Quantifying Parking Difficulty with Transport and Prediction Models for Travel Mode Choice Modelling

Maciej Grzenda[1 [0000−0002−5440−4954]], Marcin Luckner[1 [0000−0001−7015−2956]], and Łukasz Brzozowski[1 [0000−0002−3625−3312]]

Warsaw University of Technology, Faculty of Mathematics and Information Science, ul. Koszykowa 75, 00-662 Warszawa, Poland
{Maciej.Grzenda, Marcin.Luckner, Lukasz.Brzozowski}@pw.edu.pl

Abstract. Promoting sustainable transportation necessitates understanding what makes people select individual travel modes. Hence, classifiers are trained to predict travel modes, such as the use of private cars vs bikes for individual journeys in the cities. In this work, we focus on parking-related factors to propose how survey data, including spatial data and origin-destination matrices of the transport model, can be transformed into features. Next, we propose how the impact of the newly proposed features on classifiers trained with different machine learning methods can be evaluated. Results of the extensive evaluation show that the features proposed in this study can significantly increase the accuracy of travel mode choice predictions.

Keywords: Machine learning · travel mode choices · survey data

1 Introduction

Accurately modelling travel mode choices (TMCs) is an important part of transportation planning [2, 7]. It makes it possible to predict, based on multiple features such as a person's age, journey distance, and whether the person owns a car, whether the journey is likely to be made by e.g. walking, or by private car. An overview of factors impacting which mode of transportation is selected by individuals to move around can be found *inter alia* in [7]. However, the data used so far to predict travel modes do not include parking-related features even in comparative studies of different classifiers [2], or include a very limited set of features such as the parking permit feature used in [7]. Importantly, mobility is expected to be influenced, among other things, by the financial costs of mobility alternatives, which include but are not limited to parking costs [1]. Another factor related to comfort and total travel time is the time it takes to find a parking space. However, how to estimate parking difficulty for mode choice modelling is an open issue.

Hence, this work focuses on how the features capturing parking difficulty can be calculated. The proposed methods rely on the demand matrices of a

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_40

traffic model and predicted time of finding a parking space and getting from it to the ultimate journey destination. All these methods provide features used by machine learning (ML) models. To evaluate these features, we investigate their impact on mode choice models trained with different classification methods. The results obtained with three datasets documenting real journeys show that the features proposed in this study can help predict transport modes used for the analysed journeys.

The remainder of this work is organised as follows: in Sect. 2 the use of ML for TMC modelling is summarised. Novel parking features are proposed in Sect. 3, and evaluated in Sect. 4. Conclusions are made in Sect. 5.

2 Related works

The travel mode prediction problem is frequently stated as a classification problem [4, 2]. A multitude of methods has been used for the task, including but not limited to Support Vector Machine (SVM), Random Forest (RF), and XGBoost (XGB). Systematic reviews of ML methods for modelling passenger mode choice can be found in [4, 7, 2]. An important aspect of studies on TMC modelling is the data used to evaluate the models. Salas et al. in their recent work on TMC modelling with ML methods [7] observe that most of the previous studies focus on a single empirical dataset. Moreover, datasets typically used in the works, such as Dutch National Travel Survey data [8], include features such as distance travelled, age, education, and land use index, but not parking features related to journey destination. Salas et al. in [7] consider four datasets, out of which only one refers to parking issues by including the parking permit feature. An extensive review of datasets made in [4] shows that most datasets rely on trip diaries. Still, the issue of parking-related factors and their impact on mode choices is gradually being addressed. In [1], the impact of parking fees on TMC in urban environments is analysed. In [9], a fixed average parking time for the entire city is assumed. Many other works do not consider parking difficulties in TMC prediction [10, 6, 3] or consider features possibly related to parking difficulties such as population density [10], global traffic congestion [5] or trip density [8].

The impact of individual features on TMC models is frequently analysed. Trip distance, travellers' age, number of cars/bicycles owned, and trip density were among the predictors influencing the predictions of the models in [8]. We aim to extend extant TMC research by proposing and evaluating parking features.

3 Estimating parking difficulty and costs

The goal of this work is to provide a proposal for parking-related features that would help explain some TMCs. Let $X = \{(\mathbf{x}_1, M_1), \ldots, (\mathbf{x}_N, M_N)\}$ denote the set of journeys $\mathbf{x}_i$ for which travel mode M_i actually used in the past is known. Our objective is to extend the $\mathbf{x}$ vectors by appending parking-related features i.e. use for the training and evaluation of travel mode prediction models vectors $[\mathbf{x}_i, f_1(\mathbf{x}_i), \ldots, f_F(\mathbf{x}_i)]$ including both original and F parking-related features.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_40

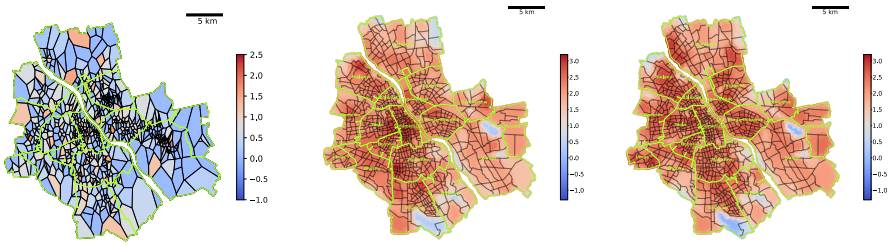
Fig. 1: Points, areas, and distances considered for parking difficulty estimation.

The estimation of parking time For some cities, data documenting the *parking time* t_p, which we define as the approximate time needed to park a car and get from its location to the actual journey destination P is available. The higher the parking time at location P is, the more reluctant travellers may be to use private cars to travel to this destination. Hence, we propose building a regression model predicting the parking time based on records documenting t_p for individual locations (x, y). Let T be the set of parking time tuples (x, y, h, t_p), where x and y stand for the geocoordinates of the journey destination, h denotes the hour of the day at which the destination was reached and t_p denotes the parking time reported by a person reaching the destination by car. Next, a regression model $\mathcal{M}_{\text{REG}}(x, y, h)$ can be developed to predict parking time t_p and provide the value of the `PRED_PTIME` feature based on input data $(x, y, h(t))$, where t denotes the approximate time of reaching the destination present in journey record $\mathbf{x}$ and $h(t)$ denotes the hour of the day.

The estimation of parking difficulty and cost We propose three methods to estimate parking difficulty, which we will explain using Figure 1. Let $P = (x, y)$ be a point at which we wish to estimate the parking difficulty parameter, i.e. the journey destination in which a car ideally would be parked if used for the journey. Let $B(P, R)$ be a circle around the point P with radius R, which defines the approximate area likely to be considered by a driver to park a car. Let $Z_1, Z_2, \ldots, Z_n$ be the transport zones (polygons) that have non-empty intersections with the ball $B(P, R)$. For all zones, we denote the average number of arrivals to the zone Z at hour $h \in \{0, 1, \ldots, 23\}$ by $A_h(Z)$. Next, let $I_1, I_2, \ldots, I_n$ denote the area of the intersection of zone Z_i with the ball $B(P, R)$. Moreover, let $c_{Z_1}, c_{Z_2}, \ldots, c_{Z_n}$ denote centroids of the corresponding zones, and $c_{I_1}, c_{I_2}, \ldots, c_{I_n}$ centroids of the intersections with the corresponding zones. Let $d(\circ, \diamond)$ denote the distance between two points. Using the above notation, we define the parking difficulty feature $f_m()$ for the point $P = (x, y)$ at time t using the method m as $f_m(P, t) = \frac{1}{n} \sum_{i=1}^{n} W_m\big(Z_i, B(P, R)\big) \cdot A_h(t)(Z_i)$.

We propose three methods m to calculate the weight $W_m(Z_i, B(P, R))$. The first provides the `PDIFF_IS_AREA` feature in which the weight is proportional to the area of the intersection I_i, i.e., $W_{IA}(Z_i, B(P, R)) = I_i$. The second feature is `PDIFF_IS_CENTR`, where the weight is proportional to the area of the

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_40



(a) Voronoi diagram of reported parking time.

(b) Parking difficulty at hour $h = 8$.

(c) Parking difficulty at hour $h = 17$.

Fig. 2: Spatial distribution of parking time and difficulty values. $\log_{10}$ scale

intersection I_i and inversely proportional to the distance between the parking point and the intersection centroid, i.e., $W_{IC}(Z_i, B(P,R)) = \frac{I_i}{d(P,c_{I_i})}$. Finally, in `PDIFF_ZONE_CENTR`, the weight is proportional to the area of the intersection I_i and inversely proportional to the distance between the parking point and the zone centroid, i.e., $W_{ZC}(Z_i, B(P,R)) = \frac{I_i}{d(P,c_{Z_i})}$.

These features can be calculated for any point P in the entire city area, covered by the transport model. The features differ in how they quantify the influence of different zones. Most likely the decisions of travellers will be more affected by expected parking difficulties in the area closer to the journey destination. This is why features giving different weights to data from different zones are proposed. To illustrate both the data used to calculate the parking time features and selected `PDIFF_IS_CENTR` values, sample values for the City of Warsaw are provided in Figure 2.

The feature estimating parking cost takes as an input arrival date and time s_a, departure date and time s_d, and the journey destination (x, y) i.e. the location at which a car could ideally be parked, all coming from a journey record $\mathbf{x}$. Based on the parking pricing policies of the city and these input data, the value of the `PCOST` feature is calculated. This yields only estimated parking costs, as factors such as the use of private parking spaces could influence actual parking costs.

4 Results

Reference data We selected three journey datasets to evaluate the methods proposed in this study. The datasets were collected in the City of Warsaw in 2022 and document journeys made by a representative sample of the parents of primary school children (PAR_W1 and PAR_W2 datasets) and a representative sample of all citizens of Warsaw (CIT_W1 dataset). The PAR_W1 includes journeys made by parents of children from three reference schools, while PAR_W2 includes journeys of a representative sample of parents of children from all schools in the city. All the datasets were prepared based on the travel diaries of respondents and included all journeys of respondents within the City of Warsaw irrespective of origin and destination i.e. also including non-school journeys.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_40

Table 1: Summary of datasets used for training and evaluation of TMC models.

Dataset	Respond.	Journeys	CAR	PT	BIKE	WALK
PAR_W1	523	1861	895 (48.54%)	344 (18.66%)	89 (4.83%)	516 (27.98%)
PAR_W2	316	798	323 (41.15%)	232 (29.55%)	18 (2.29%)	212 (27.01%)
CIT_W1	1170	2961	1044 (36.43%)	1181 (41.21%)	109 (3.80%)	532 (18.56%)

Table 1 presents the number of instances and the share of transport mode classes in the datasets. There are four modes considered in this work including public transport (PT). The survey answers were not limited to these modes, but other modes were rarely reported, which resulted in an insufficient number of examples. The raw journey records describe a respondent, i.e. education, gender, and year of birth; and information about reported journeys, such as origin, destination, departure time, and aim. Additional features were calculated using OpenTripPlanner[1] and include, separately calculated for each of the analysed travel modes, features such as distance, duration, waiting time (for PT), or estimated travel duration considering street congestion (for CAR). In this way, the journey records $\mathbf{x}_i$ used in the remainder of this work were obtained.

The calculation of parking features During the surveys, respondents were asked how long it took them to find a parking space. Answers to this question were given only if someone travelled by car. The direct usage of this data would create data leakage and bias the results. To avoid this, the missing parking time for the remaining journeys $\mathbf{x}_i$ was imputed using estimation from the k nearest neighbours (kNN) model and multivariate imputation made by the MICE algorithm overwriting the original values. Moreover, as a reported parking time equal to zero may mean not providing true data or not leaving a car at the destination at all, two attempts were used to treat zeros. While the first one used $t_p = 0$ in parking time tuples as correct values, the second one considered zeros as missing values to be imputed with proper values. The kNN-based regression model was used to predict parking time based on the set T. During optimisation, $k = 9$ was selected as an optimal value for all datasets. The average mean absolute error (MAE), calculated on the known values, reached 5.07 and 5.39 minutes for the estimation with and without zero parking times, respectively.

The MICE algorithm cannot be used directly to create an estimation model. The algorithm calculates only new values to replace missing data. This is done using multiple imputations, in our case, based on the parking attempt's longitude, latitude and hour. Therefore, the algorithm had to be modified. In the first step, the predictive mean matching was used 5 times to calculate new values. Next, all original values were removed, and the imputation was applied again (with the same parameters) to overwrite the original data. The errors obtained by the MICE algorithm are higher and more diverse than for kNN. For the dataset with zeros considered as proper values, the MAE reached 6.46 minutes;

[1] https://www.opentripplanner.org/

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_40

Table 2: Summary of parking feature sets

Feature set	Features used
BASELINE	$\mathbf{x}$
C_TIME	$[\mathbf{x}, \texttt{PCOST}(), \texttt{PRED\_PTIME}()]$
C_DIFF	$[\mathbf{x}, \texttt{PCOST}(), \texttt{PDIFF\_IS\_CENTR}(), \texttt{PDIFF\_IS\_AREA}(), \texttt{PDIFF\_ZONE\_CENTR}()]$
C_TIME_DIFF	BASELINE ∪ C_TIME ∪ C_DIFF

after removing such instances, i.e. once zero parking time values were removed and imputed with MICE, the error exceeded 8.15 minutes. Thus, four parking time features `PRED_PTIME` were provided, i.e. two by kNN and two by MICE. Finally, radius R was set to 1000 meters to calculate parking difficulty features. This value reflects the distance for which walking is used most of the time.

Algorithm 1: The evaluation of the importance of parking features

Input: D - matrix of n feature vectors, $P \in \mathbb{R}^n$ - vector of corresponding n transport modes, K - the number of CV folders, r - the number of runs

begin
  for $i = 1, \ldots r$ **do**
    $\{D_j, P_j\}_{j=1,\ldots,K}$ = DivideSetUsingStratifiedCrossValidation(D,P,K);
    for $k = 1 \ldots K$ **do**
      $D_{\mathrm{T}} = D_k$; $D_{\mathrm{V}} = D_{(k+1) \bmod K}$; $D_{\mathrm{L}} = D \setminus D_{\mathrm{T}} \setminus D_{\mathrm{V}}$;
      $\mathbf{h} = FindBestHyperParameterValues(D_{\mathrm{L}}, P(D_{\mathrm{L}}), D_{\mathrm{V}}, P(D_{\mathrm{V}}))$;
      $M = TrainWithBestHyperParams(i, D_{\mathrm{L}}, P(D_{\mathrm{L}}), D_{\mathrm{V}}, P(D_{\mathrm{V}}), \mathbf{h})$;
      $E_{\mathrm{T}}((i-1) * K + k) = E(M(D_{\mathrm{T}}), P(D_{\mathrm{T}}))$;
  $E_{\mathrm{T}} = [mean(E_{\mathrm{T}}()), median(E_{\mathrm{T}}())]$;

The evaluation of features Alg. 1 was applied separately for each dataset described in Sect. 4 and each feature set listed in Table 2. It was executed with $r = 10$ and $k = 10$. For higher diversity, a different ML technique was used for each $i = 1, \ldots, r$. Moreover, the best hyperparameter values were determined first for each of the following methods: kNN, multi-layer perception, SVMs with linear and radial kernels, XGBoost and XGBDart, ranger, naive Bayes, decision tree, and RF.

Table 3 shows the mean and median accuracy (ACC) E_{T} obtained for all ML methods considered together, i.e. based on 100 tests per one feature set-dataset pair. Next, for each dataset, the ML method yielding the highest median accuracy for C_TIME_DIFF was determined.

For the PAR_W1 dataset, extending the features to C_TIME_DIFF increases the mean and median ACC by about 3%, similarly to C_TIME. For the PAR_W2

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_40

Table 3: The accuracy (ACC) of mode choice predictions on testing subsets

		(a) All methods		(b) Best method		
Dataset	Feature set	mean [%]	median [%]	method	mean [%]	median [%]
PAR_W1	BASELINE	63.95	66.00	XGBoost	67.42	67.67
	C_DIFF	64.02	66.67	XGBoost	70.43	70.33
	C_TIME	66.80	69.00	XGBoost	77.26	77.33
	C_TIME_DIFF	66.82	68.67	XGBoost	75.82	76.33
PAR_W2	BASELINE	57.62	58.93	XGBDart	64.18	63.39
	C_DIFF	57.78	60.38	XGBDart	63.93	63.39
	C_TIME	57.74	59.65	XGBDart	61.45	60.96
	C_TIME_DIFF	57.96	58.93	XGBDart	65.02	67.25
CIT_W1	BASELINE	57.03	58.56	XGBDart	62.14	62.67
	C_DIFF	57.12	58.46	XGBDart	60.55	60.00
	C_TIME	56.41	57.92	XGBDart	60.10	60.04
	C_TIME_DIFF	56.91	58.78	XGBDart	61.77	61.71

dataset, C_TIME_DIFF provides the highest mean ACC, but the median ACC is better for C_DIFF. The ACC changes for CIT_W1 are only minor ones.

However, the best classifier for each dataset is sought in practice. Applying the best classifier – XGBoost – and the C_TIME set increases the mean ACC for PAR_W1 to 77.26% i.e. by nearly 10 per cent points. For PAR_W2, while the mean ACC can be slightly improved using XGBDart and all parking time features, the median ACC was improved to 67.25% when all features were included. The results for the PAR_W2 dataset show that the best classification method may yield a much higher ACC benefit arising from exploiting C_TIME_DIFF features, than suggested by mean ACC. The reason for the difference in the ACC gains between PAR_W1 and PAR_W2 may be higher MAE for parking time predictions for PAR_W2 than for PAR_W1.

The best predictor, i.e. XGBDart, cannot improve the average ACC on the CIT_W1 dataset. The method overfits training data and adding more features reduces the ACC of the models. These results show that whether additional features are helpful should be decided separately for each dataset.

5 Conclusions

In this work, we propose two novel categories of parking-related features to be used for travel mode choice modelling. The first category necessitates the use of data showing how much time it took drivers to find a parking space in different areas of the city of interest. The second group of features transforms data describing travel demands into parking difficulty features. Both feature groups contribute to the development of travel choice models. Which of them should be used depends inter alia on the available data sources. The case of the city-wide data for primary school parents (the PAR_W2 dataset) shows that the use of both feature types together may be needed to help model development.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_40

In the case of one of the datasets, the introduction of parking-related features negatively affected some ML methods. Still, the case of the two remaining datasets shows that significant accuracy gains can be expected once these features are used together with survey-based features. This suggests that the features proposed in this work should be considered in future travel mode choice studies.

Acknowledgements This research has been supported by the CoMobility project. The CoMobility benefits from a 2.05 million€ grant from Iceland, Liechtenstein and Norway through the EEA Grants. The aim of the project is to provide a package of tools and methods for the co-creation of sustainable mobility in urban spaces.

References

1. Ding, L., Yang, X.: The response of urban travel mode choice to parking fees considering travel time variability. Advances in Civil Engineering **2020**, 1–9 (7 2020). https://doi.org/10.1155/2020/8969202
2. Hagenauer, J., Helbich, M.: A comparative study of machine learning classifiers for modeling travel mode choice. Expert Systems with Applications **78**, 273–282 (7 2017). https://doi.org/10.1016/j.eswa.2017.01.057
3. Hasnine, M.S., Habib, K.N.: What about the dynamics in daily travel mode choices? A dynamic discrete choice approach for tour-based mode choice modelling. Transport Policy **71**(August), 70–80 (2018). https://doi.org/10.1016/j.tranpol.2018.07.011
4. Hillel, T., Bierlaire, M., Elshafie, M.Z., Jin, Y.: A systematic review of machine learning classification methodologies for modelling passenger mode choice. Journal of Choice Modelling **38**, 100221 (2021). https://doi.org/10.1016/j.jocm.2020.100221
5. Li, M., Zou, M., Li, H.: Urban travel behavior study based on data fusion model. Elsevier Inc. (2018). https://doi.org/10.1016/B978-0-12-817026-7.00005-9
6. Lu, Y., Kawamura, K.: Data-mining approach to work trip mode choice analysis in Chicago, Illinois, area. Transportation Research Record **2156**(1), 73–80 (2010). https://doi.org/10.3141/2156-09
7. Salas, P., la Fuente, R.D., Astroza, S., Carrasco, J.A.: A systematic comparative evaluation of machine learning classifiers and discrete choice models for travel mode choice in the presence of response heterogeneity. Expert Systems with Applications **193**, 116253 (5 2022). https://doi.org/10.1016/j.eswa.2021.116253
8. Tamim Kashifi, M., Jamal, A., Samim Kashefi, M., Almoshaogeh, M., Masiur Rahman, S.: Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. Travel Behaviour and Society **29**(July 2021), 279–296 (2022). https://doi.org/10.1016/j.tbs.2022.07.003
9. Tenkanen, H., Toivonen, T.: Longitudinal spatial dataset on travel times and distances by different travel modes in Helsinki Region. Scientific Data **7**(1) (dec 2020). https://doi.org/10.1038/s41597-020-0413-y
10. Xie, C., Lu, J., Parkany, E.: Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks. Transportation Research Record **1854**(1), 50–61 (2003). https://doi.org/10.3141/1854-06

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36030-5_40