LogForum

> Scientific Journal of Logistics <

2023, 19 (3), 331-345

http://doi.org/10.17270/J.LOG.2023.845

http://www.logforum.net p-ISSN 1895-2038 e-ISSN 1734-459X

ORIGINAL PAPER

ON THE USE OF MACHINE LEARNING TECHNIQUES AND DISCRETE CHOICE MODELS IN MODE CHOICE ANALYSIS

Ouissam Benjdiya, Naoufal Rouky, Othmane Benmoussa, Mouhsene Fri

Euromed Polytechnic School, Euromed Research Center, Euromed University of Fes, Fes, **Morocco**

ABSTRACT. Background: The mode choice stage is a critical aspect that transportation experts rely on to develop a robust transportation system for a particular region. Various techniques are utilized to model mode choice behavior, including Discrete Choice Models (DCMs) and Machine Learning (ML) techniques. However, existing reviews typically focus on either DCMs or ML techniques, and reviews that cover both categories often concentrate on one category while merely mentioning some techniques from the other. This paper aims to address this gap by examining the principal DCMs and ML techniques published over the past four years, differentiating between models based on the granularity level, namely aggregate and disaggregate models. Additionally, a comprehensive discussion is conducted on the accuracy of the different models used in the reviewed articles.

Methods: This paper provides a thorough and enhanced analysis of travel mode choice models and analysis techniques used in articles published on "ScienceDirect" from 2020 to 2023. To ensure a comprehensive coverage of the subject, a meticulous search strategy was employed, utilizing targeted keywords. As a result, a total of 38 articles were carefully selected for detailed examination and analysis.

Results: The findings of this study highlight the suitability of different modeling approaches for varying levels of analysis. Discrete Choice Models demonstrate effectiveness in aggregate-level analyses, whereas Machine Learning Techniques prove more appropriate for disaggregate-level analyses. Moreover, the study suggests that employing hybrid models can potentially yield a promising solution to attain enhanced prediction accuracy without compromising interpretability.

Conclusions: The examination of selected articles revealed several key points. Firstly, there is a concentration of studies on travel mode choice in European countries, China, and the USA, indicating a need for more research in developing countries. Secondly, the reviewed articles often lack in-depth analysis of individual behavior and fail to consider external factors like weather or seasons when employing disaggregate models. Thus, future studies should leverage technological advancements and explore new factors influencing mode choice behavior. Additionally, there is a need for further research on hybrid models that combine Discrete Choice Models (DCMs) with Machine Learning (ML) techniques or deep learning approaches. This research can provide guidance for practitioners unfamiliar with these methods and aid in the design of effective transportation policies. Lastly, considering the variety of models available, it is crucial to understand the extent to which these models can be generalized to different contexts, emphasizing the importance of studying model applicability and generalizability in diverse settings.

Keywords: travel mode choice, machine learning, discrete choice model, hybrid models

INTRODUCTION

The transportation industry has undergone rapid development in recent years, providing cities with a variety of modes of transportation including trains, cars, buses, and two-wheel vehicles. This has resulted in a multitude of options for individuals to choose from when deciding how to travel. To ensure the safe, efficient, and environmentally friendly movement of people and goods, it is necessary to understand the characteristics that consumers consider when selecting their preferred mode of transportation, and how they prioritize these factors in order to design a well-optimized transportation system.

To address this challenge, researchers have employed numerous techniques, including the Discrete Choice Models (DCMs) family introduced by Ben-Akiva and Lerman [1985].

Citation: Benjdiya O., Rouky N., Benmoussa O., Fri M., 2023. On the use of Machine Learning Techniques and Discrete Choice Models in Mode Choice Analysis. LogForum 19 (3), 331-345, http://doi.org/10.17270/J.LOG.2023.845

Received: 22.05.2023, Accepted: 23.07.2023, on-line: 01.10.2023

These models attempt to predict travel demand and individual behaviors when presented with various alternatives and have long been the main tools for predicting individual choice behavior. However, with the development of computers, new techniques have emerged, such as Machine Learning (ML) classifiers that have taken the prediction of travel mode choice to new heights.

Several studies have compared ML techniques and DCMs [Xie et al., 2003; Zhang and Xie, 2008; Wang and Ross, 2018; Cheng et al., 2019; Wang et al., 2020]. They have found that ML techniques generally outperform discrete choice models in terms of robustness and accuracy, but they may lack interpretability [Mohammadian and Miller, 2002; Wang et al., 2020]. The advantages and disadvantages of the two different models' categories have led researchers to try to combine both, creating new hybrid models [Wong and Farooq, 2021] or at least using different models from the two categories to benchmark results or reach the accuracy of ML techniques while ensuring the interpretability of DCMs [Wang et al., 2020].

In fact, most of the existing reviews mainly focus on discrete choice models [Barff et al.,1982; Jing et al., 2018]. Ratrout et al., [2014] and Sekhar [2014] enumerate artificial intelligence approaches as well as discrete choice models, with a focus on artificial neural network and fuzzy logic approaches, while not covering ML techniques. Hillel et al., [2021], on the other hand, deal with ML techniques and artificial intelligence approaches used for passenger choice modelling but limits the DCMs only to the logit-based models. Additionally, neither of the previous reviews distinguishes between the models based on the granularity level, namely aggregate versus disaggregate, even though predicting the behavior of one individual or a group of people differs in many ways.

This review aims to comprehensively cover the various Discrete Choice Models (DCMs) and Machine Learning (ML) techniques used for individual and aggregate-level travel mode choice modeling. Specifically, it focuses on studies published within the last four years and examines the dataset collection techniques, granularity levels, and model types used. The second section outlines the methodology used for identifying relevant studies and the data extraction process. In the third section, a concise overview of existing data collection techniques is provided, followed by an enumeration of DCMs and ML techniques used for travel mode choice modeling. The fourth section presents the findings, and the fifth section discusses the accuracy of the reviewed models. Finally, the paper concludes by identifying research trends for passenger choice modeling in each category.

METHODOLOGY

Searching Approach

The study was based on peer reviewed journal articles published in "Science direct" over the four past years. This database includes a high number of publications, and it is one of the most used.

Due to the purpose of covering both the traditional models (DCMs) and the new ones (ML techniques) while trying to identify the relevance of every category for researchers, the following keywords were selected: *Mode choice, travel mode, discrete choice, machine learning, neural network, and fuzzy logic.*

In order to select papers that mainly discuss applications of both DCMs and ML techniques in travel choice behavior, the previous terms were combined as follows:

Science direct: (TITLE-ABS-Key: ("mode choice" OR "travel mode") AND ("machine learning" OR "neural network" OR "discrete choice "OR "fuzzy logic")).

The search resulted in 94 articles, out of which 56 were excluded after analysis. Ten of the excluded articles were conference papers, while 28 were found to be irrelevant to our study. Additionally, 16 articles used DCMs or ML techniques for topics related to transportation other than travel choice behavior. One article focused on a literature review regarding the value of time concept applied to freight transportation, and another article discussed social sciences. As a result, 38 articles were included in the final analysis, and are listed in Table 1.

Table 1. Selected articles for review

Id.	Paper	Id.	Paper
A1	Chang et al., [2020]	A20	Ali et al., [2022]
A2	Li et al., [2020]	A21	Alta de waal and Joubert [2022]
A3	Nguyen and Armoogum [2020]	A22	Bari et al., [2022]
A4	Sanko [2020]	A23	Feng et al., [2022]
A5	Ton et al., [2020]	A24	Gupta et al., [2022]
A6	Wang et al., [2020]	A25	Harz and Sommer [2022]
A7	Yu [2020]	A26	Kapitza [2022]
A8	Zhoa et al., [2020]	A27	Kashifi et al., [2022]
A9	Andani et al., [2021]	A28	Okami et al., [2022]
A10	Jochem et al., [2021]	A29	Saiyad et al., [2022]
A11	L. Yang [2021]	A30	Salas et al., [2022]
A12	Ilahi et al., [2021]	A31	Tao and Nass [2022]
A13	Mo et al., [2021]	A32	Varghese et al., [2022]
A14	Nasrin and Bunker [2021]	A33	Wu et al., [2022]
A15	Song et al., [2021]	A34	Guo et al., [2023]
A16	Sun and Wandelt [2021]	A35	Hamadneh and jaber [2023]
A17	Tu et al., [2021]	A36	Parmar et al., [2023]
A18	Wong and Farooq [2021]	A37	Xia et al., [2023]
A19	Yang et al., [2021]	A38	Zhang et al., [2023]

Data Extraction

After identifying the main articles for the review, we extracted relevant data based on predetermined criteria. Our methodology involved setting up filters for data extraction based on two primary criteria: the type of data collection technique used and the level of aggregation considered. The first category of data collection techniques was further subdivided into three techniques: GPS-based data, the interviewing method, and the web-based interviewing method. The second category was divided into three subcategories: DCMs, ML techniques, and other models. The "other models" category included several methods. Figure 1 illustrates the chosen data extraction process.

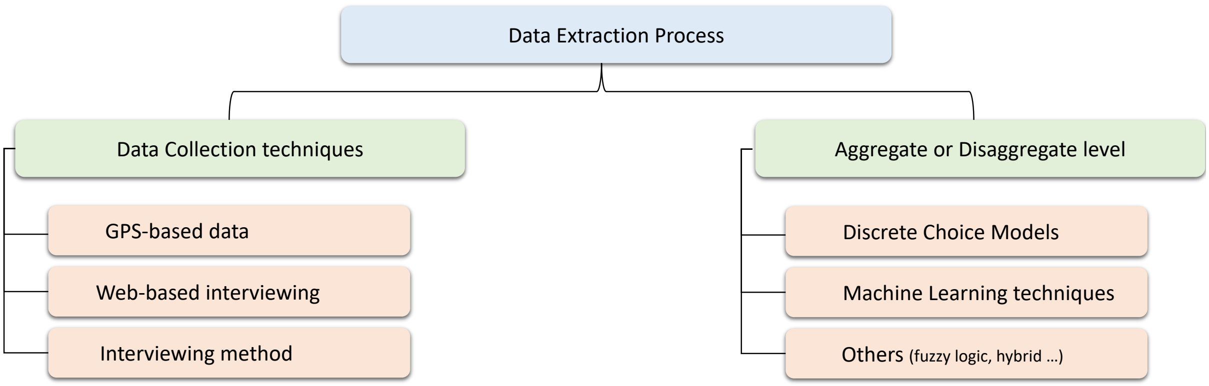

Fig. 1. Data extraction process

MODE CHOICE MODELS AND TECHNIQUES

Data Collection Techniques

As we have previously discussed, Discrete Choice Models (DCMs) and Machine Learning (ML) techniques are commonly used to model either the choice behavior of individuals (disaggregate model) or the choice behavior of a group (aggregate model). However, to obtain accurate outputs from these models, high-quality inputs are required. In this case, the inputs consist of datasets containing information about individuals' preferences for different modes of transportation. To collect this data, various methods are available, each with its own advantages and limitations. Advanced techniques like GPS-based methods or online surveys are commonly used in developed countries, while paper and pencil interviews (Classic interviewing method) are more common in developing countries due to cost constraints and limited internet access.

Regardless of the specific method used, data collection techniques can be broadly classified into three categories, each with its advantages and disadvantages.

GPS-based data (GPS): This category englobes all the techniques that are based on determining the position of the items. The tools usually used could be smartphones, dataloggers, smart cards, etc.

Interviewing method (IM): This technique is based on conducting and filling a preference survey (Stated-SP-or Revealed one-RP) administered to individuals.

Web-based interviewing method (WIM): This technique is simply based on a web questionnaire administered through websites and social networks. This technique could be especially useful when several exogenous restrictions are taking place due to extraordinary contexts or long periods of crisis, such as the pandemic episodes.

Discrete Choice Models

Discrete choice models are a class of statistical models used to analyze and predict the choices made by individuals among a set of discrete alternatives, such as choosing a mode of transportation, a brand of product, or a type of housing. These models assume that individuals choose the alternative that provides the highest utility function, based on the characteristics or attributes of each alternative and their personal preferences. Discrete choice models are widely used in transportation planning and other fields where choice behavior is important. They can provide valuable insights into the factors that influence individual choices and help predict the outcomes of policy interventions. The main types of discrete choice models are multinomial logit, multinomial probit, mixed logit, and nested logit, alongside other more advanced models (see Figure 2).

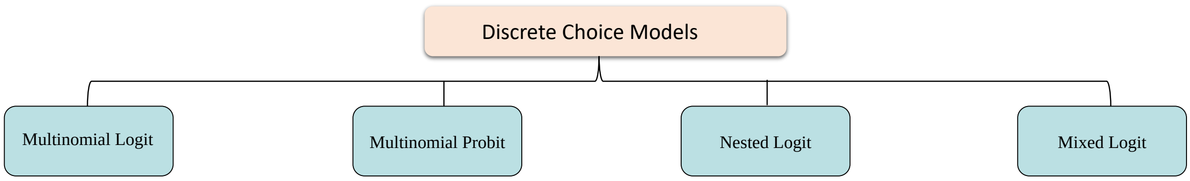

Fig. 2. Discrete Choice Models

Multinomial logit: The multinomial logit is used when the dependent variable is a categorical variable with more than two categories. It is a generalization of the binary logit model, which is used when the dependent variable has only two categories. In the multinomial logit model, the dependent variable is assumed to be drawn from a multinomial distribution, which is a generalization of the binomial distribution to more than two categories. The model estimates the probability of each category by using a logistic function, which is a sigmoid curve that maps any real-valued number to a value between 0 and 1. The multinomial logit model is often used in marketing research and other fields where researchers are interested in predicting which of several alternatives a respondent will choose. It is also used in political science and economics to predict voting behavior and consumer choice, respectively.

Multinomial probit: The multinomial probit model is a type of regression model that is used when the dependent variable is a categorical variable with more than two categories. It is similar to the multinomial logit model, but it assumes that the dependent variable is drawn from a normal distribution instead of a multinomial distribution. The multinomial probit model estimates the probability of each category by using a probit function, which is a cumulative distribution function that maps any real-valued number to a value between 0 and 1. The probit function is the inverse of the standard normal distribution function, and it is used to calculate the probability that a normally distributed random variable will fall below a given threshold. The multinomial probit model is often used in marketing research and other fields where researchers are interested in predicting which of several alternatives a respondent will choose. It is also used in political science and economics to predict voting behavior and consumer choice, respectively.

Mixed logit: The mixed logit model is an extension of the multinomial logit model that allows for more flexibility in modeling the choice behavior of individuals. It is used when the dependent variable is a categorical variable with more than two categories and the independent variables include both continuous and categorical variables. The mixed logit model accounts for unobserved heterogeneity in the data by including random effects. These random effects allow for the modeling of individual-level differences in preferences and decision-making processes that are not captured by the observed independent variables.

Nested logit: The nested logit model is used to analyze data in which the dependent variable is a categorical variable with more than two categories and there is a hierarchical structure to the choices being made. It is a generalization of the multinomial logit model, which is used when there is no nesting in the data. In the nested logit model, the choices are divided into disjoint groups called nests, and the choice within each nest is modeled using a logistic function, which is a sigmoid curve that maps any real-valued number to a value between 0 and 1. The probability of choosing a particular nest is then modeled using a second logistic function. The nested logit model is useful for modeling data in which the choices being made are not independent, such as when there is a hierarchical structure to the choices or when the choices are correlated.

Machine Learning Techniques

The use of machine learning techniques in modeling travel mode choice behavior is becoming increasingly popular. One approach involves using these techniques to estimate the parameters of the utility function. In this approach, a machine learning model is trained on a dataset containing observations of individual choices and relevant explanatory variables. The machine learning algorithm then adjusts the parameters of the utility function to minimize the difference between the predicted and observed choices in the dataset. Another approach involves using machine learning to develop more complex models that incorporate multiple utility functions or allow for non-linear relationships between variables. For instance, neural networks can estimate the utility function for each alternative separately and then combine them to predict the final choice. Overall, machine learning techniques offer a flexible and powerful approach to estimating utility functions in discrete choice models, which can enhance the accuracy and robustness of choice predictions. In Figure 3, we have illustrated the commonly used machine learning techniques for modeling travel mode choice behavior.

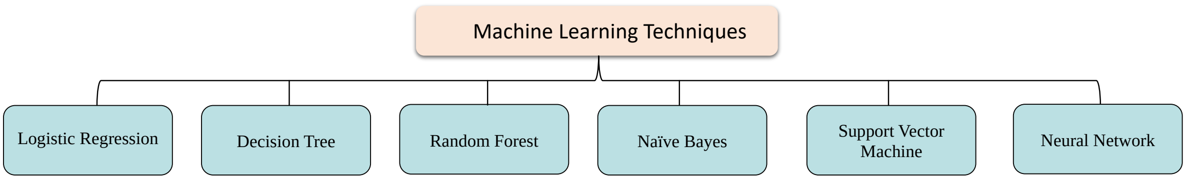

Fig. 3. Machine Learning Techniques for Travel Mode Choice Behavior Modelling

Decision Tree (DT): A decision tree is a tree-like model used for classification and regression tasks. It is a supervised learning algorithm that can be used to predict a target variable by learning decision rules inferred from features of the data. The tree is constructed by making decisions based on the features of the data, starting at the root node and working down the tree until a decision or prediction is made at a leaf node. The decisions made at each node are based on the values of the features and the target variable, and the tree is trained using labeled data. Once trained, the decision tree can be used to make predictions about new, unseen data by following the decision rules learned during training.

Random Forest (RF): A random forest is an ensemble machine learning algorithm used for classification and regression. It consists of a collection of decision trees, trained on randomly selected subsets of the training data, with the goal of reducing the variance and improving the predictive accuracy of the model. During the training process, each tree in the random forest makes a prediction based on the features of the input data, and the final prediction is made by averaging the predictions of all the trees. The random forest algorithm is a popular choice for many machine learning tasks because it is easy to implement, can handle a large number of features, and generally produces good performance on a wide range of tasks.

Logistic Regression (LR): Logistic regression is a statistical model that is used for binary classification, i.e., to predict a binary outcome (such as success or failure, 0 or 1, etc.) based on one or more predictor variables. It is a linear model that is based on the assumption that the relationship between the dependent variable and the predictors is linear, and that the outcome is a binary variable that follows a logistic distribution. During the training process, the logistic regression model estimates the coefficients of the predictor variables, and these coefficients are used to make predictions about the outcome for new data. Logistic regression is widely used in a variety of applications, including image and speech recognition, natural language processing, and social media analysis.

Naïve Bayes (NB): Naive Bayes is a classification algorithm based on the Bayes theorem, which states that the probability of an event is equal to the prior probability of the event multiplied by the likelihood of the event given some evidence. In the context of classification, the event is the class label, and the evidence is the feature values of the input data. The "naive" part of the algorithm comes from the assumption that all the features are independent of each other, given the class label. This assumption is often unrealistic, but the algorithm works well in practice and is particularly useful when there are many features and the relationships between them are not well understood.

Support Vector Machine (SVM): Support vector machines work by finding a hyperplane in a high-dimensional space that maximally separates the data points of different classes. The hyperplane is chosen in such a way that it has the largest distance (called the margin) to the nearest data points of any class, which are called support vectors. The main advantage of SVMs is that they can handle high-dimensional data and data that is not linearly separable. They do this using the kernel trick, which allows them to map the data into a higher-dimensional space where it becomes linearly separable. In addition, SVMs can give probabilities for classification tasks by using Platt scaling.

Artificial Neural Network (ANN): A neural network is a type of machine learning model inspired by the structure and function of the human brain. It consists of layers of interconnected nodes (also called neurons), which process and transmit information. There are several types of neural networks, including Feedforward Neural Networks (FFNN), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN). Neural networks have been widely used in behavior prediction, such as predicting consumer behavior or travel mode choice. They are able to learn patterns and relationships in data and can make predictions based on these patterns.

RESULTS ANALYSES

The articles selected in this literature review present the characteristics of techniques used for analyzing travel mode choice. In an effort to capture this concept, this section uses evidence from the 38 studies to explore each of the three criterions identified to develop a broader understanding, namely: the level of aggregation considered, the data collection techniques used, and the analyzing techniques employed.

Classification by Level of aggregation

We refer to the level of aggregation as the degree of detail at which the preferences are grouped together or classified. It involves deciding how to categorize the data and the prediction into broader categories or classes for the purpose of analysis. An aggregate model is a type of model that uses summary data to make predictions or draw conclusions about a group of individuals having the same characteristics. In contrast, a disaggregate model is a type of model that uses individual-level data to make predictions or draw conclusions about the behavior or choice of individual actors. The level of aggregation can have an impact on the complexity of the model, as well as the usefulness of the results for different purposes.

Based on the findings presented in Figure 4, it can be observed that out of the total number of articles analyzed, 32 articles used an aggregate model, while only 6 articles employed a disaggregate model. The results suggest that disaggregate models are particularly suited for addressing specific research objectives, such as analyzing travel behavior in response to cost reductions [Li et al., 2020; Bari et al., 2022], identifying travel modes [Nguyen and Armoogum 2020; Yu 2020], or building effective recommendation systems [Sun and Wandelt 2021; Wu et al., 2022].

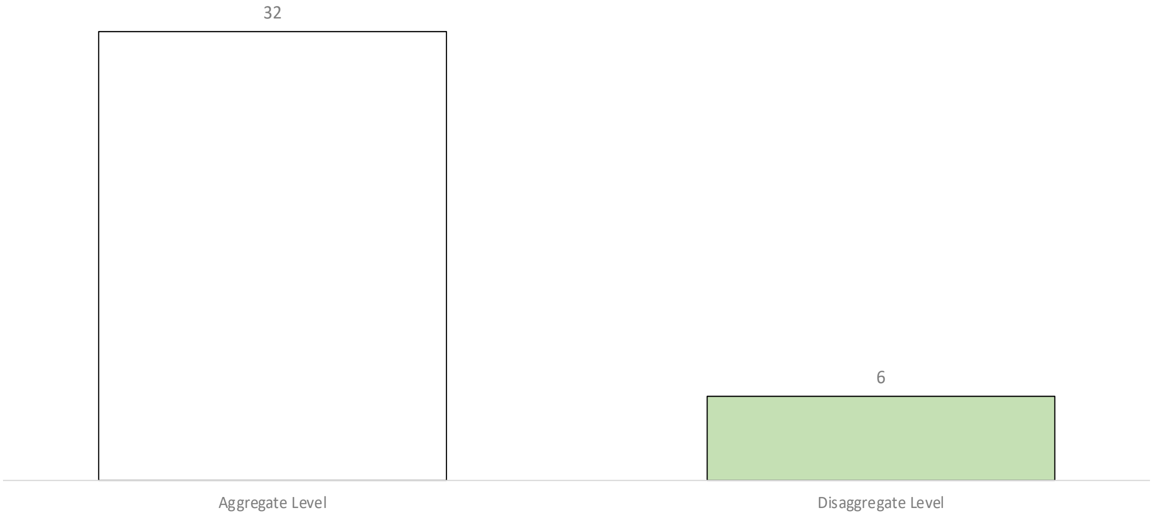

Fig. 4. Paper by level of aggregation

The Relationship Between Data Sources and Levels of Aggregation

Table 2 presents an overview of the data collection methods used in the analyzed studies. The results show that among the studies analyzed, 28 studies used a classic interviewing method, 7 studies used a GPS-based method, and 3 employed a web-based interviewing method. These findings indicate that researchers have used a variety of data collection methods in their studies, with classic interviewing being the most commonly used method and GPS-based and web-based interviewing method being less common but still utilized in some studies.

Table 2. Relationship between the level of aggregation and the data source

	GPS-based data	Interviewing method	Web-based interviewing method
Aggregate Level	A1, A17	A5, A6, A8, A9, A10, A11, A12, A14, A15, A18, A19, A20, A21, A23, A24, A25, A26, A27, A28, A29, A30, A31, A32, A35, A36, A37, A38	A4, A13, A34
Disaggregate Level	A2, A3, A7, A16, A33	A22	
Total	**7**	**28**	**3**

Furthermore, Table 2 highlights the close relationship between the data source and the level of aggregation in mode choice modeling. We can observe that the choice of data source can influence the level of aggregation in several ways. For instance, if the data source is based on group-level data, such as survey responses from a sample of individuals, then the resulting model will likely be an aggregate model that represents the behavior of the group as a whole [Ilahi et al. 2021; L. Yang 2021; Nasrin and Bunker 2021; Harz and Sommer 2022; Hamadneh and Jaber 2023]. This is because group-level data provide information about the overall behavior of a population or sample, rather than the behavior of individual actors within that population. On the other hand, if the data source is based on individual-level data, such as data collected from GPS devices, then the resulting model will likely be a disaggregate model that represents the behavior of individual actors [Li et al., 2020; Nguyen and Armoogum 2020; Yu 2020; Sun and Wandelt 2021]. This is because individual-level data provide information about the unique characteristics and behaviors of each individual, which can be used to predict their individual preferences in terms of mode choice. Therefore, the choice of data source is an important consideration in mode choice modeling, as it can influence the level of aggregation and the accuracy of the resulting model.

The Relationship Between Levels of Aggregation and Analyses techniques

Table 3 provides a summary of the articles reviewed in terms of the type of analyses techniques used for mode choice modeling. It reveals that a similar number of studies utilized Discrete Choice Models (DCMs) and Machine Learning Techniques (ML), with 17 studies utilizing DCMs and 18 studies employing ML. Some studies explored both methods, either to compare their performances [Saiyad et al., 2022] or their interpretability [Zhoa et al., 2020, Salas et al., 2022]. Additionally, the "Others" category encompasses articles that adopted alternative methods, including fuzzy logic [Nguyen and Armoogum 2020], rule-based systems [Zhang et al., 2023], data mining [Chang et al., 2020], or hybrid models [Andani et al., 2021; Wong and Farooq 2021, Gupta et al., 2022, Parmar et al., 2023].

Table 3. Relationship between levels of aggregation and analyses techniques

	Discrete Choice Models	Machine Learning Techniques	Others
Aggregate Level	A4, A5, A8, A9, A10, A12, A13, A15, A19, A23, A25, A26, A28, A29, A30, A32, A35	A6, A8, A11, A14, A17, A20, A21, A27, A29, A30, A31, A35, A37	A1, A9, A15, A18, A24, A34, A36, A38
Disaggregate Level		A2, A3, A7, A16, A22	A3, A33
Total	**17**	**18**	**10**

The relationship between the levels of aggregation in mode choice and analysis techniques, such as Discrete Choice Models and Machine Learning Techniques, can be complex and dependent on various factors. Nevertheless, we can observe from Table 3 that Discrete Choice Models are used in aggregate-level analyses, while Machine Learning Techniques are more commonly preferred in individual-level analyses. At the aggregate level, Discrete Choice Models typically use aggregate-level data, such as survey responses, to estimate the relative importance of different factors influencing mode choice, such as travel time, cost, and convenience. Examples of Discrete Choice Models used at the aggregate level include multinomial logit, nested logit, and mixed logit models. At the individual level, Machine Learning Techniques often use individual-level data, such as GPS data, to predict the mode choice behavior of individuals. Examples of Machine Learning Techniques used in mode choice analysis include neural networks, decision trees, random forests, and support vector machines.

Overall, the choice of analysis technique depends on the research question, the available data, and the level of aggregation considered. Discrete Choice Models are typically better suited for aggregate-level analyses, while Machine Learning Techniques are more suitable for disaggregate-level analyses. However, there may be cases where a combination of both techniques is necessary to obtain a comprehensive understanding of mode choice behavior.

Figure 5 presents the utilization of different machine learning techniques and discrete choice models in the analyzed studies. It reveals that under the Machine Learning Techniques category, the Random Forest/Decision Trees (RF/DT) and Neural Networks (NN) were the most frequently employed techniques, appearing in 11 and 10 studies respectively. Support Vector Machines (SVM) and Naive Bayes (NB) were employed in 3 studies each, and Logistic Regression (LR) was utilized in 2 studies, indicating their relatively less frequent usage. Turning to the Discrete Choice Models category, the figure shows that the Multinomial Logit (MNL) model was the most commonly employed, appearing in 14 studies. The Mixed Logit (ML) model was utilized in 3 studies, while the Nested Logit (NL) and Mixed Logit (MNP) models were employed in 2 and 1 study, respectively.

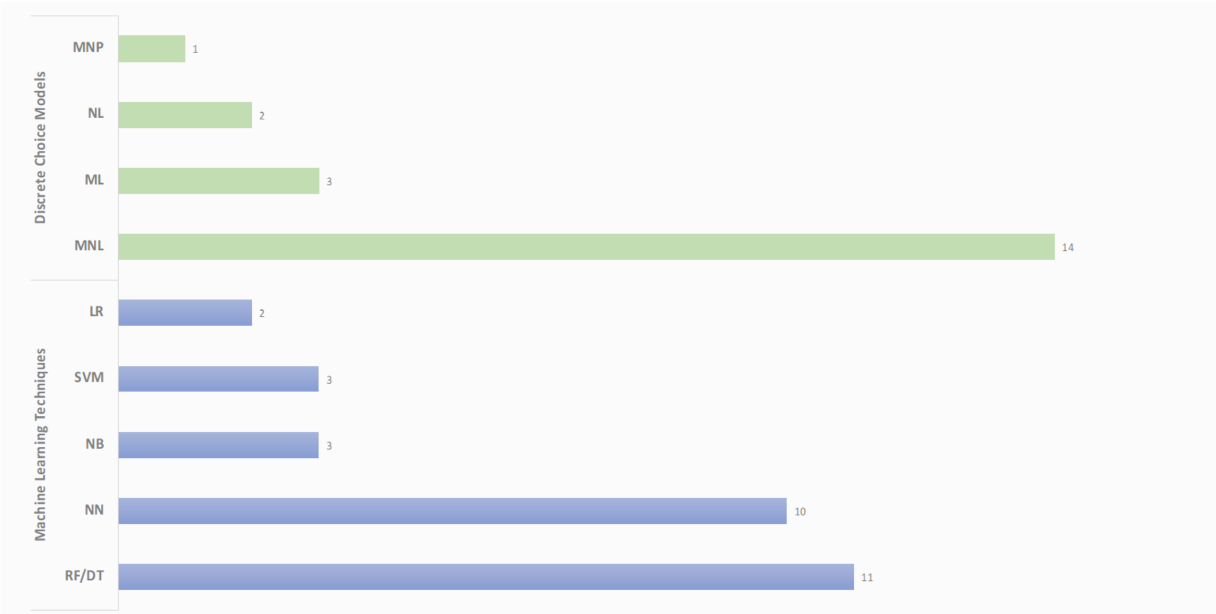

Fig. 5. DCMs and ML models in the selected articles

Figure 5 highlights the prevalence of Random Forest/Decision Trees and Multinomial Logit models in their respective categories, indicating their popularity among researchers in mode choice modeling. The usage of Neural Networks also demonstrates the growing interest in applying machine learning techniques to this field. These findings provide valuable insights for researchers and practitioners seeking to understand the techniques employed in mode choice modeling and the distribution of their usage across different models.

Models Performance

In this section we discuss the performance of the adopted models in the selected articles by comparing the accuracies of the Discrete Choice Models, Machine Learning Techniques, and the hybrid models for each granularity category, namely, the aggregate and disaggregate ones. Six articles based their study on disaggregate models, but only four of them computed the accuracy of their models and compared them with other models: [Li et al., 2020, Nguyen and Armoogum 2020, Yu 2020, Wu et al., 2022]. Secondly, 13 articles used aggregate models; 9 of them based their evaluation on the accuracy criteria: [Wang et al., 2020; Zhoa et al., 2020; L. Yang 2021; Ali et al., 2022; Kashifi et al., 2022; Saiyad et al., 2022; Salas et al., 2022, Xia et al., 2023]. Tables 4 and 5 show the minimum, average, and maximum accuracies (Acc) of the different techniques used in these studies.

Table 4. Accuracies of disaggregate models

Models	Min Acc	Average Acc	Max Acc
Machine Learning	86,7%	89,4%	92%
Hybrid models	91,5%	92,75%	94%
Discrete Choice Models	-	-	-

Table 5. Accuracies of aggregate models

Models	Min Acc	Average Acc	Max Acc
Machine Learning	61%	73,4%	86,3%
Discrete Choice Models	54,2%	62,01%	72%
Hybrid models	63,62%	79,2%	89,87%

Table 5 shows that Machine Learning (ML) techniques outperform Discrete Choice Models (DCMs) in terms of accuracy at the aggregate level. However, it is noteworthy that several studies in this review still opted for DCMs instead of ML techniques, as indicated in Table 3. The reasons behind this choice can be explained as follows: (i) Objectives: ML techniques primarily focus on making predictions, whereas DCMs are not only used for prediction but also for studying the causal factors that influence mode choice behavior. DCMs provide insights into the underlying relationships and factors affecting mode choice, in addition to predictive capabilities. (ii) Interpretability: DCMs offer greater interpretability compared to ML techniques. The meaning of variables and the steps involved in DCMs are relatively clear, making it easier to understand and interpret the results. On the other hand, the interpretability of ML models, especially complex ones like neural networks, can be challenging. The inner workings and meaning of different layers in neural networks may not be readily transparent. (iii) Generalization: DCMs are generally more suited for generalization beyond the specific study. ML techniques, if not properly controlled for overfitting, may suffer from limited generalizability. Overfitting occurs when the model becomes too specific to the training data, leading to poor performance on new unseen data. In contrast, DCMs are designed to capture general patterns and trends in mode choice behavior. In summary, while ML techniques may outperform DCMs in terms of accuracy at the aggregate level, the choice of DCMs in many studies can be attributed to their focus on causal analysis, interpretability, and generalizability.

Gao et al., [2021] proposed a solution to the generalization issue of Neural Networks by developing an extrapolation-enhanced model with knowledge-based decision-making theory. The model was trained on one dataset and extrapolated on two different datasets, giving satisfying results. Hybrid models that combine different techniques could represent the middle ground by acquiring both higher accuracy and maintaining decent interpretability. Li et al., [2020] developed such a hybrid model by combining the generative adversarial model and the convolutional neural network (CNN). The resulting model was compared to other ML techniques, including Random Forest (RF) and Support Vector Machine (SVM), and achieved an accuracy of 86.70%. Nguyen and Armoogum [2020] employed a combination of Random Forest (RF) and fuzzy logic in a hierarchical process to detect travel modes from GPS data. This approach achieved an accuracy of 89.10%. By combining different models' categories, higher accuracies were attained compared to using a single technique. In another study, Wang et al. [2020] developed a novel hybrid model by integrating the utility function with a deep neural network. This hybrid model was then compared to two discrete choice models (multinomial logit and nested logit) and five machine learning techniques (logistic regression, support vector machine, Naïve Bayesian, decision tree, and k-nearest neighbors). The findings revealed that the hybrid model outperformed all the other techniques, with a mean accuracy of 66.30%. These studies demonstrate the benefits of combining different techniques in a hybrid model.

CONCLUSION

This paper presents a comprehensive review of travel mode choice models utilized in articles published within the last four years. A systematic search strategy, employing targeted keywords, was employed resulting in the selection of 38 articles for analysis. A rigorous data extraction process was then applied to these articles, examining the data collection techniques employed, determining whether the models were aggregate or disaggregate, and classifying them into relevant categories, primarily Discrete Choice Models (DCMs) or Machine Learning (ML) models.

The findings reveal that among the selected articles, 32 utilized aggregate models, while the remaining 6 opted for disaggregate models. Additionally, a comparable number of studies employed Discrete Choice Models and Machine Learning Techniques, with some adopting alternative methods. Notably, Discrete Choice Models were predominantly used for aggregate-level analyses, whereas Machine Learning Techniques were more commonly applied in disaggregate-level analyses. A critical discussion on the accuracy of the models utilized in the reviewed articles is presented, suggesting that hybrid models may offer a promising solution for achieving higher prediction accuracy while maintaining interpretability. Furthermore, an examination of the selected articles revealed a concentration of studies in European countries, China, and the USA, highlighting the need for more research in developing countries. Moreover, it is observed that the disaggregate models used in the reviewed articles often lack in-depth analysis of individual behavior and fail to consider external factors such as weather or seasons. Therefore, future studies should leverage technological advancements and explore new factors that influence individual mode choice behavior. Additionally, there is a need for further research on the strengths and potential of hybrid models combining DCMs with ML techniques or deep learning techniques to provide clear guidance for practitioners unfamiliar with these methods, thereby facilitating the design of well-suited transportation policies. Lastly, considering the variety of models available, an important question arises regarding the extent to which the models employed in specific studies can be generalized to other contexts. This calls for a deeper understanding of model applicability and generalizability to ensure their effectiveness in diverse settings.

ACKNOWLEDGMENTS

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Ali N. F. M., Sadullah A. F. M., Majeed A. P. A., Razman M. A. M., Musa R. M., 2022, The identification of significant features towards travel mode choice and its prediction via optimised random forest classifier: An evaluation for active commuting behavior, Journal of Transport & Health, 25, 101362. https://doi.org/10.1016/j.jth.2022.101362

Andani I.G.A., La Paix Puello L., Geurs K., 2021, Modelling effects of changes in travel time and costs of toll road usage on choices for residential location, route and travel mode across population segments in the Jakarta-Bandung region, Indonesia, Transportation Research Part A: Policy and Practice, 145, pp. 81–102. https://doi.org/10.1016/j.tra.2020.12.012

Barri E. Y., Farber S., Jahanshahi H., Beyazit E., 2022, Understanding transit ridership in an equity context through a comparison of statistical and machine learning algorithms, Journal of Transport Geography, 105, 103482. https://doi.org/10.1016/j.jtrangeo.2022.103482

Barff R., Mackay D., Olshavsky R.W., 1982, A Selective Review of Travel-Mode Choice Models, Journal of Consumer Research, 8, 4, p. 370. https://doi.org/10.1086/208877

Ben-Akiva M.E., Lerman S.R., 1985, Discrete choice analysis: theory and application to travel demand, Cambridge, Mass: MIT Press (MIT Press series in transportation studies, 9).

Chang X. et al., 2020, Understanding user's travel behavior and city region functions from station-free shared bike usage data, Transportation Research Part F: Traffic Psychology and Behaviour, 72, pp. 81–95. https://doi.org/10.1016/j.trf.2020.03.018

Cheng L. et al., 2019, Applying a random forest method approach to model travel mode choice behavior, Travel Behaviour and Society, 14, pp. 1–10. https://doi.org/10.1016/j.tbs.2018.09.002

De Waal, Alta et al., 2022, Explainable Bayesian networks applied to transport vulnerability, Expert Systems with Applications, vol. 209, p. 118348. https://doi.org/10.1016/j.eswa.2022.118348

Feng Y., Zhao J., Sun H., Wu J., Gao, Z., 2022, Choices of intercity multimodal passenger travel modes. Physica A: Statistical Mechanics and its Applications, 600, 127500. https://doi.org/10.1016/j.physa.2022.127500

Gao K. et al., 2021, Extrapolation-enhanced model for travel decision making: An ensemble machine learning approach considering behavioral theory, Knowledge-Based Systems, 218, p. 106882. https://doi.org/10.1016/j.knosys.2021.106882

Gupta A., Bivina G. R., Parida M., 2022, Does neighborhood design matter for walk access to metro stations? An integrated SEM-Hybrid discrete mode choice approach, Transport policy, 121, 61-77. https://doi.org/10.1016/j.tranpol.2022.03.010

Guo D., Yao E., Liu S., Chen R., Hong J., Zhang J., 2023, Exploring the role of passengers' attitude in the integration of dockless bike-sharing and public transit: A hybrid choice modeling approach, Journal of Cleaner Production, 384, 135627. https://doi.org/10.1016/j.jclepro.2022.135627

Kashifi M. T., Jamal A., Kashefi M. S., Almoshaogeh M., Rahman S. M., 2022, Predicting the travel mode choice with interpretable machine learning techniques: A comparative study, Travel Behaviour and Society, 29, 279-296. https://doi.org/10.1016/j.tbs.2022.07.003

Kapitza J., 2022, How people get to work at night, A discrete choice model approach towards the influence of nighttime on the choice of transport mode for commuting to work, Journal of Transport Geography, 104, 103418. https://doi.org/10.1016/j.jtrangeo.2022.103418

Hamadneh J., Jaber A., 2023, Modeling of intra-city transport choice behaviour in Budapest, Hungary, Journal of Urban Mobility, 3, 100049. https://doi.org/10.1016/j.urbmob.2023.100049

Hagenauer J., Helbich M., 2017, A comparative study of machine learning classifiers for modeling travel mode choice, Expert Systems with Applications, 78, pp. 273– 282.

https://doi.org/10.1016/j.eswa.2017.01.057

Harz J., Sommer C., 2022, Mode choice of city tourists: Discrete choice modeling based on survey data from a major German city, Transportation Research Interdisciplinary Perspectives, 16, 100704. https://doi.org/10.1016/j.trip.2022.100704

Hillel T. et al., 2021, A systematic review of machine learning classification methodologies for modelling passenger mode choice, Journal of Choice Modelling, 38, p. 100221. https://doi.org/10.1016/j.jocm.2020.100221

Ilahi A. et al., 2021, Understanding travel and mode choice with emerging modes; a pooled SP and RP model in Greater Jakarta, Indonesia, Transportation Research Part A: Policy and Practice, 150, pp. 398–422. https://doi.org/10.1016/j.tra.2021.06.023

Jing, P. et al., 2018, Travel Mode and Travel Route Choice Behavior Based on Random Regret Minimization: A Systematic Review, Sustainability, 10, 4, p. 1185. https://doi.org/10.3390/su10041185

Jochem P., Lisson C., Khanna A.A., 2021, The role of coordination costs in mode choice decisions: A case study of German cities, Transportation Research Part A: Policy and Practice, 149, pp. 31–44. https://doi.org/10.1016/j.tra.2021.04.001

Li L. et al., 2020, Coupled application of generative adversarial networks and conventional neural networks for travel mode detection using GPS data, Transportation Research Part A: Policy and Practice, 136, pp. 282–292. https://doi.org/10.1016/j.tra.2020.04.005

Mo B. et al., 2021, Impacts of subjective evaluations and inertia from existing travel modes on adoption of autonomous mobility-on-demand, Transportation Research Part C: Emerging Technologies, 130, p. 103281. https://doi.org/10.1016/j.trc.2021.103281

McFadden D., Reid F., 1975, Aggregate travel demand forecasting from disaggregated behavioral models, Institute of Transportation and Traffic Engineering, University of California.

Mohammadian A., Miller E.J., 2002, Nested Logit Models and Artificial Neural Networks for Predicting Household Automobile Choices: Comparison of Performance, Transportation Research Record: Journal of the Transportation Research Board, 1807, 1, pp. 92– 100. https://doi.org/10.3141/1807-12

Nasrin S., Bunker J., 2021, Analyzing significant variables for choosing different modes by female travelers, Transport Policy, 114, pp. 312–329. https://doi.org/10.1016/j.tranpol.2021.10.017

Nguyen M.H., Armoogum J., 2020, Hierarchical process of travel mode imputation from GPS data in a motorcycle-dependent area, Travel Behaviour and Society, 21, pp. 109–120. https://doi.org/10.1016/j.tbs.2020.06.006

Okami S., Matsuyuki M., Sarmiento-Ordosgoitia I., Nakamura F., 2022, Residents' travel behavior in a low-income settlement with aerial cable cars in Medellin, Colombia. Case studies on transport policy, 10, 2, 1332-1342. https://doi.org/10.1016/j.cstp.2022.04.016

Parmar J., Saiyed G., Dave S., 2023, Analysis of taste heterogeneity in commuters' travel decisions using joint parking–and mode–choice model: A case from urban India, Transportation Research Part A: Policy and Practice, 170, 103610. https://doi.org/10.1016/j.tra.2023.103610

Salas P., De la Fuente R., Astroza S., Carrasco J. A., 2022, A systematic comparative evaluation of machine learning classifiers and discrete choice models for travel mode choice in the presence of response heterogeneity, Expert Systems with Applications, 193, 116253. https://doi.org/10.1016/j.eswa.2021.116253

Ratrout N. T., Gazder U., Al-Madani H. M., 2014, A review of mode choice modelling techniques for intra-city and border transport, World Review of Intermodal Transportation Research, 5, 1, 39-58. https://doi.org/10.1504/WRITR.2014.065055

Saiyad G., Srivastava M., Rathwa D., 2022, Exploring determinants of feeder mode choice behavior using Artificial Neural Network: Evidences from Delhi metro, Physica A: Statistical Mechanics and its Applications, 598, 127363. https://doi.org/10.1016/j.physa.2022.127363

Sanko N., 2020, Activity-end access/egress modal choices between stations and campuses located on a hillside, Research in Transportation Economics, 83, p. 100931. https://doi.org/10.1016/j.retrec.2020.100931

Sekhar Ch. R., 2014, Mode choice analysis: The data, the models and future ahead, International Journal for Traffic and Transport Engineering, 4, 3, pp. 269–285. http://dx.doi.org/10.7708/ijtte.2014.4(3).03

Song Y., Li D., Cao Q., Yang M., Ren G., 2021, The whole day path planning problem incorporating mode chains modeling in the era of mobility as a service, Transportation Research Part C: Emerging Technologies, 132, 103360. https://doi.org/10.1016/j.trc.2021.103360

Sun X., Wandelt S., 2021, Transportation mode choice behavior with recommender systems: A case study on Beijing, Transportation Research Interdisciplinary Perspectives, 11, p. 100408. https://doi.org/10.1016/j.trip.2021.100408

Tao T., Næss P., 2022, Exploring nonlinear built environment effects on driving with a mixed-methods approach, Transportation research part D: transport and environment, 111, 103443. https://doi.org/10.1016/j.trd.2022.103443

Ton D. et al., 2020, The experienced mode choice set and its determinants: Commuting trips in the Netherlands, Transportation Research Part A: Policy and Practice, 132, pp. 744–758. https://doi.org/10.1016/j.tra.2019.12.027

Tu M. et al., 2021, Exploring nonlinear effects of the built environment on ridesplitting: Evidence from Chengdu, Transportation Research Part D: Transport and Environment, 93, p. 102776. https://doi.org/10.1016/j.trd.2021.102776

Varghese V., Chikaraishi M., Jana A., 2022. Joint analysis of mode and travel-based multitasking choices in Mumbai, India, Travel Behaviour and Society, 27, 148-161. https://doi.org/10.1016/j.tbs.2022.01.006

Wu F., Lyu C., Liu Y., 2022, A personalized recommendation system for multi-modal transportation systems. Multimodal transportation, 1, 2, 100016. https://doi.org/10.1016/j.multra.2022.100016

Wang F., Ross C. L., 2018, Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model, Transportation Research Record: Journal of the Transportation Research Board, 2672, 47, pp. 35– 45. https://doi.org/10.1177/0361198118773556

Wang S., Mo B., Zhao J., 2020, Deep neural networks for choice analysis: Architecture design with alternative- specific utility functions, Transportation Research Part C: Emerging Technologies, 112, pp. 234–251. https://doi.org/10.1016/j.trc.2020.01.012

Wong M., Farooq B., 2021, ResLogit: A residual neuralnetwork logit model for data-driven choice modelling, Transportation Research Part C: Emerging Technologies, 126, p. 103050. https://doi.org/10.1016/j.trc.2021.103050

Xia Y., Chen H., Zimmermann R., 2023, A Random Effect Bayesian Neural Network (RE-BNN) for travel mode choice analysis across multiple regions, Travel Behaviour and Society, 30, 118-134. https://doi.org/10.1016/j.tbs.2022.08.011

Xie C., Lu J., Parkany E., 2003, Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks, Transportation Research Record: Journal of the Transportation Research Board, 1854, 1, pp. 50–61. https://doi.org/10.3141/1854-06

Yang, Liya et al., 2021, Driving as a commuting travel mode choice of car owners in urban China: Roles of the built environment, Cities, 112, p. 103114. https://doi.org/10.1016/j.cities.2021.103114

Yang, Linchuan et al., 2021, To walk or not to walk? Examining non-linear effects of streetscape greenery on walking propensity of older adults, Journal of Transport Geography, 94, p. 103099. https://doi.org/10.1016/j.jtrangeo.2021.103099

Yu J.J.Q., 2020, Semi-supervised deep ensemble learning for travel mode identification, Transportation Research Part C: Emerging Technologies, 112, pp. 120–135. https://doi.org/10.1016/j.trc.2020.01.003

Zhang Y., Xie Y., 2008, Travel Mode Choice Modeling with Support Vector Machines, Transportation Research Record: Journal of the Transportation Research Board, 2076, 1, pp. 141–150. https://doi.org/10.3141/2076-16

Zhang J., Feng T., Timmermans H. J., Lin Z., 2023, Association rules and prediction of transportation mode choice: Application to national travel survey data. Transportation Research Part C: Emerging Technologies, 150, 104086. https://doi.org/10.1016/j.trc.2023.104086

Zhao X. et al., 2020, Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models, Travel Behaviour and Society, 20, pp. 22–35. https://doi.org/10.1016/j.tbs.2020.02.003

Ouissam Benjdiya
Euromed Polytechnic School, Euromed Research Center,
Euromed University of Fes, Fes, **Morocco**
e-mail:o.benjdiya@ueuromed.org

Naoufal Rouky
Euromed Polytechnic School, Euromed Research Center,
Euromed University of Fes, Fes, **Morocco**
e-mail: n.rouky@ueuromed.org

Othmane Benmoussa
Euromed Polytechnic School, Euromed Research Center,
Euromed University of Fes, Fes, **Morocco**
e-mail: o.benmoussa@ueuromed.org

Mouhsene Fri
Euromed Polytechnic School, Euromed Research Center,
Euromed University of Fes, Fes, **Morocco**
e-mail: m.fri@ueuromed.org