DTU Library

Best Of Both Worlds

System Thinking Approach for Transportation Data-Driven Decision-Making

Grinberg-Rosenbaum, Keren Or; Shiftan, Yoram; Pereira, Francisco Camara; Nahmias-Biran, Bat Hen

Published in:
Transportation Research Procedia

Link to article, DOI:
10.1016/j.trpro.2024.12.003

Publication date:
2025

Document Version
Publisher's PDF, also known as Version of record

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 82 (2025) 3943–3959

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2023 Montreal 17-21 July 2023

Best Of Both Worlds: System Thinking Approach for Transportation Data-Driven Decision-Making

Keren-Or Grinberg-Rosenbaum[1,2] *, Yoram Shiftan[1], Francisco Camara Pereira[2], Bat Hen Nahmias-Biran[3]

[1]Technion, Technion City, Haifa 32000, Israel
[2]Technical University of Denmark, Anker Engelunds Vej 101, Kongens Lyngby 2800, Denmark
[3]Tel Aviv University, Ramat Aviv 6997801, Israel

Abstract

A great interest of Transport Management Centers (TMCs) operators is to use data to make their decisions accordingly. Despite increase in accessible data, current methods has various gaps from imposing strong constraints (e.g., parametric function form) in traditional statistical methods to relying on statistical associations in Machine Learning (ML) tools. Defining causal knowledge from the transportation domain for ML models can potentially overcome those gaps yet it is done implicitly without a formal framework. This interdisciplinary research proposes a Hybrid Dynamical Systems Thinking Approach (HDSTA), using systems thinking for causality interface implementation for data-driven decisions in transportation. HDSTA provide guidelines on how different parties can work together to define a knowledge graph for the transportation system model. The graphical and text description outputs will serve experts in choosing and defining variables' cause-effect relationship; data scientists in defining a causal function; and TMCs in making data-driven decisions for the public benefit.

Peer-review under responsibility of the scientific committee of the 16th World Conference on Transport Research

* Corresponding author
E-mail address: kerenor@campus.technion.ac.il

2352-1465
Peer-review under responsibility of the scientific committee of the 16th World Conference on Transport Research
10.1016/j.trpro.2024.12.003

Keywords: Causality; Systems Thinking, Object Process Methodology (OPM); Machine Learning, Decision-Making; Bus Delays;

1. Introduction

Understanding the value of data in operational decisions improves Transport Management Centers (TMCs) in surface transportation systems (Edelstein, 2001). However, finding the appropriate methodology for both extracting the relevant variables and representing the causal relations from purely observational data is a fundamental problem in science (Pearl, 2009). The challenge is greater for transportation systems that may show a nonlinear nature: the components within one system intertwine with each other, and the dynamics of each component cannot be separated from one another. Furthermore, transportation systems also involve a substantial number of variables and data streaming from diverse sources and sub-systems, for example, weather, traffic lights, and public transportation.

In recent years, new Machine Learning (ML) tools were developed to process big data and to allow complex nonlinearities, such as Neural Networks (Basu et al, 2010) and gaussian processes (Rasmussen & Nickisch, 2010). ML has many advantages in the volume, variety, and velocity of big data processing, yet to prevent a false representation of the relationship between variables, which is not necessarily a causal relationship (Guo, Cheng, Hahn & Liu, 2020), an ML model requires some sort of causality formalism.

While many models used in transportation are explicitly causal, particularly when they are used to evaluate the impact of external interventions on travel outcomes, extracting those causal relations require human intelligence, referred to as experts' judgment. Current causality formalisms have limitations regarding the details and dynamics they can represent and rarely reflect in TMC's currently used descriptive tools, such as heat maps, graphs, and numeric service values measured by sensors or apps. Representing those cause-effects in other ways, such as written reports (Babar, & Arif 2019) or pure mathematical models, are not effective tools either for nontechnical stakeholders such as TMCs.

To bridge those limitations and benefit the data collected, this interdisciplinary research presents a Hybrid Dynamical Systems Thinking Approach (HDSTA) that uses systems thinking approach to formalize causality known in transport theory and integrate it for data-driven decisions, as can be seen in Fig. 1.

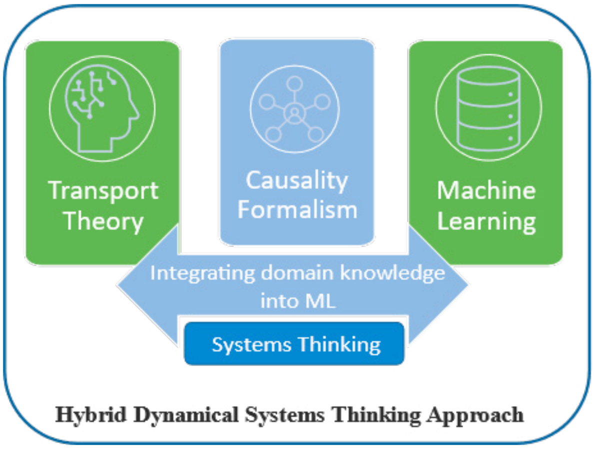


Fig. 1. Research interdisciplinary gap

HDSTA support a holistic view of the transportation system, representing a "bird's eye view" of the system's different layers. Object Process Methodology (OPM) formalized by Dori (1995) was selected to apply the Systems thinking approach. OPM can encode assumptions and domain knowledge about a problem and thereby enable the examination of interventions, without the need to build and run a full simulation. HDSTA outcomes describes the system's complex dynamics and causality structures by: (1) A graphical component of the model named Object Process Diagram (OPD); and (2) A text component named Object-Process Language (OPL). The suitability of OPM is compared with alternatives for causality interface, such as causal directed acyclic graphs (DAGs), while resolving long-standing problems, including intervention, direct and indirect effects, confounding, and attribution. The knowledge graph and text developed by the HDSTA in this research bridge this gap in ML models and yield casually consistent behavior. The use of ML models enables us to create a flexible transport model in which some parts/equations of it are linear, and some could be nonlinear or nonparametric. Individually, those ML models may lack domain knowledge, but they are structured together to reflect the system relationships of the domain.

This paper offers a different perspective on how various data can be used to solve transportation problems and improve transportation systems by developing a framework for integrating causality modeling, ML, and Experts' Insight. As a starting point, we will demonstrate our approach by using a case study of bus delays. It is then explained how HDSTA promotes an iterative process of using the help of experts to improve the model in accordance with their feedback.

HDSTA can serve as a framework for different experts, data scientists, and non-technical stakeholders to work together, by creating a shared language. It will be beneficial as: (1) Serving in workshops involving experts and non-technical parties to define a knowledge graph; (2) Defining causal functions to evaluate cause-effect equations using ML models which will be less correlated. These improved ML models will allow the creation of a flexible transport model in which some parts of its equations are linear, and some could be nonlinear or nonparametric. (3) Support data-driven decisions (interventions) made by stakeholders such as TMCs.

2. Literature review

2.1. Data-driven modeling

The emergence of various technologies in transportation including computer vision, apps, and sensors to name a few, has increased the already massive volume of data acquired and the need to be able to extract only the relevant information (Habibzadeh, 2018).

Data is used to notify about events, such as road accidents, road blockage, traffic intensity, high speed of vehicles (Rathore, et al 2016). Data-driven decision systems mostly represent the information with maps or graphs representing trends. Text is also provided. for example about proposed actions for accident~~s~~ descriptions, alongside a map with shortest routes to the incident location based on the current traffic situation (Babar, & Arif 2019). These texts, maps and graph representations lack clear references for the cause-effect relationships which led to the actions recommended. They are open to interpretations that limit their use for discussions and comparisons. Other methods such as simulation are known to have significant computational costs and time, making such simulations impractical for real-time prediction and control (Athavale, Joshi & Yoda, 2018).

In recent years, new Machine Learning (ML) tools were developed to process big data and to allow complex nonlinearities. While ML has many advantages in the volume, variety, and velocity of the big data processed, there is much to do regarding the value ML can give in extracting meaningful insights and predictions. Nonparametric ML algorithms such as Neural Networks (Basu et al, 2010) and gaussian processes (Rasmussen & Nickisch, 2010) do not make specific assumptions about the function form. They are prepared to learn any functional form directly from the training data, by propagating observed correlations across a defined neighborhood (often called kernel). Thus, to prevent a false representation of an influence effect between variables, which is not necessarily a causal relationship (Guo, Cheng, Hahn & Liu, 2020), the ML model requires some sort of causality formalism to learn the knowledge gained in Transport Theory.

2.2. Cause-effect modeling

Causality is a generic relationship between an effect and the cause that gives rise to it (Williamson, 2009). Learning causal effects is done by investigating to what extent manipulating the value of a potential cause would influence an effect. The action of manipulation is called "intervention", the variable to be manipulated is called "treatment," and the variable for which it observes the response is the "outcome".

A causal model is a mathematical abstraction that quantitatively describes the causal relations between variables. First, causal assumptions or prior causal knowledge can be represented by an incomplete causal model. Then, what is missing can be learned from the data. The two most well-known causal models are the Structural Causal Models (SCMs) (Pearl, 2009) and the potential outcome framework (Rubin, 1974). They are considered as the foundations because they enable a consistent representation of prior causal knowledge, assumptions, and estimates. While "model-free" statistics support describing data and inferring distributional parameters from a sample, causal inference requires a science-friendly language for articulating causal knowledge, and mathematical machinery for processing that knowledge, combining it with data, and drawing new causal conclusions about a phenomenon (Pearl, 2010; Scheines, 1997; Pearce, 2016).

Experts' judgments form the causality assumptions for transportation. It is done to explain how transportation demand is influenced by perceptions, attitudes, and travel behavior (Reibstein, Lovelock & Dobson, 1980; Tardiff, 1977). This judgment was applied to explore causal relations in parking and automobile use in cities (McCahill et al 2016), road accidents (Liang et al, 2022) and traffic congestion (Meng, & Han, 2018) and for interventions related to issues as fatigue (Anund et al, 2015) or public lighting (Elvik, 1995). Those are a few examples of the rich literature about causality in transportation models, investigating what caused what, that can help a modeler to build a causal graph for transportation problems (as ~~done~~ for the example shown in results section). The challenge is in gathering all the inputs together especially when new types of data are defined.

There are other inherent limitations to name a few: Complex behavioral influences such as social norms, learning by doing and by observation, choice as a strategy for search and learning, and choice as strategic behavior in games. While aspects of these factors could be built into the model structure, the increased complexity quickly reaches a dimensionality (and therefore computation) barrier. Another gap is the vague definition of behavior captured by the model, which is often left up to interpretation. For example, is the observed behavior a result of changing perceptions or changing preferences? A specification may appear to be capturing one form of behavior when it is capturing something else (Ben-Akiva, et al 2002).

2.3. Structural Equation Modeling

Structural Equation Modeling (SEM) is a family of statistical techniques permitting researchers to test a hypothesized relationship~~s~~ between constructs that are thought to be causally and structurally related to one another (Anderson & Gerbing,1988; Kahn, 2006). Variations among the styles of latent causal connections, observed variables measuring the latent variables, and in the statistical estimation strategies result in the SEM toolkit including confirmatory factor analysis, confirmatory composite analysis, path analysis, multi-group modeling, longitudinal modeling, partial least squares path modeling, latent growth modeling and hierarchical or multilevel modeling (Kaplan, 2008).

SEM causal connections are depicted with arrows. Those are equivalently represented by equations that are implied by the model and its structural features and then estimated with statistical algorithms. In SEM, researchers must evaluate multiple test statistics to determine whether the model fits the data (Weston & Gore, 2006).

In transportation, SEM is used for many aspects as planning and operation, which can determine relationships amongst unobserved constructs (i.e., independent, mediator, moderator, control, and dependent variables (Hew et al, 2018)); to model the traveler's choices, attitudes, and mode choice (Najaf et al 2018); the information needed to make those calls (Mo, Niu & Fu, 2008); and the relationship between issues as driving safety and driver characteristics (Zhao et al, 2019) or weather conditions (Lee et al, 2018). SEM is used to evaluate the perceived quality of traffic information systems (Khoo, & Ong,2013), and to identify factors that affect the user's satisfaction with the road network (Shaaban, Shakeel, Rashidi & Kim, 2022). SEM is capable to explain linear and compensatory causal relationships amongst constraints, however, SEM cannot explore nonlinear relationships (Binsawad, 2020) which were related to transportation.

2.4. Cause-effect graphs

The common method for modeling causality is using Bayesian networks, which are Directed Acyclic Graphs (DAG) where nodes represent random variables, and arrows represent a direct causal effect that defines conditional dependence between variables (Pearl, 1985). The strengths of these dependencies are expressed in conditional probabilities. These networks represent conditional dependency structures, prior assumptions about variable systems, and observational events used to update these assumptions. The typical structures of DAG diagrams, as seen in Fig. 2 include a chain diagram: Variable X causes variable Z which causes outcome Y. A fork diagram: Variable Z leads to both outcomes X and Y. A collider diagram: X and Y both lead to outcome Z (Guo et al, 2020). Since a cause must precede a result, cyclicity cannot be defined and that is why the graph is called acyclic (no loop from one node back to itself following the arrows).

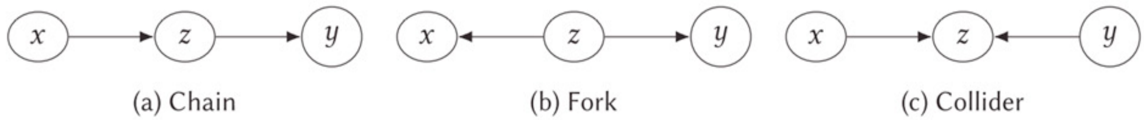


Fig. 2. Typical structures for DGA, adapted from (Guo et al, 2020).

Nodes are used in other diagramming languages such as cognitive maps, concept maps, causal maps, and causal diagrams. While these are more intuitive to construct and read, they suffer from limitations such as use of broad terms without the need to specify variables and not generating quantitative behavioral information (Schaffernicht, 2007). This research framework will be compared to those methods along the way.

2.5. Gaps and Challenges

The main problem today is not a lack of domain knowledge or data. It is the lack of models that can process them efficiently, considering the whole system's dynamic interactions. There is a growing demand for a method that can assess the benefits of big data in transportation planning (Neilson, Daniel, & Tjandra, 2019). This method should be designed to utilize data for actions and insights of practical value so that later they could provide further value than superior model performance statistics (Laña, Ibai, et al, 2021) and be useful in the decision-making process (Mandinach, 2012; Urbanek, 2017; Torre‐Bastida et al, 2018).

Causes and effects are often known only intuitively, making it hard to define them (Guo et al 2020). Thus, the causality embedded in transportation is often not considered when ML models are developed. The use of expert judgment to do so require quantitative variables to be expressed in operational terms (what observable actions can be undertaken) and objectively quantifiable units (same meaning for expert A and expert B). As an example, the variable “safety culture” is often stated to be of prime importance for the safety of an operation. However, there is no operational definition of safety culture available, thus every expert can interpret it differently and results from different experts cannot be combined (Schaffernicht, 2007).

Current causal graphs, such as DAG, support prediction and concluding interventions. However, DAGs contain limitations, primarily their non-parametric nature, which is limited to a qualitative representation; using the same arrow for all types of relationships (Vander, Weele & Robins, 2007); inability to model cycles; vague interpretations (Williams, Bach, Matthiesen, 2018) in a sense that two researchers may ask the same research question, using the same data for analysis, but each chose different connections between variables because they have different views about the underlying causal relationship. Sometimes for this reason there is ambiguity of the intent of the graph model by other stakeholders. This issue is particularly influential when a causal graph is validated by experts from different disciplines such as transportation, weather, psychology, economics, and policymakers. Although other graphical methods may oversimplify reality, for DAG the basic assumptions are the simplified ones

3. Method

This interdisciplinarity research designs a framework to integrate different tools from transportation, and ML, based on the assumption that only the knowledge of the phenomenon may be a barrier in solving problems, rather than the tool chosen as a method. To do so HDSTA framework is defined to adapt causality concepts about the transportation system into systems thinking terminology by using OPM. This supports the creation of a causal graph as SCM and to estimate of its equations using ML. The steps described in Fig.3 are explained in the results sections as well, and are specified as follows:.

Step 1: Identify a problem based on available data. Then gather prior knowledge and causality assumptions based on case studies and literature. Data will deliver the observed and collected variables in a growing database. As part of the systems thinking approach, we will associate the variables from the different data sets with objects and processes.

Step 2: Model a knowledge graph based on the data and knowledge collected. We ~~will~~ explore methods of systems thinking such as Object-Process Methodology (OPM), a conceptual language for modeling complex dynamic systems (Dori, 1995). OPM Models are created by defining objects and processes connected by several kinds of relations, expressed graphically as links. Objects and processes (OPM things) can be physical or informatical, and objects can have states. Processes transform objects by creating or consuming them, or by changing their state. All OPM models in this research will be constructed by ISO 19450:2015 (Dori, 2015), using OPCloud (Dori, Jbara., Levi & Wengrowicz, 2018), the OPM cloud-based modeling environment.

Step 3: Verify the models using Transportation experts' knowledge regarding links that represent causality or behavioral differences. This step includes the feedback needed to represent the domain knowledge and to add confounders.

Step 4: Estimate the causal relations for improved ML models. While OPM can bridge gaps in current causality graph methods, it lacks the analytical ability to create the forecast modeling. Statistical applications such as Scikit-learn environment for python (Jain, 2011) are used to code and run the data to evaluate the cause-effect models' equations (like SCM or SEM) and make predictions based on using statistical methods (like Gaussian Processes or Linear Regression). The role of ML in the SCM model, is to estimate individual ML models for each cause-effect relationship. The ML model receives a database of inputs (causes) and outputs (effects) and learns the respective functional relationship. The use of ML models will allow us to create a flexible transport model in which some parts/equations of the model are linear, and some could be nonlinear or nonparametric. Individually, those ML models may lack domain knowledge, but they are structured together to reflect the system relationships of the domain. We believe the knowledge graph and text developed by the HDSTA in this research can bridge this gap in ML models and yield more casual and less, correlated ML models.

Step 5: Examine the results with experts and improve the functions accordingly if needed. We will compare our ML model results compared to ML with no causality in terms of: (a) Ease of solving and modeling problems and scenarios; (b) Efficiency of results.

Step 6: Test interventions on variables and policies to improve service values, for example, waiting time for the bus. As other causality graph methods, OPM cannot automatically recognize cause-effect links and possible interventions in variables systematically and efficiently, which becomes much more complicated in large databases. Yet we believe it will support a systematic way to do so in the future.

Step 7: Make a data-driven decision: the HDSTA outputs (OPD, OPL, and predictions) will be examined with experts and non-technical stakeholders and compare our model to current causality graph methods for decision-making in terms of (a) Clarity and understanding of the model for decision-making; (b) Level of details; (c) Ease of modeling problems and scenarios.

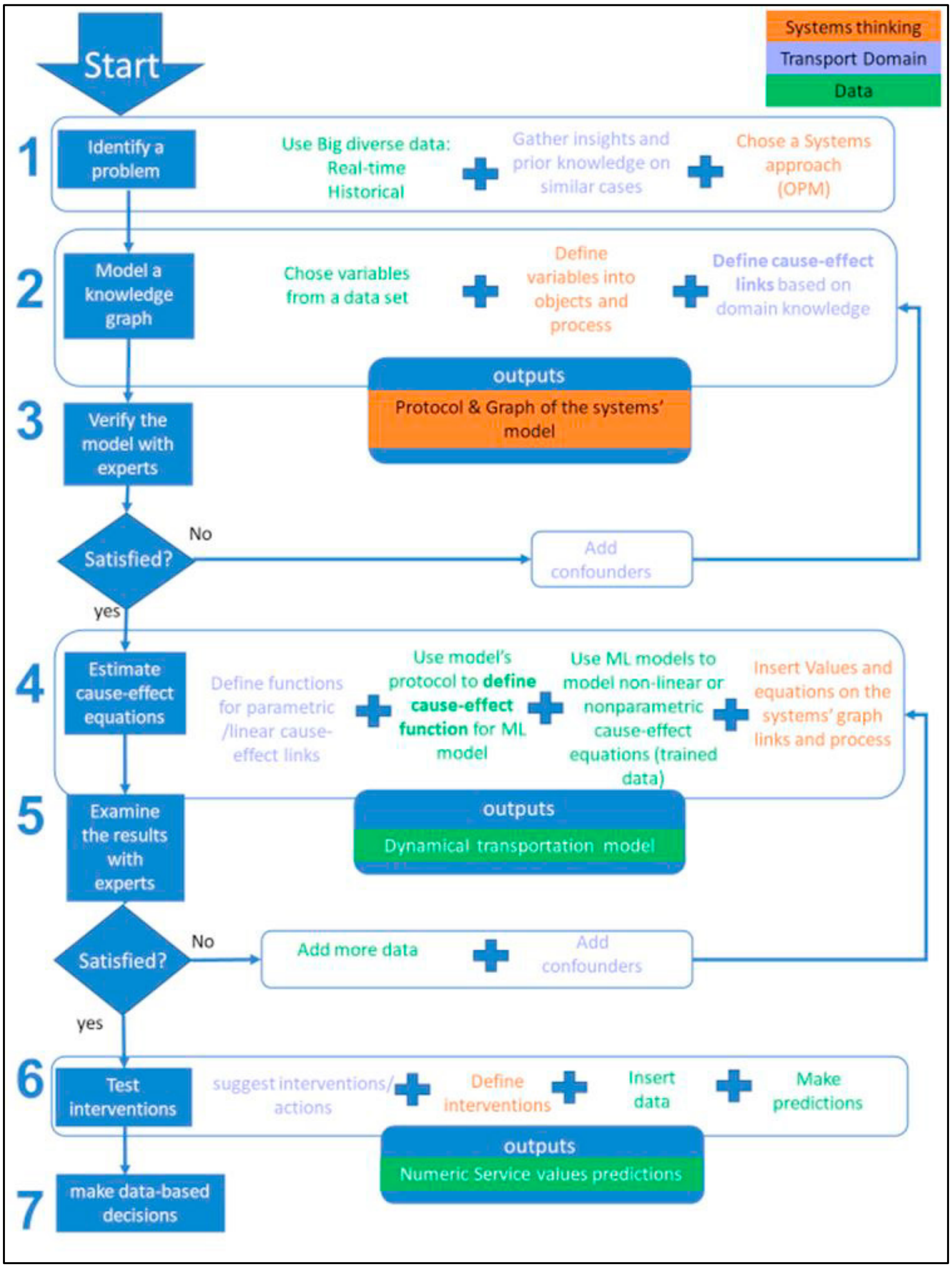


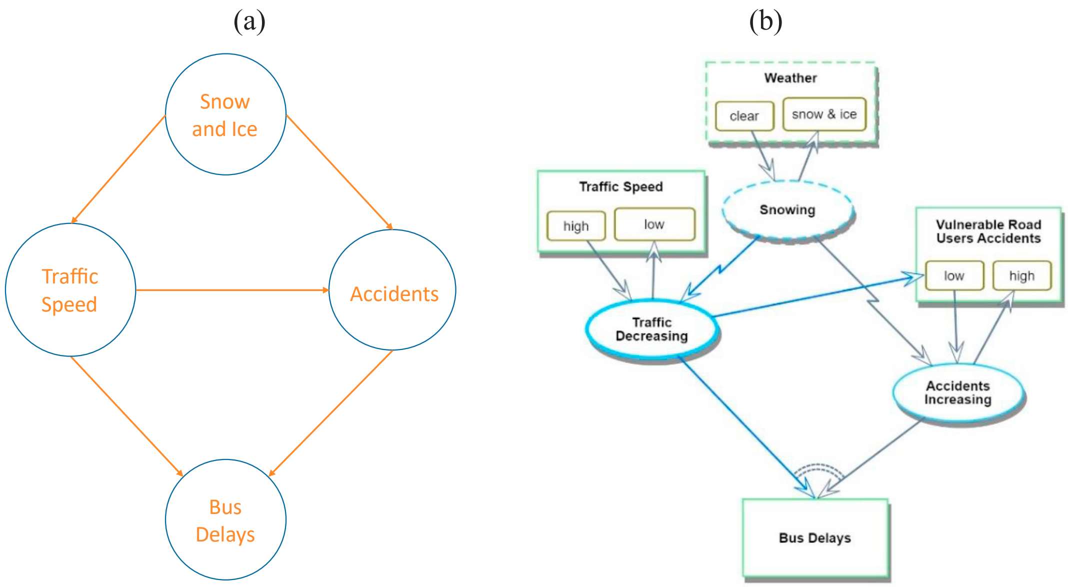
Fig. 3. Method Process Steps

4. Results

To explain HDSTA use for integration of expert's knowledge of causality for database modeling, a scenario was created reviewing the number of daily accidents and total daily delays per month in Toronto (Diab, Feng & Shalaby, 2018). During the winter season, there is a clear trend of more accidents (Åkerstedt, Kecklund & Hörte, 2001; Lépy et al, 2016) and lengthier delays (Diab, Feng & Shalaby, 2018). Following Step 1 in Fig.3 an explanatory variable was added for the causality model – and that is the season – winter, or more precisely the effects of snow and ice during the winter. This is only one explanatory variable of many possible ones (as described later), and it is chosen as a simple example.

4.1. Basic knowledge graph

A knowledge graph was created based on Step 2. In Fig.3. As a baseline, a DAG graph was created following the principles: (1) Confounder chain with a fork cause-effect link for snow (confounder) causing traffic speed which causes bus delays and vulnerable road users' accidents. (2) A confounder chain where snow causes vulnerable road users' accidents which causes bus delays. Fig.4 shows the DAG graph (a) compared to the OPM model represented with OPD and OPL (b).



OPL
Weather can be **clear** or **snow & ice**.
Traffic Speed can be **high** or **low**.
Vulnerable Road Users Accidents can be **high** or **low**.
Snowing changes **Weather** from **clear** to **snow & ice**.
Snowing invokes **Accidents Increasing** and **Traffic Decreasing**.
Traffic Decreasing changes **Traffic Speed** from **high** to **low**.
Traffic Decreasing yields **Vulnerable Road Users Accidents** at state **low**.
Accidents Increasing changes **Vulnerable Road Users Accidents** from **low** to **high**.
At least one of **Accidents Increasing** and **Traffic Decreasing** yields **Bus Delays**.

Fig.4. Snow and bus delays model using (a) DAG; (b) OPM.

While the DAG's model in Fig.4 (a) may be more intuitive, the OPM model (b) provides more insights and represents a more detailed version of this case:

Paths: The most significate difference is that the OPM model captured snow being an environmental process of "snowing" (environmental is represented by dashed line). It defined the snowing effect as invoking "Traffic Decreasing" which yields both "Bus delay" and "Vulnerable road users' accidents" in state "low". This cause-effect lane is colored in blue and help to notice a counter effect of "snowing" which also invokes the process of "accidents increasing". Accidents involve several types including car accidents and "Vulnerable road users' accidents". An increase in Accidents change the "Vulnerable road users' accidents" state from "low" to "high" and yields "bus delays". While these are discrete states they can also receive continuous numeric values defined by us.

Conditions on causes: we can model that **at least one** of the processes "Snowing" and "Traffic increasing" yields "Bus Delays" by the double bow. Later we can also add a promoter on each link to indicate which cause is stronger.

Interventions: By defining environmental entities ("weather" and "snowing") OPM allows us to notice which variable in the system cannot be controlled by us.

Accuracy: OPM forces the modeler to be accurate with definitions: Snow a state of "Weather"' which can later be expanded to various states as fog, rain, wind, and heat wave. Snow is environmental in the sense that we do not control it in the system although it effect the system. "Snowing" is not just affect "traffic speed" rather it invoke a process of "traffic decreasing".

Difference in cause-effect: OPM does not have only one arrow for all. "Traffic increasing" **changes** the amount of traffic (measured by Speed) and **yields** "bus delays" and "accidents". There are other types of links not used in this example.

Details: Not all variables are the same in a diagram. OPM does not only divide the model into processes and objects (variables) but also gives details of attributes, states, and environmental entities that give information about what is not controlled by the model (such as weather) and more. Those details are important for both human factors in decision-making but not the least to ML to learn data's structure and cause-effect complexity.

All these features detailed above demonstrate the power of the HDSTA framework in providing an objective expression of system components, dynamics, and interrelationships.

4.2. Systems view knowledge graph

Both graphs in Fig.4 are based on causality used in transportation research and methods but remain too narrow as they are focused on investigating a few variables' effects regarding predictions for delays and do not represent the system view of the decision process based on insights of the causal graph. Additional relationships and processes can be added as seen in Fig.5 that integrate the cause graph from Fig.4 as part of a border "TMC operating" process. By modeling the TMC system new entities can be explored as "policy" of "Dynamical Road allocation" that changes from the state "parking lane" to the state "car lane" aimed to decrease "Bus delays" from "high" to "low" by changing "traffic speed" from "low" to "high". The change in "traffic speed" marked in yellow to emphasize it yields another result: "high Traffic Speed" initiates "Accidents Increasing" (which consumes "Traffic Speed" in the formula later to be added). This yellow addition is important as once a TMC chose an action, it should realize other potential effects of it. In this case changing lanes and increasing the speed to improve the bus delays in the snow may increase accidents. We were able to model this thanks to the integrated causal graph as seen in Fig. 4 which is circled in red dotted lines as part of the system view in Fig.5. This yellow addition was also modeled thanks to verifying the model with Experts' knowledge regarding the effects of the proposed intervention.

4.3. Extended System View Knowledge Graph

While Fig.5 demonstrates a system view integrated with causal knowledge, it is still too flat as it does not represent other variables influencing "bus delays" other than "traffic speed" and "Accidents". The challenge in making decisions based on data in transportation operations stems from the various layers that the system contains. The response that should regulate the expected load or delays should refer to measurable parameters such as time, place, infrastructure, existing technologies, the environment, the weather, and to more unpredictable complex parameters. OPM enhances our ability as those layers expand the cause-effect representation. Fig.6 expands "Bus Delaying Model building" process which was described in a bold blue circle line in Fig.5 (the Boldness stands for a process that can be zoomed in). Based on previous research (Furth & SanClemente,2006) for predictions regarding delays new variables were added: bus Stop station location (on a hill or not), closeness to an intersection, reserved bus lanes, the vehicle Engine (diesel), and the cycle length. Those are marked in green.

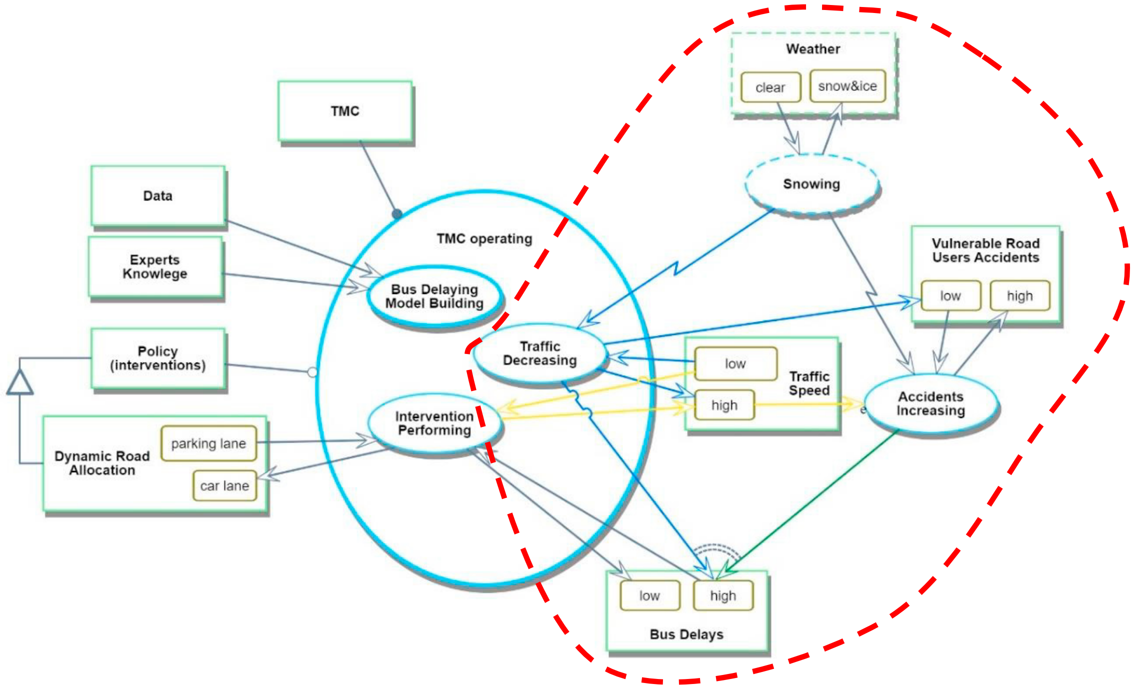


Fig.5 System view of snow and bus delay graph as part of TMC operating.

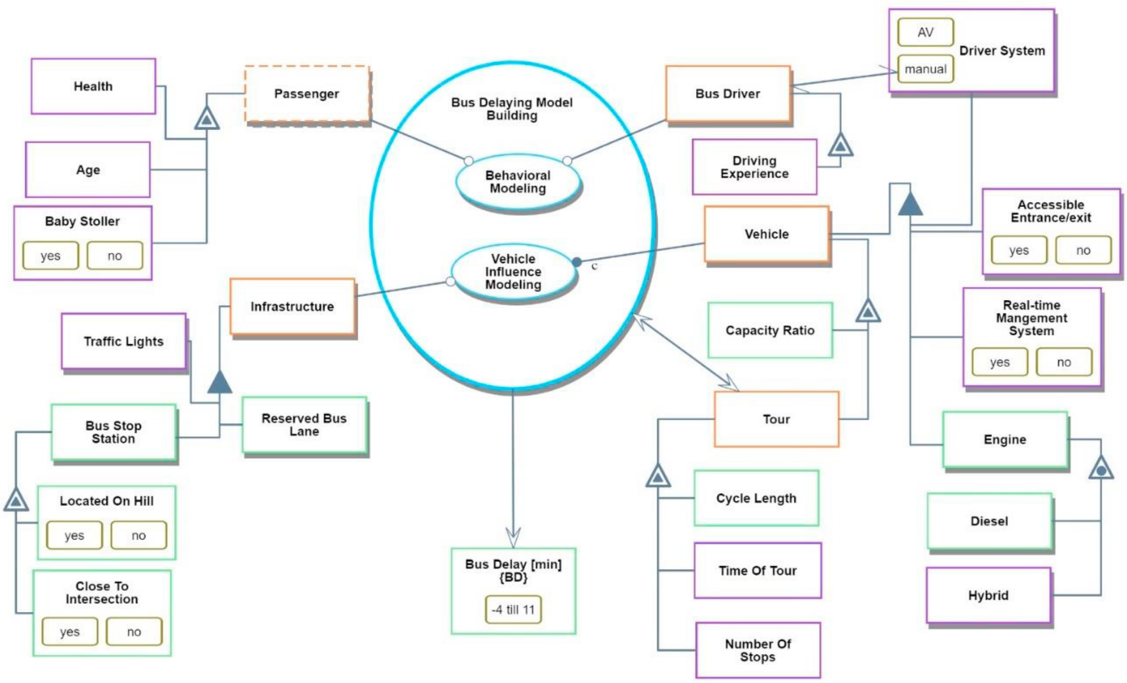


Fig.6. Zoom-in Bus Delaying model building process influences using OPM.

Another benefit of this paper approach is the ability to include more data and more insights. To name a few of the potential variables: (1) infrastructure of traffic lights ;(2) The time of day: peak hours or not, workdays or weekends; (3) New technologies: improving visibility during the snow, or automatic driver system can reduce slowdown of vehicles. (4) Road users' behavior; (5) bus priorities (6) real-time management of the buses. Objects marked in purple are additions to the ones used in Furth & SanClemente, 2006. The objects in orange colour represent the groups we defined to relate the variables: infrastructure, vehicle, tour, bus driver, and passenger– all but the passenger can be controlled/changed by the TMC. Like the weather, the passenger is an environmental entity that is not changed by the system but still influences it. The framework offered here allows representing conclusions of different past research that together can reflect a big picture of influences on the transportation system.

4.4. Causality in ML model results

To demonstrate the potential of integrating the model with ML as described in Step 4 in Fig.3, the causal graph Fig.4 was used to define causal links in a function to teach the causality model. The function at this stage is based on SCM. Based on Fig.4 "speed" is only dependent on "snow". "Accidents" are dependent on "Ice and snow" and "speed", and "delays" are dependent on "speed" (the more speed, the fewer delays) and on "Accidents" (the more accidents the more delays). Following this inference for the SCM, we used a function that, given a dataset, will propagate across the graph (from "causes" to "effects") to predict all nodes. A synthetic dataset to generate the variables according to the principles.

(1) Based on Toronto weather data[†] for January Snowfall in cm/day is normal distributed with a mean of 38.2 and a standard deviation of 7.

$$Snow_{cn\ per\ day} \sim N\ (\mu = 38.2, \sigma = 7).$$

(2) According to previous research (Liang et al, 2022) there is a linear negative effect between snowfall and speed: snow decreases speed as demonstrated in Fig.6 (a). Based on this principle[‡] , we changed in our scenario the speed limit to 70 km/h. Although the speed limit can be lower, drivers do not always follow this limitation. As we could not find a function, we created one by that principle.

$$Speed = -0.6 * Snow + 50\ ;\ \varepsilon_1 \sim RN\left(0, \left|\frac{Speed}{4}\right|\right)$$

$$Speed_{with\ noise, avrage\ per\ day} = Max(0, speed + \varepsilon_1)$$

(3) "Accidents" include all types of accidents including incidents of vulnerable road users. Accidents are a function of both speed and snow density. The snow increases the probability of accidents, as seen in Fig.7 (b). It is interesting to view Fig.7 (d) showing Accidents Versus Speed reflecting the lower the speed the more accidents which

[†] www.eldoradoweather.com/canada/climate2/Toronto.html
[‡] www.researchgate.net/figure/Average-speed-of-car-under-different-snow-depth_fig2_32330050

contradict basic intuition. Yet, as this scenario describes winter observations, it is true that lower speed caused by more snow which increases the probability for accidents.

$$Accidents_{per\ day} = \log\left(Speed_{with\ noise,avrage\ per\ day} + 1\right) * \sqrt{Snow}$$

(4) Bus delay is measured in minutes. The Bus Delays are measured by minutes per day as seen in Fig. 7 (c).

$$Delays_{minutes} = log\left(\frac{1}{Speed_{with\ noise,avrage\ per\ day} - 20}\right) + \sqrt{Accidetns} + 10 + \varepsilon_2$$

$$with\ \varepsilon_2 \sim RN(0, \left|\frac{(80 - Speed)}{100}\right|$$

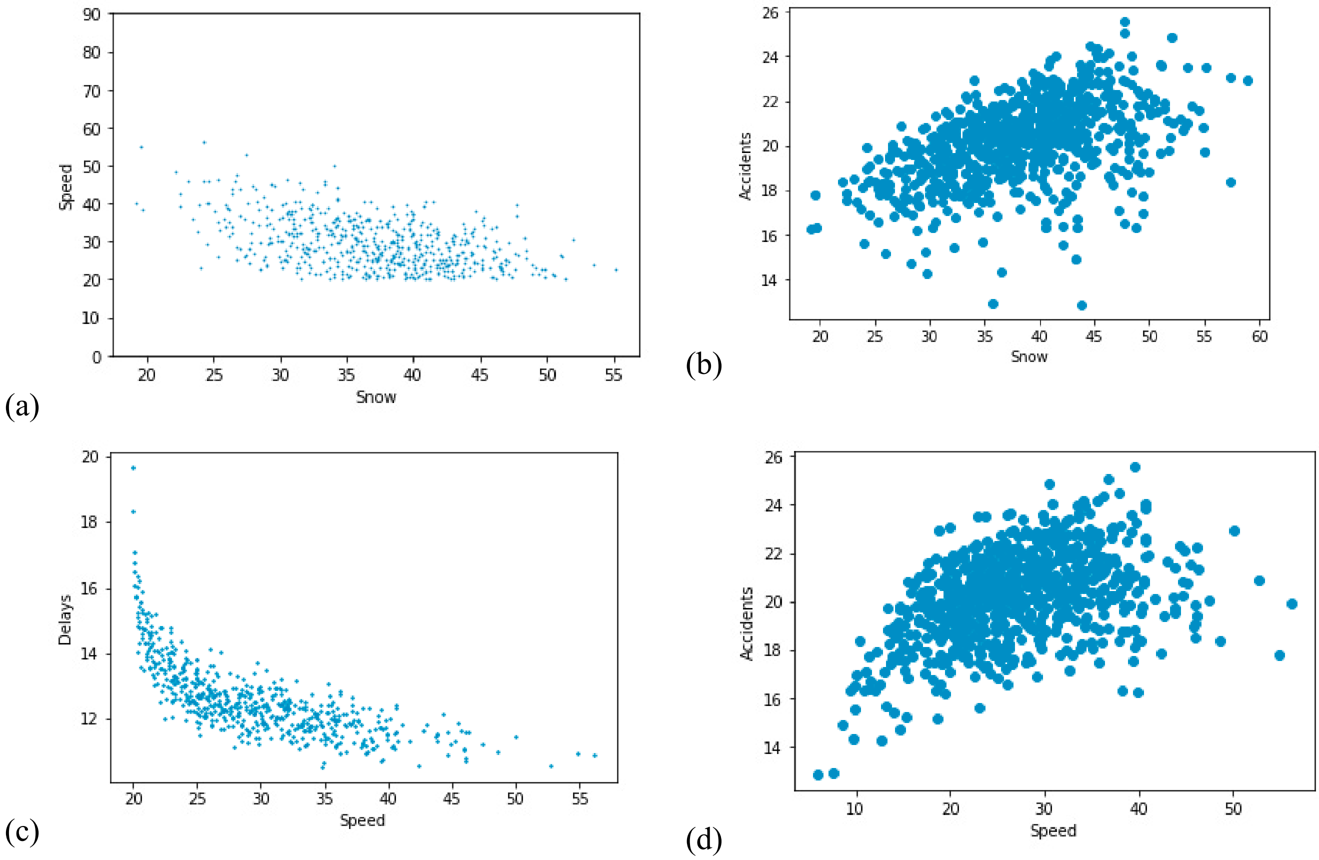


Fig.7 Graphs represent the principles of the data generated used in this paper

Using those data sets, the SCM as well as standard ML models for accidents were trained with different methods such as Multilayer Perceptron, Gaussian Processes, Gradient Boosting, Random Forest, and Linear Regression for accidents. The different models were then tested on a different data distribution which was created according to well-known climate change impacts predicting that snow precipitation may decrease on average yet increase in variance (snows less, but occasionally snowing in abnormal amounts).

Then, the "snow" data was used to predict the "speed" and those results were used to predict the "delays" and the "Accidents". The same steps were followed on the original data and on the new data (climate change). The results of using MultiLayer Perceptron were the best for the prior model without causality, and thus it was chosen to predict the record of accidents. Fig.8 shows the ML model with SCM results for observed Verse Predicated accidents.

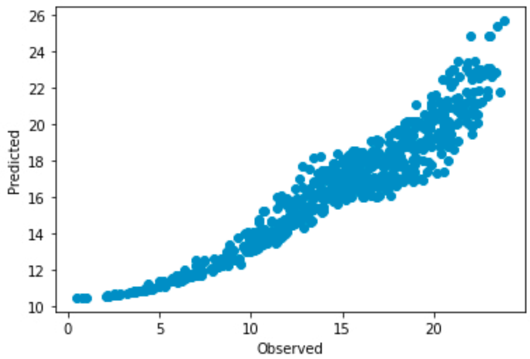


Fig.8. Accidents prediction for ML model with SCM: observed Verse Predicated

Comparison of the accident records results indicated an advantage of integrating causality for the data set following a climate change as seen in Table 1. The value of R^2 is 0.44 without the SCM and 0.64 with it.

Table 1. prediction results for Accidents for the ML model with causality (SCM) and without

Model	MultiLayer Perceptron:	Prior Climate change	Post Climate change
ML	R^2 test set performance	0.986	0.441
	MAE: test set performance:	0.130	2.228
SCM-ML	R^2	0.942	0.647
	MAE	0.323	2.435

5. Conclusions and Discussion

This paper presents the use of HDSTA as a novel framework for data-driven decision-making in transportation. Causality is part of transportation research and methods, but ML models using big data in transportation tend to ignore the knowledge gained from traditional methods in transportation research. This research attempts to build models that use knowledge of the domain by the right causality graph rather than using ML models for a correlation learning.

By using HDSTA the modeling of common knowledge of experts regarding variables is given as ground roles for improved ML algorithms. This proposed approach was demonstrated by using a simple case study which addresses the different causality links between snow, traffic, accidents and bus delays. Results showed a benefit for using ML models that learned the causality by SCM function.. Even if the results for both ML models were the same, the ML model with causality would still be a superior method for several reasons: (1) In a climate change scenario snow may be predicted, but the traffic speed, bus delays, and accidents cannot. Yet by giving the ML model a function of the causality and the values of snow we can predict the accidents; (2) Using the SCM function the code chooses the best model that suits the data and is not limited to a linear or parametric model; and (3) For discussion regarding which actions (decisions) should be taken for the sake of reducing accidents that are not related to bus delays. Based on the causal graph in Fig.4, it is clear that if the traffic does not slow down during snow (as before the intervention), the probability of accidents during the snow may increase. This clarification is important for the choice of a future intervention.

SCM graph was examined both as a graphical tool and for application in practice for ML. Yet, it was shown that SCM is limited in the representation of the different influences "snow and ice" compared to "Speed" has on "accidents". Are these causes equal in their effect on "accidents"? HDSTA is offered here to solve this gap.

The results support both the move from correlation to causality and the integration of tools from ML, human (experts) insight on transportation, systems thinking approaches, and statistical methods. At the practical level, this research methodology contributes to the creation of new tools for modeling complex data-based transportation operation models that support policymakers and TMC's managers decisions by providing a graph, a text and some numeric values gained in models based on HDSTA.

To date, to the knowledge of the authors, no research has examined the application of causality at a system level to improve data-driven transportation operation modeling. In doing so this novel approach adds to knowledge regarding: (1) Casualty graphs modeling by systems thinking approaches; (2) Integration of transport knowledge with ML models; and (3) Causality application for transportation system view.

The proposed approach creates a shared language to evaluate actions offered in different scenarios leading to a standard in the representation of experts' knowledge and refining their data-based recommendations.

6. Limitations and Future Work

This research suggests using available data by explaining its relation to the actions taken in real-time. The interactions between different values of the variables were not addressed here and should be explored in future work. For example, the interaction between a time variable (weekend) and a human factor variable ("lack of driving experience") is essential for examining the desired intervention on traffic as a response to snow. While at this stage the concepts of causality were adapted using OPM, other systems thinking tools can be explored such as systems Dynamics (Stave, 2002). The more cases modeled by HDSTA guidelines, the more systems thinking approaches can be explored for transportation operation needs of both operation and behavioral causes effects. In transportation there is an importance of using variables in the right context, cultural aspects, and behavioral. These factors are considered in the causality knowledge domains and can be validate using HDSTA. This was explored in this paper and should be further expanded to include a wider range of past insights about variables contribution and cause-effect links. By doing so, future work can create templates regarding cause-effect links in transportation.

Further explorations of this methodology can be effective for other needs as well, such as long-term planning of transportation systems. Integration of a causality structure between variables may follow an improvement in other issues connected to the data-based modeling flow, including collecting and storing variables, merging different databases, and creating simulations for policy examination.

Acknowledgments

This work was partially supported by the Technion, EuroTech alliance, .Ph.D fellowship from the Israel Scholarship Education Foundation (ISEF), and Israel Smart Transportation Research Center.

References

Åkerstedt, T., Kecklund, G., & Hörte, L. G. (2001). Night driving, season, and the risk of highway accidents. Sleep, 24(4), 401-406.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. Psychological bulletin, 103(3), 411.

Anund, A., Fors, C., Kecklund, G., Leeuwen, W. V., & Åkerstedt, T. (2015). Countermeasures for fatigue in transportation: a review of existing methods for drivers on road, rail, sea and in aviation.

Athavale, J., Joshi, Y., & Yoda, M. (2018, May). Artificial neural network-based prediction of temperature and flow profile in data centers. In 2018 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm) (pp. 871-880). IEEE.

Babar, M., & Arif, F. (2019). Real-time data processing scheme using big data analytics in internet of things based smart transportation environment. Journal of Ambient Intelligence and Humanized Computing, 10(10), 4167-4177.

Basu, J. K., Bhattacharyya, D., & Kim, T. H. (2010). Use of artificial neural network in pattern recognition. International journal of software engineering and its applications, 4(2). Ben-Akiva, M., Bowman, J.L. (1998), Integration of an activity-based model system and a residential location model, Urban Stud. 35(7):1131-1153

Ben-Akiva, M., McFadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., ... & Munizaga, M. A. (2002). Hybrid choice models: Progress and challenges. Marketing Letters, 13(3), 163-175.

Binsawad MHJIA (2020) Corporate social responsibility in higher education: a PLS-SEM neural network approach. IEEE Access 8:29125–29131

Booth, C., & Richardson, T. (2001). Placing the public in integrated transport planning. Transport policy, 8(2), 141-149.

Dori, D. (1995). "Object-Process Analysis: Maintaining the Balance between System Structure and Behavior". Journal of Logic and Computation. 5 (2): 227–249Dori D. (2002), Object-Process Methodology: A Holistic Systems Paradigm. Berlin Heidelberg: Springer-Verlag.

Dori D. (2015), Automation systems and integration - Object-Process Methodology, ISO/PAS 19450, ISO 2015.

Dori D., Jbara A., Levi N., and Wengrowicz N., (2018), Object-Process Methodology, OPM ISO 19450 – OPCloud and the Evolution of OPM Modeling Tools. Systems Engineering Letters, Project Performance International (PPI) SyEN 61.

Edelstein, R. (2001, August). TRANSPORTATION MANAGEMENT CENTERS: A MULTIMODAL APPROACH. In ITE 2001 Annual Meeting and Exhibit Institute of Transportation Engineers (ITE) (No. CD-013).

Elvik, R. (1995). Meta-analysis of evaluations of public lighting as accident countermeasure. Transportation Research Record, 1485(1), 12-24.

Furth, P. G., & SanClemente, J. L. (2006). Near side, far side, uphill, downhill: impact of bus stop location on bus delay. Transportation research record, 1971(1), 66-73. Gilman, S,, and Cahill P., (2021). Data Dashboards at State DOTs: Case Studies of Select Transportation Agencies. No. DOT-VNTSC-FHWA-21-03. United States. Federal Highway Administration.

Guo, R., Cheng, L., Li, J., Hahn, P. R., & Liu, H. (2020). A survey of learning causality with data: Problems and methods. ACM Computing Surveys (CSUR), 53(4), 1-37.

Habibzadeh, Hadi, et al. (2018), Soft sensing in smart cities: Handling 3Vs using recommender systems, machine intelligence, and data analytics. IEEE Communications Magazine 56.2: 78-86.

Hew J-J, Leong L-Y, Tan GW-H, Lee V-H, Ooi K-BJTM (2018) Mobile social tourism shopping: a dual-stage analysis of a multi-mediation model. Tour Manag 66:121–139

Jain, A. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res, 12, 2825-2830.

Kahn, J. H. (2006). Factor analysis in counselling psychology research, training, and practice: Principles, advances, and applications. The counselling psychologist, 34(5), 684-718.

Khoo, H. L., & Ong, G. P. (2013). Evaluating perceived quality of traffic information system using structural equation modeling. KSCE Journal of Civil Engineering, 17(4), 837-849.

Laña, I., Sanchez-Medina, J. J., Vlahogianni, E. I., & Del Ser, J. (2021). From Data to Actions in Intelligent Transportation Systems: A Prescription of Functional Requirements for Model Actionability. Sensors (Basel, Switzerland), 21(4), 1121. https://doi.org/10.3390/s21041121

Lee, J., Chae, J., Yoon, T., & Yang, H. (2018). Traffic accident severity analysis with rain-related factors using structural equation modeling–A case study of Seoul City. Accident Analysis & Prevention, 112, 1-10.

Liang, M., Min, M., Guo, X., Song, Q., Wang, H., Li, N., ... & Sun, Y. (2022). The relationship between ambient temperatures and road traffic injuries: a systematic review and meta-analysis. Environmental Science and

Pollution Research, 1-14.

Lépy, É., Rantala, S., Huusko, A., Nieminen, P., Hippi, M., & Rautio, A. (2016). Role of winter weather conditions and slipperiness on tourists' accidents in finland. International journal of environmental research and public health, 13(8), 822.

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. Educational Psychologist, 47(2), 71-85.

Meng, X., & Han, J. (2018). Roads, economy, population density, and CO2: A city-scaled causality analysis. Resources, Conservation and Recycling, 128, 508-515.

McCahill, C. T., Garrick, N., Atkinson-Palombo, C., & Polinski, A. (2016). Effects of parking provision on automobile use in cities: Inferring causality. Transportation Research Record, 2543(1), 159-165.

Mo, Y., Niu, Y., & Fu, Y. (2008). Structural equation modeling of traveller information needs. In 2008 International Conference on MultiMedia and Information Technology (pp. 233-236). IEEE.

Najaf, P., Thill, J. C., Zhang, W., & Fields, M. G. (2018). City-level urban form and traffic safety: A structural equation modeling analysis of direct and indirect effects. Journal of transport geography, 69, 257-270.

Neilson, A., Daniel, B., & Tjandra, S. (2019). Systematic review of the literature on big data in the transportation domain: Concepts and applications. Big Data Research, 17, 35-44.

Kaplan, D. (2008). Structural equation modeling: Foundations and extensions (Vol. 10). SAGE publications.

Rathore, M. M., Ahmad, A., Paul, A., & Thikshaja, U. K. (2016). Exploiting real-time big data to empower smart transportation using big graphs. In 2016 IEEE Region 10 Symposium (TENSYMP) (pp. 135-139). IEEE.

Rasmussen, C. E., & Nickisch, H. (2010). Gaussian processes for machine learning (GPML) toolbox. The Journal of Machine Learning Research, 11, 3011-3015.

Schaffernicht, M. (2007, July). Causality and diagrams for system dynamics. In Actas de la 50th International Conference of the System Dynamics Society, Boston (pp. 147-154).

Scheines, R. (1997). An introduction to causal inference.

"OPCloud" [Online]. Available: https://opcloud-client.firebaseapp.com

Pearl, J. (2009). Causality. Cambridge university press.

Pearl, J. (2010). Causal inference. Causality: objectives and assessment, 39-58.

Pearl, J. (1985). Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning.

Pearce, N. & Lawlor, D. A. (2016). Causal inference-so much more than statistics. Int. J. Epidemiol. 45, 1895–1903

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology, 66(5), 688.

Shaaban, K., Shakeel, K., Rashidi, T. H., & Kim, I. (2022). Measuring users' satisfaction of the road network using structural equation modeling. International Journal of Sustainable Transportation, 16(9), 792-803

Stave, K.A. (2002), Using system dynamics to improve public participation in environmental decisions. Syst. Dyn. Rev., 18: 139-167.

Tardiff, T. J. (1977). Causal inferences involving transportation attitudes and behavior. Transportation Research, 11(6), 397-404..

Torre‐Bastida, A. I., Del Ser, J., Laña, I., Ilardia, M., Bilbao, M. N., & Campos‐Cordobés, S. (2018). Big Data for transportation and mobility: recent advances, trends, and challenges. IET Intelligent Transport Systems, 12(8), 742-755.

Urbanek, A. (2017). Big data-a challenge for urban transport managers. Communications-Scientific letters of the University of Zilina, 19(2), 36-42.

VanderWeele, T. J., & Robins, J. M. (2007). Four types of effect modification: a classification based on directed acyclic graphs. Epidemiology, 18(5), 561-568.

Wang, J., & Mueller, K. (2015). The visual causality analyst: An interactive interface for causal reasoning. IEEE transactions on visualization and computer graphics, 22(1), 230-239.

Weston, R., & Gore Jr, P. A. (2006). A brief guide to structural equation modeling. The counselling psychologist, 34(5), 719-751.

Williamson, J. (2009). Probabilistic theories of causality. The Oxford handbook of causation, 185-212.

Williams, T.C., Bach, C.C., Matthiesen, N.B. et al. (2018). Directed acyclic graphs: a tool for causal studies in paediatrics. Pediatr Res 84, 487–493.

Zhao, X., Xu, W., Ma, J., Li, H., & Chen, Y. (2019). An analysis of the relationship between driver characteristics and driving safety using structural equation models. Transportation research part F: traffic psychology and behaviour, 62, 529-545.