Travel Behaviour and Society 39 (2025) 100983

Contents lists available at ScienceDirect

Travel Behaviour and Society

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

Assessing the impact of walkability indicators on health outcomes using machine learning algorithms: A case study of Michigan

Musab Wedyan, Fatemeh Saeidi-Rizi *

School of Planning, Design and Construction, Michigan State University, East Lansing, MI, USA

A B S T R A C T

Urban planning and public health are increasingly interlinked in efforts to shape healthier communities. To build a healthier community, walkability has shown positive outcomes for population health. This study employs machine learning to analyze the impact of walkability indicators such as intersection density, proximity to transit stops, employment mix, and employment and household mix and social vulnerability factors on health outcomes in Michigan. Data from the Environmental Protection Agency (EPA) and Centers for Disease Control and Prevention (CDC) were used to evaluate health outcomes including obesity, blood pressure, cholesterol, and depression. The analysis also incorporated the Social Vulnerability Index (SVI) to examine the influence of socioeconomic and demographic factors. Different supervised machine-learning algorithms were applied to assess these relationships. Among the algorithms, the Random Forest algorithm showed the best performance. The results indicate that there is a variation in the impact of walkability indicators on health outcomes. Key findings reveal that among walkabiltity indicators, intersection density is the most significant predictor of all health outcomes, while the other indicators have less impact. In addition, it was found that variables such as Socioeconomic Status, Household Composition & Disability, Minority Status, Housing Type and Transportation have also impact of health outcomes. In conclusion, this research shows the relationship between walkability and human health by providing an evidence-based guidance for building healthier, more walkable communities.

1. Introduction

Regular walking as part of daily routines is recognized for its broad health benefits across all age groups, contributing to an active lifestyle (Wang et al., 2016). Additionally, using public transportation often involves some level of walking or cycling, further contributing to overall health (Rissel et al. 2012). Areas with higher walkability, characterized by densely populated residential areas, diverse land uses, and interconnected streets, report more frequent walking for transportation (Gunn et al. 2017). Thus, urban planning that enhances walkability and provides robust public transportation options plays a crucial role in fostering healthier lifestyles.

Although elements such as active transport and public facilities are integral to walkability and are widely recognized, a deeper exploration is required to quantify the relationship between walkability's various dimensions and their health benefits. However, most existing research lacks precise quantification of how specific walkability dimensions independently affect health outcomes. These factors are typically considered collectively, yet further exploration is needed to isolate and understand the distinct influence of each dimension on physical and mental health. This is essential for creating more accurate forecasts of the health impacts as outlined in existing frameworks (Zuniga-Teran et al. 2017) and therefore, refining urban planning strategies (Westenhöfer et al. 2023). This research investigates the impact of walkability indicators on several health outcomes by prioritizing the most important factors of walkability.

1.1. Measuring walkability

Measuring walkability involves assessing various aspects of the urban environment that contribute to facilitating walking. Current walkability indices are constructed based on different factors such as road features, accessibility to amenities, land use characteristics, safety, and comfort. For example, (Gu et al. 2018) used open-source data to measure walkability based on convenience, comfort, and safety on the roads. Other researchers, such as (Sugiyama et al. 2013) explored how green space attributes influence adults' likelihood of starting or continuing recreational walking. While green space characteristics didn't significantly affect the initiation of walking, factors such as positive perceptions of green space presence and proximity were associated with a higher likelihood of maintaining regular walking over four years. (Frank et al., 2010) created a walkability index using z-scores based on net residential density, retail floor area ratio, intersection density, and land use mix. Walkability indices are often calculated using z-scores of

* Corresponding author.
E-mail addresses: wedyanmu@msu.edu (M. Wedyan), saeidiri@msu.edu (F. Saeidi-Rizi).

https://doi.org/10.1016/j.tbs.2025.100983
Received 22 July 2024; Received in revised form 25 December 2024; Accepted 2 January 2025
Available online 6 January 2025

these indicators, divided by the number of indicators, and expressed in various formats, including quintiles (Creatore et al. 2016), quartiles (Frank et al. 2007), or continuous scores (Lathey et al. 2009). In the U.S., a National Walkability Index (NWI) was developed to assess the relative walkability of areas (Thomas, 2017). The Index rates block groups based on design, distance, and diversity in the urban environment. It was constructed by considering factors that influence individuals' preferences for walking as a transportation mode. The factors of the index include street intersection density, the proximity of population centers to the nearest transit stops in meters, and land use types. Overall, these studies collectively underscore the significant role of urban planning in enhancing walkability.

1.2. Walkability and health outcomes

Walkability's impact on health is increasingly recognized, with studies linking higher walkability to lower obesity rates, improved mental health, and reduced cardiovascular risks (Makhlouf et al. 2023). Also, (Frank et al. 2010) argued that factors such as density and connectivity improve cardiovascular health.

In addition, neighborhood physical layout can affect mental health through perceived aesthetics, safety, and accessibility issues, or even by geographic regional variations in the built environment (Melis et al. 2015, Wu et al. 2017). Besides, studies have indicated that community centers, bus stops, libraries, restaurants, and high-quality public places are positively associated with cognitive performance (Guo et al. 2019, Finlay et al., 2020). Similarly, environments with a variety of options and access to nearby businesses and services, parks with suitable structures for use, regular and well-maintained sidewalks, and well-connected streets can increase physical activity by up to 57 % (Balcetis et al. 2020). The association between walkability and the cognitive and emotional status that support components of mental health was explained by leisure-time physical activity (Solis-Urra et al. 2020).

1.3. Research significance and gap

City planning and policy-related strategies focused on enhancing street network connectivity, land-use diversity, and housing density can create supportive environments that promote health. Based on that, understanding those factors that contribute to walkability allows for eliminating health issues. Previous studies, such as (Watson et al. 2020) highlighted the general benefits of walkable neighborhoods on human health by employing NWI. (Watson et al. 2020) used National Walkability Index (NWI) variables, including street connectivity, land-use mix, and proximity to public transit, along with health-related measures such as walking frequency and demographic factors. Using NHIS data, it found that higher walkability scores were linked to a greater likelihood of walking, especially for transportation, indicating that walkable environments encourage active transportation behaviors. Other studies examine the relationship between walkability and health outcomes using NWI and Centers for Disease Control and Prevention (CDC) data. For instance, (Makhlouf et al. 2023) examined NWI-related walkability variables alongside cardiovascular risk factors (hypertension, cholesterol, obesity, diabetes, and coronary artery disease) and demographic controls in a cross-sectional analysis. It found that neighborhoods with higher walkability scores were associated with lower prevalence rates of cardiovascular disease. Another recent study explored the influence of built and natural environmental features, such as walkability, bike infrastructure, and greenspace, on neighborhood-level prevalence of hypertension and obesity across the United States. Leveraging data from the CDC's PLACES database, the National Environmental Database (NED), the Social Vulnerability Index (SVI), and detailed bike infrastructure and safety metrics, the research examined these relationships across varied socio-geographic contexts. The National Walkability Index (NWI), which integrates measures of residential density, street connectivity, and transit access, was employed to quantify walkability. By using advanced quantile regression methods, the study captured heterogeneity in these associations, identifying areas where interventions could yield the greatest impact.

Although the current research has demonstrated the relationship between walkability and health outcomes, existing studies primarily focus on aggregate measures of walkability, such as the National Walkability Index (NWI), without isolating the specific impacts of individual walkability indicators. Furthermore, most studies rely on traditional statistical methods, which may not effectively capture the complex and nonlinear relationships between urban features and health. Additionally, prior research often overlooks spatial and social heterogeneity in these relationships, limiting their applicability across diverse urban and regional contexts. Also, gaps remain in extending research to varying urban layouts and integrating dynamic temporal factors, such as changing urban policies or demographic shifts, that could influence walkability-health relationships over time.

Unlike conventional models, ML techniques excel at capturing intricate, nonlinear patterns in data, which is essential for understanding complex factors affecting health (Liaw 2002). ML's predictive accuracy is also highly valuable in public health contexts, where reliable forecasts of health outcomes can drive effective interventions and strategic resource allocation (Obermeyer and Emanuel 2016). Moreover, the adaptability of ML models across different times shows its effectiveness for wide-scale, practical applications in walkability and health research (Wang et al. 2024).

This research contributes to the body of knowledge in two aspects. First, the novelty of this approach lies in the use of machine learning to identify and rank the most influential features of walkability on the following health outcomes; obesity, blood pressure, mental health, cholesterol and depression. The study hypothesizes that not all walkability and SVI factors have the same impact on health outcomes. To illustrate, we hypothesize that (1) Proximity to Transit Stops is hypothesized to affect obesity, blood pressure, cholesterol levels, and depression; (2) Higher intersection density is expected to reduce obesity, blood pressure, and depression; (3) A diverse employment mix is anticipated to lower obesity rates, blood pressure, depression, and cholesterol levels; and (4) Employment and Household Mix affects obesity, blood pressure, and depression, mental health and cholesterol levels.

2. Research methodology

Fig. 1 outlines a methodical approach to our study. It begins with a data overview, examining indicators of walkability such as employment diversity and proximity to transit, alongside health outcomes. The data processing phase involves preparation, such as matching datasets by census tracts and cleaning steps including normalization and outlier removal. Then, different machine learning algorithms were selected. Different tuning parameters and evaluation metrics were applied. The best-performing algorithm is then used to rank the importance of the walkability features. The process concludes with the generation of partial dependence plots, which visually depict the relationship between each walkability factor and health outcomes in more detail.

2.1. Study region characteristics

This study was conducted within the state of Michigan, a region with diverse geographic, social, and economic landscapes. Michigan's urban areas, particularly Detroit, Lansing, and Grand Rapids, feature varying levels of population density, industrial and commercial development, and socioeconomic diversity, which are key factors in assessing walkability. The state also encompasses suburban and rural areas with lower density and limited access to public transportation, creating a range of walkability scores across the study region. Michigan's economy, historically centered around the automotive industry, has shifted toward a

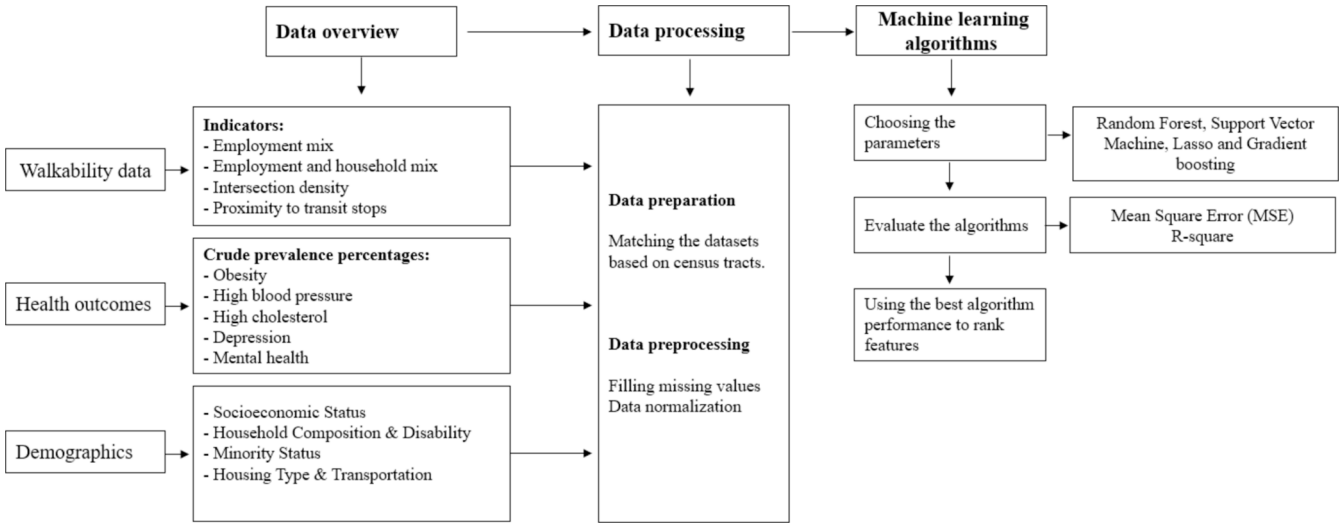

Fig. 1. Research framework.

more diversified landscape including healthcare, education, and technology sectors, influencing employment mix and accessibility. These geographic and economic characteristics provide a unique setting for analyzing the relationship between walkability indicators and health outcomes. This variability makes Michigan an informative case for studying walkability in regions with both highly urbanized and more rural settings, though further studies in differing contexts would be needed to confirm the broader applicability of these results (Michigan, 2019, MSU 2022).

2.2. Data sources and variables definitions

2.2.1. Walkability data

To study the walkability, the National Walkability Index (NWI) (National Walkabiltiy Index, 2021) was employed which assigns a score to each block group in the United States. This index is derived from built environment metrics such as street intersection density, proximity to transit stops, and land-use diversity, leverages data from the Smart Location Database (SLD). Every block group in the U.S. is assigned a NWI score and each of these four components. These variables were selected not only due to their relevance to active transportation but also because consistent and accessible data at the block group level is available. This set of variables ensures that the Index remains straightforward to comprehend. The index is quantified through an equation incorporating these diverse factors.

$$\text{Final National Walkability Index score} = (w/3) + (x/3) + (y/6) + (z/6)$$

where w = block group's ranked score for intersection density x = block group's ranked score for proximity to transit stops y = block group's ranked score for employment mix z = block group's ranked score for employment and household mix. The block groups are assigned their final National Walkability Index scores on a scale of 1 to 20. This score is then used to classify block groups into four categories of walkability: 'Least walkable' with scores from 1 to 5.75, 'Below average walkable' from 5.76 to 10.5, 'Above average walkable' from 10.51 to 15.25, and 'Most walkable' from 15.26 to 20. These classifications help to indicate the relative ease of walking in an area based on these weighted factors, which are indicative of the availability and proximity of infrastructure and amenities that support pedestrian activity (National Walkabiltiy Index, 2021).

Also, according to (National Walkabiltiy Index, 2021), the block group's ranked score for each component (e.g., intersection density) in the National Walkability Index (NWI) is calculated by ranking each block group's value relative to all other block groups within the dataset. For instance, intersection density is measured by calculating the number of intersections per unit area for each block group. These values are then sorted, and each block group is assigned a percentile rank based on its position in the distribution. This ranking standardizes the scores across different metrics, allowing for the combination of intersection density, proximity to transit stops, employment mix, and employment and household mix into a single NWI score. Each ranked score is weighted in the final formula to reflect its contribution to walkability. This percentile ranking approach ensures that the index is comparable across diverse urban and rural environments, as it accounts for variations in each walkability component across the nation.

Proximity to transit stops is the distance from the center of the population to the nearest transit stop in meters. It plays a role in promoting walking, with shorter distances encouraging more walking trips. Similarly, the density of street intersections within an area, with higher densities associated with more walking trips, is another important factor. The variety in employment types and the level of occupied housing is referred to as the employment/household occupancy mix. Areas with a diverse range of employment options and high occupancy rates tend to have more walking. Finally, the employment mix within a block group, which considers the variety of job types (such as retail, office, industrial, and service sectors), is linked to increased walking activity, more diverse employment options correlate with more walking trips (National Walkabiltiy Index, 2021). The walkability data focuses on metrics related to walking for transportation. Table 1 shows the formulation of each variable. Table 2 shows the soical valunerability index variables related to the population in the State of Michgian.

Understanding the context and specific areas where the Index is associated with walking can be beneficial for public health and other relevant professionals when evaluating the walkability of their communities (Mayne et al. 2013, Stockton et al. 2016). These findings emphasize the NWI's value as a tool for researchers and policymakers to assess and enhance walkability in urban settings (Watson et al. 2020).

2.2.2. Health data

The health data utilized in this study are derived from the Population Level Analysis and Community Estimates (PLACES) Project, an expansion of the original 500 Cities Project initiated by the Centers for Disease Control and Prevention (CDC) in collaboration with the Robert Wood Johnson Foundation (RWJF) and the CDC Foundation (PLACES 2019). The PLACES Project generates small area estimations using a sophisticated multi-level statistical modelling framework. This framework integrates primary data from several key sources: the CDC Behavioral Risk Factor Surveillance System (BRFSS), the Census 2010 population data,

Table 1
Walkability variables formulations.

Variable name	Formulation
Residential density	This measure is calculated as number of housing units per acre of unprotected land. Gross residential density = number of housing units (both single family and multi-family) / area of developable land in acres (water bodies and conservation areas excluded).
Street network intersection density	The street intersection density is calculated by taking a weighted sum of the densities of different types of intersections, where three-leg multi-modal intersections and three-leg pedestrian-oriented intersections are each weighted at 0.667, and four-leg multi-modal and pedestrian-oriented intersections are fully weighted, then dividing this total by the total land area in acres.
Employment mix	This measure is based on the 8 employment categories to calculate employment mix (entropy).
Access to transit	This measure captures the number of transit stops within a census block group.
Walkability Index	This variable is calculated as a composite measure of the above built environment features related to connectivity, diversity, density, transit, and automobile use.

Table 2
Social vulnerability index variables.

Category	Variable	Definition
Socioeconomic Status	Below poverty	Percent individuals below poverty
	Unemployed	Percent of civilians unemployed
	No high school diploma	Percent of individuals with no high school diploma
Household Composition & Disability	Civilian with a disability	Percent of civil individuals of age 5 years or older with a disability
	Single-parent household	Percent of single-parent households with children under 18 years old.
Minority Status	Minority	Percent minority (all individuals except white, non-Hispanic)
Housing Type & Transportation	Multi-unit structures	Percent of housing units with 10 or more units in a structure
	Mobile homes	Percent of housing units that are mobile homes
	Crowding	Percent of total housing units occupied with > 1 person per room
	No vehicle	Percent of households with no vehicle available
	Group quarters	Percent of individuals who are in group quarters, both institutionalized (e.g., correctional institutions, nursing homes) and non-institutionalized (e.g., college dormitories, military quarters).

and estimates from the American Community Survey (ACS). These comprehensive data sources enable detailed health and demographic analyses at the county and census tract levels, providing valuable insights into community health metrics and disparities. The methodology and statistical approaches employed in the PLACES Project are designed to produce reliable and actionable public health data, which are crucial for informed decision-making and targeted interventions at the local level.

The health dataset used in this study includes adults aged 18 and over at the census tract level, covering the prevalence of high cholesterol, mental health issues, clinical depression, and obesity rates. High cholesterol data exclude pregnant individuals and those diagnosed with borderline hypertension within the last five years. Mental health data includes the prevalence of mental health issues, characterized by frequent reports of poor mental health over 30 days, and clinical depression confirmed by health professionals. Obesity rates were calculated using a body mass index (BMI) threshold of 30.0 kg/m^2, derived from self-reported height and weight. The measure for obesity is calculated using self-reported weight and height from respondents, excluding those with extreme values (height < 3ft or ≥ 8ft, weight < 50 lbs or ≥ 650 lbs, BMI < 12 kg/m^2 or ≥ 100 kg/m^2, and pregnant women). The numerator includes adults with a BMI of 30.0 kg/m^2 or higher, while the denominator consists of all adults for whom BMI can be calculated, excluding those with unknown or refused weight/height data (PLACES 2019).

2.2.3. Social vulnerability data

The Social Vulnerability Index (SVI), developed by the Centers for Disease Control and Prevention (CDC). The SVI measures the social and demographic factors that influence a community's ability to prepare for, respond to, and recover from external stresses, including natural disasters, public health crises, and economic challenges (CDC, 2023). By incorporating data on poverty levels, unemployment rates, age demographics, disability prevalence and vehicle access, SVI provides a comprehensive view of the social context that shapes health and mobility behaviors. In this study, the SVI variables were selected to assess their impact on the relationship between walkability and health outcomes. For example, variables such as poverty levels (EP_POV150) and vehicle access (EP_NOVEH) can directly impact an individual's ability to benefit from walkable environments, while factors like disability status (EP_DISABL) highlight populations that might require specific infrastructure to facilitate safe and comfortable walking.

2.3. Analysis approach

The use of machine learning techniques in walkability research has seen notable growth. Techniques like the random forest algorithm have been utilized to analyze how built-environment characteristics influence body weight status (Kong et al. 2022). Moreover, studies such as those by (Guo et al. 2023) leverage random forest models to analyze the nonlinear effects of social characteristics and the built environment on choices of walking and cycling, considering both home and workplace environments. In addition, algorithms such as support-vector machines (Cesare et al. 2019), and fuzzy logic (Giabbanelli et al. 2013) demonstrated how sophisticated data analysis can significantly enhance our understanding of the obesity-environment relationship. So, employing machine learning (ML) in this study allows for the analysis of complex and nonlinear relationships between walkability indicators and health outcomes, which traditional statistical methods may not adequately capture.

The analysis employed a systematic approach to ensure robust model evaluation and generalizability. The dataset was initially divided into 80 % for training and 20 % for testing, preserving the test set for unbiased final validation. Within the training set, a 3-fold cross-validation was applied during the hyperparameter tuning phase using RandomizedSearchCV. This step identified the optimal parameters for each model by evaluating performance across three distinct data splits, ensuring that the model could generalize well to unseen subsets of the training data. After determining the best model parameters, the performance of each model was assessed on the reserved test set. Key metrics such as Mean Squared Error (MSE) and R^2 values were calculated to evaluate the accuracy and explanatory power of the models. To further confirm the robustness of the models, a 10-fold cross-validation was conducted across the entire dataset using the best-tuned model configurations. This additional validation step provided insights into the stability of the models' performance and their ability to generalize to different subsets of the data. This multi-step validation process, combining train-test splitting with nested cross-validation, ensured that overfitting was minimized and the reported metrics accurately reflected the models' predictive capabilities.

2.3.1. Data preprocessing

Before the data preprocessing phase for machine learning, a significant number of cases in the walkability data with missing values and nonsensical figures were identified. For instance, the variable

representing the distance to the nearest transit stop contained cases like −9999, which is not plausible. These cases have been excluded from the analysis. According to the (NWI, 2021) blanks or missing data may occur because data was not collected, not provided, or did not meet quality standards. After removing those cases, the sample size was 603. All these cases were at the census tract level. In addition, both the health data and walkability data were initially provided at different spatial levels. To integrate these datasets and ensure consistency, we used ArcGIS to align them at the census tract level, following a multi-step process. First, we imported both datasets into ArcGIS, ensuring that each dataset included geographic identifiers (such as census tract IDs) to facilitate spatial alignment. Next, we used the 'Spatial Join' tool to overlay the walkability data with the health data, assigning each health data point to the appropriate census tract based on geographic location. When health data covered larger areas than the walkability data, we used ArcGIS's 'Dissolve' tool to aggregate data to the census tract level, recalculating averages or summing values as needed. Finally, we validated the resulting dataset to ensure that each census tract contained one consistent set of walkability and health indicators. This integration allowed for a uniform dataset at the census tract level across the same spatial scale.

2.3.2. Machine learning parameters and evaluation metrics

The Python programming language was employed for data analysis. Initially, Pandas library was used to import and preprocess the data. We applied *train_test_split* for dividing the dataset, and *StandardScaler* for feature scaling. A variety of regression algorithms, including Linear Regression, Random Forest Regressor, Lasso, Gradient Boosting Regressor, and Support Vector Machine (SVR) were employed. Model performance was quantitatively assessed using Mean Squared Error (MSE) and R-squared. After that, we estimated feature importance and model coefficients based on the highest model accuracy. Lastly, after choosing the best model, the *Matplotlib* library facilitated the visualization of partial dependence plots, enabling us to graphically determine the effect of each feature.

The selection of algorithm parameters, a process known as hyperparameter tuning, is a critical step in the model development phase (Luo 2016). Grid Search for hyperparameter tuning was employed previously (Schratz et al. 2018) where by systematically testing specified ranges for each parameter to identify the optimal settings for each model as illustrated in Table 3. For the Random Forest Regressor, we tested different numbers of trees (50 to 500), depths (5 to 20, and None), minimum samples required for splitting, and features per split. In Lasso regression, we varied the alpha (regularization) values, while Gradient Boosting Regressor was tested across a range of estimators, learning rates, depths, and subsampling rates. Support Vector Regression (SVR) was tuned by adjusting the regularization parameter (C), epsilon, and kernel types. Following extensive experimentation, we identified the optimal parameters for each algorithm, representing the best-performing models. Specifically, we set the number of trees to 100 for Random Forest, an alpha of 0.1 for Lasso regression, 100 estimators for Gradient Boosting, and, for SVR, a regularization parameter (C) of 1.0 and epsilon of 0.2. These final parameter choices, informed by prior testing, achieved the highest accuracy and efficiency for each model.Table 4 represents the performance of all the algorithms.

Table 3
Machine learning algorithms parameters.

Algorithm	Parameter	Values
Random Forest Regressor	Number of Trees	[50, 100, 200, 300, 500]
	max_depth	[5, 10, 15, 20, None]
	min_samples_split	[2, 5, 10]
	max_features	['auto', 'sqrt', 'log2']
Lasso	alpha	[0.01, 0.1, 0.5, 1, 5, 10]
	max_iter	[500, 1000, 5000]
Gradient Boosting Regressor	n_estimators	[50, 100, 200, 300, 500]
	learning_rate	[0.01, 0.05, 0.1, 0.2, 0.3]
	max_depth	[3, 5, 10]
	min_samples_split	[2, 5, 10]
	subsample	[0.6, 0.8, 1.0]
Support Vector Regression	C	[0.1, 1, 10, 100]
	epsilon	[0.01, 0.1, 0.2, 0.5]
	kernel	['linear', 'poly', 'rbf', 'sigmoid']

Table 4
Machine learning algorithms performance.

Algorithm	Outcome	Best Parameters	Test MSE	Test R^2	Mean CV MSE
Random Forest	Cholesterol	[max_depth: 15, min_samples_split: 5, n_estimators: 500]	4.64	0.73	4.94
	Depression	[max_depth: 20, min_samples_split: 2, n_estimators: 500]	2.29	0.68	2.33
	Obesity	[max_depth: 20, min_samples_split: 2, n_estimators: 500]	7.68	0.81	8.22
	Blood pressure	[max_depth: None, min_samples_split: 10, n_estimators: 200]	2.54	0.86	2.79
	Mental Health	[max_depth: 10, min_samples_split: 5, n_estimators: 100]	1.55	0.80	1.64
SVR	Cholesterol	[kernel: linear, epsilon: 0.01, C: 100]	5.87	0.65	3.12
	Depression	[kernel: rbf, epsilon: 0.01, C: 10]	2.55	0.61	4.28
	Obesity	[kernel: rbf, epsilon: 0.01, C: 10]	7.93	0.80	3.01
	Blood pressure	[kernel: linear, epsilon: 0.5, C: 100]	3.04	0.79	5.06
	Mental Health	[kernel: linear, epsilon: 0.2, C: 1]	1.74	0.77	4.68
GBR	Cholesterol	[subsample: 0.6, n_estimators: 200, min_samples_split: 5, max_depth: 7, learning_rate: 0.05]	4.89	0.71	5.04
	Depression	[subsample: 0.8, n_estimators: 300, min_samples_split: 2, max_depth: 7, learning_rate: 0.01]	2.28	0.65	2.24
	Obesity	[subsample: 0.8, n_estimators: 300, min_samples_split: 2, max_depth: 7, learning_rate: 0.01]	2.29	0.65	2.33
	Blood pressure	[subsample: 0.6, n_estimators: 200, min_samples_split: 5, max_depth: 7, learning_rate: 0.05]	2.55	0.83	2.83
	Mental Health	[subsample: 0.6, n_estimators: 200, min_samples_split: 5, max_depth: 7, learning_rate: 0.05]	1.60	0.79	1.69
Lasso	Cholesterol	[max_iter: 500, alpha: 0.01]	5.7	0.66	5.68
	Depression	[max_iter: 500, alpha: 0.01]	2.71	0.58	2.71
	Obesity	[max_iter: 500, alpha: 0.01]	10.3	0.74	10.41
	Blood pressure	[max_iter: 500, alpha: 0.01]	2.92	0.79	3.26
	Mental Health	[max_iter: 500, alpha: 0.01]	1.71	0.78	1.8

3. Results

3.1. Models performance

The analysis revealed significant differences in model performance across outcomes, with Random Forest emerging as the most consistent performer. For example, Random Forest achieved the highest R2R^2 value of 0.819 for predicting Blood Pressure and showed robust performance across other outcomes, as evidenced by low test set MSEs and favorable cross-validation results. This model's flexibility and ensemble approach likely contributed to its strong performance in capturing complex relationships within the data. Support Vector Regression (SVR) also performed well in some cases, particularly for predicting Obesity (R2 = 0.636R^2 = 0.636). However, its performance was weaker for other outcomes, such as Cholesterol, where it recorded an R2R^2 of only 0.358, indicating limited explanatory power for this variable. Gradient Boosting Machines (GBM) displayed consistent performance, with notable success in predicting Mental Health (R2 = 0.697R^2 = 0.697) and Obesity (R2 = 0.644R^2 = 0.644). These results suggest that GBM's ability to model non-linear relationships is effective for certain health-related variables. Lasso regression demonstrated reasonable performance for predicting Mental Health (R2 = 0.692R^2 = 0.692) but struggled with outcomes like Cholesterol, where its R2R^2 value dropped to 0.386. This may be due to Lasso's tendency to shrink coefficients, which can limit its flexibility in capturing complex patterns when predictors are not highly informative. Overall, Random Forest was the strongest performer across outcomes, while GBM and Lasso showed promise for specific cases. The models with high R2R^2 values and low MSEs indicated substantial explanatory power and predictive accuracy, demonstrating their utility for understanding the relationships between predictors and health outcomes. These findings underscore the importance of selecting model types that align with the characteristics of the data and the predictive goals of the analysis.

3.2. Features importance for each health outcome

As shown in Fig. 2, the analysis reveals the substantial impact of socioeconomic, demographic, and environmental factors on various health outcomes. Among socioeconomic indicators, poverty stands out as a critical factor. A higher percentage of individuals living below the poverty line is most strongly associated with mental health challenges (0.30) and depression (0.29), underscoring the toll of financial hardship. Similarly, systemic inequities are evident in the influence of minority status (non-White, non-Hispanic individuals), which is significantly linked to obesity (0.55) and mental health (0.27). Educational attainment also plays a role, with lower levels of education measured by the percentage of individuals without a high school diploma—contributing to mental health issues (0.09) and depression (0.07). Furthermore, unemployment adds to these burdens, showing smaller yet noteworthy associations with mental health (0.05) and depression (0.03). Household composition and disability factors provide additional insights. The proportion of individuals aged five or older living with a disability is linked to cholesterol levels (0.06) and depression (0.04). Single-parent households with children under 18 years are also associated with adverse mental health outcomes (0.04) and, to a lesser extent, depression (0.02). In the realm of housing and transportation, mobile home living emerges as a notable factor, influencing both mental health (0.08) and cholesterol levels (0.06). Overcrowding in homes, defined as more than one person per room, also shows an association with mental health (0.03). Meanwhile, households without vehicle access and individuals residing in group quarters exhibit smaller effects, contributing less prominently to conditions such as depression and obesity.

Walkability-related features demonstrate varying degrees of influence on different health outcomes. For example, intersection density, an indicator of urban connectivity, plays a notable role in mental health (0.08) and obesity (0.05). This suggests that well-connected urban environments may encourage physical activity and contribute to improved mental well-being. The employment mix within neighborhoods is particularly relevant for cholesterol (0.03) and depression (0.02), pointing to the health benefits of diverse local economic opportunities. Similarly, the employment and household mix, though with lower importance scores for mental health (0.02) and depression (0.01), reflects how the balance of residential and employment spaces shapes environments that support overall well-being. Proximity to transit stops

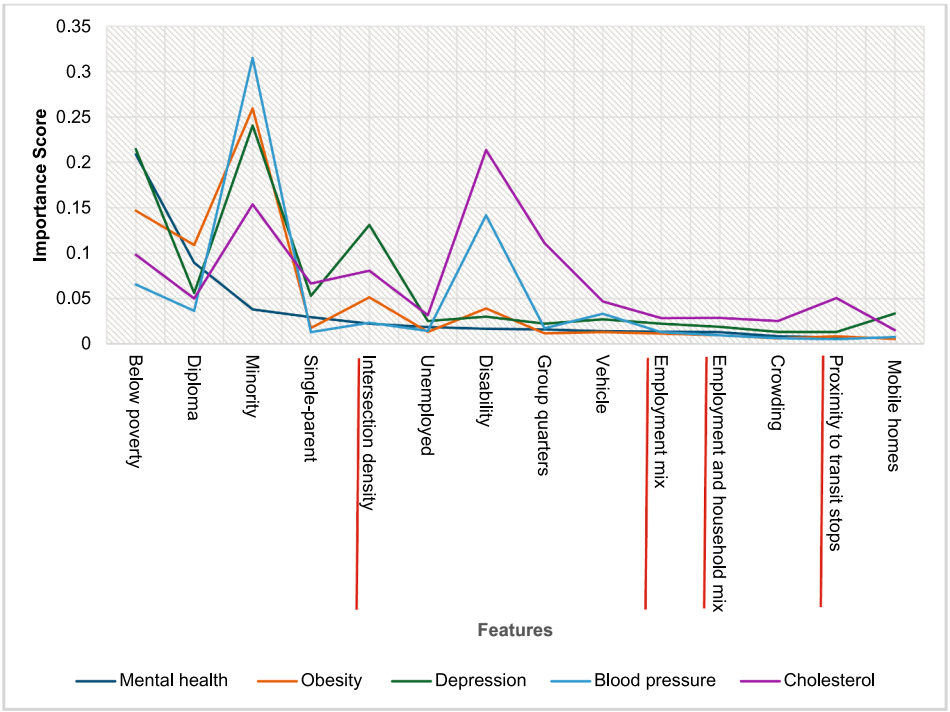

Fig. 2. Feature importance scores for predicting health outcomes.

also emerges as a factor, with small yet significant importance for depression (0.01) and mental health (0.02). This highlights how access to public transportation can enhance connectivity, reduce social isolation, and potentially improve mental health.

4. Discussion

According to the current literature, studies have emphasized the significant impact of neighborhood walkability on health outcomes. Increased walkability has been linked to higher levels of physical activity (Rundle et al. 2019), lower Body Mass Index (Creatore et al. 2016), better mental health (James et al. 2017, Li et al. 2021) and lower blood pressure (Sarkar et al. 2018, Howell et al. 2019). Therefore, improving walkability whether for transportation or leisure is highly encouraged. However, it is less clear to what extent walkability indicators have an impact on these health outcomes. The main contribution of this study is highlighting the effect of the walkability factors on various health outcomes using a machine learning approach by focusing on the State of Michigan as a case study. According to the results, intersection density emerged as the most influential walkability factor. Proximity to transit stops and employment mix showed relevance across outcomes. Additionally, socioeconomic factors such as poverty level and minority status were particularly impactful for mental health and depression, reflecting the critical influence of social disparities. Besides, features such as employment and household mix, crowding and disability status also contributed to specific outcomes. Therefore, addressing both walkability and broader social factors to enhance public health outcomes effectively impacts health outcomes.

Our study found that the RF algorithm has the highest accuracy level among other algorithms in studying the relationship between walkability and health outcomes. RF has shown the best performance in previous studies as well (Tribby et al. 2017, Yang et al. 2021, Hatami et al. 2023). Similarly, (Lotfata et al. 2023) applied geographical random forest models to assess the impact of socioeconomic neighborhood factors on hypertension prevalence in Chicago, revealing the varying importance of these factors across different areas. Overall, the Random Forest algorithm showed its effectiveness in unraveling nonlinear associations in these studies.

The results of the analysis emphasize the critical influence of socioeconomic, demographic, and environmental factors on health outcomes. Economic factors, as captured by the Percent individuals below poverty, consistently emerge as one of the strongest predictors across various health outcomes. Similarly, the high importance of Percent minority (all individuals except white, non-Hispanic) reveals the systemic inequities faced by marginalized populations, which contribute significantly to health disparities, particularly in obesity and mental health outcomes. Education, employment, and housing stability further shape these outcomes. The results also show the significant role of walkability and urban design in promoting health. Features like Intersection density, Employment mix, Employment and household mix, and Proximity to transit stops demonstrate that well-connected and economically diverse communities with accessible public transportation can create healthier lifestyles and mitigate risks associated with conditions such as depression (Berke et al. 2007, Gibney et al. 2020, Guo YingQi et al. 2020), obesity (Lee 2012), and mental health disorders. In addition, in assessing the relationship between neighborhood walkability and factors related to heart and metabolic health, it was found that enhancing neighborhood walkability could contribute to better cardiovascular health and cholesterol levels (Ruppar et al. 2014, del Pozo-Cruz et al. 2018). Similar results were found in the following studies in Canada as well (Loo et al. 2017, Howell et al. 2019). However, our results contradict with the findings of another recent study indicating that residential density as a parameter of walkability has no impact (de Courrèges et al. 2021).

4.1. Practical implications

The results point to several actionable strategies to improve health outcomes and reduce disparities. First, socioeconomic variables are essential to addressing the strong association between poverty and health risks, such as mental health disorders and depression. Policymakers can prioritize poverty reduction through financial assistance programs, affordable housing initiatives, and raising the minimum wage. Additionally, fostering local economic development and offering job training programs can reduce unemployment, thereby addressing a significant determinant of health. Second, health disparities faced by minority populations call for community-based and equity-focused initiatives. Outreach programs that are culturally tailored to the needs of these communities, coupled with equity-focused healthcare services, can address the disproportionate risks of obesity, depression, and other conditions. Furthermore, enforcing anti-discrimination policies in education, housing, and employment can help mitigate systemic inequities and support better health outcomes for marginalized populations. Third, urban planning and infrastructure improvements also have significant potential to enhance public health. Encouraging mixed-use development that increases employment mix and employment and household mix in neighborhoods can foster economic diversity and convenience, reducing stress and improving overall well-being. Expanding public transit access by ensuring proximity to transit stops can enhance connectivity, particularly for individuals without vehicles, and enable access to jobs, healthcare, and other essential resources. Fourth, housing instability, including issues such as overcrowding and reliance on mobile homes, presents another area for intervention. Providing affordable, stable housing options and improving living conditions can alleviate the stress associated with housing insecurity, leading to better health outcomes. Finally, integrated public health programs should also prioritize support for vulnerable groups. For instance, disability-inclusive services such as accessible healthcare and transportation can address the needs of individuals with disabilities.

4.2. Study limitations

This study has several limitations that warrant consideration. First, the research focused solely on Michigan, limiting the generalizability of the findings to regions with different urban layouts, socioeconomic conditions, and climates that may influence walkability and health outcomes differently. The analysis was conducted at the census tract level, potentially overlooking individual-level variations in health outcomes and their interaction with the urban environment. Furthermore, the study utilized the National Walkability Index (NWI) as the sole measure of walkability, which may exclude other critical walkability dimensions or indicators that could provide a more comprehensive understanding. While machine learning techniques captured non-linear relationships, the study may not fully account for the complex interactions between walkability, socioeconomic, and demographic factors. Temporal changes, such as evolving urban policies or demographic shifts, were also not considered, limiting insights into dynamic changes over time. Lastly, cases with missing or implausible data were excluded, which reduced the sample size and may have constrained the scope of the analysis. These limitations highlight the need for future research to include diverse regions, incorporate additional walkability metrics, and adopt finer spatial and temporal scales to better understand the intricate relationships between walkability and health outcomes.

5. Conclusion

In conclusion, this study employed a machine learning approach to study the relationship between walkability indicators and health outcomes, with a focus on Michigan. The findings emphasize the critical role of intersection density as the most influential walkability factor across several health metrics, particularly mental health and obesity.

Proximity to transit stops and employment mix also emerged as significant contributors to health outcomes, particularly for blood pressure, cholesterol, and depression. Additionally, socioeconomic and demographic factors, such as poverty and minority status, demonstrated a strong influence on mental health and depression, underscoring the interplay between urban design and social inequities. These results emphasize the necessity of integrating walkability improvements with targeted socioeconomic interventions to enhance public health. While the study advances our understanding of walkability's impact on health, its findings are subject to limitations related to regional scope, data completeness, and temporal dynamics.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Musab Wedyan: Writing – original draft, Methodology, Formal analysis, Data curation. **Fatemeh Saeidi-Rizi:** Writing – review & editing, Supervision, Investigation, Conceptualization.

Declaration of competing interest

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

Acknowledgements

This paper has greatly benefited from collaborative discussions with members of the HealthScape Lab at Michigan State University. We wish to express our gratitude to Michigan State University for providing the necessary infrastructure and resources to conduct this study.

References

Balcetis, E., et al., 2020. Keeping the goal in sight: Testing the influence of narrowed visual attention on physical activity. Pers. Soc. Psychol. Bull. 46 (3), 485–496.
Berke, E.M., et al., 2007. Protective association between neighborhood walkability and depression in older men. J. Am. Geriatr. Soc. 55 (4), 526–533.
CDC (2023). "Social Vulnerability Data." https://www.atsdr.cdc.gov/place-health/php/svi/index.html.
Cesare, N., et al., 2019. Use of social media, search queries, and demographic data to assess obesity prevalence in the United States. Palgrave Commun. 5 (1), 1–9.
Creatore, M.I., et al., 2016. Association of neighborhood walkability with change in overweight, obesity, and diabetes. J. Am. Med. Assoc. 315 (20), 2211–2220.
de Courrèges, A., et al., 2021. The relationship between neighbourhood walkability and cardiovascular risk factors in northern France. Sci. Total Environ. 772, 144877.
del Pozo-Cruz, J., et al., 2018. Replacing sedentary time: meta-analysis of objective-assessment studies. Am. J. Prev. Med. 55 (3), 395–402.
Finlay, J., et al., 2020. Fast-food for thought: Retail food environments as resources for cognitive health and wellbeing among aging Americans? Health Place 64, 102379.
Frank, L.D., et al., 2007. Stepping towards causation: do built environments or neighborhood and travel preferences explain physical activity, driving, and obesity? Soc. Sci. Med. 65 (9), 1898–1914.
Frank, L.D., et al., 2010. The development of a walkability index: application to the Neighborhood Quality of Life Study. Br. J. Sports Med. 44 (13), 924–933.
Giabbanelli, P.J., et al., 2013. Modelling the Joint Effect of Social Determinants and Peers on Obesity Among Canadian adults. Theories and Simulations of Complex Social Systems. Springer, pp. 145–160.
Gibney, S., et al., 2020. Age-friendly environments and psychosocial wellbeing: A study of older urban residents in Ireland. Aging Ment. Health 24 (12), 2022–2033.
Gu, P., et al., 2018. Using open source data to measure street walkability and bikeability in China: A case of four cities. Transp. Res. Rec. 2672 (31), 63–75.
Gunn, L.D., et al., 2017. Designing healthy communities: creating evidence on metrics for built environment features associated with walkable neighbourhood activity centres. Int. J. Behav. Nutr. Phys. Act. 14, 1–12.
Guo, Y., et al., 2019. Neighborhood environment and cognitive function in older adults: A multilevel analysis in Hong Kong. Health Place 58, 102146.
Guo, L., et al., 2023. Examining the nonlinear effects of residential and workplace-built environments on active travel in short-distance: A random forest approach. Int. J. Environ. Res. Public Health 20 (3), 1969.
Guo YingQi, G. Y., et al. (2020). Association of neighbourhood social and physical attributes with depression in older adults in Hong Kong: a multilevel analysis.
Hatami, F., et al., 2023. Non-linear associations between the urban built environment and commuting modal split: A random forest approach and SHAP evaluation. IEEE Access 11, 12649–12662.
Howell, N.A., et al., 2019. Association between neighborhood walkability and predicted 10-year cardiovascular disease risk: The CANHEART (Cardiovascular Health in Ambulatory Care Research Team) Cohort. J. Am. Heart Assoc. 8 (21), e013146.
James, P., et al., 2017. Built environment and depression in low-income African Americans and Whites. Am. J. Prev. Med. 52 (1), 74–84.
Kong, L., et al., 2022. How do different types and landscape attributes of urban parks affect visitors' positive emotions? Landsc. Urban Plan. 226, 104482.
Lathey, V., et al., 2009. The impact of subregional variations in urban sprawl on the prevalence of obesity and related morbidity. J. Plan. Educ. Res. 29 (2), 127–141.
Lee, K.-H., 2012. A Study on the Correlation between City's Built Environment and Residents' Health-A Case study of small and medium-sized cities in korea. J. Korea Acad.-Indust. Coop. Soc. 13 (7), 3237–3243.
Li, X., et al., 2021. Pathways between neighbourhood walkability and mental wellbeing: A case from Hankow, China. J. Transp. Health 20, 101012.
Liaw, A. (2002). "Classification and regression by randomForest." R news.
Loo, C.J., et al., 2017. Association between neighbourhood walkability and metabolic risk factors influenced by physical activity: a cross-sectional study of adults in Toronto, Canada. BMJ Open 7 (4) e013889.
Lotfata, A., et al., 2023. Using geographical random forest models to explore spatial patterns in the neighborhood determinants of hypertension prevalence across chicago, illinois, USA. Environ. Plann. B: Urban Anal. City Sci. 50 (9), 2376–2393.
Luo, G., 2016. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. Network Model. Anal. Health Inf. Bioinf. 5, 1–16.
Makhlouf, M.H., et al., 2023. Neighborhood walkability and cardiovascular risk in the United States. Curr. Probl. Cardiol. 48 (3), 101533.
Mayne, D.J., et al., 2013. An objective index of walkability for research and planning in the Sydney Metropolitan Region of New South Wales, Australia: an ecological study. Int. J. Health Geogr. 12, 1–10.
Melis, G., et al., 2015. The effects of the urban built environment on mental health: A cohort study in a large northern Italian city. Int. J. Environ. Res. Public Health 12 (11), 14898–14915.
Michigan, C.R.C.o., 2019. Exploring Michigan's Urban/Rural Divide. Michigan Public Policy Research - Citizen's Research Council.
MSU (2022). "Michigna Demographic Trends " https://ippsr.msu.edu/sites/default/files/LLP/22/MI_Demographic_Trends.pdf.
National Walkabiltiy Index (2021). "National Walkabiltiy Index ". from https://www.epa.gov/smartgrowth/national-walkability-index-user-guide-and-methodology.
NWI (2021). "National Walkability Index Methodology and User Guide." https://www.epa.gov/sites/default/files/2021-06/documents/national_walkability_index_methodology_and_user_guide_june2021.pdf.
Obermeyer, Z., Emanuel, E.J., 2016. Predicting the future—big data, machine learning, and clinical medicine. N. Engl. J. Med. 375 (13), 1216–1219.
PLACES (2019). https://www.cdc.gov/places/index.html.
Rissel, C., et al., 2012. Physical activity associated with public transport use—a review and modelling of potential benefits. Int. J. Environ. Res. Public Health 9 (7), 2454–2478.
Rundle, A.G., et al., 2019. Development of a neighborhood walkability index for studying neighborhood physical activity contexts in communities across the US over the past three decades. J. Urban Health 96, 583–590.
Ruppar, T.M., et al., 2014. Lipid outcomes from supervised exercise interventions in healthy adults. Am. J. Health Behav. 38 (6), 823–830.
Sarkar, C., et al., 2018. Neighbourhood walkability and incidence of hypertension: Findings from the study of 429,334 UK Biobank participants. Int. J. Hyg. Environ. Health 221 (3), 458–468.
Schratz, P., et al. (2018). "Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data." arXiv preprint arXiv: 1803.11266.
Solis-Urra, P., et al., 2020. The mediation effect of self–report physical activity patterns in the relationship between educational level and cognitive impairment in elderly: a cross-sectional analysis of chilean health national survey 2016–2017. Int. J. Environ. Res. Public Health 17 (8), 2619.
Stockton, J.C., et al., 2016. Development of a novel walkability index for London, United Kingdom: cross-sectional application to the Whitehall II Study. BMC Public Health 16, 1–12.
Sugiyama, T., et al., 2013. Initiating and maintaining recreational walking: a longitudinal study on the influence of neighborhood green space. Prev. Med. 57 (3), 178–182.
Thomas, J. and L. Zeller (2017). "National walkability index user guide and methodology." Environ. Prot. Agency: Washington, DC, USA.
Tribby, C.P., et al., 2017. Analyzing walking route choice through built environments using random forests and discrete choice techniques. Environ. Plann. B: Urban Anal. City Sci. 44 (6), 1145–1167.
Wang, Y., et al., 2016. A review on the effects of physical built environment attributes on enhancing walking and cycling activity levels within residential neighborhoods. Cities 50, 1–15.
Wang, Q., et al., 2024. Unraveling the dynamic relationship between neighborhood deprivation and walkability over time: A machine learning approach. Land 13 (5), 667.
Watson, K.B., et al., 2020. Associations between the national walkability index and walking among US adults—National Health Interview Survey, 2015. Prev. Med. 137, 106122.

Westenhöfer, J., et al., 2023. Walkability and urban built environments—a systematic review of health impact assessments (HIA). BMC Public Health 23 (1), 518.
Wu, Y.-T., et al., 2017. The built environment and cognitive disorders: results from the cognitive function and ageing study II. Am. J. Prev. Med. 53 (1), 25–32.
Yang, L., et al., 2021. To walk or not to walk? Examining non-linear effects of streetscape greenery on walking propensity of older adults. J. Transp. Geogr. 94, 103099.
Zuniga-Teran, A.A., et al., 2017. Designing healthy communities: Testing the walkability model. Front. Archit. Res. 6 (1), 63–73.