Travel Behaviour and Society 36 (2024) 100789

Contents lists available at ScienceDirect

Travel Behaviour and Society

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

Determining causality in travel mode choice

Rishabh Singh Chauhan [a,*], Christoffer Riis [b], Shishir Adhikari [c], Sybil Derrible [a], Elena Zheleva [c], Charisma F. Choudhury [d], Francisco Câmara Pereira [b]

[a] Department of Civil, Materials, and Environmental Engineering, University of Illinois Chicago, Chicago 60607, United States
[b] DTU Management, Transport Division, Technical University of Denmark, Kgs. Lyngby 2800, Denmark
[c] Department of Computer Science, University of Illinois Chicago, Chicago 60607, United States
[d] Institute for Transport Studies School of Civil Engineering University of Leeds, Leeds LS2 9JT, United Kingdom

ARTICLE INFO

Keywords:
Causality
Travel behavior
Activity based models
Travel mode choice models

ABSTRACT

This article presents one of the pioneering studies on causal modeling in travel mode choice decision-making using causal discovery algorithms. These models are a major advancement from conventional correlation-based techniques. We propose a novel methodology that combines causal discovery with structural equation modeling (SEM). This modeling approach overcomes some of the limitations of SEM by combining the strengths of both causal discovery and SEM. Causal discovery algorithms determine causal graphs from observational data and domain knowledge, and SEMs estimate direct causal effects and test the performance of causal discovery algorithms. In this study, we test four causal discovery algorithms: Peter-Clark (PC), Fast Causal Inference (FCI), Fast Greedy Equivalence Search (FGES), and Direct Linear Non-Gaussian Acyclic Models (DirectLiNGAM). The results show that DirectLiNGAM based SEM model best captures causality in mode choice behavior. It passes several goodness-of-fit tests, including Root Mean Square Error of Approximation (RMSEA) and Goodness-of-Fit Index (GFI), and it achieves the lowest Bayesian Information Criterion (BIC) value. The analyses are conducted on data collected from the 2017 National Household Travel Survey in the New York Metropolitan area.

1. Introduction

Understanding and modeling travel mode choice behavior is a classic problem in transportation (Derrible, 2019). The long-established practice for conducting these studies is by statistical models based on utility maximization theory, such as multinomial logit models, and nested logit models (Zhao et al., 2020; Cheng et al., 2019; Xie and Waller, 2010; Ma et al., 2017). These models make strong assumptions about the underlying relationships between the variables that are often considered unrealistic (Cheng et al., 2019). To overcome some of the limitations of conventional models, several machine learning-based methods for travel choice modeling have been proposed, including neural networks, extreme gradient boosting, random forest, and decision trees to name a few (Zhao et al., 2020; Cheng et al., 2019; Xie et al., 2003; Lee et al., 2019; Lee et al., 2018).

Despite their popularity, statistical and machine learning modeling techniques are still based on correlations (Pearl and Mackenzie, 2018) but as is well known, correlation does not imply causation. Two variables can have a mathematical relation (or correlation) between them without having a causal relation—a condition which is called 'spurious correlation' (Hujoel et al., 2006; Listl and Chiavegatto Filho, 2021). This is the reason behind the high correlation between extremely unlikely pairs of variables like cheese consumption versus fatal bedsheet tangling accidents and beef consumption versus deaths by lightning (Vigen, 2015). These examples highlight that correlations can be misleading and, thus, making policies based on correlations can be flawed.

Another limitation of commonly used travel mode choice models is that they assume that the explanatory variables affecting mode choice are independent of one another; i.e., these models do not consider the interaction between the factors influencing mode choice. This assumption is often unrealistic since the decision-making process is likely complex, with several variables directly or indirectly affecting mode choice decisions. For example, variables like household income and number of household members or age and education level are generally not independent to one another.

To address these issues, this article proposes a causality-based modeling approach for travel mode choice. Such models provide a complex graphical representation of the causal (or generative)

* Corresponding author.
E-mail address: Rchauh6@uic.edu (R.S. Chauhan).

https://doi.org/10.1016/j.tbs.2024.100789
Received 16 May 2023; Received in revised form 22 December 2023; Accepted 19 March 2024
Available online 4 April 2024

mechanisms involved in mode choice decision-making. To estimate these causal models, we introduce the following novel procedure:

- **First, the causal links are determined using causal discovery algorithms:** Causal discovery is the process of extracting causal structures from observational and/or experimental data (Shen et al., 2020). It is grounded in the complex concepts of causality and rooted in fields of statistics, economics, epidemiology, computer science, philosophy, and others (Nogueira et al., 2022). Since causal discovery deals with causation, and not simply correlations, they are well suited to provide policy implications. Several causal discovery algorithms have been proposed in the literature, which differ in methodology and assumptions (Heinze-Deml et al., 2018). We selected four popular algorithms for this analysis: Peter-Clark (PC), Fast Causal Inference (FCI), Fast Greedy Equivalence Search (FGES), and Direct Linear Non-Gaussian Acyclic Models (DirectLiNGAM). These algorithms represent a diverse set of causal discovery algorithms where each follows a unique methodology, as explained in the methods section.
- **Second, Structural Equation Models (SEM) are used to (1) determine the most suitable causal model and (2) compute the quantitative measure of the direct causal effects between the variables:** SEMs have been previously used to model mode choice (Wang et al., 2017; Golob et al., 1997; Golob and Hensher, 1998; Levine et al., 1999). They possess some distinct advantages over most other techniques. In particular, SEMs: (i) are capable of handling exogenous, endogenous, and latent variables; (ii) can account for indirect, multiple, and reverse relationships; (iii) accept non-normal data; (iv) offer easier visualization of the modeled network; and (v) can potentially provide causal estimates (Wang et al., 2017; Pearl, 2012). However, despite these advantages, SEMs have their limitations as well. A key limitation is their inherent property to be confirmatory tools instead of exploratory. Therefore, the modeler needs to provide a hypothesized structural graph before creating the SEM (Golob, 2003). As Bollen and Pearl (2013) pointed out, "…the SEM represents and relies upon the causal assumptions of the researcher. These assumptions derive from the research design, prior studies, scientific knowledge, logical arguments, temporal priorities, and other evidence that the researcher can marshal in support of them. The credibility of the SEM depends on the credibility of the causal assumptions in each application" (Bollen and Pearl, 2013). Given the importance of these causal assumptions, constructing a reasonable, hypothesized causal graph can be challenging. This might be particularly true in the case of large and complicated networks, like the mode choice decision-making process. Existing studies have usually had to rely only on domain knowledge and literature to build causal graphs. The lack of a dependable mechanism to generate a credible causal graph can create a substantial barrier in estimating a credible SEM. A standard, yet flawed, process to circumvent this limitation in SEM building is to alter a graph until the SEM achieves reasonable accuracy. This process is deemed controversial and can lead to an overfitted, unstable, unreliable, and incorrect model (Tarka, 2018; McCoach et al., 2007; Kelloway, 1995; MacCallum et al., 1992). Therefore, we propose using causal discovery (from step 1) as a precursor to the SEM. The causal structure obtained from causal discovery can be fed as an input to a SEM model. This new step helps overcome the above-mentioned limitation of the SEMs.

Ultimately, this union between causal discovery and SEM is mutually beneficial. Causal discovery gains the quantitative measures of the causal effect from the SEM, while the SEM benefits from the data-driven causal graphs extracted by the causal discovery algorithms to be used as its input.

Overall, the objectives of this study are to:

i. Apply causal discovery algorithms to determine a graphical causal model of travel mode choice decision-making;
ii. Compare the performance of the various causal discovery algorithms to determine the most appropriate algorithm for studying travel mode choice;
iii. Estimate quantitative causal effects among the variables affecting the travel mode choice;
iv. Introduce a novel methodology that combines causal discovery and an SEM to model travel mode choice.

Causal modeling of travel mode choice has been rarely attempted. Brathwaite and Walker (2018) have advocated for the incorporation of causal inference in travel demand modeling provided a theoretical framework for causal mode choice models. Xie and Waller (2010) applied a Bayesian Network and used both observational information along with cause-effect hypotheses to learn a causal graph for mode choice prediction. Similarly, Ma et al. (2017) used Structure Learning and unsupervised Bayesian Networks with domain knowledge to infer a causal graph. They explored three different classes of learning algorithms: constrained-based algorithms, score-based algorithms, and model averaging. More recently, Monteiro (2020) used a constrained-based causal discovery algorithm, Find One-Factor Cluster (FOFC), to estimate a causal graph for travel satisfaction as well as mode choice. The inferred graph was compared with a graph constructed based on domain knowledge. It was found that FOFC recovered many of the cause-effect relations but had an undesirable property of being dependent on the order of the inputs. They also saw some potential in FOFC to contribute to the hypothesis generation for an SEM. Previously, studies like Golob and Hensher (1998) have used an SEM to model mode choice and provided causal interpretations to their findings. These studies, however, had to rely on hypothesized causal graphs derived from domain knowledge and previous studies. It must be noted that there are studies within the field of transportation but not related to travel mode choice that have estimated causal models (Karwa et al., 2011; Dastjerdi, 2018), however, the discussion on such studies is beyond the scope of this paper.

Our study seeks to fill the gap in the research in two important ways. First, it applies, analyzes, and compares four causal discovery algorithms that have never been used before for mode choice modeling. These algorithms can not only confirm well established causal hypotheses, but they can also reveal new causal relationships in the data. Thus, this study contributes an important analysis of the usage of causal discovery in transportation. Second, our proposed approach of combining causal discovery algorithms and SEMs can support modelers in making causal assumptions that are data driven, more substantiated, and more reliable than those usually made in the past studies.

2. Methods

2.1. Key concepts

This subsection explains the main concepts relevant to this study.

Structural causal model (SCM): A structural causal model (Pearl, 2009) is used to model the causal assumptions in a domain by representing the relevant features and their interactions. A SCM, represented by $M(V, U, f)$, models how nature assigns values to the variables of interest using a set of variables U, V, and a set of functions f that assign values to each variable in V using other variables. The variable set U is termed as exogenous variables, which are external to the causal model and often considered errors or disturbances. The value of an endogenous variable $V_i \in V$ is explained by a function $f_i \in f$ of at least one exogenous variable $U_i \in U$ and optionally other endogenous variables; i.e., $V_i = f_i(V_{pa}, U_i)$, where $V_{pa} \subset V \setminus V_i$ is a set of direct causes of V_i, i is a specific unit, and V_{pa} is a set of parents of V_i (explained later). Therefore, a variable is a function of its known direct causes and unknown disturbances.

Causal graphical model: The causal relationship among the variables in an SCM can be represented using a directed acyclic graph (DAG) represented as *G(V, E)*, where *V* and *E* are a set of nodes (a.k.a. vertices) and links (a.k.a. edges) respectively. Each node (or vertex) has incoming links from its direct causes. $V_i \rightarrow V_j$ denotes a link from V_i to V_j, where V_i is the parent, V_j is the child, and j is another unit (Fig. 1). Two nodes are adjacent if there is a link between them. A directed path is a sequence of nodes obtained following the direction of the link. A graph is directed acyclic if there are no directed paths with repeated nodes. The nodes preceding the tail node of the directed path are the ancestors of the tail node. Similarly, the nodes following the head node of the directed path are descendants of the head node. Let the symbols $pa(V_i, G)$, $anc(V_i, G)$, and $des(V_i, G)$ indicate the sets of parents, ancestors, and descendants of V_i for graph G respectively. In the Markovian case, the exogenous variables are assumed to be independent of one another and are not explicitly shown in the graph.

Conditional independence relations: The data generated by an SCM should adhere to the conditional independence relations that the causal graphical model entails. The joint probability distribution *P* described by *G(V, E)* factorizes to the product of the conditional probability of each random variable given its parents according to the causal Markov assumption, i.e.,

$$P(V_1, V_2, ..., V_n) = \prod_{i=1}^{n} P(V_i|pa(V_i, G)) \tag{1}$$

The factorization in Eq. (1) follows the chain rule of probability theory and conditional independence (CI) relations entailed from the Causal Markov assumption. With the causal Markov assumption, a random variable is independent of all other variables except its parents and its descendants conditioned on the parents, i.e., $V_i \perp V \setminus \{V_i \cup pa(V_i, G) \cup des(V_i, G)\} \mid pa(V_i, G)$. The Markov conditions are not all the conditional independence relationships captured by the causal model. The notion of *d*-separation (Pearl, 1988) is used to read off all the conditional independencies that hold for any data distribution that is generated by the mechanism described by a graphical model. The rules of *d*-separation are formally defined with the help of three sub-graph structures: (1) a chain, $V_i \rightarrow V_j \rightarrow V_k$, with an unidirectional path, (2) a fork, $V_i \leftarrow V_j \rightarrow V_k$, with a common cause, and (3) a collider, $V_i \rightarrow V_j \leftarrow V_k$, with a common effect. An undirected path is said to be blocked by a node V_j with a conditioning set S of observed variables if one of two conditions hold: (i) $V_j \in S$ and V_j is not a collider or (ii) V_j is a collider and $V_j \in/ S \wedge des(V_j, G) \in/ S$. Two nodes are said to be *d*-separated by a conditioning set S if all the paths between the nodes are blocked by S. The *d*-separated nodes are independent of one another conditioned on set S (Pearl, 2009).

Causal structure learning (CSL): Causal discovery is transformed to the problem of CSL (Heinze-Deml et al., 2018) that concerns learning the adjacencies of nodes and the orientation of the edges in *G(V, E)* using an observational distribution. The idea of CSL is to utilize conditional independence in the data distribution to infer the structure of the causal graphical model. However, the same conditional independence relation

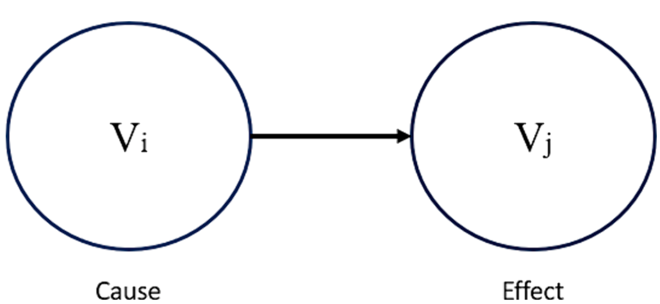

Fig. 1. Diagrammatic representation of a simple causal graph where V_i is a direct cause of V_j, and the edge denote the direction of causality. V_i is also the parent and ancestor of V_j, while V_j is a descendant of V_i.

can be satisfied by multiple causal models belonging to a Markov equivalence class. For example, the conditional independence relation $V_i \perp V_k | V_j$ is satisfied by the fork sub-graph $V_i \leftarrow V_j \rightarrow V_k$ as well as two chain sub-graphs $V_i \rightarrow V_j \rightarrow V_k$ and $V_i \leftarrow V_j \leftarrow V_k$. The causal structures entailing the same set of conditional independence relations belong to the Markov equivalence class. CSL generally concerns learning a Markov equivalence class of the underlying causal model. Additional parametric assumptions and domain knowledge are needed to identify a causal model within the Markov equivalence class.

Assumptions: In general, CSL methods make assumptions about the underlying data generating mechanism to learn a causal structure from observational data. Two common assumptions are: (1) *causal faithfulness* that implies that all the conditional independencies observed from the data distribution are entailed by the *d*-separation conditions of an underlying causal graph, and (2) *causal sufficiency* that refers to the absence of any unmeasured common causes of variables in *V*. Additionally, most CSL algorithms make assumptions of no selection bias and infinite sample size. These common assumptions may be relaxed by some algorithms. The next subsection briefly explicates the four causal discovery algorithms used in this study.

2.2. Causal discovery algorithms

This subsection describes the four causal discovery algorithms used.

PC (Spirtes et al., 2000; Colombo and Maathuis, 2014) is a constraint-based algorithm that uses conditional independencies in the data as constraints to estimate an equivalence class of the underlying SCM. It starts with fully connected undirected graph and progressively removes edges between conditionally independent variables (Glymour et al., 2019). It makes the causal sufficiency and causal faithfulness assumptions for the correctness of edge adjacencies. Then, *v*-structure discovery followed by Meek rules (Meek, 1995) produces an equivalence class graph, also known as a completed partial directed acyclic graph (CPDAG). An undirected edge $V_i - V_j$ in a CPDAG suggests both orientations $V_i \leftarrow V_j$ and $V_i \rightarrow V_j$ are possible for given data which can be oriented using the domain knowledge.

FCI (Spirtes et al., 2000) is another constraint-based algorithm that assumes causal faithfulness. The algorithm begins by assuming a fully connected undirected graph and then goes on removing the edges between variables that are conditionally independent (Shen et al., 2020). Specific details about the orientation of the edges can be found in Spirtes et al. (2000). FCI outputs a partial ancestral graph (PAG) to incorporate the hidden confounders using a bidirectional arrow, i.e., $V_i \leftrightarrow V_j$. In a PAG, $V_i \rightarrow V_j$ is interpreted as V_i being an ancestor of V_j and V_j not being an ancestor of V_i. Similarly, $V_i\, o\!\rightarrow V_j$ indicates either V_i is an ancestor of V_j or there is a hidden confounder between V_i and V_j.

FGES is an optimized and parallelized version of the Greedy Equivalence Search (GES) (Chickering, 2002) algorithm. FGES is a score-based method that approaches CSL as the problem of fitting a causal graph that best describes the conditional independencies in the data using a relevant score function. GES starts with an empty graph and greedily keeps adding edges that increase the goodness-of-fit score. The algorithm then removes the edges until the score does not improve to return the equivalence class of DAGs with the maximum score. FGES makes the causal sufficiency assumption but allows some violation in the faithfulness assumption. FGES returns a CPDAG as output similar to PC.

DirectLiNGAM (Shimizu, 2011) is a variation of the causal discovery algorithm called Linear Non-Gaussian Acyclic Models (LiNGAM). LiNGAM (Shimizu et al., 2006) is a functional causal model-based (or equivalently structural equation model-based) CSL algorithm. The functional causal models (Glymour et al., 2019) use SCM with additional assumptions on the distribution of U and V to distinguish between different DAGs in the same equivalence class and thus determine *cause* and *effect* from the observed data. It assumes causal sufficiency, linear continuous data generating process, and exogeneous variables with non-Gaussian distributions of non-zero variance. The non-Gaussian nature of noise enables asymmetric cause-effect relationships that help in identification beyond the equivalence class.

Detailed explanation of the algorithms can be found in additional resources (Heinze-Deml et al., 2018; Glymour et al., 2019; Tetrad Single HTML Manual, xxxx).

2.3. Structural equation model (SEM)

A SEM is a subclass of a structural causal model (SCM). The relationship between the target and explanatory variables are often represented as nonlinear and nonparametric functions in a SCM, whereas the functions in a SEM are often represented by linear relationships—this type of SEM is more precisely referred to as a *linear* SEM. In that sense, a SCM consists of a set of equations of the form $V_i = f_i(V_{pa}, U_i)$, where f_i is a nonparametric, nonlinear generalization of the linear SEM $V_i = \alpha_i + B V_{pa} + U_i$, where α_i is vector of intercept terms for the equation and B is the matrix of coefficient. Each equation in a SEM represents an autonomous mechanism, and if each variable has a distinct equation, then we can call the SEM an SCM (Pearl, 2009).

A SEM has two components: a structural model and a measurement model. The former is related to the hypothetical assumptions about the relations between the latent variables, while the latter deals with connecting latent variables to observed variables (Bollen and Noble, 2011). Mathematically, the linear structural model can be represented as:

$$\eta_i = \alpha_n + B\eta_i + \Gamma\xi_i + \zeta_i \tag{2}$$

where η_i refers to vector of latent endogenous variable for unit i; α_n is vector of intercept terms for the equation; B is the matrix of coefficient giving expected effects of latent endogenous variables (η_i) on each other; Γ denotes the coefficient matrix giving the expected effects of the latent exogenous variable (ξ) on latent endogenous variables (η); ζ_i is vector of disturbances (Bollen and Noble, 2011).

The measurement model can mathematically be represented by the following equations:

$$y_i = \alpha_y + \Lambda_y\eta_i + \varepsilon_i \tag{3}$$

$$x_i = \alpha_x + \Lambda_x\xi_i + \delta_i \tag{4}$$

where y_i is the vector of observed indicator η_i; x_i is the vector of observed indicator ξ_i ; Λ_y denotes matrix of factor loading or regression coefficients giving the impact of the latent variable η_i on y_i; Λ_x denotes matrix of factor loading or regression coefficients giving the impact of the latent variable ξ_i on x_i; ξ_i is the unique factors of y_i; δ_i is the unique factors of x_i (Bollen and Noble, 2011).

In the literature, several metrics have been proposed to evaluate the performance of SEMs. These include Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Goodness-of-Fit Index (GFI), Adjusted Goodness-of-Fit index (AGFI), Normed Fit Index (NFI), Tucker-Lewis Index (TLI), Akaike information criterion (AIC), Bayesian Information Criterion (BIC), and Chi-Square (χ^2) test. RMSEA is the indicator of the discrepancy of the model per degree of freedom (Wang et al., 2017). CFI indicates the amount of variance accounted for in a covariance matrix (Fan, 2016). GFI, AGFI, NFI, and TLI are all goodness-of-fit measures with variations in their formulations. AIC and BIC are the relative measures of the information lost when a model generates data and the extent of a model to be parsimonious respectively (Fan, 2016). Chi-square test indicates the discrepancy in the model and hence it should preferably be non-significant (Fan, 2016).

In this study, we estimated the SEMs using polychoric correlation based Unweighted Least Square (ULS) estimation method which is the recommended method for ordered categorical data (Xia and Yang, 2019). The reader is referred to additional resources (Wang et al., 2017; Bollen and Noble, 2011; Fan, 2016; Xia and Yang, 2019) to know more about the estimation methods and goodness of fit measures.

2.4. Data

In this study, we used the 2017 National Household Travel Survey (NHTS) data (National Household Travel Survey, 2017). These data are collected from a stratified random sample of U.S. households in all the 50 U.S. states and the District of Columbia. The data consists of information about each trip made by each household member on the household's travel day. Out of all the variables in the dataset, 14 variables were selected since these were suspected to affect the mode choice decision based on domain knowledge and previous studies (Xie and Waller, 2010; Ma et al., 2017). These variables could be grouped into three categories, namely trip characteristics, trip attributes, and socio-demographic information. The variables were discretized and converted to binary and ordinal variables to fit the requirements of some of the causal discovery algorithms used in this study.

To narrow the scope of this study, only the trips that were made using cars, public transport, or walking were studied. Additionally, any respondents of age less than 18 years were removed. Further, any trip data with unknown values (for example, 'not ascertained', 'I don't know', 'I prefer not to answer', 'appropriate skip') were removed from the dataset. Since the availability of transportation infrastructure varies substantially throughout the country which could affect travel mode choices, we reduced the dataset to only the trip that occurred in New York Metropolitan area (also called the New York-Newark-Jersey City, NY-NJ-PA metropolitan statistical area). The cleaned dataset used in the analysis consisted of a total of 21,618 observations. Further, the data were scaled between 0 and 1 before applying the causal discovery algorithms. Table S1 in the supplementary materials presents the list of the variables, along with their description, coding, and percentage distribution.

2.5. Proposed methodology

In this study, we propose to use a combination of causal discovery and SEMs to model travel mode choice. This methodology involves the following steps:

Step 1: Survey data together with (obvious) domain knowledge is inputted into a causal discovery algorithm.
Step 2: Causal graphs are obtained as an output from the causal discovery algorithm.
Step 3: The causal graph and the survey data are fed into an SEM.
Step 4: The SEM estimates the direct causal effects between the variables.
Step 5: The performance of the SEM is judged by a set of goodness-of-fit measures.
Step 6: Steps 1–5 are repeated for a different causal discovery algorithm.
Step 7: The algorithm that provides the best performance compared to the rest of the algorithms in step 5 is selected as the final causal graph and the model results are interpreted.

Py-causal (Wongchokprasitti et al., 2019) and Lingam (LiNGAM - Discovery of non-gaussian linear causal models, 2022) python libraries were used for causal discovery algorithms, CausalNex (Beaumont et al., 2021) library was used to draw causal graphs, and the semopy (Igolkina and Meshcheryakov, 2020) library was used for the SEMs. Fig. 2 illustrates the proposed methodology.

In step 1, domain knowledge was added to the causal discovery algorithms since they are known to improve the performance of the causal algorithms (Ma et al., 2017; Shen et al., 2020). Therefore, some graphical restrictions were added to the causal graphs based on the domain expertise. Care was taken to minimize the number of such restrictions and to add only the ones that are obvious. It must be noted that the goal of adding such obvious domain knowledge is only to restrict some search space for causal discovery algorithms by eliminating clearly

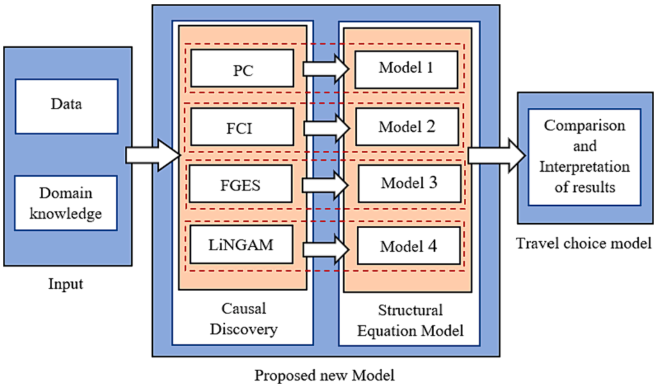

Fig. 2. Diagrammatic representation of the proposed new methodology.

illogical causal connections. The following domain knowledge restrictions were added:

- The three travel modes (car, public, and walk) were set as the target variable and hence were not allowed to cause any other variables.
- The trip characteristics were not allowed to cause the trip attributes and the socio-demographic variables.
- The trip attributes were assumed not to cause the socio-demographic variables.
- A few variables were assumed to be exogenous variables for the scope of this study. These were place type, gas price, race, age, and gender.
- Respondents' education level was assumed not to be caused by their worker status, vehicle ownership, and household income.

Since several causal discovery algorithms used in this study cannot detect latent confounders, for the sake of fair comparison, it was assumed that the causal mechanism behind mode choice decision making is not affected by any variables other than those mentioned in Table S1. In other words, we assume the causal Markov property and causal sufficiency for all the causal discovery algorithms.

3. Results

As a preliminary analysis, the correlations between the variables were computed. Since the variables are ordinal and binary, Spearman rank-order correlation was used. We found that the correlations were quite low, except for those between the travel modes. Only one correlation exceeds 0.5, between trip distance and walking (with a correlation value of −0.56). Thus, there were no highly correlated variables in the study data. Fig. S1 in the supplementary materials shows the heat map of the correlations between the variables.

Several model evaluation metrices were used to compare the SEMs developed from each of the four causal models as shown in Table 1. The results show that DirectLiNGAM-based SEM outperforms other models based on the evaluation metrices. DirectLiNGAM also achieves accepted levels of CFI, GFI, AGFI, NFI, TLI, and RMSEA. None of the models obtained the preferred non-significant p-value for chi-square test. However, this could perhaps be due to the large size of the study dataset (Fan, 2016).

Based on the results from the evaluation metrices, it can be concluded that DirectLiNGAM-based SEM are the most reliable out of the other models tested in this study. Fig. 3 shows the causal graph obtained from the DirectLiNGAM-based SEM model. The blue and red edges correspond to the positive and negative values of the corresponding path coefficients respectively. The thickness of the edges is made proportional to the magnitudes of the path coefficients. Due to space limitation, the graph is simplified by removing any edges with the value of path coefficient between 0.25 and −0.25. Figs. S2-4 in the supplementary materials show the causal graphs obtained for the PC-based, FCI-based, and FGES-based SEM models. Tables S2-4 show the model output from each of these models. Fig. S5 and Table S5 present the complete graphical causal model and model results from the DirectLiNGAM-based SEM which is found to be most accurate.

Table 1
Modeling performance of the various SEMs. Bolded figures show the best result.

	PC-based	FCI-based	FGES-based	Direct LiNGAM-based	Standard accepted levels
Comparative fit index (CFI)	0.922	0.803	0.942	**0.986**	≥0.95
Goodness-of-fit index (GFI)	0.922	0.803	0.941	**0.986**	Closer to 1.0 is preferred.
Adjusted goodness-of-fit index (AGFI)	0.896	0.755	0.918	**0.960**	Closer to 1.0 is preferred.
Normed fit index (NFI)	**0.922**	0.803	**0.941**	**0.986**	>0.90
Tucker-Lewis index (TLI)	0.897	0.755	**0.918**	**0.960**	>0.90
Root Mean Square Error of Approximation (RMSEA)	0.094	0.155	0.084	**0.055**	<0.06
Akaike information criterion (AIC)	71.99	10.10	84.26	$\mathbf{-3.15}\cdot 10^{16}$	Lower the better
Bayesian information criterion (BIC)	399.22	209.64	467.36	$\mathbf{-3.15}\cdot 10^{16}$	Lower the better
Log likelihood	5.01	19.95	5.87	$1.57\cdot 10^{16}$	
Degree of freedom	95	66	88	50	
Degree of freedom baseline	126	82	124	141	
Chi-squared test	18,111	34,326	13,408	3326	
p-value for chi-squared test	0.000	0.000	0.000	0.000	>0.05
Chi-squared test baseline	231,024	174,027	229,090	231,594	
Number of links/edges	31	16	36	91	

Given that the DirectLiNGAM-based SEM model was built on causal structural graph and has passed the goodness-of-fit tests, the interpretation of the result from this model can be done causally. Noted computer scientist Judea Pearl has advocated for interpreting SEMs causally (Pearl, 2012). On SEMs, he explained that "The 'path coefficient,' β, quantifies the (direct) causal effect of X on Y. Once we commit to a particular numerical value of β, the equation claims that a unit increase for X would result in β units increase of Y regardless of the values taken by other variables in the model, regardless of the statistics of U_X and U_Y, and regardless of whether the increase in X originates from external manipulations or variations in U_X." (Pearl, 2012) Here U_X and U_Y denotes exogenous variables in the model.

The rest of this section discusses the results from the selected DirectLiNGAM-based SEM, focusing on mode choice decision making.

Based on the values of path coefficients, choosing car as a travel mode has strong positive direct causal effect from number of vehicles owned (1.357) and trip distance (0.181). It has strong negative causal effects from household size (-0.573), white race (-0.319), and household income (-0.215). Further, small direct causal effects (path coefficients between 0.1 and −0.1) are caused by place type, education level, peak hour, age, gender, home-based trip, worker status, and weekday.

Similarly, choosing public transit as travel mode has strong positive direct causal effect from trip distance (1.018), household size (0.951), white race (0.742), and weekday (0.123). It has strong negative causal effect from number of vehicles owned (-1.858), place type (-0.356), and home-based trip (-0.103). Small direct causal effects are caused by

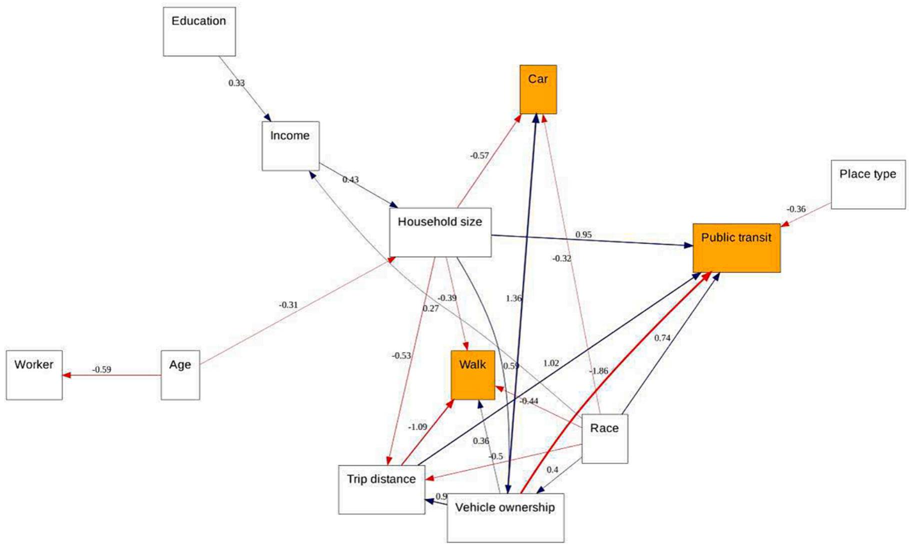

Fig. 3. Simplified results from the DirectLiNGAM-based SEM.

household income, gas price, peak hour, age, and worker status.

Lastly, choosing walking has strong positive direct causal effect from number of vehicles owned (0.361), place type (0.243), and home-based trip (1.72). It has strong negative causal effect from trip distance (−1.093), white race (−0.436), and household size (−0.387). Small direct causal effects are caused by household income, worker status, gender, education, weekday, gas place, and age. These results seem to be logically correct.

4. Discussion

Over recent years, the field of transportation has seen a growing popularity in using advanced techniques from statistics, machine learning, and deep learning (Chauhan, 2019; Chauhan et al., 2020; Taghipour et al., 2022; Lin, et al., 2019; Parsa et al., 2020). Almost all these techniques are based on correlations. Introducing the concept of causality in transportation is a step forward (Xin et al., 2022) and to do so, this study introduces a novel methodology for causal modeling of travel modes. Several noteworthy observations were made during this analysis which highlights the challenges and opportunities in the application of causality in transportation modeling. These include:

i. **Comparison between the causal discovery algorithms:** The four causal discovery algorithms tested in this study differ in their assumptions and methodology. These differences are reflected in the causal graphs they produced. These causal graphs varied drastically in terms of their complexity. DirectLiNGAM produced the most complex causal graph while FCI generated the simplest of all. The number of edges in the causal graph from DirectLiNGAM are more than five times larger than that from FCI. The algorithms agreed on some while disagreed on other cause-and-effect pairs. For example, all four algorithms agreed that number of vehicles owned and trip distance have a direct causal effect on choosing public transit as well as on choosing to walk. In contrast, no single common variable was found to be a direct cause of choosing car from all four algorithms. All the algorithms were found to benefit from the domain knowledge that was inputted to them.

ii. **Further potential in FCI:** The FCI algorithm performed the worst in almost all the evaluation metrices. Nevertheless, this might be because the full potential of FCI was not explored in this study. FCI does not hold the causal sufficiency assumption and hence can suggest the presence of unobserved cofounders. In our analysis, FCI discovered three unobserved latent variables (Fig. S3) and indicated the possibility of some more. For comparison purposes, we assumed that the causal mechanism behind mode choice decision making is not affected by any other variables than the 14 variables selected in this study (Table S1). This limited the usefulness of FCI. Future studies can explore the potential of FCI in detecting unobserved latent variables in transportation modeling.

iii. **Uncovering complex societal issues using causal discovery:** Unless restricted, the causal discovery algorithms find causal relations between all the variables in the dataset. This could lead to the detection of some nontrivial causal connections. For instance, all four algorithms found a causal link from race to number of vehicles owned, and from race to household size. This highlights the potential of causal discovery in discovering complex socio-economic issues. While the analysis of the validity of these causal links is beyond the scope of this study, we do not suggest that race is a cause of any of the travel behavior associated variables. The variable 'race' was kept in this study because of a long-established practice of including race in mode choice modeling, which is expected to serve as a proxy for a complex mix of socio-cultural, political, and economical factors, instead of racial/ethnical differences. Further, it was found that removing race from the analysis decreases the performance of the models substantially. The causal effects from variable 'race' also emphasizes on the care that must be taken while interpreting causal discovery models which might be based on the assumption of causal sufficiency. Causal sufficiency implies that there exist no unobserved confounders between the variables. However, it is highly likely that there are some unobserved exogenous variables that exist between race and other variables.

iv. **Domain knowledge:** The use of domain knowledge proposed in our methodology is very different from that used in SEMs alone. The SEM model estimation require the input of the whole graphical structure derived from domain knowledge. This is a major challenge since it is virtually impossible to come up with a complete and accurate causal graph from domain knowledge alone. In contrast, the goal of adding domain knowledge in our methodology is to restrict the search space of the algorithms by eliminating any illogical causal relations. Causal algorithms provide the option of removing/adding certain causal relationships. Previous research has found domain knowledge to improve the performance of causal discovery (Ma et al., 2017). Though we recommend that only obvious (minimum) domain knowledge should be added to avoid adding bias.

v. **Presence of non-manipulable variables:** Some researchers believe in 'no causation without manipulation' and are skeptical about determining causal effects from non-manipulable variables like gender and race (Bollen and Pearl, 2013; Pearl, 2018). Other researchers have found this practice legitimate (Bollen and Pearl, 2013; Pearl, 2018). However, based on the arguments proposed by Pearl (Pearl, 2018), we decided to keep non-manipulable variables in our models. Pearl (Pearl, 2018) suggested that even if some variables may be non-manipulable, knowing their causal effects can be useful for guiding policy making.

vi. **Caveats of the proposed methodology:** In this study, we generated causal graphs from causal discovery algorithms and (minimum) domain knowledge. These causal graphs were used to estimate several SEMs. The SEM that fits the data best was selected as the final graph (which in our analysis was the DirectLiNGAM-based SEM). Our methodology makes an important contribution in laying out the procedure for causal modeling. However, it must be noted that the final graph selected may not be the ultimate perfect causal graph and there is always the possibility to find a more accurate graph. Bollen and Pearl (Bollen and Pearl, 2013) explained this property of SEM as "Fitting the data does not 'prove' the causal assumptions, but it makes them tentatively more plausible. Any such positive results need to be replicated and to withstand the criticisms of researchers who suggest other models for the same data" (Bollen and Pearl, 2013). Despite this caveat, we argue that the same is true with any other transportation data modeling technique. Next, it must also be noted that causal discovery algorithms and SEMs are based on certain assumptions and limitations. One of these is that the algorithms used in this study cannot detect bidirectional cause-and-effect links; i.e., the causal graph do not have a provision of feedback loops. This could be an interesting research topic for future studies. Another important assumption that we made was that there are no variables other than those selected by the authors (Table S1) that affects mode-choice decision. However, it is likely that there can be more variables in the final causal graph (simplified in Fig. 3 and full graph in Fig. S5) that were out of the scope of this study.

vii. **Contribution of the proposed methodology:** The long-established approach for mode choice modeling is correlation based. SEM modeling has been proposed as an alternative approach (Wang et al., 2017; Golob et al., 1997; Golob and Hensher, 1998; Levine et al., 1999). SEMs could be considered a causal model; however, its accuracy depends on the accuracy of the hypothesized causal graph supplied to them (Bollen and Pearl, 2013). These hypothesized causal graphs have almost always been deduced from domain knowledge. Our proposed methodology focuses on causal modeling by advancing the SEM approach by combining it with causal discovery algorithms. This combination is mutually beneficial. Causal discovery algorithms can extract causal graphs directly from the observational data, which can be used as an input to the SEM. Subsequently, the SEM can estimate the quantitative direct causal effects for the causal graph. In addition, the goodness of fit measures obtained from the SEM models can be used to conduct a comparative evaluation of the performance of the various causal discovery algorithms. To do such a comparison, previous studies had to rely on the proximity of a causal graph with the 'true' causal graph. This true causal graph is created based on the expert's knowledge or from the knowledge of data creating process (e.g., in simulation studies) (Shen et al., 2020; Heinze-Deml et al., 2018). Yet, creating a true causal graph is often challenging, if not impossible, in complex real-world scenarios. Our suggested methodology, therefore, provides a data-driven approach to compare the performance of the various causal discovery algorithms. Further, this methodology advances mode choice modeling and is a step forward towards developing reliable causal models as opposed to the traditional correlation-based approach. The resulting causal graph presents the visual representation of the complexity involved in mode choice decision making. The variables known to be affecting mode choice were found to have mutual causal connections. This contradicts the assumption made by most of the statistical and machine learning approaches to mode choice modeling. Our modeling results suggest that the novel methodology is intensive and dependable.

5. Conclusion

Overall, the causal discovery-based SEM is a dependable methodology. The biggest advantage of this approach is that it helps SEMs by providing them with data-driven causal graphs making them more objective. Our study found that DirectLiNGAM is the most accurate algorithm for mode choice modeling out of the four algorithms tested. In the comparative analysis, the DirectLiNGAM-based SEM passes several goodness-of-fit tests like RMSEA, CFI, GFI, AGFI, NFI, and TLI, and achieves the lowest AIC and BIC values. The implementation of this DirectLiNGAM-based SEM method provided insights into the complex processes involved in travel decision making. The study identified several logically reasonable variables that cause travel mode choice.

The study has potential for improvement: (1) The most notable one could be to include latent or unobserved confounding variables. (2) The level-of-service attributes of the different modes were not included in the model due to data limitations. (3) Non-linear causal relations could be studied. (4) Testing the proposed framework for other mobility choice contexts, for instance residential location, can help to generalize the findings. (5) The model performance can be tested in different scenarios like more/less number of variables, more/less correlated data, more/less noisy data, or different levels of domain knowledge. (6) Causal models could be developed using continuous variables. These improvements can be part of future research.

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.

Data availability

The dataset analysed during the current study were conducted on NHTS data which are available at https://nhts.ornl.gov/.

Acknowledgments

This work was supported in part by the United States National Science Foundation (NSF) under Grant No. 1551731. Professor Charisma Choudhury's time was supported by the UKRI Future Leader Fellowship [MR/T020423/1].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tbs.2024.100789.

References

Beaumont, P., Horsburgh, B., Pilgerstorfer, P., Droth, A., Oentaryo, R., Ler, S., Nguyen, H., Ferreira, G.A., Patel, Z., Leong, W., 2021. CausalNex [Computer software]. https://github.com/quantumblacklabs/causalnex.

Bollen, K.A., Pearl, J. Eight myths about causality and structural equation models. in Handbook of causal analysis for social research 301–328 (Springer, 2013).

Bollen, K.A., Noble, M.D., 2011. Structural equation models and the quantification of behavior. Proc. Natl. Acad. Sci. 108, 15639–15646.
Brathwaite, T., Walker, J.L., 2018. Causal inference in travel demand modeling (and the lack thereof). J. Choice Model. 26, 1–18.
Chauhan, R., Shi, Y., Bartlett, A., Sadek, A.W., 2020. Short-term traffic delay prediction at the Niagara frontier border crossings: Comparing deep Learning and statistical modeling approaches. J. Big Data Anal. Transp. 2, 93–114.
Chauhan, R. S. Short-Term Traffic Delay Prediction at the Niagara Frontier Border Crossings Using Deep Learning. (2019).
Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F., 2019. Applying a random forest method approach to model travel mode choice behavior. Travel Behav. Soc. 14, 1–10.
Chickering, D.M., 2002. Optimal structure identification with greedy search. J. Mach. Learn. Res. 3, 507–554.
Colombo, D., Maathuis, M.H., 2014. Order-independent constraint-based causal structure learning. J Mach Learn Res 15, 3741–3782.
Dastjerdi, A.M., 2018. ATIS and commuters behavior: factors affecting behavioral intentions toward the use of. Travel Information Technology.
Derrible, S., 2019. Urban engineering for sustainability. MIT Press.
Fan, Y., et al., 2016. Applications of structural equation modeling (SEM) in ecological studies: an updated review. Ecol. Process. 5, 1–12.
Glymour, C., Zhang, K., Spirtes, P., 2019. Review of causal discovery methods based on graphical models. Front. Genet. 10, 524.
Golob, T.F., 2003. Structural equation modeling for travel behavior research. Transp. Res. Part B Methodol. 37, 1–25.
Golob, T. F., Kitamura, R. & Supernak, J. A panel-based evaluation of the San Diego I-15 Carpool Lanes Project. in Panels for Transportation Planning 97–128 (Springer, 1997).
Golob, T.F., Hensher, D.A., 1998. Greenhouse gas emissions and australian commuters' attitudes and behavior concerning abatement policies and personal involvement. Transp. Res. Part Transp. Environ. 3, 1–18.
Heinze-Deml, C., Maathuis, M.H., Meinshausen, N., 2018. Causal structure Learning. Annu. Rev. Stat. Its Appl. 5, 371–391.
Hujoel, P., Cunha-Cruz, J., Kressin, N., 2006. Spurious associations in oral epidemiological research: the case of dental flossing and obesity. J. Clin. Periodontol. 33, 520–523.
Igolkina, A.A., Meshcheryakov, G., 2020. semopy: a python package for structural equation modeling. Struct. Equ. Model. Multidiscip. J. 27, 952–963.
Karwa, V., Slavković, A. B. & Donnell, E. T. Causal inference in transportation safety studies: Comparison of potential outcomes and causal diagrams. (2011).
Kelloway, E.K., 1995. Structural equation modelling in perspective. J. Organ. Behav. 16, 215–224.
Lee, D., Derrible, S., Pereira, F.C., 2018. Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling. Transp. Res. Rec. 2672, 101–112.
Lee, D., Mulrow, J., Haboucha, C.J., Derrible, S., Shiftan, Y., 2019. Attitudes on autonomous vehicle adoption using interpretable gradient boosting machine. Transp. Res. Rec. 2673, 865–878.
Levine, J., Park, S., Wallace, R.R., Underwood, S.E., 1999. Public choice in transit organization and finance: the structure of support. Transp. Res. Rec. 1669, 87–95.
Lin, L. et al. Developing Predictive Border Crossing Delay Models. (2019).
LiNGAM - Discovery of non-gaussian linear causal models. (2022).
Listl, S., Chiavegatto Filho, A. D. Big data and machine learning. in Oral Epidemiology 357–365 (Springer, 2021).
Ma, T.-Y., Chow, J.Y., Xu, J., 2017. Causal structure learning for travel mode choice using structural restrictions and model averaging algorithm. Transp. Transp. Sci. 13, 299–325.
MacCallum, R.C., Roznowski, M., Necowitz, L.B., 1992. Model modifications in covariance structure analysis: the problem of capitalization on chance. Psychol. Bull. 111, 490.
McCoach, D.B., Black, A.C., O'Connell, A.A., 2007. Errors of inference in structural equation modeling. Psychol. Sch. 44, 461–470.
Meek, C., 1995. Causal inference and causal explanation with background knowledge in uncertainty in artificial. Intelligence 11.
Monteiro, M. M. Adaptation of transnational short-term residents: understanding the factors influencing residential location choice and travel behavior. (2020).
National Household Travel Survey. https://nhts.ornl.gov/.
Nogueira, A.R., Pugnana, A., Ruggieri, S., Pedreschi, D., Gama, J., 2022. Methods and tools for causal discovery and causal inference. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 12, e1449.
Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S., Mohammadian, A.K., 2020. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accid. Anal. Prev. 136, 105405.
Pearl, J., 1988. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann.
Pearl, J., 2009. Causality. Cambridge University Press.
Pearl, J., 2018. Does obesity shorten life? Or is it the soda? On non-manipulable causes. *J. Causal Inference* 6.
Pearl, J., Mackenzie, D., 2018. The book of why: the new science of cause and effect. Basic Books.
Pearl, J. The causal foundations of structural equation modeling. (2012).
Shen, X., Ma, S., Vemuri, P., Simon, G., 2020. Challenges and opportunities with causal discovery algorithms: application to Alzheimer's pathophysiology. Sci. Rep. 10, 1–12.
Shimizu, S., et al., 2011. DirectLiNGAM: a direct method for learning a linear non-gaussian structural equation model. J. Mach. Learn. Res.-JMLR 12, 1225–1248.
Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A., Jordan, M., 2006. A linear non-Gaussian acyclic model for causal discovery. J. Mach. Learn. Res. 7.
Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D., 2000. Causation, prediction, and search. MIT Press.
Taghipour, H., Parsa, A.B., Chauhan, R.S., Derrible, S., Mohammadian, A.K., 2022. A novel deep ensemble based approach to detect crashes using sequential traffic data. IATSS Res. 46, 122–129.
Tarka, P., 2018. An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. Qual. Quant. 52, 313–354.
Tetrad Single HTML Manual. https://cmu-phil.github.io/tetrad/manual/#alpha.
Vigen, T. Spurious correlations. (Hachette UK, 2015).
Wang, Y., Yan, X., Zhou, Y., Xue, Q., 2017. Influencing mechanism of potential factors on passengers' long-distance travel mode choices based on structural equation modeling. Sustainability 9, 1943.
Wongchokprasitti, C., Hochheiser, H., Espino, J., Maguire, E., Andrews, B., Davis, M., Inskip, C.. bd2kccd/py-causal v1.2.0 (v1.2.0). Zenodo. https://doi.org/10.5281/zenodo.3364590.
Xia, Y., Yang, Y., 2019. RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: the story they tell depends on the estimation methods. Behav. Res. Methods 51, 409–428.
Xie, C., Waller, S., 2010. Estimation and application of a bayesian network model for discrete travel choice analysis. Transp. Lett. 2, 125–144.
Xie, C., Lu, J., Parkany, E., 2003. Work travel mode choice modeling with data mining: decision trees and neural networks. Transp. Res. Rec. 1854, 50–61.
Xin, Y., Tagasovska, N., Perez-Cruz, F. & Raubal, M. Vision paper: causal inference for interpretable and robust machine learning in mobility analysis. in 1–4 (2022).
Zhao, X., Yan, X., Yu, A., Van Hentenryck, P., 2020. Prediction and behavioral analysis of travel mode choice: a comparison of machine learning and logit models. Travel Behav. Soc. 20, 22–35.