Estimating the Potential Modal Split of Any Future Mode Using Revealed Preference Data

de Clercq, G.K.; van Binsbergen, A.J.; van Arem, B.; Snelder, M.

DOI
10.1155/2022/6816851

Publication date
2022

Document Version
Final published version

Published in
Journal of Advanced Transportation

Citation (APA)
de Clercq, G. K., van Binsbergen, A. J., van Arem, B., & Snelder, M. (2022). Estimating the Potential Modal Split of Any Future Mode Using Revealed Preference Data. *Journal of Advanced Transportation*, *2022*, Article 6816851. https://doi.org/10.1155/2022/6816851

Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 6816851, 11 pages
https://doi.org/10.1155/2022/6816851

Research Article

Estimating the Potential Modal Split of Any Future Mode Using Revealed Preference Data

Gijsbert Koen de Clercq, Arjan van Binsbergen, Bart van Arem, and Maaike Snelder

Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, 2628 CN Delft, Netherlands

Correspondence should be addressed to Gijsbert Koen de Clercq; g.k.declercq@tudelft.nl

Received 28 April 2022; Revised 1 November 2022; Accepted 23 November 2022; Published 16 December 2022

Academic Editor: Arkatkar Shriniwas

Mode choice behaviour is often modelled by discrete choice models, in which the utility of each mode is characterized by mode-specific parameters reflecting how strongly the utility of that mode depends on attributes such as travel speed and cost, and a mode-specific constant value. For new modes, the mode-specific parameters and the constant in the utility function of discrete choice models are not known and are difficult to estimate on the basis of stated preferences data/choice experiments and cannot be estimated on the basis of revealed preference data. This paper demonstrates how revealed preference data can be used to estimate a discrete mode choice model without using mode-specific constants and mode-specific parameters. This establishes a method that can be used to analyze any new mode using revealed preference data and discrete choice models and is demonstrated using the OViN 2017 dataset with trips throughout the Netherlands using a multinomial and nested logit model. This results in a utility function without any alternative specific constants or parameters, with a rho-squared of 0.828 and an accuracy of 0.758. The parameters from this model are used to calculate the future modal split of shared autonomous vehicles and electric steps, leading to a potential modal split range of 24–30% and 37–44% when using a multinomial logit model, and 15–20% and 33–40% when using a nested logit model. An overestimation of the future modal split occurs due to the partial similarities between different transport modes when using a multinomial logit model. It can therefore be concluded that a nested logit model is better suited for estimating the potential modal split of a future mode than a multinomial logit model. To the authors' knowledge, this is the first time that the future modal split of shared autonomous vehicles and electric steps has been calculated using revealed preference data from existing modes using an unlabelled mode modelling approach.

1. Introduction

In the last decade, numerous mobility systems, such as shared bicycles and scooters, automated cars, ride-hailing services, electric bicycles, and other personal light electric vehicles have been developed. Such new mobility systems could potentially change the way our urban areas look substantially in terms of spatial use, sustainability, health, equity, safety, and economic opportunities [1–4]. For instance, it is estimated that the yearly impact of automated vehicles (AV) alone could approach 4,000 generalized US dollars per person per year, including economic benefits, crash cost savings, travel time reductions (due to a reduction in congestion), and lower parking costs [1].

A commonly accepted definition of new mobility systems does not exist in the literature. To define new mobility systems, it is important to consider what a mobility system entails and when such a system can be considered new. In our research, we define a mobility system as a set of components that, as such and as a whole, provide a means of transport for people and/or goods. Mobility systems are highly integrated into society and, therefore, challenging to analyze and describe due to their complex, large-scale, interconnected, open, and sociotechnical nature [5]. Systems can be differentiated and categorized on the basis of a multitude of attributes [6]. A mobility system can be considered a new mobility system in a specific area, if it substantially differs from already implemented mobility systems, such that mode choice changes can be expected when introduced. The novelty of a system is, therefore, relative and depends on the context: a system can already exist somewhere in the world but can be new for a specific area if its implementation differs from the implementation in other areas. For example, a metro system can be considered a new mobility system in one city, encouraging people to start using the metro instead of cars, whereas increasing the metro frequency of an already existing metro system in another city leads to a stronger competitive position and a (further) modal shift to the metro system in that city and is not considered a new mobility system. Another example is introducing shared bikes in an area where there is no local public transportation, allowing for last-mile trips, and enabling new public transport trips and tours. In this paper, a new mobility system is defined as follows: new mobility systems add value, such that mode choice changes significantly compared to already existing, implemented mobility systems in the researched area.

Introducing new mobility systems might lead to a change in accessibility (e.g., changing travel times and congestion), and this, in turn, can lead to a change in land use and activities. A conceptual model to describe this development has been proposed (see Figure 1). The model structures the dominant relationships found in the literature and is an adaptation of the LUT feedback cycle from Wegener [8]. New mobility systems are placed in the centre of the conceptual framework to represent the main source of effects on mode choice. When a new mobility system is deployed, the available transport options change, which can change mode choice behaviour (e.g., people use shared AVs instead of buses, trams, and metros) and thus modal split. A change in mode choice indicates improved accessibility (e.g., due to the high use of shared AVs, the average travel time decreases). If accessibility improves, urban areas might become more attractive (e.g., more people might move there) and be used more intensively in the long term as well, which will again put pressure on the transport systems and might evoke the need for new improvements. Note that this study focuses on the orange part of the conceptual framework and that the grey part is outside the scope.

Analyzing how mode choice behaviour could change when new mobility systems become available is challenging since potential users are not familiar with such systems yet. Mode choice is determined by numerous attributes that can roughly be separated into three categories: mobility system (e.g., costs), personal (e.g., age, gender, and income), and trip (e.g., origin and destination locations, trip purpose, and precipitation) attributes. Traditionally, attributes such as transport cost and transport speed are used to describe mobility systems. Additional attributes, such as type of ownership (e.g., buying, leasing), protection against weather, space for luggage, and availability in time, play a role in describing mobility systems as well [7]. New mobility systems can change the values of already identified attributes, but they can also introduce new attributes and, therefore, be described by appending and/or replacing attributes (e.g., car availability instead of car ownership) [9].

A way to avoid introducing implicit preferences towards existing modes when describing existing (and future) modes is to not use any mode-specific constants and parameters [10, 11]. Quandt and Baumal developed the so-called unlabelled mode modelling approach, in which mode choice is assumed to be explained only by attributes such as speed, frequency of service, comfort, and cost [10]. Their model does not include a mode-specific constant related to the perceived (partly unexplained) overall utility of a mode. The model describes a mode by merely looking at the type of service that travellers get for an unlabelled mode (e.g., "mode A" instead of labelling the mode as "a car", thereby avoiding the implicit inclusion of unidentified "car" attributes). Quandt and Baumal's exploratory study considers different modes of choice situations, which are characterized by different combinations of attributes such as speed, frequency of service, comfort, and cost. Their approach aims to expose the "true" trade-offs made by travellers between the attribute's levels. The unlabelled mode modelling approach has been applied in several papers. DeSalvo and Hug implemented Quandt and Baumal's approach to analyze the mode choice of existing modes and urban household behaviour by considering costs, commuting time, speed, and distance [12]. Malalgoda and Lim used a similar approach to research the use of existing public transit in the U.S. by considering the variables passenger miles, unlinked passenger trips, vehicle hours, operating employees, fuel, fare, income, and population [13]. Malalgoda and Lim used this approach because of its ability to consider continuous modes and, therefore, find mathematical optimums.

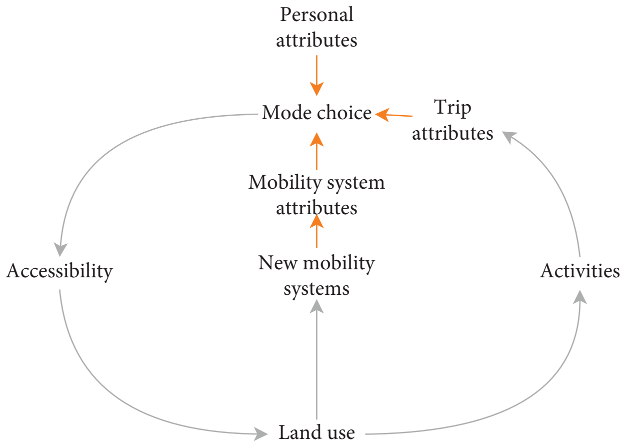

Figure 1: Dominant relationships of the effects of new mobility systems on mode choice. Adaptation from Wegener [7].

Based on the literature review, we expect that unlabelled mode modelling can be particularly useful to expose tradeoffs based on objectifiable attributes to travellers, allowing them to make choices between existing and non-existing modes for which the mode-specific constant cannot be known, such that the future modal split can be estimated. An important requirement for leaving out mode-specific constants and parameters is the availability of a complete and coherent set of attributes that can represent both existing and new mobility modes, and the assumption that travellers' valuation of the modes' attributes will not change when a new mobility system is introduced.

This paper demonstrates how revealed preference data and discrete choice models without mode-specific constants and parameters can be used to give insight into how new mobility systems could change mode choice. This method is demonstrated by calculating the future modal split of shared autonomous vehicles and electric steps. To the authors' knowledge, this is the first time that the future modal split of shared autonomous vehicles and electric steps using revealed preference data of existing modes is calculated. It also identifies knowledge gaps and possible pathways for future research on theories and methods to assess the impact of new mobility systems on mode choice.

2. Literature

A common practice to model the way people choose their transport mode is to use discrete choice models using a generalized utility function covering mobility system attributes and personal attributes [14], as shown in the following equation:

$$U_{ij} = \alpha_{ij} + \sum_{k=1}^{nk} \beta_{ijk}\chi_{jk} + \varepsilon_{ij}, \tag{1}$$

where α =mobility system and person-specific constant; β =estimated parameter; χ =mobility system attributes; i =persons (or clusters); j =mobility system; k =mobility system attributes; nk =number of mobility system attributes; and ε =error-term (excluding the mobility system constant α).

Understanding the generalized utility of new mobility systems is of vital importance to understanding how new mobility systems affect mode choice. For instance, a multitude of studies use nested logit models to model mode choice in the context of automated driving [15, 16], shared driving [17–21], and multimodal trips [22, 23]. These studies all make assumptions about mobility system-specific parameters (e.g., time is often valued differently in an automated car than in a conventional car) and mobility system-specific constants, which are used to capture effects that cannot be explained by the used mobility system attributes (e.g., a car has a higher level of status compared to taking the bus). These mode-specific constants can only be calibrated when using data for modes in which choice data is available, so they cannot be used when aiming to predict the modal share of new mobility systems. The multinomial logit model assumes that the attributes of all alternatives are orthogonal (no correlation between attributes). If this does not hold, then the so-called "red/blue box paradox" occurs when two alternatives are too similar, which leads to an overestimation of those alternatives. To overcome this overestimation, other types of discrete choice models, such as a nested logit model, can be used [24]. This introduces other model-specific scaling parameters that need to be estimated and (manually) estimated when adding a new mobility system. For a nested logit model, one scaling parameter to define to what extent the alternatives within a nest have independence from irrelevant alternatives (IIA) outside of the said nest needs to be defined [24]. This parameter is based on the similarity between the attributes of two alternatives and defines to what extent the nest behaves as a nest or as two alternatives (as in a multinomial logit model). The similarity between all attributes of the two alternatives can be defined by taking the normalized multidimensional distance between the two alternatives [25]. The similarity is defined as 1 minus the multidimensional distance. Subsequently, the mode with the highest similarity to the future mode and the future mode are put in one nest in a nested logit model. The following formula is given as follows:

$$\text{Dist} = \frac{\sum_{k=1}^{nk} \left|\chi_{\text{norm},1} - \chi_{\text{norm},2}\right|}{nk}, \tag{2}$$

where Dist =normalized multidimensional distance; k =mobility system attributes; nk =number of mobility system attributes; and $\chi_{\text{norm},i}$ =normalized mobility system attribute.

To create overlapping nests when alternatives do not fit in one nest in the nested logit model, a cross-nested model or a paired combinatorial logit model can be used. For a cross-nested logit, between two and the number of alternative scaling parameters that need to be estimated [24, 26, 27], which becomes complex quickly. When using a paired combinatorial logit model, $1 + 2 *$ number of alternatives (including future alternatives) and their scaling parameters need to be estimated, which comes down to 13 scaling parameters in the case of five existing alternatives and is computationally extremely difficult without preference data about the future mobility system [27]. The mobility system-specific parameters and constants and model-specific scaling parameters are ideally estimated using stated or revealed preference data. This is, however, challenging for new mobility systems, as explained in the next section.

Empirical research uses large-scale stated and revealed preference surveys to estimate the relevant parameters for modelling mode choice. Stated preference research can help to understand mode choice, but it can be challenging to determine how results from stated preference studies translate to the real world. This is because stated preference research is, by definition, based on a representation of reality, where certain (unknown) attributes are not taken into account in the research [28, 29]. Instead, revealed preference research helps to understand how people make choices in the real world, but it can only test how existing mobility systems are used. Revealed preference research helps to find out how and when people start to use new mobility systems, such that the change in mode choice and travel behaviour that can be analyzed is limited [28]. Although studies often try to analyze mode choice using pilots with sometimes limited implementations, they already give insight into how a new mobility system might be used in the real world [30, 31].

Stated and revealed preference research can be combined to analyze how new mobility systems might be used. Extrapolating revealed preferences (read: values of mobility system-specific constants and parameters) to a new set of mobility systems with a new (unused) alternative and, subsequently, normalizing these results using stated preference research is a way to combine stated and revealed preference research [29, 32]. This approach, however, includes implicit preferences by including values of mobility system-specific constants and parameters of the analyzed mobility systems to model mode choice, so when extrapolating this to new mobility systems, assumptions about implicit preferences are also carried over and influence the predicted modal split of the newly added mobility system.

3. Methodology

This paper first describes the way a utility function of a discrete choice model without mode-specific constants and parameters can be estimated on the basis of revealed preference data from OViN 2017 [33]. The paper then demonstrates that the modal split of a subset of current systems can be estimated on the basis of that function. The paper subsequently demonstrates how this approach can be used to estimate the modal share of additional (also new) modes, insofar as the main choice-determining characteristics of such a mode can already be experienced in current transport systems.

For the revealed preference dataset, it is assumed that all attributes are orthogonal (no correlation between attributes). Furthermore, since generalized utility functions are used without mobility system-specific parameters or constants, it must be assumed that people are familiar with all mobility systems and that initial familiarization and adoption have occurred. Therefore, we assume that, if a future new transport system can be described as a combination of already known transport system characteristics, we can calculate its mode choice and modal share.

Two studies demonstrating the method are performed to estimate the potential modal split of new mobility systems with (1) synthetic data and (2) revealed preference data. The algorithm below describes all the steps involved in using revealed preference data. The algorithm is the same for synthetic data, except for the first step of importing the dataset, which has been generated (Algorithm 1).

3.1. Synthetic Data. To demonstrate the method, synthetic data with a utility function with two main attributes is created. First, a utility function (see equation (3)) is defined to create a training (80%) and test (20%) dataset with 5 modes. All permutations of age, income, and distance are used to create the datasets and define the cost and time of each mode (see Tables 1 and 2), with 147,460 entries. After the training dataset is inserted into Biogeme [34], Biogeme estimates the two parameters $((\beta_{\text{time}}, \beta_{\text{cost}}))$ using a logit model where the probability of a certain mode choice is calculated (see equation (4), [34]). Subsequently, the parameters can be filled in the utility function to calculate the modal split using the test dataset. This calculated modal split with 5 modes can be compared to the training dataset. This comparison can be conducted by looking at how well the mode choices of the original synthetic dataset match the mode choices in the test set using rho-squared (see equation (5), [34]) and modal split, where it is expected that the performance of both indicators is (almost) perfect due to the synthetic nature of the data. Now, the attributes of a future mode can be added when calculating the modal split using the test dataset because the utility function is the same for each mode and the parameters are already estimated. When filling in the utility function for a future mode, the modal split including this future mode can be calculated. To verify this method and check if the code is behaving as expected, the calculated modal split based on the test dataset with 6 modes can be compared with the modal split of the synthetically generated test dataset with 6 modes.

$$U_{ij} = \beta_{\text{time}} * \text{age}_i * \text{time}_j + \beta_{\text{cost}} * \frac{200.000}{\text{income}_i} * \text{cost}_j, \tag{3}$$

where β = parameters; i = persons; and j = mobility system.

3.2. Revealed Data. A study demonstrating the method is performed to estimate the potential modal split of new mobility systems with revealed preference data, enriched with precipitation by TNO from OViN [33]. This labelled dataset was restructured to add 9 more mode attributes (see Table 3). The labelled dataset contains 75,043 entries with 11 personal attributes, 9 trip attributes, and 11 mode attributes, including 5 modes (car, carpool, transit (BTM), bicycle, and walk) and the mode choice for each entry. This dataset is shuffled and separated into a training (80% of entries) and a test (20% of entries) dataset. It was decided that a minimum acceptable performance (e.g., minimum rho-squared or accuracy) for the discrete choice model was to be defined by inserting the dataset into a latent class analysis (so without alternative specific constants or parameters). This was carried out to benchmark the minimum (and added) accuracy of a discrete choice model compared to a latent class analysis. Any performance lower than a latent class analysis was assumed to indicate that more "information was still embedded in the dataset that could predict mode choice."

Next, a k-means cluster analysis is performed to take into account personal and trip attributes by grouping similar entries into one cluster [35]. This dataset is fed to a multinomial logit model where the probability of a certain mode choice is calculated (see equation (3)) in Biogeme [34] using a predefined utility function with mode attributes from the dataset (see equation (1), where the mobility and person-specific constants are equal to 0) with randomized initial values of the parameters between −0.5 and 0.5. Note that this is performed for each cluster. In this way, personal and trip attributes (read: dummy variables) do not need to be included in the generalized utility function since similar attributes are already clustered [36].

Subsequently, the modal split of the 5 modes can be calculated by filling in the parameters of the utility function to calculate the modal split of the test dataset. Rho-squared (see equation (4)), precision (see equation (5)), recall (see equation 6), f1-score (see equation (7)), and accuracy (see equation (8)) were used to analyze the performance of the estimation.

$$E_{ij} = \frac{e^{U_{ij}}}{\sum_{k=1}^{nk} e^{U_{ik}}}, \tag{4}$$

```
Initialize
(1) Import full OViN dataset
(2) Perform latent class analysis to define a "minimum performance benchmark"
(3) Define clusters based on personal and trip attributes using $k$-means and elbow function in the full dataset
(4) Retrieve train (80%) and test (20%) dataset
(5) Define general utility function
    Estimate current modal split (with and without alternative-specific constant)
(6) Estimate parameters of the utility function of a discrete choice model with 5 modes per cluster using the train dataset
(7) Calculate the modal split of 5 modes per cluster in the test dataset
(8) Compare calculated modal split with recorded modal split in the full test dataset
    Estimate future modal split
(9) Define the attributes of future mode, incl. variations of ± 20% for sensitivity analysis (SA)
(10) Calculate the similarity of all modes and the future mode to estimate the scaling parameter in a nest (only for nested logit), see
     equation (2)
(11) Calculate modal split ranges (SA) of 6 modes per cluster in the test dataset using results of the modal split of step 6 (without
     alternative-specific constant)
(12) Create a Sankey diagram (excl. variations of ± 20%)
```

Algorithm 1: Estimate future modal split using revealed preference data.

Table 1: Synthetic dataset person and trip attributes.

Attribute	Range
Age (year)	18–90, steps of 1
Income (€/year)	10.000–200.000, steps of 10.000
Distance (km)	0.5–100.5, steps of 1

where U = generalized utility; i = persons; and j = mobility system; nk = number of mobility system; k = mobility system.

$$\rho^2 = 1 - \frac{\mathscr{L}^*}{\mathscr{L}^i}, \tag{5}$$

where $\mathscr{L}^*$ = final log-likelihood; and $\mathscr{L}^i$ = initial log-likelihood.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{6}$$

where TP = True positive; and FP = False positive.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{7}$$

where TP = True positive; and FN = False negative.

$$f1\,\text{score} = 2 * \frac{\text{precision} * r\text{ecall}}{\text{precision} + r\text{ecall}}, \tag{8}$$

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{TP} + \text{FN}}, \tag{9}$$

where; TP = True positive; TN = True negative; FP = False positive; and FN = False negative.

The modal split of the 5 modes is also calculated with an alternative-specific constant to see whether the performance changes and whether the approach without an alternative-specific constant is appropriate. Note that the rho-squared and the calculated modal split are based on the log-likelihood of a certain choice (the outcome of equation (3)), whereas precision, recall, f1-score, and accuracy do not consider the probability of a choice but merely the choice with the highest "generalized random utility." This calls for a thorough analysis and interpretation of each metric since the comparison of metrics is not trivial (e.g., accuracy cannot be compared with rho-squared).

When using a nested logit, the similarity between each mode and the future mode is calculated by normalizing the values of all attributes and calculating the so-called multidimensional distance between each mode (see equation (2), [25]). The distance between two modes for cost and time is calculated by taking the normalized squared difference, and for all other attributes, the absolute normalized difference is taken. Then, this value is divided by the number of attributes to determine the multidimensional distance. The similarity is defined as 1 minus the multidimensional distance. Subsequently, the mode with the highest similarity to the future mode and the future mode are put in one nest in a nested logit model.

Subsequently, a new mobility system is added, and the modal split of this mobility system is calculated using the same utility function and parameters as the estimated discrete choice model without the new mobility system. The values of the attributes of the new mobility system are varied within reasonable ranges (see Table 4) to find the ranges of the modal split when a new mobility system is introduced. This is to account for uncertainties and see which attributes of new mobility systems will affect the modal split.

4. Results

To define the minimum acceptable performance as described in the previous section, a latent class analysis was performed in *R* using the *mclust* package. The latent class analysis used the full dataset to estimate mode choice. The accuracy was 0.41, and the Brier score was 0.53. This will serve as a baseline to compare the accuracy of the discrete choice model. Note that the accuracy is based on the final mode choice without

Table 2: Synthetic dataset mobility system attributes.

Attribute	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	Future*
Average speed (km/hour)	60	50	40	30	20	25
Cost (€)	Distance/2	Distance/4	Distance/8	Distance/16	0	Distance/32
Time (hour)	Distance/average speed					

Table 3: Mode attributes based on OViN [33] and assumptions.

Mode attribute	Source and determination
Cost (€)	Car, transit, cycle, and walk from data; carpool = costs of car/2
Time (min)	Car, carpool, transit, cycle, and walk from data
Driving task (–)	Car, cycle = 1; carpool, transit, and walk = 0
Skills (–) (i.e., drivers license)	Car = 1; carpool, transit, cycle, and walk = 0
Weather protection (–)	Car, carpool, and transit = 1; cycle, walk = 0
Luggage (–)	Car, carpool = 1; transit = 0.5; cycle, walk = 0
Shared (–)	Car, carpool, transit, cycle, and walk from data
Availability (–)	Car = 1; carpool = 0.1; transit = urban density origin (from 1 to 5) ∗ urban density destination (from 1 to 5)/25; cycle = 1, walk = 1
Reservation (–)	Car, carpool, cycle, and walk = 1; transit = 0
Active (–)	Car, carpool, and transit = 0; cycle, walk = 1
Accessible (–)	Carpool, transit = 1; car, cycle, and walk = 0

taking into account probabilities (i.e., variations in individual choice behaviour), but can serve as a basis to compare performance.

In mode choice research, a wide range (0.20–1.00) of rho-squared (see equation (4)) seems to be acceptable as a result [37–39]. Using the standard in the field and the findings of the latent class analyses, it was decided that in this research, a rho-squared of 0.60 or higher and an accuracy of at least 0.45 will serve as the minimum performance requirements.

The results of the synthetic data are shown in Table 5. The estimation of the parameters in Biogeme resulted in a rho-squared of 0.998. As can be observed, the calculated modal split (columns 3 and 4) is the same as the modal split in the dataset (columns 1 and 2). The accuracy of the calculated modal split with 6 modes is 1.000. Therefore, it can be concluded that estimating future modal splits can work with a synthetic dataset.

The estimation of the parameters with a utility function without alternative specific constants and two parameters scaling the utility of cost and time in Biogeme using Python resulted in a rho-squared of 0.265. Since this rho-squared is considered too low, all mode attributes and the personal information of having a driving license in the dataset have been added as input as well, increasing the total number of parameters to 12. This resulted in a rho-squared of 0.540 and an accuracy of 0.663. To account for socioeconomic and trip-specific attributes without complicating the utility function by adding dummy variables and enhancing accuracy [36], 6 clusters were identified based on personal and trip attributes. This was carried out using a k-means clustering algorithm and the elbow method to determine the optimal number of clusters [40]. Three out of 6 clusters were based on trip purpose (business, home, and work). The three other clusters had a trip purpose of "other," where one cluster only contained trips with people that do not own a car and the other two clusters contained trips with people that own a car. These two final clusters were differentiated by the information that people are or are not the main car users.

Table 4: Ranges of mode attributes of new mobility systems.

Mode attribute	Range
Cost (€)	(0, 4, 8, 12, 16, 20) (0.05, 0.15, 0.25) ∗ distance (km)
Time (min)	Distance (km)/(10, 20, 40, 60, 80, 100)/60
Driving task (–)	(0, 1)
Skills (–) (i.e., drivers license)	(0, 1)
Weather protection (–)	(0, 1)
Luggage (–)	(0, 1)
Shared (–)	(0, 1)
Availability (–)	(0, 0.5, 1)
Reservation (–)	(0, 1)
Active (–)	(0, 1)
Accessible (–)	(0, 1)

Estimating the parameters of the utility function for each cluster resulted in a rho-squared of 0.828 and an overall accuracy of 0.758 (see Table 6). It can be observed that the performance metrics in Table 6 for modes with a larger modal split (i.e., car, cycle, and walk) are higher compared to modes with a smaller modal split (i.e., carpool and transit). Moreover, it can be observed that the total macro average f1-score is lower than the total weighted average f1-score, indicating the discrete choice model is optimized more for modes that have a larger modal split in the dataset. Note that the modal split in Table 7 is based on probabilities that a mode was chosen, and that the metrics in Table 6 are based on the final mode choice with the highest utility.

This study also demonstrates that the exclusion of an alternative specific constant in the utility function leads to a comparable result using the current 5 modes. Using a utility function with 12 parameters and 1 alternative specific constant (with the alternative specific constant of the car set to 0) leads to a rho-squared of 0.823 and an overall accuracy of 0.740; this is similar to the performance without an alternative specific constant. It should be noted that the values of the alternative specific constants vary between −1.57 and 1.27. Because of the similar performance between the discrete choice models with and without alternative specific constants, it was concluded that the effect of an alternative specific constant in this case, even including the mentioned outliers, is negligible, and therefore we can use the results without the alternative specific constant to calculate the future modal split.

Before estimating the future modal split with a multinomial logit model, the so-called "red/blue bus paradox" is tested by adding each mode as a future mode and subsequently calculating the total modal split for each mode (see Table 7). The largest difference is observed for the mode "cycling" (7.5 percentage point difference). A nested logit model is also estimated to overcome the "red/blue bus paradox." The modal shares of each model can be compared with each other to see whether the attributes of the modes are orthogonal, and a nested logit is needed to calculate the future modal split.

The estimation of these parameters is used to calculate the future modal split by calculating the modal split of each permutation of a future mode according to Table 4. The modal split of future modes ranges between 4.7% and 88%, with an average modal split of 45%. This means that by varying all attributes, a wide range of modal splits is found, which is to be expected since all possible combinations are included. From these results, one can find the modal split for any future mode by defining the attributes of this mode.

In this paper, two example future modes were defined to demonstrate the consequences of using a multinomial logit and a nested logit model. The first one is a shared autonomous car, and the second one is a rented electric step; their properties are defined in Table 8. Estimating the mode choices and modal split in the setting with the additional modes results in modal shares of 24% for the shared autonomous car and 37% for the electric step when using the multinomial logit model. When applying a nested logit model, first the nests are determined by taking the highest similarity index of an existing mode compared to both of the future modes (see Table 9). This resulted in putting the future mode-shared autonomous car in one nest with the carpool and the electric step in the same nest with the cycle. Application of the thus defined nested logit model resulted in an estimated modal share of 15% for the shared autonomous car and 33% for the electric step. Sensitivity analyses are performed to get a better understanding of how robust the calculated modal splits are. The sensitivity analysis is performed by varying all mode attributes that can be varied by ±20%. The results can be found in Tables 10 and 11. Sankey diagrams (see Figure 2) visualize how people's mode choice changes from the currently available modes and the future available modes using the standard values (i.e., not the varied mode attributes of the sensitivity analysis) for the nested logit model.

Table 5: Results synthetic data.

Mode	Modal split synthetic dataset		Calculated modal split	
	5 modes (%)	6 modes (%)	5 modes (%)	6 modes (%)
Mode 1	3.6	3.6	5.6	5.6
Mode 2	46	46	43	43
Mode 3	37	37	36	36
Mode 4	8.3	5.6	9.4	9.0
Mode 5	5.2	3.8	5.7	3.2
Future	—	4.1	—	3.2

Table 6: Results modal split estimation OViN data for 6 clusters.

Mode	Precision	Recall	f1-score
Car	0.670	0.971	0.792
Carpool	0.492	0.321	0.369
Transit	0.934	0.232	0.358
Cycle	0.857	0.691	0.728
Walk	0.898	0.944	0.917
Total macro average	0.708	0.514	0.533
Total weighted average	0.756	0.758	0.716
Overall accuracy		0.758	

5. Discussion

This study presents an approach for calculating the mode choice and modal split of new transport modes in a future situation in which such modes are well established using a discrete choice model without alternative specific constants, whose parameters are estimated based on revealed preference data. This study uses the examples of an electric step and a shared autonomous car to explore this method. First, the accuracy of this method is discussed. Then, it is discussed if a multinomial logit or nested logit model can better calculate the modal share of a future mode by taking into account the so-called "red/blue bus paradox." Finally, some assumptions and computational challenges are scrutinized.

As expected, the accuracy is higher (0.76) for the final estimation with 12 parameters and 6 user clusters than when performing the latent class analysis (0.41). For currently known modes, it is demonstrated that using an alternative specific constant in the utility function does not produce significantly different results than our approach. Therefore, it can be concluded that the unlabelled mode choice modelling approach is valid for this dataset.

When applying the estimated utility logit function to predict future mode choice, it can be observed that the future modal shares of the new modes seem to be relatively high when using a multinomial logit model. This could be an overestimation caused by a violation of the IIA assumption, i.e., some modes in the model are regarded as being completely different while in fact there are partially overlapping characteristics. Due to its formulation, the model tends to overestimate the mode choice of such overlapping modes.

Table 7: Results modal split double mode (“red/blue bus paradox”).

	Original (%)	Car 2x (%)	Carpool 2x (%)	Transit 2x (%)	Cycle 2x (%)	Walk 2x (%)
Car	33	40	30	32	30	33
Carpool	14	11	20	13	12	13
Transit	4.9	4.0	4.4	7.8	3.9	4.8
Cycle	30	27	28	29	37	29
Walk	18	18	18	18	18	19
Max. difference between original and 2x		6.7	6.4	2.8	7.5	1.3

Table 8: Attributes shared by autonomous cars and electric steps are listed with sensitivity analysis values between square brackets based on a 20% variation of all mode attributes.

Mode attribute	Shared autonomous car	Electric step
Cost (€)	0.05 (0.04 0.045 0.05 0.055 0.06) per km	4 (3.20 3.60 4.00 4.40 4.80)
Time (min)	Distance (km)/60 (48 54 60 66 72) (km/hr)/60	Distance (km)/10 (8 9 10 11 12) (km/hr)/60
Driving task (−)	0	1
Skills (−) (i.e., drivers license)	0	0
Weather protection (−)	1	0
Luggage (−)	1	0
Shared (−)	0	0
Availability (−)	0.5 (0.4 0.45 0.5 0.55 0.6)	1 (0.8 0.9 1)
Reservation (−)	1	0
Active (−)	0	1
Accessible (−)	1	0

Table 9: Similarity index for each mode compared to the future mode.

Future	Car	Carpool	Transit	Cycle	Walk
Shared autonomous car	0.62	0.82	0.66	0.48	0.52
Electric step	0.49	0.22	0.35	0.88	0.75

Table 10: Calculated modal split electric step with sensitivity analysis ranges between brackets based on a 20% variation of all mode attributes.

Mode	Current modal split (%)	Future modal split	
		MNL (%)	NL (in nest with bicycle) (%)
Car	33	20 (17–20)	24 (20–24)
Carpool	14	9.5 (8.8–9.5)	9.7 (8.9–9.7)
Transit	4.9	2.1 (1.8–2.1)	2.4 (2.0–2.4)
Cycle	30	19 (17–19)	18 (16–18)
Walk	18	13 (12–13)	14 (13–14)
Shared electric step	—	37 (37–44)	33 (33–40)

Table 11: Calculated modal split for a shared autonomous car with sensitivity analysis ranges between brackets based on a 20% variation of all mode attributes.

Mode	Current modal split (%)	Future modal split	
		MNL (%)	NL (in nest with carpool) (%)
Car	33	25 (22–25)	31 (29–31)
Carpool	14	10 (9.5–10)	9.0 (8.8–9.0)
Transit	4.9	3.6 (3.3–3.6)	4.1 (3.9–4.1)
Cycle	30	23 (21–23)	25 (23–25)
Walk	18	15 (14–15)	16 (15–16)
Shared autonomous car	—	24 (24–30)	15 (15–20)

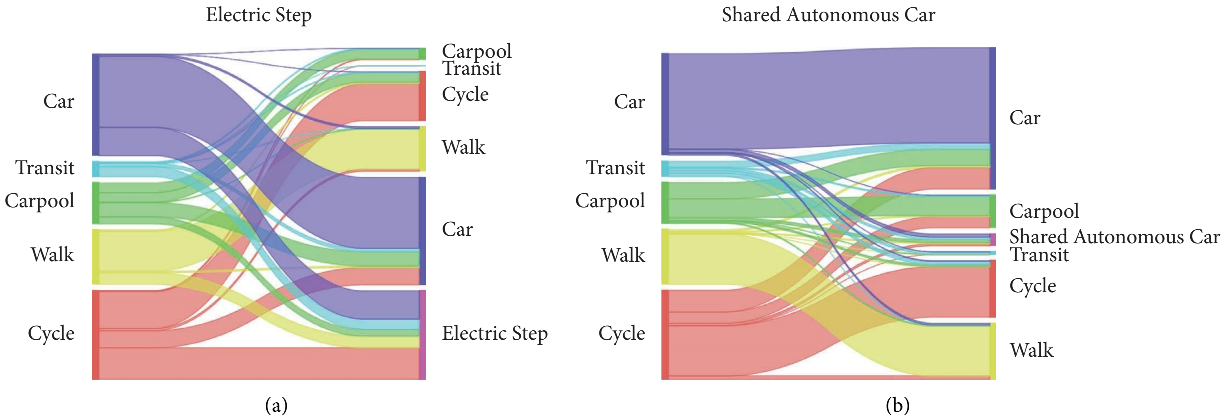

Figure 2: Sankey diagrams of current mode choice (a) and future mode choice (b) on each diagram for the nested logit model. (a) Electric step; (b) a shared autonomous car.

The overestimation when using "double modes" is quite substantial, up to 7.5 percentage points in the multinomial logit model with this dataset.

To overcome the similarity issue, a nested logit was also implemented. Using that approach, the future modal shares seem to be more modest, with up to 9 and 4 percentage points lower modal splits for shared autonomous cars and electric steps, respectively. The nest in this nested logit model consists of the future mode and the most similar existing mode, which are determined by calculating the multidimensional distance of each pair of modes [25]. Knowing from the "red/blue bus paradox" that an overestimation of the future modal split occurs when using a multinomial logit model, it can be concluded that the nested logit is the preferred discrete choice model in this case.

What should be noted as well is that the attributes (values) that can be derived from empirical data from the current mobility systems do not necessarily properly represent the attributes for future systems (e.g., what exactly is shared) and that new attributes might become significant that are not currently measured (e.g., the fear of autonomous driving). Moreover, preferences are changing over time; one example would be the changing trend that people start leasing more cars instead of owning them. This should be taken into account when interpreting these results, extrapolating them to other modes, or changing traveller preferences.

Calculating the future modal splits using all of the presented combinations (41,472) requires a lot of calculations and can take a lot of computation time (up to 7 days) on a MacBook Pro with a 2,4 GHz Quad-Core Intel Core i5 and 8 GB of RAM. For the presented future modes, fewer combinations (up to 125) are tested, and the computation times remain relatively limited (up to 30 minutes). To use this approach in workshops with policymakers or stakeholders, it is recommended to implement a Monte Carlo estimation instead of a test set to reduce the computation time even more. To achieve this, the distribution of variables (i.e., personal, trip, and future mode attributes) in the training set needs to be determined to create the input for the Monte Carlo simulation.

6. Conclusions and Future Research

This study successfully explores an approach for calculating the mode choice and modal split of new transport modes in a future situation when such modes are well established. This is achieved by calculating the modal split of two future modes (shared autonomous car and electric step). This is carried out by estimating a multinomial logit model and a nested logit model without alternative specific constants and parameters, such that this utility function can be used to calculate the modal split of a future mode. Note that the main characteristics determining the choice of future transport modes are already experienced in current transport systems. This study demonstrates that using a utility function without any alternative specific constants or parameters resulted in a rho-squared of 0.828 and an overall accuracy of 0.758 when using clusters to group similar people and similar trips. The approach is applied to a dataset based on empirical data (OVIN [33]) with 5 existing modes and 2 future modes, where each future mode is analyzed separately.

When predicting the modal split of a future mode using a multinomial logit model, it might be concluded that an overestimation of the future modal split occurs due to the partial similarities between different transport modes. For this reason, this study also implemented a nested logit model, which can solve this challenge and be generalized by automatically nesting the future mode in a nest with the "most similar" existing mode. It can be concluded that a nested logit model is better suited for estimating the potential modal split of a future mode than a multinomial logit model.

Mixed logit models can overcome the methodological shortcomings (assumption of IIA, unobserved preferences, and individual preferences over time) of both MNL and NL. The main aim of this study is to demonstrate that revealed data preferences can be used to calculate the potential modal share of a future mode using a discrete choice model without an alternative-specific constant. The distributions for each mode attribute coefficient would need to be assumed in order to cope with the open-form expression of a mixed logit. Future studies can extend this approach by comparing a mixed logit model with the multinomial and nested logit models.

Further exploration can be carried out with other types of discrete choice models (e.g., cross-nested logit, paired combinatorial logit) to get a better grasp on the calculation of the modal split of future modes. The main challenge with modelling these more detailed discrete choice models is that multiple scaling parameters need to be simultaneously estimated for the future mode, for which there is no revealed preference data available.

As demonstrated in this study, different future modes can be analyzed based on their attributes alone. This also means this approach has a practical application in policymaking. Specifically, subsidies and tax reductions can be analyzed for existing and future modes by reducing, e.g., the value of the cost attribute for future autonomous cars, increasing the cost for conventional cars, or calculating the needed capacities for (new) modes and their infrastructure. Several combinations of policies and available modes can be analyzed and combined into multiple scenarios to help policymakers make effective policies.

And lastly, it is recommended to connect this modal split model to a traffic assignment model to see how the second- and third-order aspects change (e.g., activities, accessibility, and land use).

Data Availability

The used OViN 2017 dataset with trips in the Netherlands can be requested for free at https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:103498. The data source can be cited as the Centraal Bureau voor de Statistiek (CBS); Rijkswaterstaat (RWS) (2017): Onderzoek Verplaatsingen in Nederland 2017—OViN 2017. DANS. https://doi.org/10.17026/dans-xxt-9d28.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research is part of the project SUMMALab, which was supported by the research program Sustainable Living Labs from the Dutch Research Council (NWO) (Grant no. 439.18.460 B).

References

[1] D. J. Fagnant and K. Kockelman, “Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations,” *Transportation Research Part A: Policy and Practice*, vol. 77, pp. 167–181, 2015.

[2] S. Shaheen, A. Cohen, N. Chan, and A. Bansal, “Sharing strategies: carsharing, shared micromobility (bikesharing and scooter sharing), transportation network companies, microtransit, and other innovative mobility modes,” in *Transportation, Land Use, and Environmental Planning*, pp. 237–262, Elsevier, Amsterdam, Netherlands, 2020.

[3] D. Milakis, B. Van Arem, and B. Van Wee, “Policy and society related implications of automated driving: a review of literature and directions for future research,” *Journal of Intelligent Transportation Systems*, vol. 21, no. 4, pp. 324–348, 2017.

[4] B. van Arem, A. A. Ackerman, T. Chang et al., *Building Automation into Urban and Metropolitan Mobility Planning*, vol. 1, Springer International Publishing, Heidelberg, Germany, 2019.

[5] J. M. Sussman, P. A. Mostashari, N. Stein, S. J. Carlson, and R. Westrom, *The CLIOS Process*, Vol. 88, Special Edition for the East Japan Railway Company, Yoyogi, Shibuya, Tokyo, 2014.

[6] M. V. I. En Waterstaat, “Kerncijfers mobiliteit,” *Mobiliteitsbeeld 2017-Mobiliteitsbeeld en Kerncijfers Mobiliteit-Kennisinstituut voor Mobiliteitsbeleid*, vol. 20, 2018.

[7] M. N. Bagley and P. L. Mokhtarian, “The impact of residential neighborhood type on travel behavior: a structural equations modeling approach,” *The Annals of Regional Science*, vol. 36, no. 2, pp. 279–297, 2002.

[8] M. Wegener, “Overview of land use transport models,” *In Handbook of transport geography and spatial systems*, Emerald Group Publishing Limited, vol. 5, pp. 127–146, Bingley, England, 2004.

[9] A. Soteropoulos, M. Berger, and F. Ciari, “Impacts of automated vehicles on travel behaviour and land use: an international review of modelling studies,” *Transport Reviews*, vol. 39, no. 1, pp. 29–49, 2019.

[10] R. E. Quandt and W. J. Baumal, “The abstract mode model: theory and measurement,US Department of Commerce,” *Office of the Under Secretary of Commerce for Transportation*, Transport Systems Planning Division, Northwestern University in Evaston, Illinois, 1966.

[11] W. Jin, H. Jiang, Y. Liu, and E. Klampfl, “Do labeled versus unlabeled treatments of alternatives’ names influence stated choice outputs? Results from a mode choice study,” *PLoS One*, vol. 12, no. 8, Article ID e0178826, 2017.

[12] J. S. DeSalvo and M. Huq, “Mode choice, commuting cost, and urban household behavior,” *Journal of Regional Science*, vol. 45, no. 3, pp. 493–517, 2005.

[13] N. Malalgoda and S. H. Lim, “Do transportation network companies reduce public transit use in the U.S,” *Transportation Research Part A: Policy and Practice*, vol. 130, pp. 351–372, 2019.

[14] D. Ortuzar and L. Willumsen, *Modelling transport*, John wiley & sons, Hoboken, N J, USA, 2011.

[15] M. Snelder, I. Wilmink, J. van der Gun, H. J. Bergveld, P. Hoseini, and B. van Arem, “Mobility impacts of automated driving and shared mobility,” *European Journal of Transport and Infrastructure Research*, vol. 19, no. 4, 2019.

[16] S. Puylaert, M. Snelder, R. van Nes, and B. van Arem, “Mobility impacts of early forms of automated driving – a system dynamic approach,” *Transport Policy*, vol. 72, pp. 171–179, 2018.

[17] N. S. Daisy, H. Millward, and L. Liu, “Trip chaining and tour mode choice of non-workers grouped by daily activity patterns,” *Journal of Transport Geography*, vol. 69, pp. 150–162, 2018.

[18] F. Zhou, Z. Zheng, J. Whitehead, R. K. Perrons, S. Washington, and L. Page, “Examining the impact of car-sharing on private vehicle ownership,” *Transportation Research Part A: Policy and Practice*, vol. 138, pp. 322–341, 2020.

[19] C. F. Choudhury, L. Yang, J. de Abreu e Silva, and M. Ben-Akiva, “Modelling preferences for smart modes and services: a

case study in Lisbon," *Transportation Research Part A: Policy and Practice*, vol. 115, pp. 15–31, 2018.

[20] K. Ikezoe, E. Kiriyama, and S. Fujimura, "Car-sharing intention analysis in Japan by comparing the utility of car ownership for car-owners and non-car owners," *Transport Policy*, vol. 96, pp. 1–14, 2020.

[21] K. Winter, O. Cats, K. Martens, and B. van Arem, "Identifying user classes for shared and automated mobility services," *European Transport Research Review*, vol. 12, no. 1, p. 36, 2020.

[22] A. Fan, X. Chen, and T. Wan, "How have travelers changed mode choices for first/last mile trips after the introduction of bicycle-sharing systems: an empirical study in beijing, China," *Journal of Advanced Transportation*, vol. 2019, Article ID 5426080, 16 pages, 2019.

[23] P. H. L. Bovy and S. Hoogendoorn-Lanser, "Modelling route choice behaviour in multi-modal transport networks," *Transportation*, vol. 32, no. 4, pp. 341–368, 2005.

[24] K. E. Train, *Discrete choice methods with simulation*, Cambridge university press, Cambridge, England, 2003.

[25] E. Baikousi, G. Rogkakos, and P. Vassiliadis, "Similarity measures for multidimensional data," in *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, pp. 171–182, Hannover, Germany, April 2011.

[26] M. Bierlaire, "A theoretical analysis of the cross-nested logit model," *Annals of Operations Research*, vol. 144, no. 1, pp. 287–300, 2006.

[27] C. G. Prato, "Route choice modeling: past, present and future research directions," *Journal of Choice Modelling*, vol. 2, no. 1, pp. 65–100, 2009.

[28] E. Cherchi and J. d. D. Ortúzar, "On fitting mode specific constants in the presence of new options in RP/SP models," *Transportation Research Part A: Policy and Practice*, vol. 40, no. 1, pp. 1–18, 2006.

[29] A. Daly and C. Rohr, *Forecasting demand for new travel alternatives*, TheorY Found. Travel Choice Model, pp. 451–471, The Hague Consulting Group in the Hague, Netherlands, 1998.

[30] E. Anagnostopoulou, J. Urbancic, E. Bothos et al., "From mobility patterns to behavioural change: leveraging travel behaviour and personality profiles to nudge for sustainable transportation," *Journal of Intelligent Information Systems*, vol. 54, no. 1, pp. 157–178, 2020.

[31] N. Mundorf, C. A. Redding, and A. L. Paiva, "Sustainable transportation attitudes and health behavior change: evaluation of a brief stage-targeted video intervention," *International Journal of Environmental Research and Public Health*, vol. 15, no. 1, p. 150, 2018.

[32] A. Polydoropoulou and M. Ben-Akiva, "Combined revealed and stated preference nested logit access and mode choice model for multiple mass transit technologies," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1771, no. 1, pp. 38–45, 2001.

[33] B. Wouters and J. V. Brakel, "Simulatiestudie naar methodebreuken in het onderzoek verplaatsingen in nederland," *InColloquium Vervoersplanologisch Speurwerk (CVS)*, vol. 39, 2017.

[34] M. Bierlaire, "A short introduction to Pandas Biogeme, Technical report TRANSP-OR 200605," Transport and Mobility Laboratory, ENAC, EPFL, Toulouse, France, 2020.

[35] A. H. Shoabjareh, A. R. Mamdoohi, and T. Nordfjærn, "Analysis of pedestrians' behaviour: a segmentation approach based on latent variables," *Accident Analysis & Prevention*, vol. 157, Article ID 106160, 2021.

[36] L. Ding and N. Zhang, "A travel mode choice model using individual grouping based on cluster analysis," *Procedia Engineering*, vol. 137, pp. 786–795, 2016.

[37] M. B. Haque, C. Choudhury, S. Hess, and R. C. dit Sourd, "Modelling residential mobility decision and its impact on car ownership and travel mode," *Travel Behavior Society*, vol. 17, pp. 104–119, 2019.

[38] P. Ashkrof, G. Homem de Almeida Correia, O. Cats, and B. van Arem, "Impact of automated vehicles on travel mode preference for different trip purposes and distances," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2673, no. 5, pp. 607–616, May 2019.

[39] D. Ton, D. C. Duives, O. Cats, S. Hoogendoorn-Lanser, and S. P. Hoogendoorn, "Cycling or walking? Determinants of mode choice in The Netherlands," *Transportation Research Part A: Policy and Practice*, vol. 123, pp. 7–23, May 2019.

[40] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-means clustering method and elbow method for identification of the best customer profile cluster," *IOP Conference Series: Materials Science and Engineering*, vol. 336, no. 1, Article ID 012017, 2018.