Explaining Walking in Cities – a Machine Learning Approach

Rasha Bowirrat*[1], Karel Martens[2], Yoram Shiftan[3]

[1] Architect and Urban Planner, Faculty of Architecture and Town Planning, Technion, Israel

[2] Professor, Faculty of Architecture and Town Planning, Technion, Israel

[3] Professor, Faculty of Civil and Environmental Engineering, Technion, Israel

SHORT SUMMARY

A large body of research has developed on walking and walkability, in part in response to increasing concerns over people's health, climate change, livability, and social cohesion. Literature shows that some built environment and socio-demographic characteristics influence walking rates more than others.

Different approaches and methods have been used to study the relationship between the built environment characteristics, socio-demographic variables and walking patterns. Yet, so far very few studies have applied machine learning tools to study and explore these relationships. This research aims to start filling this void.

The study draws on a dataset contains details about trips made by over 37,000 respondents in the Tel-Aviv metropolitan area. The detailed data allow us to differentiate between walk-only trips and walk trips that are combined with other modes of transport. Our results show that the built environment shapes walk-only trips more than walking as an access or egress mode.

Keywords: Data analysis, Machine learning, Mobility, Urban planning, Walkability, Walking.

1. INTRODUCTION

Walking is a mode of transport that enables getting from one place to another, and it is the most prevalent form of physical activity. Walking is a fundamental constituent of nearly all trips, as it enables physical access to different kinds of facilities. Transportation means such as trains, buses, and private transport, require walking both for access and egress (Wigan, 1995).

Walking is not shaped solely by dedicated infrastructure (e.g., pavements and crossings) but is also highly dependent on other features of the built environment, as these can promote or constrain walking (Forsyth & Krizek, 2010; Lee & Moudon, 2004). The act of walking is shaped by the city, infrastructures, its built environment characteristics, and the sociodemographic variables of people.

Different approaches and tools are used to study walking behavior and investigate its relationship with personal and sociodemographic variables and built environment variables. Very few studies have applied data analysis and machine learning tools to study and explore the (non-linear) relationship between the different variables.

This study aims to disentangle the potential of the built environment effects on walking in urban areas and to determine the relative importance of built environment and socio-demographic variables in shaping walking patterns employing a machine learning and data analysis approach. This research is conducted among a diverse population in terms of their characteristics using a large data set that includes nearly 37,100 participants. The research question is: “What type of variables most strongly shape walking patterns?”.

2. METHODOLOGY

This research applies a Random Forest (RF) multiclass classification algorithm to identify the set of (walkability) parameters that most strongly shape walking in urban areas. The two compared groups of parameters include the sociodemographic and the built environment variables. RF algorithm can easily handle a large number of variables as it weighs the contribution of each variable according to how dependent it is on other variables (Breiman, 1996; T. Shi & Horvath, 2006).

We developed a model where we distinguish between four possible trip types that compile for the dependent variables in the model: walk only trips, walk trips in combination with public transport, walk trips in combination with car use, and trips that do not include walking at all (Table 1). We distinguish between these trip types, because we expect that effect of the built environment and of people’s socio-demographic characteristics may vary between the trips. We may expect that built environment factors may particularly shape walk only trips, while being less important for the other two trip types that include a walking leg. By distinguishing between the trip types, we can test this expectation, which has not yet been done in the literature.

We hypothesize that it is more likely that the built environment shapes walk-only trips, since walking as part of public transport or car trip is largely unavoidable. Yet, we also hypothesize that the built environment is more likely to shape the choice of walking integrated with public transport than the choice of walking and car use, as literature shows that the decision to use public transport is partly shaped by built environment factors (An et al., 2022). The dataset contains 288,555 trips. These trips are described by 40 different parameters including built environment variables (Table 2), and socio-demographic variables (Table 3).

Table 1. Dependent Variables; Key characteristics of the four Trips Types Distinguished in the Research

Class index	Class	Description	Percentage of trips of all walk trips	Mean length of travel distance	Mean travel time of trips
1	Walk-only trips	Trips that consist solely of one or more walk legs	24%	0.79 km	9.34 min
2	Walk + public transportation trip	Public transport means that were taken into consideration in this category are: bus, taxi, train, and organized shared transit Every trip with at least one walk leg and one PT leg, irrespective of whether the entire trip chain also includes other modes of transport for some trip legs (e.g. car, bicycle) are included 3. Trips that include public transport, but for which no walk trip was reported, were also added to this category, since we hypothesized that to get to a public transport stop a walk would be necessary in virtually all cases.	9%	8.09 km	35.78 min
3	Walk + car trip	Every trip with at least one walk leg and one leg by car (as driver or passenger) or motorbike, unless the trip chain includes PT for one or more trip legs (and irrespective whether the trip included yet other modes of transport for some trip legs (e.g., bicycle).	2%	7.74 km	32.96 min
4	Trips without a walk leg	All trips that do not belong to one of the categories mentioned above (bicycle trips are included).	65%	3.66 km	12.89 min

Table 2. Built Environment Variables in the Dataset & included in RF model

	Variable	**Description**	**Share**	**Mean**	**SD (Standard deviation)**
Built Environment variables (Zonal level)	Residential density (HHdens)	Total number of households in zone divided by zone surface area (number/m^2)		3.8	4.0
	Population density (popdens)	Total population divided by zone surface area (persons/m^2)		10.1	10.1
	Land use mix in terms of jobs employment	Number of jobs in zone divided by population (jobs/persons)		0.4	0.1
	Employees	Number of employees in area		1389.7	1146.1
	Parking capacity	Parking capacity in area		3119.2	5883.1
	Urban area type (share of zones that belong to each category)	Metropolitan CBD	3.36%		
		Urban Residential - Low-density: zones under 5,000 inhabitants per sq. km.	12.80%		
		Urban Residential - High-density: zones over 5,000 inhabitants per sq. km.	57.76%		
		Major public institutions: educational / legal / hospitals.	0.40%		
		Commercial: include city centers, shopping centers and markets.	8.48%		
		Major Employment centers: employment centers over 4,000 employees.	4.72%		
		Medium Employment centers: employment centers under 4,000 employees.	2.00%		

		Mixed Use Areas: areas with a mix of residential, commerce and employment.	3.60%		
		Major Transport facilities: airports, ports and bus stations.	0.24%		
		Sports and Tourism: areas with concentrations of hotels, beaches and sport facilities.	2.56%		
		Rural Areas: rural settlements (like moshavim and kibutzim), agricultural land and isolated developments.	3.12%		
		Open Areas: empty or non-built areas with no special use.	0.16%		
		Military Areas: zones used by the army.	0		
		Cemetery	0.48%		
		Small Settlements: isolated development of urban residential uses outside the urban core.	0		
	Students	Number of students studying in zone		110.79	1105.88
	Socio economic status	Socio-economic level of zone		11.30	4.76
	Parking availability at employment place (EmpPark)	EmpPark_1: Parking available for free for workers	22.2%		
		EmpPark_2: Parking available only near workplace	0.8%		
		EmpPark_3: Unavailable parking spaces	69.3%		

Table 3: Sociodemographic Variables in the Dataset & included in RF model

	Variable	**Description**	**Share**	**Mean**	**SD (Standard deviation)**
Socio-demographic variables	Age	Respondent`s age		33.2	22.9
	Gender	Male respondents	51.5%		
		Female respondents	48.5%		
	Sector	Secular Jew	70.2%		
		Religious Jew	12.6%		
		Orthodox Jew	14.8%		
		Arab	2.3%		
	Education level	Highly educated respondent – undergraduate & graduate studies	26.0%		
		Medium educated respondent – high school certificate	24.8%		
		Low educated respondent – adult without high school certificate	19.0%		
		School students and other	29.9%		
	Employment status	Employed (full/part time job)	46.8%		
		Unemployed	23.3%		
		Other (unknown, irrelevant)	28.8%		
	Car license holding (Clic)	Clic_1: The respondent holds a driving license	62.9%		
		Clic_2: The respondent does not hold a driving license	13.7%		
		Clic_9: Unknown whether the respondent holds a driving license	0.003%		
		Clic_99: Irrelevant	23.3%		
	Household size (HHsize)	Person\household		2.9	1.76
	Children under age 8 in household	Per household		0.4	0.9

	Car ownership (HHVeh)	Households with at least one car in their ownership	76.7%		
	Bicycle ownership	Households with at least one bicycle in their ownership	31.4%		
	Household composition (HHType)	HHType_1: Households with one individual person	22.0%		
		HHType_2: Households for a couple	24.2%		
		HHType_3: Households for parents and their children	51.8%		
		HHType_4: Households for disabled person and assistant	0.1%		
		HHType_5: Shared Households (between partners)	1.6%		

Techniques for dealing with the imbalanced dataset

As can be seen in Table 1, the data is imbalanced since the data set has skewed class proportions. The trip type with the least observation is the Walk+car class, as it accounts for only 2% of all trips. In contrast, Trips without any walk leg make up the majority class with 65% of the data. In this case, the RF algorithm will mainly relate to the majority class and treat the minority class features as noise in the data and ignore them. SMOTE technique was used in order to overcome this issue.

Hyperparameters Tuning

Tuning is the task of finding optimal hyperparameters for a RF model for a given dataset (Probst et al., 2018), thus optimizing the model in terms of its performance and running time. RF models works reasonably well with the default values of the hyperparameters specified in software packages. Nevertheless, tuning the hyperparameters can improve the performance of RF. The technique that was used in this research in order to overcome the imbalanced dataset was fitted to the hyperparameters that were accepted after 30 iterations (Table 4).

Oversampling

SMOTE (Synthetic Minority Oversampling Technique) was applied on the training set, alongside with hyper-tuning the model, which is an oversampling technique where synthetic samples are generated for the minority classes to rebalance the original training set.

After running this technique, an evaluation of the results should be done. Since the number of observations in each class is initially unequal, a so-called confusion matrix is needed to describe the performance of a classification algorithm.

In the confusion matrix, the number of correct and incorrect predictions is described with counted values for each class. Prediction for each class can be described by its precision and recall. Precision measures the share of data cases signaled by the model that are real predictions. Recall measures the share of data cases occurring in the domain that are “captured” by the models (Torgo & Ribeiro, 2009).

F1 score is the harmonic mean of the precision and recall of the model (Equation 1) which delivers the best value at 1 and worst score at 0 (Lipton et al., 2014).
Equation 1: F1 score

$$F1 = 2 * (precision * recall) / (precision + recall)$$

Examining the first model`s performance (4 classes model) based on the F1-score shows a relatively high prediction accuracy score for all of the classes apart from the walk + car class. The low F1 score for the walk + car trip class indicates that the model did not predict the true labels in this class, as it predicted correctly only 6% of the labels. The majority of the wrong predictions was in favor to trips without a walk leg. This suggests that the model mistakenly confuses between these two classes, thus resulting in inaccurate feature importance of each one of the parameters in the model regarding the walk + car class (Figure 1.a).

In order to overcome this issue, we developed a second model (3 classes model) where we excluded the trips without a walk leg class from the model. In this case, the F1 score for the walk + car trip class was substantially higher than in the first model (62%) (Table 4).

Table 4. Hyperparameters setting and Models classification reports

	Parameter / Model		SMOTE Model 1: 4 classes model	SMOTE Model 2: 3 classes model
Hyperparameters	Parameter	Description (58)		
	n_estimators	Decision trees number being built in the forest	1000	2000
	min_sample_split	Minimum number of samples required to split an internal node	10	2
	min_sample_leaf	Minimum number of data required in node	1	2
	max_features	Maximum features number used for a node split process	"auto"	"auto"
	max_depth	Maximum depth and levels a decision tree is allowed	80	60
	bootstrap	False value: All data is used for every decision tree; else selected bootstrap samples are used when building decision trees	False	False
F1 score for different Classes	1. Walk only trips		0.78	0.91
	2. Walk trip + PT		0.61	0.75
	3. Walk trip + Car		0.08	0.62
	4. Trips without a walk leg		0.90	-

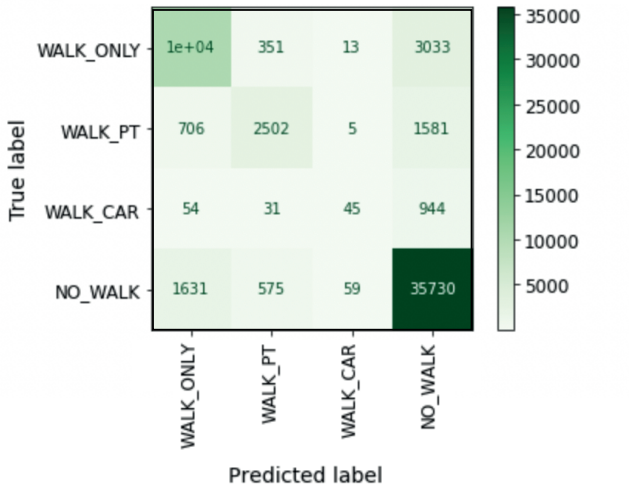

Figure 5.a: Confusion Matrix for SMOTE + Tuned Random Forest Model (4 classes model)

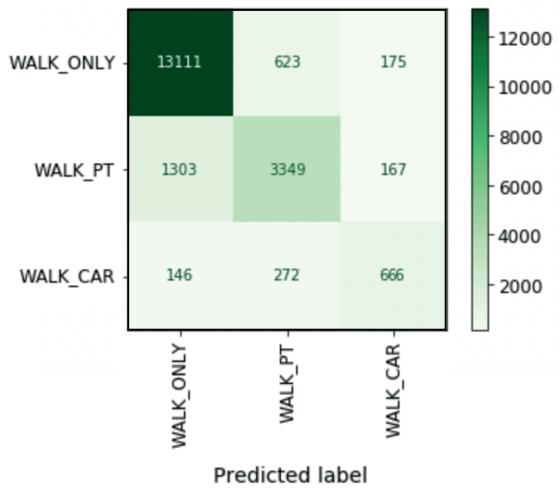

Figure 5.b: Confusion Matrix for SMOTE + Tuned Random Forest Model (3 classes model)

Figure 1. Confusion Matrix for (a) 4 Classes Model, and (b) 3 Classes Model

3. RESULTS AND DISCUSSION

The results of our analysis concern the four trip types thar were investigated in this research and depicted in figure 2. For walk-only trips, the likelihood that a person makes a walk-only trip increases with population density at origin, population density at the destination, household density at origin, household density at the destination, no parking available at workplaces, household size, and workers number in origin. In contrast, it decreases with car license holding, household vehicle, and age that have the strongest negative influence.

Walk + PT trips are related to several features. Vehicle license and number of vehicles in the household all shape the number of the walk + PT trips. Additional sociodemographic characteristics that are related to making walk + PT trips are age, gender, sector, employment status, and education level. Males and secular Jews tend to make less walk + PT trips. In terms of the built environment variables, free parking at the workplace leads to a reduction in walk + PT trips. In contrast, the unavailability of parking near the workplace has a positive impact on walk + PT trips. Furthermore, population density at the trip origin increases the choice for walk + PT trips.

For walk + car trips, the likelihood of making this trip type is influenced by holding a car license, age, and the number of cars in the household, among other factors. These three variables have strong importance for a person choosing this trip type. An additional factor is the household size. Part of the most important built environment factors that reduce the likelihood that a person will make walk + car trips is population density at origin and destination, workspace parking availability, and the number of workers at the destination area.

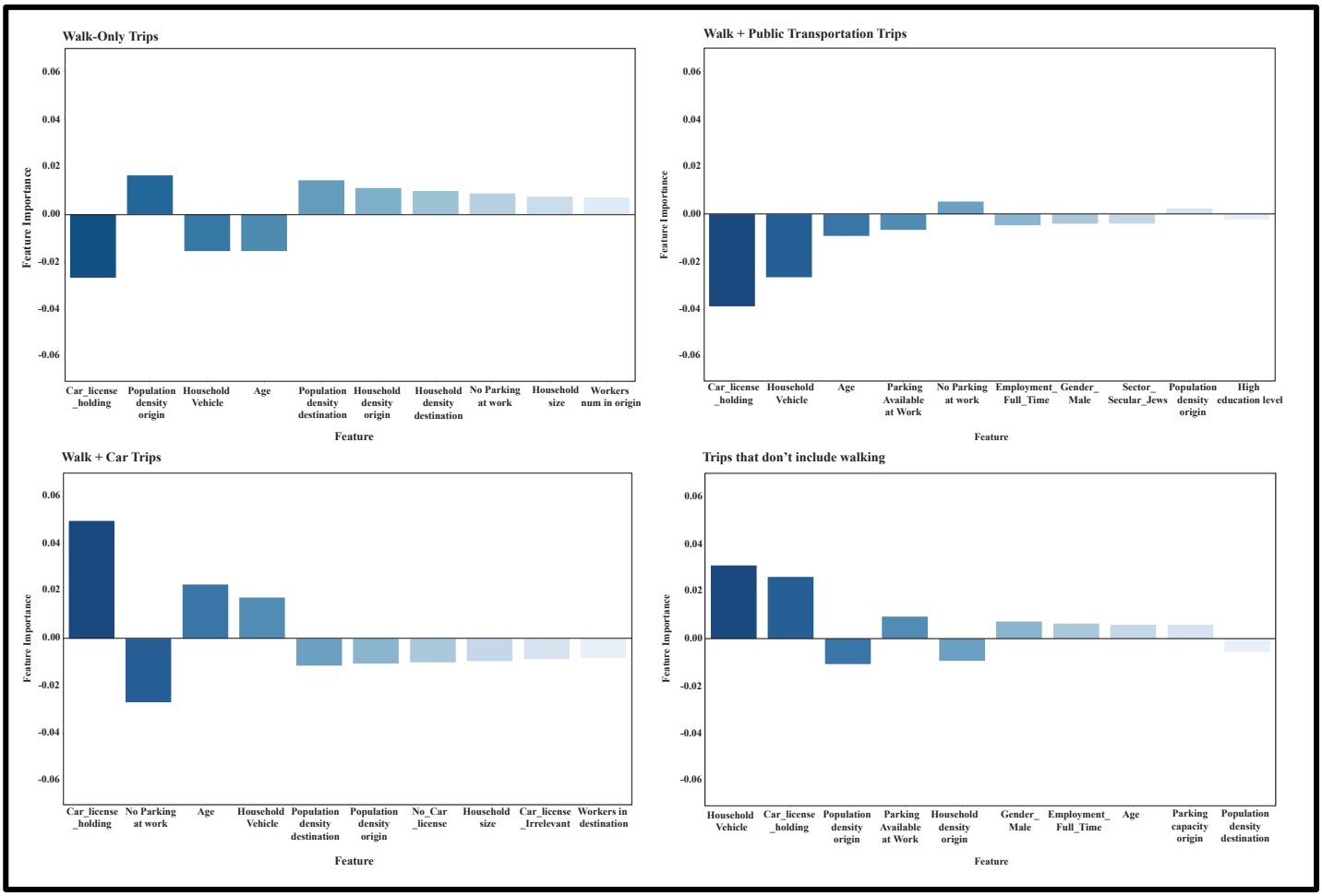

Figure 2. Model results: Feature Importance for Different Trips Types

4. CONCLUSIONS

Our results illustrate the importance of built environment variables in shaping walking-only trips more than other types of trips, with household and population densities having the strongest positive effect on walk-only trips of all built environment characteristics.

Our study shows that machine learning and data science approaches hold promise for the analysis of walking patterns, and for gaining insight into the set of variables that influence walking in cities.

The study suggests that the impacts of minimum parking norms for offices and other employment types have an impact beyond the home-to-work trip. Our findings indicate that parking at the workplace also affects the frequency with people engage in walk-only trips. This underscores the importance of abolishing parking minimums, as is gradually occurring in Israel and elsewhere (Christiansen et al., 2017; Shiftan & Burd-Eden, 2001; SimiAeviA et al., 2013).

ACKNOWLEDGEMENTS

This research was supported by the Azrieli Foundation, and the Israeli Smart Transportation Research Center.

REFERENCES

An, R., Wu, Z., Tong, Z., Qin, S., Zhu, Y., & Liu, Y. (2022). How the built environment promotes public transportation in Wuhan: A multiscale geographically weighted regression analysis. *Travel Behaviour and Society*, *29*, 186-199.

Breiman, L. (1996). Bagging predictors [Article]. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Christiansen, P., Engebretsen, Ø., Fearnley, N., & Usterud Hanssen, J. (2017). Parking facilities and the built environment: Impacts on travel behaviour [Article]. *Transportation Research. Part A, Policy and Practice*, *95*, 198–206. https://doi.org/10.1016/j.tra.2016.10.025

Forsyth, A., & Krizek, K. J. (2010). Promoting Walking and Bicycling: Assessing the Evidence to Assist Planners [Article]. *Built Environment (London. 1978)*, *36*(4), 429–446. https://doi.org/10.2148/benv.36.4.429

Lee, C., & Moudon, A. V. (2004). Physical Activity and Environment Research in the Health Field: Implications for Urban and Transportation Planning Practice and Research [Article]. *Journal of Planning Literature*, *19*(2), 147–181. https://doi.org/10.1177/0885412204267680

Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). *Thresholding Classifiers to Maximize F1 Score*. http://arxiv.org/abs/1402.1892

Probst, P., Wright, M., & Boulesteix, A.-L. (2018). *Hyperparameters and Tuning Strategies for Random Forest*. https://doi.org/10.1002/widm.1301

Shi, Q., & Abdel-Aty, M. (2015). Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways [Article]. *Transportation Research. Part C, Emerging Technologies*, *58*, 380–394. https://doi.org/10.1016/j.trc.2015.02.022

Shiftan, Y., & Burd-Eden, R. (2001). Modeling Response to Parking Policy [Article]. *Transportation Research Record*, *1765*(1), 27–34. https://doi.org/10.3141/1765-05

SimiAeviA, J., VukanoviA, S., & MilosavljeviA, N. (2013). The effect of parking charges and time limit to car usage and parking behaviour [Article]. *Transport Policy*, *30*, 125–131. https://doi.org/10.1016/j.tranpol.2013.09.007

Torgo, L., & Ribeiro, R. (2009). Precision and recall for regression. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5808 LNAI*, 332–346. https://doi.org/10.1007/978-3-642-04747-3_26

Wigan, M. (n.d.). Transportation Research Record No. 1487, Nonmotorized Transportation Research, Issues, and Use. In *TRANSPORTATION RESEARCH RECORD*.