established by Jakarta Province (2016) and Ministry of Transportation (2016), respectively. Assumptions to calculate the remaining modes were based on then-current travel costs in Greater Jakarta (Ilahi et al., 2021).

Table 1
Assumptions to calculate travel attributes of the unchosen mode.

Mode	Travel time (min)	Travel cost (x1000 IDR.km)	Availability
Walk	API Walking	–	<500 m
Bike	$\frac{APICarDistanceNoTolls}{BikeSpeed}$	–	<1400 m
Car	API Driving	2.95 per km	Has car
Motorcycle (mc)	$\frac{APICarDistanceNoTolls}{APICarSpeed+3km/h}$	0.59 per km	Has mc
Bus	API Transit: Bus	10 per 10 km	>250 m
BRT	API Transit: Bus	3.5	>250 m
Train	API Transit: Rail	3(25km) + 1 per 10 km	>1400 m
Microbus	API Driving	5 per 10 km	Always
Taxi car	API Driving	6(base) + 4.5 per km	Always
ODT car	API Driving	10(base) + 10 per km	Always
Taxi mc	$\frac{APICarDistanceNoTolls}{APICarSpeed+3km/h}$	10(base) + 3 per km	Always
ODT mc	$\frac{APICarDistanceNoTolls}{APICarSpeed+3km/h}$	10(4 km) + 2.5 per km	Always

3.4. Data cleaning

Since this study focuses on commuting behaviour, data preprocessing and cleaning were performed to obtain the relevant datasets. This includes filtering the data for first-leg trips involving work-related trips to the office which accounts for 8770 observations. The first-leg trips were then classified to single trip and multimodal trip (i.e. access trip, primary trip, and egress trip) in order to get more comprehensive understanding of the correlation between reported and Google Maps API data in trip-level. Furthermore, irrelevant trips were removed, such as trips made by respondents whose origin and destination resided outside the case study area. Subsequently, 8169 observations remained, forming the base dataset for this study.

Given that large discrepancies between Google Map API and reported travel data may exist (e.g. if there were human errors in reporting the travel time), data cleaning is required to remove outliers for which these discrepancies are implausibly large. A possible approach is to use the interquartile range (IQR) to identify trips for which the difference between Google Maps API and reported travel data is too large (Dash et al., 2023), with observations beyond a certain threshold omitted in further stages of analysis. An IQR threshold-based cutoff was employed in order to obtain different dataset with a higher correlation of travel time and travel cost between Google Maps API and respondent-reported data. Upper- and lower-whisker, as well as first quartile and third quartile, are the two thresholds used as the removal limit. Therefore, any observations that exceed this threshold will be excluded from the analysis. In addition, there is a dataset containing the intersection of all observations within the upper and lower whiskers of the filtered travel time and travel cost dataset from the prior cutoff procedure in order to gain more observations in the dataset. This results in four datasets in total: the full dataset, a dataset with a travel time cutoff based on first and third quartile, a dataset with a travel cost cutoff based on first and third quartile, and an intersection dataset.

3.5. Attribute treatment

The datasets were analysed by applying different attribute treatment approaches in regard to travel time and travel costs. In this context, attribute treatment was used to determine which source of travel time and travel cost data to be incorporated into the model. This stage was used to evaluate the output of the model and the performance of data treatments. There are three travel attributes used in this study, which are reported, Google Maps API, and adjusted travel data.

The use of adjusted API data can result in increased correlation with reported travel data. This can be done by updating the travel attribute values retrieved from Google Maps API. A previous study by Zannat et al. (2021) developed a sub-model to generate more reliable travel time estimates throughout the day by establishing relationships among the three traffic models provided by Google Maps API: best guess, pessimistic, and optimistic. Their study found that applying such an adjustment improved the correlation between reported and adjusted API travel times. In this study, the adjusted API data is updated based on power regression equation of the chosen mode for a specific trip made by individual who reported their travel time and cost (Eq. (1)). Power regression is used to update the travel attributes of the chosen mode, given that the relationship between travel data generated by Google Maps API and reported by the respondents can exhibit nonlinear relationships (Hillel et al., 2016). In the case that mode is not chosen for that specific trip, the unchosen mode is used as the dependent variable, while the Google Maps API travel data of the unchosen mode in that particular trip type, acts as the independent variable (Eq. (2)).

$$\overline{x}_{n,t,j^*} = \alpha_{j^*} \cdot {x^R_{n,t,j^*}}^{\beta_{j^*}} \tag{1}$$

$$\ln(\overline{x}_{n,t,j^*}) = \ln(\alpha_{j^*} \cdot {x^R_{n,t,j^*}}^{\beta_{j^*}})$$

$$\ln(\overline{x}_{n,t,j^*}) = \ln(\alpha_{j^*}) + \beta_{j^*} \cdot \ln(x^R_{n,t,j^*})$$

$$\beta_{j^*} \cdot \ln(x^R_{n,t,j^*}) = \ln(\overline{x}_{n,t,j^*}) - \ln(\alpha_{j^*})$$

$$\ln(x^R_{n,t,j^*}) = \frac{\ln(\overline{x}_{n,t,j^*}) - \ln(\alpha_{j^*})}{\beta_{j^*}}$$

$$x^R_{n,t,j^*} = \exp\frac{\ln(\overline{x}_{n,t,j^*}) - \ln(\alpha_{j^*})}{\beta_{j^*}}$$

$$x_{n,t,j} = \exp\frac{\ln(\overline{x}^{API}_{n,t,j}) - \ln(\alpha_j)}{\beta_j} \tag{2}$$

where,

$\overline{x}_{n,t,j^*}$ = The adjusted travel time or cost of the chosen mode j^* for specific trip t made by individual n;
$\overline{x}_{n,t,j}$ = The adjusted travel time or cost of the unchosen mode j for specific trip t made by individual n;
α_{j^*} = constant for chosen mode j^*;
β_{j^*} = exponent representing the rate of change for chosen mode j^*;
α_j = constant for unchosen mode j;
β_j = exponent representing the rate of change for unchosen mode j;
x^R_{n,t,j^*} = (reported) travel time or cost of the chosen mode j^* for specific trip t made by individual n;
$x^{API}_{n,t,j}$ = (Google Maps API) travel time or cost of the unchosen mode j for specific trip t made by individual n.

3.6. Combinations of data filtering and attribute treatment

Considering that there are four datasets, three attribute options for chosen travel cost and time, and two attribute options for unchosen travel cost and time, Table 2 presents all 24 possible dataset and attribute treatment combinations used in this study.

Table 2
Attribute treatments.

Attribute treatments	Chosen mode	Unchosen mode	Dataset
A	Reported data	Google Maps API	Full
B	Reported data	Adjusted API	Full
C	Google Maps API	Adjusted API	Full
D	Google Maps API	Google Maps API	Full
E	Adjusted API	Google Maps API	Full
F	Adjusted API	Adjusted API	Full
G	Reported data	Google Maps API	TT Cutoff
H	Reported data	Adjusted API	TT Cutoff
I	Google Maps API	Adjusted API	TT Cutoff
J	Google Maps API	Google Maps API	TT Cutoff
K	Adjusted API	Google Maps API	TT Cutoff
L	Adjusted API	Adjusted API	TT Cutoff
M	Reported data	Google Maps API	TC Cutoff
N	Reported data	Adjusted API	TC Cutoff
O	Google Maps API	Adjusted API	TC Cutoff
P	Google Maps API	Google Maps API	TC Cutoff
Q	Adjusted API	Google Maps API	TC Cutoff
R	Adjusted API	Adjusted API	TC Cutoff
S	Reported data	Google Maps API	Intersection
T	Reported data	Adjusted API	Intersection
U	Google Maps API	Adjusted API	Intersection
V	Google Maps API	Google Maps API	Intersection
W	Adjusted API	Google Maps API	Intersection
X	Adjusted API	Adjusted API	Intersection

TT Cutoff: travel time cutoff; TC Cutoff: travel cost cutoff.

3.7. Modelling framework

A multinomial logit model (MNL) (McFadden, 1972) was employed to establish the relative importance of travel time and cost variables in influencing choice of commuting mode. This model was also used to examine the relative performance after some improvements conducted using attribute treatments approaches. In this case, the random utility of mode (j) in multimodal trips (t) for an individual commuter (n) is shown by the following Eq. (3).

$$V_{n,t,j} = \alpha_j + \beta_{cost} \cdot Cost_{n,t,j} + \beta_{time} \cdot Time_{n,t,j}$$
$$U_{n,t,j} = V_{n,t,j} + \epsilon_{n,t,j} \quad (3)$$

where,

$V_{n,t,j}$ = deterministic utility of mode j made by individual n during trip t;
α_j = alternative specific constant for mode j;
$Cost_{n,t,j}$ = travel cost of mode j for trip t made by individual n;
$Time_{n,t,j}$ = travel time of mode j for trip t made by individual n;
β = estimated coefficients;
$\epsilon_{n,t,j}$ = error term specific to individual n, trip t, and mode j.

This study also measured the value of travel time savings (VTTS) of a commuter in the Indonesian Rupiah (IDR) by dividing travel time parameter and travel cost parameter (Ben-Akiva et al., 1985). VTTS evaluates an individual's willingness to pay for a decrease in travel duration.

4. Results and discussion

4.1. Discrepancy between RP and google maps API data

There are 8169 observations as the base dataset after the initial data cleaning process. Fig. 3 shows a scatter plot of travel time and travel cost values from Google Maps API against reported travel data. The regression line is clearly below the line $y = x$. This indicates that the recorded travel time and cost tend to be significantly higher than the travel attributes data obtained from the Google Distance Matrix API. Moreover, the R-squared values for travel time and travel cost are 0.374 and 0.177, respectively. This indicates a low correlation between the travel time and cost generated by the Google Maps API and those reported by the respondents.

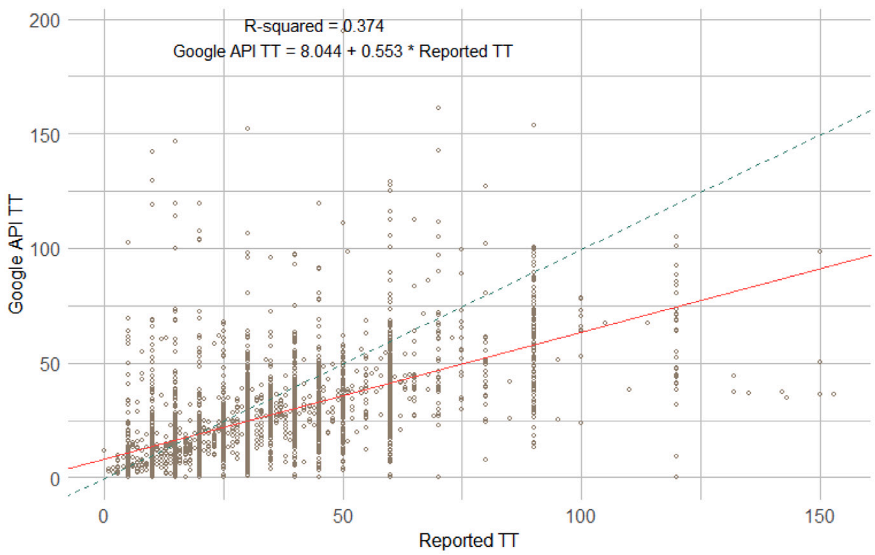
Following this, the cutoff based on interquartile range approach was applied in order to improve the correlation between travel data reported by respondents and that of the Google Maps API. Interquartile range diagram is illustrated in Fig. 4 showing that there is a wide variation of both travel time and travel cost for each mode. For example, it can be seen that for some road-based public transport, such as buses and BRT, reported times are substantially less than Google Maps API times. The Google Maps API, in contrast, underestimates private vehicles including cars and motorcycles. In regards to travel cost, the diagram displays that reported travel cost are slightly higher than those measured by Google Maps API (except for bus). Meanwhile, travel cost of taxi and on-demand transport are reported to be less than values obtained by Google Maps API.

All observations beyond the determined threshold were removed, for both travel time and travel cost. Table 3 shows the number of observations as well as the R-squared values for each dataset.

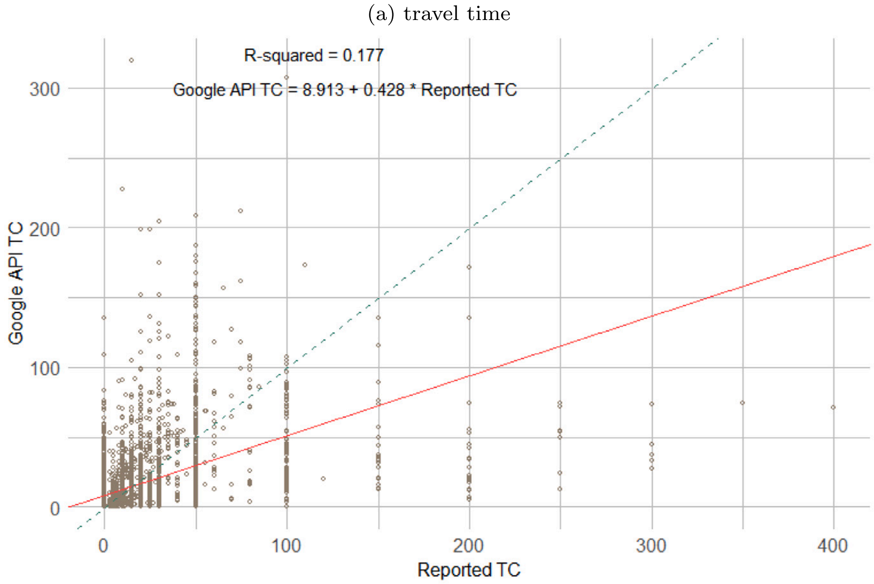
The utilisation of a cutoff approach based on the first and third quartiles produced a second and third dataset in which the responses of respondents showed a strong correlation with travel time and cost data obtained from the Google Maps API. Nevertheless, this approach significantly reduces the number of observations. In contrast, the dataset constructed utilising the intersection of all observations within the area defined by the limit of the upper and lower whisker produced a larger number of observations than the earlier method.

4.2. Characteristics of excluded observations

A total of 729; 3692; and 4053 observations were excluded when the initial dataset was segmented into intersection, travel cost cutoff, and travel time cutoff datasets, respectively. It is essential to carry out an investigation into the excluded observations so that the characteristics of the removed observations can be evaluated. This includes using a multiple regression model to examine the causal relationship between excluded variables to identify the most influential factors that led to their exclusion. The analysis is performed with respect to the spatial dimension of the respondents, in addition to their sociodemographic characteristics and travel patterns. Table 4 presents the total number of variables excluded at each steps of obtaining the final dataset, together with their proportion. Meanwhile, Table 5 shows the output of the



(a) travel time



(b) travel cost

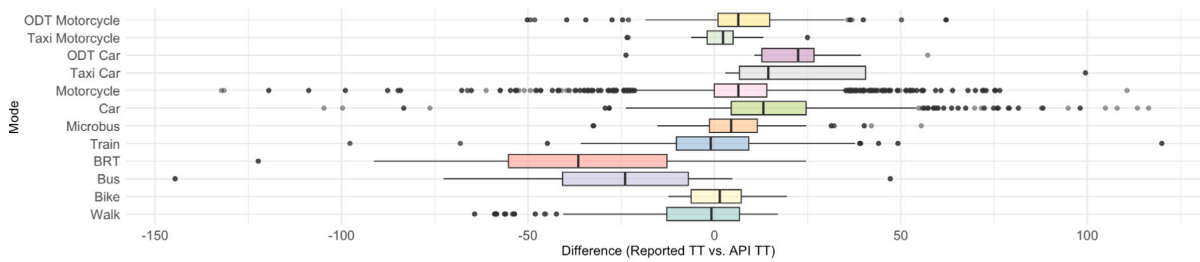
Fig. 3. Scatter plot of all modes. The closer the regression line (solid red) is to the x=y line (dashed), the smaller the average discrepancy.

Table 3
No. of observations and R-squared value for each dataset.

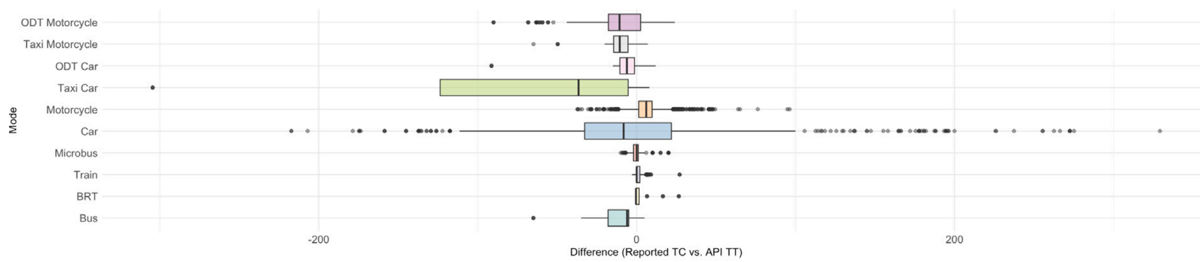
No	Datasets	No. of observations	R-squared	
			Travel time	Travel cost
1	Base Dataset	8169	0.374	0.177
2	Cutoff based on first- and third-quartile in travel time	4116	0.815	0.059
3	Cutoff based on first- and third-quartile in travel cost	4477	0.399	0.761
4	Intersection of all observations within upper- and lower-whisker of filtered travel time and cost dataset	7449	0.581	0.313

multiple regression model to provide a more informative understanding of the factors influencing the exclusion of observations.

The proportions of males and females that were excluded are essentially similar for all three datasets. Therefore, there are no gender-based differences in travel time and travel cost estimation. Meanwhile, the finding suggests that younger respondents have a tendency to provide more accurate estimations of travel time in comparison to older respondents. It may simply reflect that younger people use mobile phone navigation devices more often than older individuals. These results are supported by previous findings that older people have lower levels



(a) IQR travel time



(b) IQR travel cost

Fig. 4. Interquartile range of difference between Google Maps API and Reported by respondents.

of mobility than younger individuals (Deka, 2022; Luiu et al., 2016; Nordbakke and Schwanen, 2014), raising the possibility of inaccurate trip duration estimation.

Individuals who conduct multimodal trips are more likely to be excluded in the travel time cutoff and intersection datasets, possibly as a result of there being a greater chance of misreporting (e.g. not accurately reporting the travel time of all parts of the trip). The distribution with respect to distance reveals that higher number of observations removed for medium and long travel distance in comparison to short travel distance. This is likely a result of the fact that for a longer travel distance, there is going to be more discrepancy in terms of traffic in real life and that is going to be disproportionately impacts the travel time estimation (Christensen, 2024).

When considering the spatial dimension, residents of Jakarta and its neighbouring cities tend to have a high correlation between the travel time and cost reported by respondents and those obtained from the Google Maps API. It is reasonable that individuals of Jakarta and surrounding areas possess a better ability for forecasting travel time as a result of their higher exposure to reliable transport infrastructure, as well as excellent public transport systems that able to provide travel time estimation due to current or historical traffic conditions (Jenelius, 2018).

With respect to the mode chosen by the respondents, the result of regression model reveals that public transport users tend to provide more accurate estimates of travel time compared to other modes. This might due to public transport users heightened awareness of the reliability of public transport services (Benezech and Coulombel, 2013; Daganzo, 2009). However, their ability to provide more accurate estimates of transit fares is lower, as evidenced by the travel cost cutoff.

4.3. MNL model and VTTS

The MNL model is implemented to examine the influence of travel time and cost on mode choices. Table 6 shows the performance of the different models, the estimates for the travel time and travel cost parameters, as well as VTTS for each combination of dataset and attribute treatment.

In the table above, attribute treatments with significant estimates and meaningful VTTS results are denoted by colour coding. A red cell indicates an unreasonable or insignificant result for the parameter in travel time and cost estimates, whereas it suggests an unreasonable result for VTTS. The green cell in VTTS, meanwhile, represents realistic VTTS in the context of Greater Jakarta. It is clear that a variety of outputs can be observed depending on the attribute interventions implemented. Only attribute treatments G, M, P, Q, and V produce a significant and reasonable sign for travel attribute parameters and VTTS, while all other attribute treatments produce some counterintuitive results. Nevertheless, despite the better results, the implementation of attribute treatments G, M, P, and Q leads to a significant reduction in the number of observations, accounting for almost 50% of the entire dataset. As a contrast, attribute treatment V has the number of observations closest to the original dataset, whilst still producing favourable results for VTTS.

We also study the outputs from more complex models with mode-specific travel time coefficients (replacing β_{time} with $\beta_{time,j}$ in Eq. (3)). The mode-specific VTTS differ depending on the specific attribute treatments. Mode-specific VTTS is observed for taxi and on-demand transportation (ODT), public transport, car, and motorcycle. Table 7 reveals that the results of attribute treatments P and V for VTTS, travel time and cost parameters show meaningful results, as evidenced by the non-negative values. In contrast, the other attribute treatments generate non-intuitive sign of VTTS parameters for a particular mode. Finally, the best performing attribute treatments (P and V) are compared to prior research conducted in the Greater Jakarta context in order to validate the findings, as illustrated in Table 8 and Fig. 5.

The mode-specific VTTS for attribute treatment P is slightly higher than previous findings by Belgiawan et al. (2019) and Ilahi et al. (2021) and significantly higher than project report of the JABODETABEK Urban Transportation Policy Integration Project Phase 2 (JUPTI 2) conducted by Coordinating Ministry of Economic Affairs of the Republic of Indonesia (JICA, 2019). In contrast, mode-specific VTTS for attribute treatment V is lower in comparison to prior research that utilised the SP dataset (Belgiawan et al., 2019) and pooled SP-RP dataset (Ilahi et al., 2021). However, the outcome is similar to the results of JUTPI 2 (JICA, 2019), a project that also applied the RP dataset.

The current work found that VTTS for car is lower than for public transport; meanwhile, VTTS for Taxi & ODT is the highest. Furthermore, the mode-specific results of VTTS for attribute treatment V in this study are notably similar to those of VTTS in the JUTPI 2 report, which uses data from the most recent and extensive cross-sectional activity-travel diary survey conducted in Greater Jakarta.

Table 4
The number of each variables in each dataset.

No	Variables	Base data %(n)	TT cutoff %(n)	TC cutoff %(n)	Intersection %(n)
1	Gender				
	Male	70.71 (5776)	71.53 (2899)	71.07 (2624)	74.49 (543)
	Female	29.29 (2393)	28.47 (1154)	28.93 (1068)	25.51 (186)
2	Age				
	<15	0.51 (42)	0.57 (23)	0.43 (16)	0.41 (3)
	15–20	4.60 (376)	4.79 (194)	5.12 (189)	3.16 (23)
	21–25	16.67 (1362)	18.38 (745)	16.74 (618)	19.34 (141)
	26–30	21.19 (1731)	20.82 (844)	19.28 (712)	18.79 (137)
	31–35	11.05 (903)	9.55 (387)	10.24 (378)	10.70 (78)
	36–40	9.50 (776)	10.49 (425)	8.94 (330)	10.56 (77)
	41–45	9.28 (758)	9.77 (396)	10.51 (388)	9.74 (71)
	46–50	12.85 (1050)	12.39 (502)	12.76 (471)	13.17 (96)
	51–55	8.61 (703)	8.39 (340)	9.43 (348)	8.50 (62)
	56–60	3.75 (306)	3.08 (125)	4.50 (166)	3.43 (25)
	>60	1.98 (162)	1.78 (72)	2.06 (76)	2.19 (16)
3	Trip type				
	Single trip	87.91 (7181)	87.69 (3554)	94.93 (3505)	87.65 (639)
	Multimodal trip	12.09 (988)	12.31 (499)	5.07 (187)	12.35 (90)
4	Distance (km)				
	0–4	25 (2073)	22.80 (924)	16.79 (620)	17.97 (131)
	5–9	24 (1951)	19.37 (785)	22.37 (826)	16.46 (120)
	10–17	26 (2087)	27.41 (1111)	29.55 (1091)	26.61 (194)
	18–36	20 (1660)	25.41 (1030)	25.08 (926)	28.81 (210)
	>36	2 (190)	2.99 (121)	3.39 (125)	7.54 (55)
5	Home city				
	Central Jakarta	14.20 (1160)	12.02 (487)	10.08 (372)	12.07 (88)
	North Jakarta	6.41 (524)	6.39 (259)	6.31 (233)	7.58 (56)
	West Jakarta	8.47 (692)	6.88 (279)	9.29 (343)	6.86 (50)
	South Jakarta	10.06 (822)	8.31 (337)	10.21 (377)	5.49 (88)
	East Jakarta	14.64 (1196)	13.64 (553)	18.17 (671)	11.25 (82)
	Depok City	4.43 (362)	5.77 (234)	3.06 (113)	4.94 (36)
	Bekasi City	8.61 (703)	9.43 (382)	10.37 (383)	12.89 (94)
	Bekasi Regency	1.38 (113)	2.10 (85)	1.46 (54)	1.51 (11)
	Karawang Regency	0.18 (15)	0.37 (15)	0.41 (15)	1.78 (13)
	Tangerang City	5.55 (453)	5.67 (230)	6.85 (253)	5.76 (42)
	South Tangerang City	1.57 (128)	1.95 (79)	1.54 (57)	1.78 (13)
	Tangerang Regency	1.2 (98)	1.6 (65)	1.76 (65)	1.78 (13)
	Bogor City	7.11 (581)	7.01 (284)	5.04 (186)	9.74 (71)
	Bogor Regency	15.89 (1298)	18.33 (743)	14.92 (551)	14.54 (106)
	Sukabumi	0.18 (15)	0.37 (15)	0.41 (15)	1.51 (11)
6	Mode choice				
	Walk	6.00 (490)	5.82 (236)	0.08 (3)	5.21 (38)
	Bike	0.15 (12)	0.10 (4)	0.76 (28)	0 (0)
	Bus	0.75 (61)	0.72 (29)	0.76 (28)	0.55 (4)
	BRT	0.91 (74)	0.81 (33)	0.73 (27)	1.10 (8)
	Train	4.47 (365)	4.32 (175)	1.90 (70)	5.76 (42)
	Microbus	2.17 (177)	2.15 (87)	2.09 (77)	4.39 (32)
	Car	19.88 (1624)	20.01 (811)	21.97 (811)	18.93 (138)
	Motorcycle	57.38 (4687)	57.78 (2342)	63.41 (2341)	58.16 (424)
	Car Taxi	0.15 (12)	0.15 (6)	0.16 (6)	0.69 (5)
	Car ODT	0.29 (24)	0.27 (11)	0.33 (12)	0.27 (2)
	Motorcycle Taxi	0.69 (56)	0.62 (25)	0.68 (25)	0.55 (4)
	Motorcycle ODT	7.19 (587)	7.25 (294)	7.91 (292)	4.39 (32)

VTTS of motorcycles is higher than car; this might be because motorcycles are generally more fuel-efficient than cars, and this can contribute to cost savings as well as time savings. The lower operating costs may lead to a higher perceived value of time for motorcycle users. Further, in certain situations, motorcycles may be able to take advantage of shortcuts or access routes that are not available to cars. This can result in reduced commute times, contributing to a higher VTTS for motorcycle users.

Compared to VTTS revealed by Belgiawan et al. (2019) and Ilahi et al. (2021), therefore, it makes sense that the VTTS of attribute treatment V derived from this study is lower, considering that individuals tend to express higher values for time savings when asked directly in an SP survey. In hypothetical scenarios, people may assign a higher monetary value to their time than what is revealed through their actual behaviour in the RP datasets. This is also consistent with a recent study by Li et al. (2018), in which they found that the SP/RP ratio for cars is 0.9 and the SP/RP ratio for public transport is 2.44. Additionally, this work contributes to the current body of knowledge by employing RP data in calculating VTTS, particularly in the context of the Global South, while other studies commonly utilised SP data (Batley et al., 2017; Belgiawan et al., 2019) or RP data with a focus on developed countries (Li et al., 2018; Calastri et al., 2019). Conversely, a little discrepancy in VTTS is present between the results of JUTPI 2 and this study; this could be related to distinct survey methodologies, as

Table 5
Multiple regression coefficients of excluded observations..

Excluded observations of:	Estimate	Std. Error	t-value	p-value
Travel time cutoff				
(Intercept)	0.493	0.020	24.020	0.000***
Male	0.019	0.012	1.536	0.124
Young individuals	−0.041	0.014	−2.755	0.006**
Multimodal trip	0.061	0.027	2.239	0.025*
Long distance	0.161	0.037	4.293	0.000***
Outskirts city	0.082	0.019	4.447	0.000***
Public transport	−0.083	0.032	−2.578	0.009**
Travel cost cutoff				
(Intercept)	0.386	0.020	19.375	0.000***
Male	−0.003	0.012	−0.286	0.774
Young individuals	−0.004	0.014	−0.289	0.772
Multimodal trip	−0.274	0.026	−10.542	0.000***
Long distance	0.275	0.036	7.568	0.000***
Outskirts city	0.113	0.018	6.246	0.000***
Public transport	0.096	0.169	6.135	0.000***
Intersection				
(Intercept)	0.121	0.014	8.778	0.000***
Male	0.011	0.008	1.309	0.190
Young individuals	−0.031	0.009	−3.164	0.002**
Multimodal trip	0.261	0.018	14.463	0.000***
Long distance	0.245	0.025	9.757	0.000***
Outskirts city	0.089	0.012	7.114	0.000***
Public transport	−0.120	0.022	−5.550	0.000***

*** $p < 0.001$.
** $0.001 < p < 0.01$.
* $0.01 < p < 0.05$.

Table 6
Multinomial logit model results.

Attribute treatments	Dataset	Chosen mode	Unchosen mode	$LL(0)$	LL (ρ^2)	β travel time (t.rat)	β travel cost (t.rat)	VTTS (IDR/h)
A	Full	Rep.	API	−18,636.09	−7098.99 (0.62)	−0.00003 (−0.03)	−0.010 (−12.379)	235
B	Full	Rep.	Adj.	−18,636.09	−4050.5 (0.78)	0.174 (49.737)	0.014 (6.636)	766,035
C	Full	API	Adj.	−18,636.09	−2,462.16 (0.87)	0.408 (43.74)	0.160 (32.69)	153,133
D	Full	API	API	−18,636.09	−6664.58 (0.64)	−0.043 (−29.506)	0.014 (18.097)	−189,396
E	Full	Adj.	API	−18,636.09	−6874.07 (0.63)	−0.035 (−22.983)	0.005 (5.379)	−391,879
F	Full	Adj.	Adj.	−18,636.09	−4186.58 (0.77)	0.157 (48.953)	0.047 (30.363)	198,787
G	TT Cutoff	Rep.	API	−9434.69	−3583.76 (0.62)	−0.013 (−5.576)	−0.008 (−7.150)	90,413
H	TT Cutoff	Rep.	Adj.	−9434.69	−2576.25 (0.72)	−0.100 (−30.624)	0.014 (6.000)	−442,417
I	TT Cutoff	API	Adj.	−9,434.69	−2096.36 (0.78)	−0.106 (−27.683)	0.055 (14.460)	−115,636
J	TT Cutoff	API	API	−9434.69	−3180.46 (0.66)	−0.080 (−23.165)	0.021 (13.968)	−227,163
K	TT Cutoff	Adj.	API	−9434.69	−3417.48 (0.64)	−0.053 (−18.931)	0.006 (4.157)	−490,226
L	TT Cutoff	Adj.	Adj.	−9434.69	−2309.5 (0.75)	−0.077 (−28.645)	0.052 (19.557)	−89,158
M	TC Cutoff	Rep.	API	−10,243.09	−4389.22 (0.57)	−0.016 (−10.733)	−0.009 (−4.278)	110,379
N	TC Cutoff	Rep.	Adj.	−10,243.09	−2549.23 (0.75)	0.205 (34.664)	−0.041 (−11.602)	−298,314
O	TC Cutoff	API	Adj.	−10,243.09	−2868.1 (0.72)	0.255 (32.913)	−0.017 (−6.717)	−903,294
P	TC Cutoff	API	API	−10,243.09	−3989.37 (0.61)	−0.062 (−25.440)	−0.011 (−5.092)	332,426
Q	TC Cutoff	Adj.	API	−10,243.09	−4213.67 (0.59)	−0.036 (−19.317)	−0.015 (−7.246)	140,970
R	TC Cutoff	Adj.	Adj.	−10,243.09	−2572.2 (0.75)	0.199 (36.96)	−0.036 (−15.65)	−329,602
S	Intersection	Rep.	API	−17,033.54	−6494.95 (0.62)	−0.0082 (−6.080)	−0.0018 (−1.409)	277,039
T	Intersection	Rep.	Adj.	−17,033.54	−5037 (0.70)	0.093 (38.081)	0.06 (22.768)	92,496
U	Intersection	API	Adj.	−17,033.54	−5330.18 (0.69)	0.084 (29.412)	0.100 (35.249)	50,067
V	Intersection	API	API	−17,033.54	−5902.23 (0.65)	−0.063 (−29.737)	−0.008 (−5.131)	474,486
W	Intersection	Adj.	API	−17,033.54	−6081.17 (0.64)	−0.047 (−26.556)	−0.001 (−0.472)	3710,718
X	Intersection	Adj.	Adj.	−17,033.54	−5684.53 (0.66)	0.034 (17.970)	0.078 (32.700)	26,623

Table 7
VTTS of attribute treatments G, M, P, Q, and V.

Mode	Attribute treatments				
	G (IDR/hour)	M (IDR/hour)	P (IDR/hour)	Q (IDR/hour)	V (IDR/hour)
Public transport	−13,503	14,174	139,472	45,549	23,214
Car	−468,582	−103,983	87,476	−91,061	17,979
Motorcycle	−144,134	−3645	231,787	56,959	42,260
Taxi & ODT	−200,591	−65,465	238,507	12,315	52,287

Table 8
Comparing mode-specific VTTS with previous studies in Greater Jakarta.

Mode	Current work[a]		Belgiawan et al. (2019)[b]	Ilahi et al. (2021)[c]	JUTPI 2 (2019)[a]
	Attribute treatments P (IDR/h)	Attribute treatments V (IDR/h)	(IDR/h)	(IDR/h)[d]	(IDR/h)
Public transport	139,472	23,214	91,180	–	35,292
Car	87,476	17,979	82,704	14,186–56,744	52,478
Motorcycle	231,787	42,260	–	70,930–212,790	31,692
Taxi & ODT	238,507	52,287	–	141,860–567,440	–

[a] RP dataset.
[b] SP dataset.
[c] Pooled SP and RP.
[d] VTTS related to income and distance.

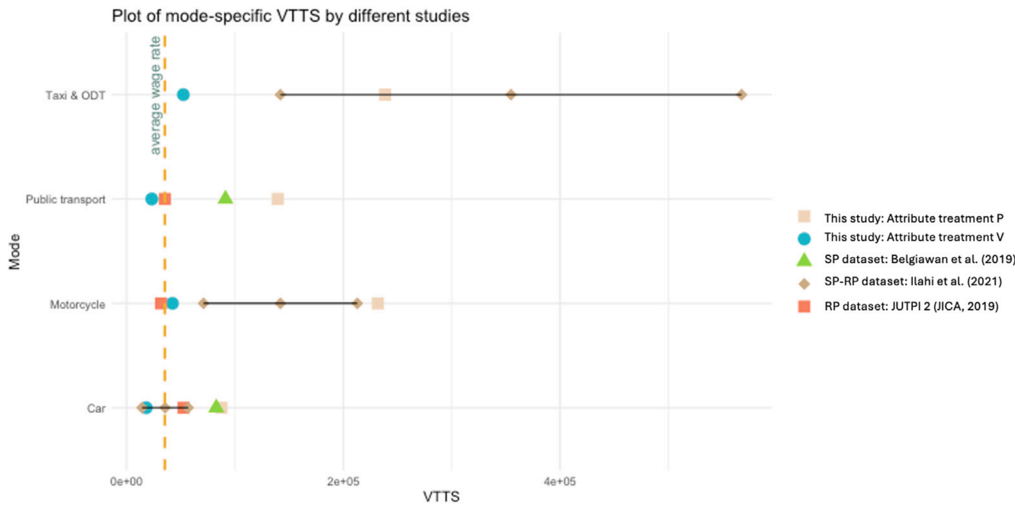


Fig. 5. Mode-specific VTTS by different studies.

a smartphone application called ADS MEILI (Prelipcean et al., 2018) was specifically used in JUTPI 2 to capture RP travel attributes, which may have reduced bias.

Furthermore, the estimation of travel time savings (VTTS) using a percentage of the average wage rate as the basis for the VTTS remains a prevalent practice. Some circulating estimates, including those used in the evaluation criteria, placed the VTTS of private travel at approximately 40% of the average wage rate (Douglas, 2018), while this number is estimated around 50 percent of after-tax wages in low- and middle-income countries (Whittington D, 2019). The current VTTS estimates of attribute treatment V is then compared with the average wage rate in the sample. Average gross income of sample is 5,629,682 IDR per month, or 35,185 IDR/hour, based on common assumption of 40 working hours per week. The percentage of VTTS to sample average wage rate for public transport, car, motorcycle, and taxi & ODT are 66%, 51%, 120%, and 148%, respectively. The results indicate that the average savings in travel time vary depending on the mode and does not align closely to the 40 or 50% rule. The varied VTTS to average wage is in line with recent findings by Hensher (2019), who examined the implications of travel time heterogeneity under different travel conditions.

5. Conclusions and policy implications

This study investigated the discrepancy between mode-specific attribute values observed in an RP dataset and their corresponding obtained values from Google Maps API data, in the context of developing countries. The discrepancy between reported travel data and API-based travel data can introduce biases that affect travel behaviour analysis and modelling. Furthermore, inaccurate estimates may lead to ineffective transport policies and inefficient transport infrastructure planning. After pre-processing travel data and creating the required base dataset of 8169 observations, the relative impact of data sorting and attribute treatment on model outputs was examined using multinomial logit model. This research provides important theoretical and practical contributions by addressing key gaps in the limited studies on discrepancies between self-reported travel attributes and those from online navigation map providers in developing countries.

5.1. Key insights

The findings indicate that there are clear discrepancies between Google Maps API travel times and costs and the respondents' reported travel times and costs in the context of Greater Jakarta. The disparities between travel time and cost of travel vary systematically across travel modes. Reported times were significantly less than those provided by the Google Maps API for certain road-based public transportation. The Google Maps API, however, underestimated times for private vehicles. Concerning travel costs, reported costs were somewhat greater than those estimated by Google Maps API. Therefore, transport planning tools that use Google Maps API data in any form should acknowledge that such systematic variations can lead to inaccuracies in predictions, potentially introducing bias into policy decisions based on the forecasts.

This results in an important follow-up issue: which observations should be treated as outliers and thus be categorised as 'unreliable', and which should be kept for modelling work. Notably, our results show that preprocessing the data and attribute treatments can completely change the VTTS estimates. Since VTTS reflects how much individuals value time savings, this coefficient plays a crucial role in assessing the economic viability of transportation projects. Accurate VTTS estimates are essential for making informed decisions about which projects should receive funding and the design of policies aimed at improving travel efficiency. Inaccurate VTTS estimates could lead to misallocation of resources, resulting in the underfunding of high-impact projects or the overemphasis on projects with lower economic returns. This shows how important it is to carefully choose which data to include and how to combine travel data attributes, especially when doing research on transport in a developing country. In this case, the VTTS only makes sense for the model with attribute treatment V, which utilised Google Maps API travel data for both chosen and unchosen alternatives and applied less strict criteria for retaining observations for modelling work.

A further analysis was performed on the excluded observations to investigate the respondent characteristics that may contribute to discrepancies between their reported travel data and the ones from the Google Maps API. The results revealed that older individuals and commuters who needed to make multimodal or medium and long distance trips were more likely to report values that have larger deviations from the API values.

5.2. Practical implications

As the first to examine discrepancies in mode-specific travel time and cost between the RP dataset and Google Maps API, this research highlights the importance of systematic data cleaning to ensure robust models for planning and policy analyses while minimising data loss. The consequences of these findings are that, where possible, Global Positioning System (GPS) surveys in a smartphone app for completion of Activity Travel Diary Surveys (ADS) should be used to minimise human perception and memory errors in estimating travel time. Where it is not possible, it is highly recommended that open-ended questions are used in place of static responses for paper-based surveys in order to prevent rounding effects when collecting travel data from respondents.

An additional recommendation is to implement specific sampling techniques, such as stratified or weighted sampling. Stratified sampling is a method to adjust for known biases in particular sub-populations of variables that lead to exclusion (Keramat and Kielbasa, 1998; Kaymaz et al., 2019). This will help provide better accuracy for different sub-populations within the variable. Meanwhile, weighted sampling allows the likelihood function for each observation to be determined using the weighted sampling method, which represents the relative real-world characteristics. Weights that were suitable for the sub-samples utilised in the different models were determined specifically (Manski and Lerman, 1977; Hess and Polak, 2005). These methods may help alleviate issues generated by biases generated through data cleaning exercises, such as those shown in this work where e.g. older respondents' data is more likely to be excluded.

5.3. Future research directions

Despite the valuable insights obtained from this study, several limitations must be acknowledged. First, it is expected that similar discrepancies might exist within the context of other datasets from other cities that have characteristics comparable to the Greater Jakarta context, such as diverse transport modes, complex urban forms, limited access to reliable public transport, and extensive urban agglomeration. Therefore, to assess the reliability of the findings of this work, it is necessary to carry out this research in other regions or countries. Second, given that this study focused solely on work-related travel, future studies may also wish to address other trip purposes (e.g., leisure, social activities, shopping, study, and eating out) and types of trips (e.g., multimodal trips, first/last-mile connectivity, and variations in time-of-day effects). The findings can contribute to providing information to model reliable travel behaviour in the context of utilising the RP dataset, particularly in developing countries. Consequently, they can lead to better planning and policy analyses tools. Third, while this study provided an analysis of the key characteristics influencing the discrepancies, future research could apply the Geographically Weighted Regression (Fotheringham et al., 2002) to gain deeper insights into the spatial heterogeneity that contributes to differences between the Google Maps API and reported travel data. This approach would help identify relationships that may vary depending on geographic location. Lastly, while this study compares the API-based travel attributes with reported travel data, future research could explore alternative validation techniques or data sources, such as direct comparisons with GPS-tracked trips, transport modelling outputs, and big data sources from mobility data apps (e.g., Uber, Lyft, or bike-sharing systems). This would help further enhance the accuracy and reliability of API-derived travel attributes for travel behaviour studies.

CRediT authorship contribution statement

Faza Fawzan Bastarianto: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Thomas O. Hancock:** Writing – review & editing, Validation, Supervision, Methodology. **Anugrah Ilahi:** Supervision, Resources. **Ed Manley:** Writing – review & editing, Validation, Supervision. **Charisma Farheen Choudhury:** Writing – review & editing, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was primarily supported by the Indonesian Education Scholarship Program (BPI), administered by the Center for Higher Education Funding and Assessment (PPAPT), Ministry of Higher Education, Science, and Technology of the Republic of Indonesia and the Indonesia Endowment Fund for Education (LPDP), Ministry of Finance of the Republic of Indonesia [Grant No. 1772/J5/KM.01.00/2021]. Charisma Choudhury and Thomas Hancock acknowledge the financial support received from the UKRI Future Leader Fellowship, UK [MR/T020423/1-NEXUS].

References

Auld, J., Karimi, B., Pourabdollahi, Z., Mohammadian, A.K., Kawamura, K., 2016. Long-distance trips and mode choice in illinois. Transportation Research Board 95th Annual Meeting.

Bastarianto, F., Irawan, M., Choudhury, C., Palma, D., Muthohar, I., 2019. A tour-based mode choice model for commuters in Indonesia. Sustainability 11 (3), 788. http://dx.doi.org/10.3390/su11030788.

Batley, R., Bates, J., Bliemer, M., Börjesson, M., Bourdon, J., Cabral, M.O., Chintakayala, P.K., Choudhury, C., Daly, A., Dekker, T., Drivyla, E., Fowkes, T., Hess, S., Heywood, C., Johnson, D., Laird, J., Mackie, P., Parkin, J., Sanders, S., Sheldon, R., Wardman, M., Worsley, T., 2017. New appraisal values of travel time saving and reliability in great britain. Transportation 46 (3), 583–621. http://dx.doi.org/10.1007/s11116-017-9798-7.

Belgiawan, P.F., Ilahi, A., Axhausen, K.W., 2019. Influence of pricing on mode choice decision in Jakarta: A random regret minimization model. Case Stud. Transp. Policy 7 (1), 87–95. http://dx.doi.org/10.1016/j.cstp.2018.12.002.

Ben-Akiva, M.E., Lerman, S.R., Lerman, S.R., 1985. Discrete Choice Analysis. MIT Press.

Benezech, V., Coulombel, N., 2013. The value of service reliability. Transp. Res. Part B: Methodol. 58, 1–15. http://dx.doi.org/10.1016/j.trb.2013.09.009.

Benita, F., 2023. Exploring non-mandatory travel behavior in Jakarta city: Travel time, trip frequency, and socio-demographic influences. Transp. Res. Interdiscip. Perspect. 21, 100896. http://dx.doi.org/10.1016/j.trip.2023.100896.

BPTJ, 2023. Laporan Pengukuran Capaian Indikator Kinerja Transportasi Jabodetabek Tahun 2023. Technical Report, Jabodetabek Transportation Management Agency, Ministry of Transportation of the Republic of Indonesia.

Calastri, C., Hess, S., Choudhury, C., Daly, A., Gabrielli, L., 2019. Mode choice with latent availability and consideration: Theory and a case study. Transp. Res. Part B: Methodol. 123, 374–385. http://dx.doi.org/10.1016/j.trb.2017.06.016.

Christensen, L., 2024. Considerations on how to conduct a survey about long-distance travel with reduced memory effect. Transp. Res. Procedia 76, 534–544. http://dx.doi.org/10.1016/j.trpro.2023.12.074.

Daganzo, C.F., 2009. A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. Transp. Res. Part B: Methodol. 43, 913–921. http://dx.doi.org/10.1016/j.trb.2009.04.002.

Dash, C.S.K., Behera, A.K., Dehuri, S., Ghosh, A., 2023. An outliers detection and elimination framework in classification task of data mining. Decis. Anal. J. 6, 100164. http://dx.doi.org/10.1016/j.dajour.2023.100164.

Deka, D., 2022. Trip deprivation among older adults in the context of the capability approach. J. Transp. Geogr. 100, 103325. http://dx.doi.org/10.1016/j.jtrangeo.2022.103325.

dell'Olio, L., Ibeas, A., de Oña, J., de Oña, R., 2018. Designing a survey for public transport users. In: Public Transportation Quality of Service. Elsevier, pp. 49–61. http://dx.doi.org/10.1016/b978-0-08-102080-7.00004-5.

Douglas, N., 2018. The Australian Transport Assessment and Planning Guidelines. Australian Transport Assessment and Planning, URL https://atap.gov.au/mode-specific-guidance/public-transport/index.aspx.

Enam, A., Choudhury, C.F., 2011. Methodological issues in developing mode choice models for Dhaka, Bangladesh. Transp. Res. Rec.: J. Transp. Res. Board 2239 (1), 84–92. http://dx.doi.org/10.3141/2239-10.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2002. Geographically Weighted Regression. Wiley, Wiley.

Fox, J., Daly, A., Hess, S., Miller, E., 2014. Temporal transferability of models of mode-destination choice for the greater Toronto and Hamilton area. J. Transp. Land Use 7 (2), 41–62. http://dx.doi.org/10.5198/jtlu.v7i2.701.

Fu, X., Kakkar, D., Chen, J., Moynihan, K.M., Hegland, T.A., Blossom, J., 2023. A comparative study of methods for drive time estimation on geospatial big data: A case study in usa. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XLVIII-4/W7-2023, 53–60. http://dx.doi.org/10.5194/isprs-archives-xlviii-4-w7-2023-53-2023.

Google Developers, 2023a. Directions API overview | google for developers. https://developers.google.com/maps/documentation/directions/overview, (Accessed 25 September 2023).

Google Developers, 2023b. Geocoding API overview | google for developers. https://developers.google.com/maps/documentation/geocoding/overview, (Accessed 25 September 2023).

Google Developers, 2023c. Google maps platform documentation | distance matrix API | google for developers. https://developers.google.com/maps/documentation/distance-matrix/, (Accessed 25 September 2023).

Guzman, L.A., Oviedo, D., 2018. Accessibility, affordability and equity: Assessing 'pro-poor' public transport subsidies in bogotá. Transp. Policy 68, 37–51. http://dx.doi.org/10.1016/j.tranpol.2018.04.012.

Hensher, D., 2019. Using the average wage rate to assess the merit of value of travel time savings: A concern and clarification. Transp. Find. http://dx.doi.org/10.32866/5772.

Hess, S., Polak, J.W., 2005. Mixed logit modelling of airport choice in multi-airport regions. J. Air Transp. Manag. 11, 59–68. http://dx.doi.org/10.1016/j.jairtraman.2004.09.001.

Hillel, T., Guthrie, P., Elshafie, M., Jin, Y., 2016. Assessing the discrepancies between recorded and commonly assumed journey times in London. In: Transforming the Future of Infrastructure Through Smarter Information. ICE Publishing, http://dx.doi.org/10.17863/CAM.7981.

Ilahi, A., Belgiawan, P.F., Balac, M., Axhausen, K.W., 2020. Mobility Jakarta Stated and Revealed Preference Survey. Technical Report, IVT, ETH Zurich.

Ilahi, A., Belgiawan, P.F., Balac, M., Axhausen, K.W., 2021. Understanding travel and mode choice with emerging modes; a pooled SP and RP model in Greater Jakarta, Indonesia. Transp. Res. Part A: Policy Pr. 150, 398–422. http://dx.doi.org/10.1016/j.tra.2021.06.023.

Indonesia, S., 2019. In: Dwisyahesti, N., Albertha, W. (Eds.), Jabodetabek Commuter Survey 2019. Badan Pusat Statistik.

Jakarta Province, 2016. DKI Jakarta Governor Regulation No. 79 of 2016 on Passenger Transport Fares by Public Bus. Technical Report, DKI Jakarta Province.

Jenelius, E., 2018. Public transport experienced service reliability: Integrating travel time and travel conditions. Transp. Res. Part A: Policy Pr. 117, 275–291. http://dx.doi.org/10.1016/j.tra.2018.08.026.

JICA, 2019. Project Completion Report, The JABODETABEK Urban Transportation Policy Integration Project Phase 2 in the Republic of Indonesia. Technical Report, Japan International Cooperation Agency, URL https://openjicareport.jica.go.jp/pdf/12356358.pdf.

Kaymaz, H., Korkmaz, H., Erdal, H., 2019. Development of a driving cycle for Istanbul bus rapid transit based on real-world data using stratified sampling method. Transp. Res. Part D: Transp. Environ. 75, 123–135. http://dx.doi.org/10.1016/j.trd.2019.08.023.

Keramat, M., Kielbasa, R., 1998. A study of stratified sampling in variance reduction techniques for parametric yield estimation. IEEE Trans. Circuits Syst. II 45, 575–583. http://dx.doi.org/10.1109/82.673639.

Li, Z., Hensher, D.A., Ho, C., 2018. An empirical investigation of values of travel time savings from stated preference data and revealed preference data. Transp. Lett. 12 (3), 166–171. http://dx.doi.org/10.1080/19427867.2018.1546806.

Luiu, C., Tight, M., Burrow, M., 2016. The unmet travel needs of the older population: a review of the literature. Transp. Rev. 37, 488–506. http://dx.doi.org/10.1080/01441647.2016.1252447.

Malokin, A., Circella, G., Mokhtarian, P.L., 2019. How do activities conducted while commuting influence mode choice? using revealed preference models to inform public transportation advantage and autonomous vehicle scenarios. Transp. Res. Part A: Policy Pr. 124, 82–114. http://dx.doi.org/10.1016/j.tra.2018.12.015.

Manski, C.F., Lerman, S.R., 1977. The estimation of choice probabilities from choice based samples. Econometrica 45, 1977. http://dx.doi.org/10.2307/1914121.

Martinez, R., Masron, I.N., 2020. Jakarta: A city of cities. Cities 106, 102868. http://dx.doi.org/10.1016/j.cities.2020.102868.

McFadden, D., 1972. Conditional logit analysis of qualitative choice behavior.

Ministry of Transportation, 2016. Regulation of The Minister of Transportation of The Republic of Indonesia PM 35 of 2016 on Fares for Transporting People By Train Economy Class Services To Carry Out Public Service Obligations. Technical Report, The Ministry of Transportation of The Republic of Indonesia.

Muñoz, J.C., Soza-Parra, J., Raveau, S., 2020. A comprehensive perspective of unreliable public transport services' costs. Transp. A: Transp. Sci. 16 (3), 734–748. http://dx.doi.org/10.1080/23249935.2020.1720861.

Nordbakke, S., Schwanen, T., 2014. Transport, unmet activity needs and wellbeing in later life: exploring the links. Transportation 42, 1129–1151. http://dx.doi.org/10.1007/s11116-014-9558-x.

Prelipcean, A.C., oz˝ o Gidófalvi, G., Susilo, Y.O., 2018. MEILI: A travel diary collection, annotation and automation system. Comput. Environ. Urban Syst. 70, 24–34. http://dx.doi.org/10.1016/j.compenvurbsys.2018.01.011.

Qi, Z., Lim, S., Rashidi, T.H., 2019. Assessment of transport equity to central business district (CBD) in sydney, Australia. Transp. Lett. 12, 246–256. http://dx.doi.org/10.1080/19427867.2019.1584694.

Schleinitz, K., Petzoldt, T., Franke-Bartholdt, L., Krems, J., Gehlert, T., 2017. The german naturalistic cycling study – comparing cycling speed of riders of different e-bikes and conventional bicycles. Saf. Sci. 92, 290–297. http://dx.doi.org/10.1016/j.ssci.2015.07.027.

Train, K., 1986. Qualitative Choice Analysis. MIT Press.

Train, K.E., Winston, C., 2007. Vehicle choice behavior and the declining market share of U.S. automakers. Internat. Econom. Rev. 48 (4), 1469–1496. http://dx.doi.org/10.1111/j.1468-2354.2007.00471.x.

Wagner, J.M., Eschbach, M., Vosseberg, K., Gennat, M., 2020. Travel time estimation by means of google API data. IFAC- Pap. 53 (2), 15434–15439. http://dx.doi.org/10.1016/j.ifacol.2020.12.2365.

Walton, D., Buchanan, J., 2012. Motorcycle and scooter speeds approaching urban intersections. Accid. Anal. & Amp; Prev. 48, 335–340. http://dx.doi.org/10.1016/j.aap.2012.02.001.

Whittington D, C.J., 2019. Valuing changes in time use in low- and middle-income countries. J. Benefit- Cost Anal. 10(Suppl 1), 51–72. http://dx.doi.org/10.1017/bca.2018.21.

Wu, H., 2018. Comparing google maps and uber movement travel time data. Transp. Find. http://dx.doi.org/10.32866/5115.

Wu, K., Chen, Y., Ma, J., Bai, S., Tang, X., 2017. Traffic and emissions impact of congestion charging in the central Beijing urban area: A simulation analysis. Transp. Res. Part D: Transp. Environ. 51, 203–215. http://dx.doi.org/10.1016/j.trd.2016.06.005.

Zannat, K., Choudhury, C.F., Hess, S., 2021. Modeling departure time choice of car commuters in Dhaka, Bangladesh. Transp. Res. Rec.: J. Transp. Res. Board http://dx.doi.org/10.1177/03611981211039840.

Zhang, Y., Cheng, L., 2023. The role of transport infrastructure in economic growth: Empirical evidence in the UK. Transp. Policy 133, 223–233. http://dx.doi.org/10.1016/j.tranpol.2023.01.017.