Transport multi-mode choice prediction using a hybrid multi-output regression modelling

Tuo Mao[1*], Seunghyeon Lee[1], Yuming Ou[1] and Adriana-Simona Mihăiţă[1†]

[1*]Faculty of Engineering and IT, University of Technology, Sydney, 15 Broadway, Ultimo, Sydney, 2007, NSW, 2007.

*Corresponding author(s). E-mail(s): tuo.mao@uts.edu.au;
Contributing authors: seunghyeon.lee@uts.edu.au;
yuming.ou@uts.edu.au; adriana-simona.mihaita@uts.edu.au;
†These authors contributed equally to this work.

Abstract

Predicting how many travellers will choose a specific transport mode for daily commuting is always a challenging problem due to separate and large data sets, lack of integration and a significant over reliance mostly on surveying approaches. This paper presents a new approach for multi-modal transport choice prediction, via a hybrid structure of regressor-chain and multi-output regression modelling, with the purpose of predicting the number of travellers choosing either a single or a combination of transport modes for a regular home-to-work journey. This is a unique attempt to leverage data-driven approaches as compared to traditional multi-nominal logit models, and is applied over 10-year worth of data for 198 local government areas from New South Wales, Australia. The results indicate that our advanced machine learning framework predicts with an excellent accuracy of MAPE and RMSE: a) below **0.3%** and **0.001** respectively, for a single mode prediction, and b) below **9.76%** and **0.0025** respectively, for a multiple mode prediction, while significantly outperforming baseline regressors considered for comparison.

Keywords: Multi-output regression, Transport Mode choice, Regressor chain regression, Machine learning, Multi-modal behavioural choice.

1 Introduction

Modelling transport behaviour and people choices has been a challenging topic for several decades due to the high complexity of interference between traditional transportation modes with several external factors, technological improvements and new on-demand transportation modes. Modelling the way people make decisions is an important step for many transport planning agencies that need to make sure the current and future traffic demand are met with the necessary supply, especially from a public transport perspective [1].

Traditionally, mode choice models are econometric discrete choice models which have been constructed on the behavioural principle of random utility maximisation; therefore, the transport mode choice behavior is interpreted from the the econometric discrete choice models [2]-[3]. In general, the discrete choice models require more intensive work to be able to define them and estimate the correct otuput. Often the transport movement is separated in several explanatory variables, e.g. travel purpose, and then for each of those segments, a different functional form of the observed part of the utility is specified, depending on the fit to the data. The unobserved part of the utility has a fixed statistical distribution imposed [2]. Additionally, the random utility models are estimated on a per-decision maker case, meaning careful consideration of biases in the data are necessary as these can affect the final prediction results [1].

Given the rise of rich data sets and numerous transport modes, more scalable and data-hungry models are needed to be able to keep up the pace with the multitude or variables and all the possible combinations of scenarios that can be generated. This aspect motivated our current approach of investigate the potential of data-driven modelling, as compared to traditional state of art works as detailed in Section 2 of this paper. The study area is NSW, Australia powered by all available datasets from the previous 10 years of census information as detailed in the Section 3.

Section 3 presents the coverage area, the problem formulation and the methodology we have applied in using the proposed advanced hybrid ML models, together with their hyper parameter tuning and evaluation. Sections 4 - presents the experimental results across all scenarios applied on seven different targeted transport modes and their performance evaluation against each other. Lastly, we conclude with a discussion on the most important features affecting the multi-modal mode choice in Australia, and the lessons learned for future applications in Section 6.

2 Literature review

The four-stage model (FSM) is a commonly used model in the urban transportation model systems (UTMS). In the FSM, one usually models the mode choice of the users as a trip-based decision making process which means that each trip is independent in the decision making process of choosing certain traffic mode. Therefore, the discrete mode choice model is always solved by

the Multinomial logit model (MNL). The MNL model is an econometric model which aims to maximized the utility of the choices [2]. Later on, the extension of the MNL model was developed into the nested logit model[4] and mixed logit model [5] in order to loose the Independence of irrelevant alternatives (IIA) assumption.

Machine Learning (ML) techniques have known an increasing success across several domains, including transportation planning and operations management. They represent a great alternative to the traditional Nested Logit Models (NL) due to the capability of adapting to large data sets, over short or longer periods of time. The discrete mode choice problem is modelled as a classification problem in most ML researches.[6–14] Authors in [8] have proposed four Machine Learning models such as artificial neural net-MLP (ANNet-MLP), ANNet-RBF, multinomial logistic regression (MNL) and support vector machines (SVMs) in order to predict the travel model choice of individuals based on their characteristics, transport mode specifications and data predicting working places and residence. However, the results only predicted the utilisation of cars, public transport in general and a soft mode (walking, biking). The authors in [11] have compared the performance of the MNL model with that of extreme gradient boosting (XGBoost) by using data from a Delaware Valley 2012 regional household travel survey. Authors claim that XGBoost has outperformed the traditional MNL model which is based on the random utility theory, with a main limitation on the strict statistical assumptions. However, while both models have presented competitive results, they were less accurate for smaller transportation modes such as biking (which was less than 1% in the dataset).

Further studies such as the one presented in [15] have applied several ML models (logistic regression, random forest, gradient boosting machines and neural networks) on the greater Melbourne area, based on the Victorian Integrated Survey of Travel and Activity (VISTA); the study compared results against the Veitch Lister Consulting's Zenith model for Victoria which is a nested logit model. However, the models have been applied as classifiers only and on a single mode at a time. More extensively, a detailed study of [16] has studied 86 ML classifiers which led to the selection of ensemble models, including bagging, random forests and deep neural networks (DNN) as performing the best; this work is a good state of art or several ML models that have been deployed so far for modelling the travel choice problem, but still for single mode prediction only. More recent studies have followed in the same style and approach as seen in [17], [18], [19].

The multi-output regression or multi-target regression models (MTR) aim to train a model to predict two or more target variables simultaneously. Significant research studies are from ecology [20, 21] and their potential is currently under-explored for transportation problems. MTR is a structured framework of combining multiple single target regression models (STR). The methods of combining multiple STRs into a MTR can be divided into two types: 1) transform one MTR problem into several STR problems and solve each STR

problem in certain structure (in a parallel structure or a chain structure); and 2) build an algorithm adaptation to predict multiple targets (such as the adaption of single-target random forest to multi-target random forests)[22].

However, to our knowledge, most studies utilized detailed survey data and designed the questionnaires to get the information about all the impacting elements during the decision making process of each individual user. We make connections between the area's attributes and the number of users in each mode instead of the individual's attribute and the individual's choice. Also, majority of studies predict classes of transport modes using classification method, rather than the total number of people using specific transport modes using regression method, which we believe is a more powerful information to have, especially under large scale disruptions or as seen recently under COVID-19 travel restrictions when entire LGAs needed to be isolated from others, or entire transportation modes needed to get reduced in capacity, frequency or in the area coverage.

2.1 Contribution

This paper represents an innovative approach to build an advanced and multi-target regression framework to predict the total number of travellers choosing any available transport mode in large suburban areas. First, we start by evaluating the performance of baseline ML single-target regressors such as the Decision Trees Regression (DTR), the K-Nearest Neighbours Regression (k-NNR), and the Linear Regression (LR). Second, we propose two new structures which can leverage the single-target models to handle the multi-output regression problem which we name the Multi-Output (MO) structure and the Regression Chain (RC) structure. In each structure, we fit the LSVR model and the XGBoost regression model and obtain four hybrid models which we name: the Multi-Output Linear Support Vector Regression model (MO-LSVR), the Multi-Output XGBoost Regression model (MO-XGBoost), the Regression Chain Linear SVR model (RC-LSVR), and the Regression Chain XGBoost regression model (RC-XGboost). To the best of our knowledge, this is among the pioneering works combining advanced ML with traditional mode choice modelling in the transportation field. Overall, the major contribution of this paper are the following:

1. Develop new hybrid data-driven models instead of the traditional Logit model to predict travellers' mode choices from home to work and vice versa. The proposed ML models are capable of running on large data sets with a mixture of multiple transport modes for large population areas.

2. In comparison with the traditional Logit model that can be applied on a single trip at a time, according to the trip attributes, our proposed multi-output regression and regressor-chain frameworks are capable of predicting the number of users/trips across multiple modes in any area in all possible combination scenarios across all available transport modes. The approach can be applied macroscopically across large areas with multiple LGAs and the results contain the number of users for any desired traffic mode.

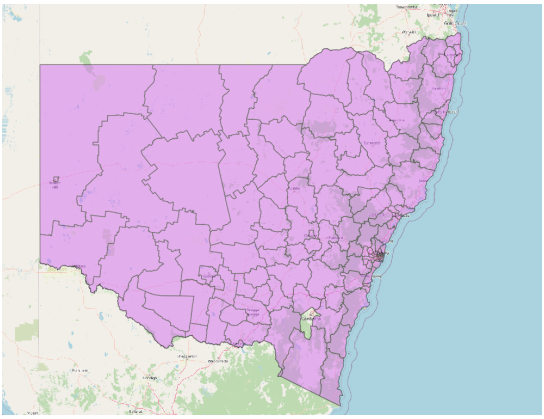

Fig. 1 NSW LGAs case study area.

3 Methodology

Our current methodology is applied over the New South Wales state in Australia, as represented in Fig. 1 which contains 198 LGA (local government areas) extending over a surface of $801,150 km^2$ with a population of 8.166 million people as of September 2020, majority of which is concentrated in the city of Sydney (5.98 million people as of 2020 [23]).

3.1 Problem formulation

The mode choice of people travelling from home to work is modelled traditionally using attraction models. One of the typical model is the Multi-nominal Logit Model which can be expressed in the following formula:

$$p_{in} = \frac{exp(\theta V_{in})}{\sum_{i=1}^{J} exp(\theta V_{in})} \quad (i = 1, 2, ..., J) \tag{1}$$

where, p_{in} is the probability of any option i being selected by a person n from the choice set J. By choice set we refer to all the available transport modes. In our case, we have a selection between: bus, train, walk only, work at home, car as driver, car as passenger, and did not go to work. θ is an unknown coefficient to be calibrated in the data training and V_{in} is called the systematic component of the utility of a transport option i. By definition, we can observe that the summation of probabilities of all options ($i \epsilon [1, J]$) is 1 as shown in Equation 2.

$$\sum_{i=1}^{J} p_{in} = \sum_{i=1}^{J} \frac{exp(\theta V_{in})}{\sum_{i=1}^{J} exp(\theta V_{in})} = \frac{\sum_{i=1}^{J} exp(\theta V_{in})}{\sum_{i=1}^{J} exp(\theta V_{in})} = 1 \tag{2}$$

In this study, we assume that in a certain area (LGA), the probability distribution of choosing any mode ($i\epsilon[1, J]$) is the same for any traveller ($n\epsilon[1, N_{total}]$) which can be expressed as the following equation:

$$Assumption : p_i = p_{in} \quad (\forall\, i\, \epsilon[1, J]\, and\, \forall\, n\, \epsilon[1, N_{total}]) \tag{3}$$

The total number of travellers choosing a certain transport option i (N_i) in a certain area can be expressed by using the following formula:

$$N_i = N_{total} \times p_i \tag{4}$$

where, N_i is the number of people choosing a transport option i, and N_{total} is the total number of people in the LGA area of our study. Therefore, we can further infer the mode choice matrix in a certain area to be as following:

$$\begin{bmatrix} N_1 \\ N_2 \\ N_3 \\ ... \\ N_i \\ ... \\ N_J \end{bmatrix} = N_{total} \times \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ ... \\ p_i \\ ... \\ p_J \end{bmatrix} \tag{5}$$

To summarise, according to Equation 5, the calculated number of travellers choosing each mode ($[N_1, N_2, N_3, ..., N_i, ..., N_J]$) become the prediction target variable of each regression model proposed in this paper. In addition, the second objective of this paper is too predict simultaneously multiple target variables by using the best regression models either trained individually or in hybrid chain structures.

3.2 Machine learning models

3.2.1 Baseline models

Once the data for each LGA has been filtered, cleaned, and selected according to the above features, we further deploy severak basekube regression models. We mainly utilize four regressors that are originally capable of conducting a multi-output regression which are: the Linear Regressor, the K-Neighbours Regressor, the Decision Tree Regressor, and the Random Forest Regressor. The definitions of these 4 regressor are:

The **Linear Regressor** (LR) fits a linear model to minimize the residual sum of squares between the observed targets in the data set, and the targets predicted by the linear approximation. They have been used widely in transport prediction problems across the years[24–27].

The **K-Nearest-Neighbors Regressor** (k-NNR) is one type of neighbors-based regression which can be used where the data labels are continuous and the label assigned to a query point is computed based on the mean

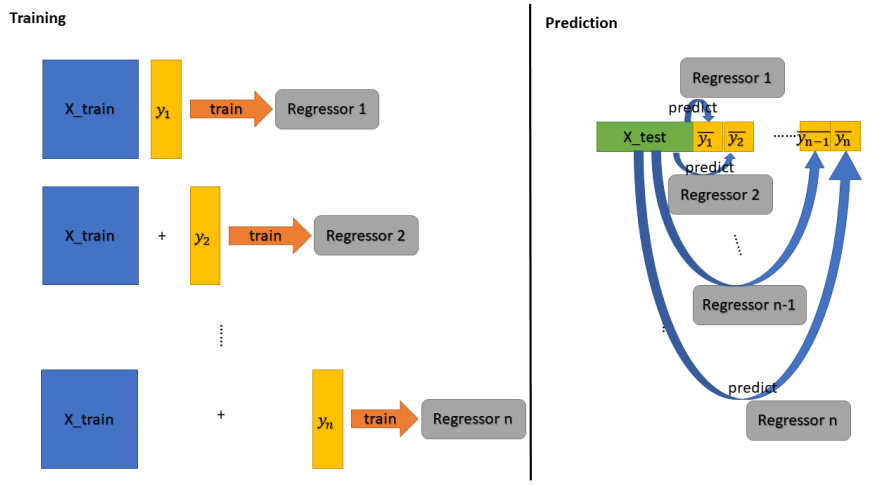

Fig. 2 Proposed multi-output (MO) regression proposed framework for this study.

of the labels of its nearest neighbors. k-NNR implements learning based on the k nearest neighbors of each query point, where k is an integer value specified by the user[28–31].

The **Decision Tree Regressor** (DTR) predicts the matrix of a target variable by learning simple decision rules inferred from the data features[32–35].

The **Random Forest Regressor** (RFR) is a meta estimator that fits a number of decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and to control over-fitting [36–40].

3.2.2 Proposed multi-output and multi-chain framework

The main idea behind our proposed framework is to use the multi-output regression strategy to extend regressors that do not nativity support multi-output regression. This strategy consists of fitting one regressor per target in a first instance. Secondly, multiple regressors are combined to predict the number of people choosing each travel mode in dedicated LGA area where they reside and presumably start the first trip of their regular day (see Fig. 2 in which each y_i represents a travel mode and the prediction is launched on all transport modes based on the training of each regressor). Thirdly, we use this strategy to extend two regressors which are the Linear Support Vector Regressor and the Extreme Gradient Boosting Regressor.

The **Linear Support Vector Regressor** (LSVR) is a subclass of Support Vector Regressor (SVR) which has a linear kernel and it has more flexibility when choosing the penalties and the loss functions. It has good scalability to a large numbers of samples[41–43].

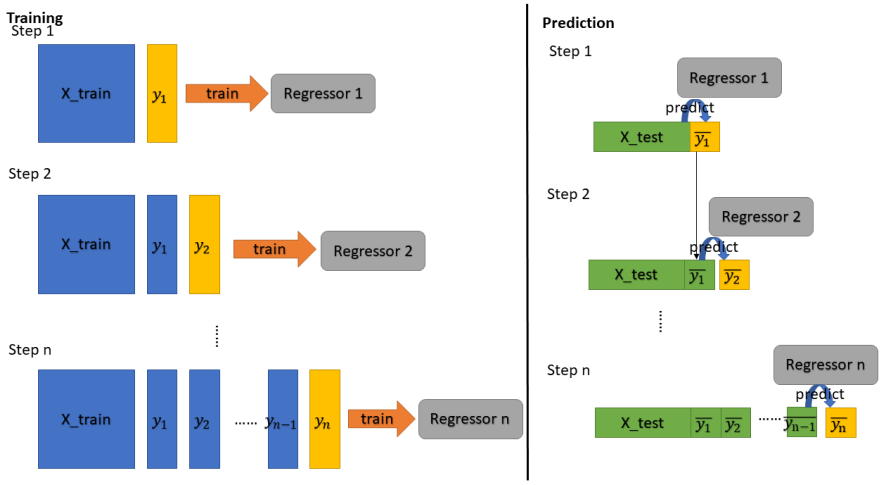

Fig. 3 Proposed regressor chain framework for the multi-modal transport choice prediction.

The **Extreme Gradient Boosting** (XGBoost) is a enhanced version of Gradient Boosted Decision Tree (GBDT) by introducing a regularization parameter in the learning objective function (to control over-fitting); it also introduces a sparsity awareness algorithm for parallel tree learning and has a better support for multi-core processing[44–46].

Lastly, we apply the **Regressor Chain structure** [47] based on Linear SVR and XGBoost in this paper. Regressor chain mechanism makes the prediction in the order specified by the chain using all of the available features provided to the model plus the prediction outcomes of all models that are earlier used in the chain. Fig. 3 shows the detailed steps of the regressor chain mechanism, in which, for example, Regressor 2 for the y_2 mode, incorporates as well in the training the initial transport mode y_1, etc. Similarly, the prediction of y_2 is based on the prediction results $\overline{y}_1$, etc.

3.2.3 Performance metrics

In order to compare the performance of each regression model and evaluate their accuracy, we considered several performance metrics such as: the Root Mean Squared Error (RMSE) and the Mean Absolute Percentage Error (MAPE). RMSE is an estimator which measures the average of squares of the errors and it's calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\bar{y}_i - y_i)^2} \tag{6}$$

where $\bar{y_i}$ is the prediction and y_i is the true value. MAPE is a measure of prediction accuracy of a forecasting method which is indicated below:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} |\frac{y_i - \bar{y_i}}{y_i}| \tag{7}$$

As the outcome of the multi-output regression is a matrix, we convert this matrix to an array in order to easily calculate the performance metrics across all models.

3.2.4 Hyper-parameter tuning

The chosen machine learning algorithms have a set of hyper-parameters, which are parameters related to the internal design of the algorithm that cannot be fitted from the training data. In order to fine tune the dozens of parameters for each regression model that we have been using in our optimization framework, we perform a five-fold cross-validation (5CV) method when deciding the training and the testing data sets. First, we randomly divide our whole data set into five folds which have the same size. Then we choose 4 folds as the training data set and use the remaining 1 fold as the testing data set. We will shuffle the folds five times and each fold serves as a testing data set once. For each regression, we tune the hyper-parameters on each training data set, at each learning fold using various random combinations, evaluated using the 5CV. The final evaluation of the prediction is being done on a hidden testing data set which was not used for model training.

3.3 Data Pre-processing

Normalization is the process of scaling individual samples to have a unit norm. Since the data of our prediction has a big variation (from 0 to about 15,000), we normalize the data set to reduce large variations between each transport mode.

4 Experiments

4.1 Data

In this study, we use the census data from the Australia Bureau of Statistics (ABS). ABS has released population data from 1788 to 2016, but the early years data (From 1788 to 2005) do not contain a detailed investigation for each LGAs. From 2006, ABS has released the detailed statistics for different zoning categories every 5 years. The coverage of the data includes population, education, employment, culture, dwelling, transport, and so on. We use the census data from 2006, 2011, and 2016 as our data source. We use New South Wales (NSW) as our investigated state, and Local Government Areas (LGA) as our zoning category. Although there are minor differences in LGA zoning

Table 1 Categories and features used in the ML training.

Category	Features examples	Number of features
1 year moving history	Number of families in which all residents in the family aged one year and over had a different address one year ago, Number of families which some residents in the family aged one year and over had a different address one year ago.	3
5 year moving history	Number of families in which all residents in the family aged one year and over had a different address five year ago, Number of families in which some residents in the family aged one year and over had a different address one year ago.	3
Education level	Number of people who have finished year 12 or an equivalent education, Number of people who have finished Year 9 or equivalent, Number of people who have finished Year 8 or below	6
Income (weekly)	Number of people who have Nil income, Number of people who have $1-$149 weekly income, Number of people who have $400-$599 weekly income, Number of people who have $600-$799 weekly income,	5
No. of children	Number of families which have no children, Number of families which have one children, Number of families which have owo children,	8
Vehicle ownership	Number of families which have no vehicle, Number of families which have two motor vehicles	5
Population	Male, Female, Total	3
Area and density	LGA area, LGA population density	2
Census year	Year of the census	1
Travel mode to work	Number of people who did not go to work, number of people who worked at home, number of people who take the bus to work	17
Place of work	Number of people who go to Waverley to work, etc.	108
Total number of features		161

among 2006, 2011, and 2016, we manage to obtain 108 LGAs and the corresponding data. The reason why we only obtain 108 LGAs is that each census has some modifications of the LGA boundaries and only 108 LGAs are in common among these three censuses. We find out that the census content among 2006, 2011 and 2016 also has some difference, therefore we select the common categories among these three years, which include population, education level, income, vehicle ownership, number of children, 1 year moving history, and 5 year moving history. For each categories, there are certain grouping criteria to divide the total population into subgroups which are presented in Table 1. We call these grouping criteria as the "features" of our prediction problem. In total we collected 161 features from the ABS census data for each LGA across the years, out of which 17 feature describe the available transport modes in each LGA.

4.2 Experiment design

We use the the 17 variables in the category "travel mode to work" from Table 1 as our prediction targets (Y matrix) and the rest of the 144 features as our given known feature matrix (X matrix). Starting with the easiest prediction problem (predicting the number of people adopting only one transport mode), we first run parameter tuning to predict a single target among all the prediction targets. Table 2 shows the detailed parameters and values for each regressor used in our paper. Both the multi-output regression strategy and the chain regressor mechanism inherit the parameters of their base estimators.

Table 2 Parameters for each baseline regressor

Regressor Name	**Parameter name**	**Typical values**
LR	fit_intercept	True , False
LR	normalize	True , False
k-NNR	weights	uniform, distance
k-NNR	algorithm	auto, ball_tree, kd_tree, brute
k-NNR	leaf_size	3, 6, 9, 12, 15, 18, 21, 24, 27, 30
k-NNR	p	1,2
DTR and RFR	criterion	mse, friedman_mse, mae
DTR and RFR	max_depth	3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32
DTR and RFR	min_samples_split	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1
DTR and RFR	min_samples_leaf	0.1, 0.2, 0.3, 0.4, 0.5
LSVR	estimator__loss	epsilon_insensitive, squared_epsilon_insensitive
LSVR	estimator__max_iter	1000,2000,3000,4000,5000
XGBoost	estimator__booster	gbtree
XGBoost	estimator__learning_rate	0.1,0.2,0.3,0.4,0.5
XGBoost	estimator__max_depth	3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32
XGBoost	estimator__min_samples_split	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1
XGBoost	estimator__min_samples_leaf	0.1,0.2,0.3,0.4,0.5

As a selection, we choose “Did not go to work”, “Car, as driver”, “Walked only”, “Car, as passenger”, “Worked at home”, “Train”, and “Bus” as our prediction targets. In total we have selected 7 predictors out of 17 travel modes because the rest of travel modes do not contain a significant number of people (less than 10% of total population). When training the models, we hyper-tune each regressor at this step. We then use the best parameters which were found in this step for the cross validation and testing. After predicting one transport mode at a time, we then predict simultaneously multiple scenarios with combinations of the selected predictors and gradually increase the number of prediction targets from 2 to 7. Table 3 shows the detailed scenarios with all the combinations of prediction targets by the given prediction length.

5 Results

5.1 Predicting a single transport mode choice

Fig. 4 and Fig. 5 showcase the performance comparison of all the regressors when predicting one travel mode at a time in terms of MAPE and RMSE. We observe that most regressors have a MAPE below 50% except for LR, and a large part of them below 10%, indicating a very good prediction accuracy. “Train” users seem to be the most difficult to predict and contain the largest MAPE (around 40%) among all regressors. “Bus” users are the second most difficult to predict and contain the second largest MAPE (around 10%) among most regressors. When predicting “Train” and “Bus” users, LR has significant worse results than the rests. This confirms that the “Train” and “Bus” users

Table 3 Detailed prediction targets combinations in each scenario in our experiments

Number of targets per prediction	Detailed prediction targets combinations	Number of prediction runs in the scenario
1	[Did not go to work] [Car, as driver] [Walked only] [Car, as passenger] [Worked at home] [Train] [Bus]	7
2	[Did not go to work , Car, as driver] [Did not go to work , Walked only] [Did not go to work , Car, as passenger] [Did not go to work , Worked at home] [Did not go to work , Train] [Did not go to work , Bus] [Car, as driver , Walked only] [Car, as driver , Car, as passenger]	21
3	[Did not go to work , Car, as driver , Walked only] [Did not go to work , Car, as driver , Car, as passenger] [Did not go to work , Car, as driver , Worked at home] [Did not go to work , Car, as driver , Train]	35
4	[Did not go to work , Car, as driver , Walked only , Car, as passenger] [Did not go to work , Car, as driver , Walked only , Worked at home] [Did not go to work , Car, as driver , Walked only , Train] [Did not go to work , Car, as driver , Walked only , Bus] [Did not go to work , Car, as driver , Car, as passenger , Worked at home] [Did not go to work , Car, as driver , Car, as passenger , Train]	35
5	[Did not go to work , Car, as driver , Walked only , Car, as passenger , Worked at home] [Did not go to work , Car, as driver , Walked only , Car, as passenger , Train] [Did not go to work , Car, as driver , Walked only , Car, as passenger , Bus]	21
6	[Did not go to work , Car, as driver , Walked only , Car, as passenger , Worked at home , Train] [Did not go to work , Car, as driver , Walked only , Car, as passenger , Worked at home , Bus]	7
7	[Did not go to work , Car, as driver , Walked only , Car, as passenger , Worked at home , Train , Bus]	1

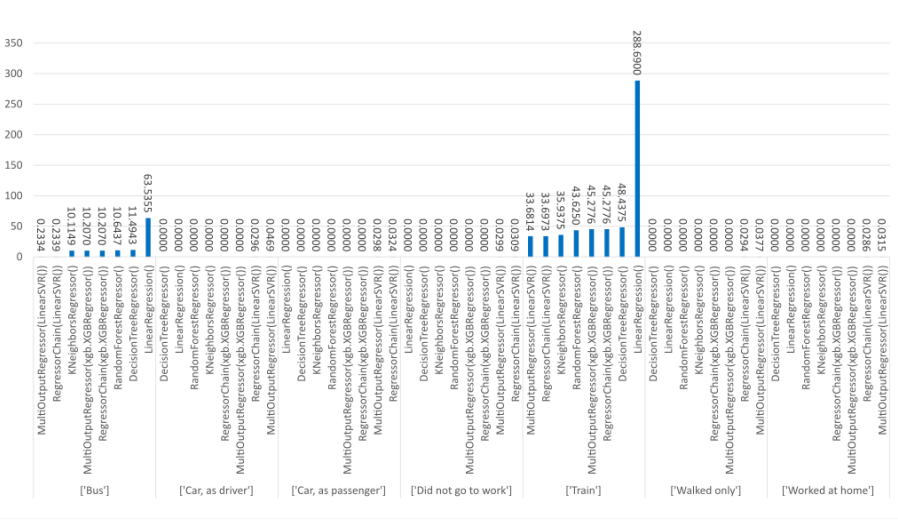

Fig. 4 MAPE in predicting one mode (unit:%)

are not linearly related to the given features. We do make the observation because that a 30-40% MAPE for prediction problems in transportation is often considered as acceptable why MAPE values below 10% are considered

to be excellent[48–51]. The results of best MAPE (33.68% when predicting "Train" and 0.23% when predicting "Bus") for the proposed MO-LSVR is outperforming the rest of the regressors, making it a great performer even for the hardest transport modes to predict "Train" and "Bus".

All the other modes such as "Car, as driver", "Car, as passenger", "Did not go to work", "Walked only" and "Worked at home" achieved a very small MAPE (below 0.04%) and RMSE (below 0.0002) which are excellent results for a transport mode prediction problem, mostly affected by large errors or noise in the dataset. This also indicates that the input features can severely impact the prediction outcome on these transport modes.

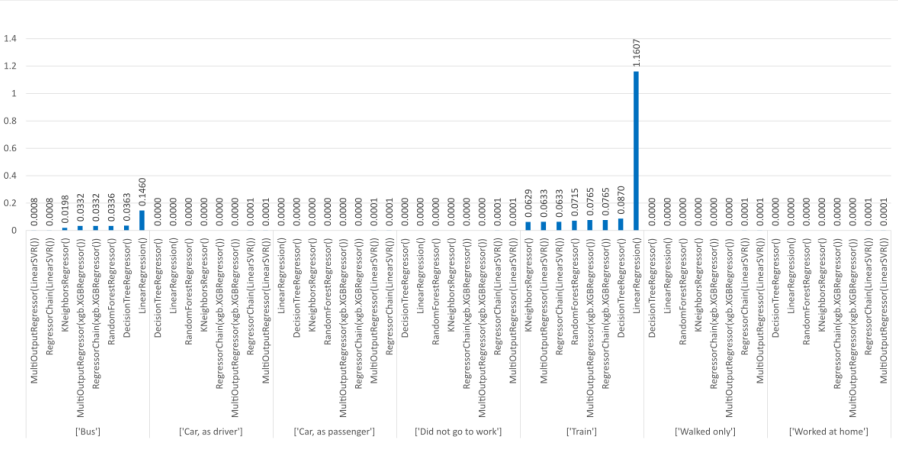

Fig. 5 RMSE in predicting one mode

Same trends are found in the RMSE results for all modes when most regressors have a RMSE below 0.05 except for LR. Besides, "Train" users are the most difficult to predict and contains the largest RMSE (around 0.07) among most regressors, while "Bus" users are the second most difficult to predict with an RMSE of around 0.04. Results which are however excellent for these two modes according to normal error levels found in the related literature [48–51]. When predicting "Train" and "Bus" users respectively, LR has significant worse results than the rests, which confirms that "Train" and "Bus" users are not linearly related to the given features; another explanation is given by the low percentage of public transport users in New South Wales Australia, falling below 7% according to recent studies.

5.2 Predict two or more modes at a time

After increasing the number of output targets/modes, the prediction accuracy drops as expected, but some of the combinations still achieved high accuracy. As detailed in the following, MAPE and RMSE results for each combination

of two modes are shown in the Fig. 6 and Fig. 7. It is obvious that, if the combinations contain either "Bus" or "Train", the MAPE and RMSE results will increase largely. All the combinations without "Bus" or "Train" have MAPEs below 22% (when predicting "Train" and "Bus" at a time), but when predicting combinations which contain "Bus" or "Train", the MAPEs increase to 167% with the largest MAPE reaching 6531.08% (when predicting "Car, as driver" and "Train" at a time). Same trends can be observed from RMSE. All the combinations without "Bus" or "Train" have MAPEs below 0.013, but when predicting combinations which contain "Bus" or "Train", the MAPEs increase to 0.0486 (when predicting "Walked only" and "Train" at a time) with the largest MAPE reaching 0.0822 (when predicting "Train" and "Bus" at a time). The best result appears to be obtained when predicting ["car, as driver", "Did not go to work"] with MAPE=6.1 and RMSE =0.01. Same trends can be observed from the combinations of 3,4,5,6, and 7 prediction targts at a time.

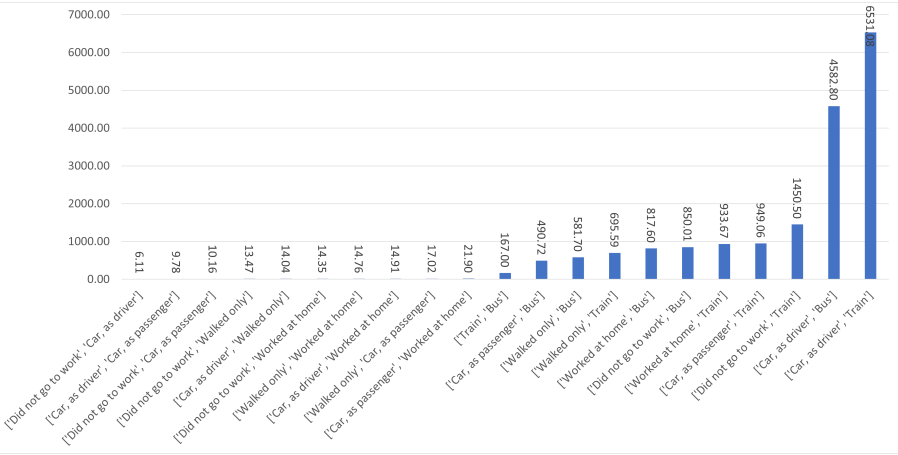

Fig. 6 Mean MAPE among all regressors in predicting two modes

5.3 Comparison of regressors for multi-target prediction problems

An obvious trend is that the baseline LR is worse than the other regressors as expected. By comparing the MAPE, the results for the RC-LSVR is the lowest (36.45%) which makes it almost 45.66% better than the worst LR model, while the other regressors are not far behind (around 37.00%). When analysing the RMSE results, the RFR seems to be the lowest (0.00377) which is almost 80.80% better than the worst LR model, while the other regressors are not far behind (around 0.004). In the following, we show the details of the average MAPE and RMSE among all regressors in predicting combinations not including "Train" or "Bus" in Fig. 8 and Fig. 9.

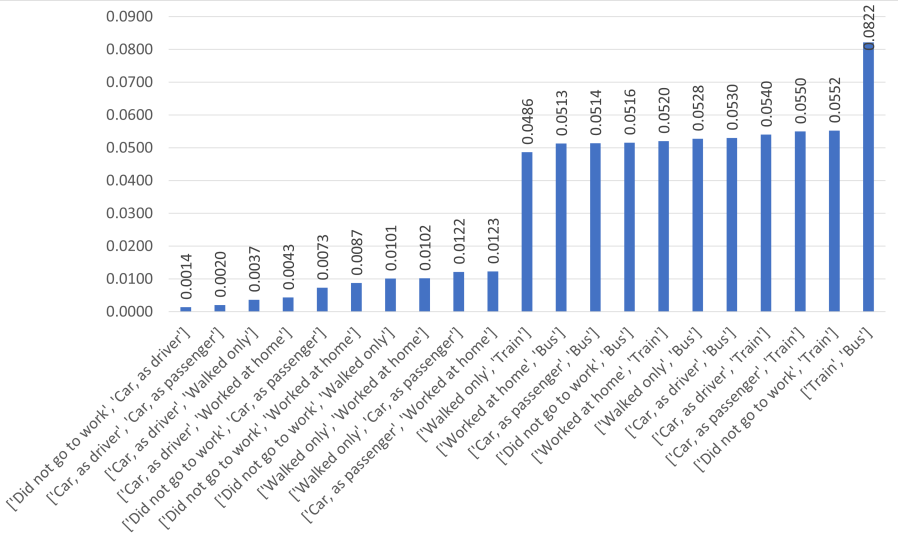

Fig. 7 Mean RMSE among all regressors in predicting two modes (%)

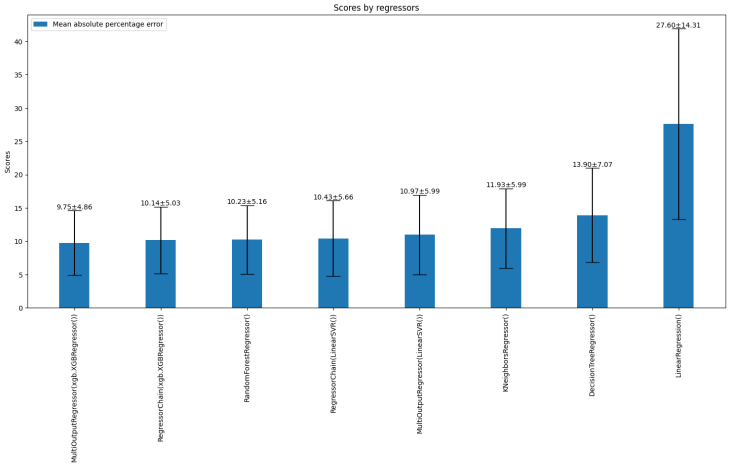

Fig. 8 MAPE comparison between single output baseline ML models versus our proposed Multi-Output-Regressor framework (MO) and Regressor-Chain framework (RC) in predicting all possible combinations not including “Train” or “Bus”.

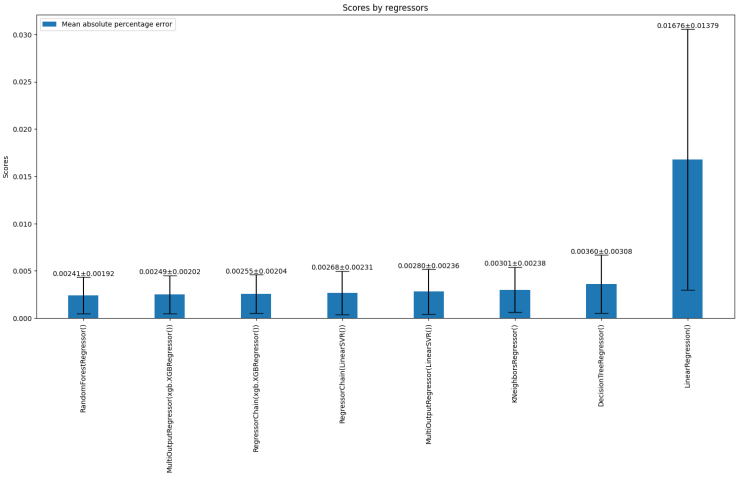

Fig. 9 RMSE comparison between single output baseline ML models versus our proposed Multi-Output-Regressor framework (MO) and Regressor-Chain framework (RC) in predicting all possible combinations not including "Train" or "Bus".

5.4 Comparison between the MO and the RC structure

According to MAPE, there is no significant trend between MO and RC structure. The RC-LSVR is 5.21% better than the MO-LSVR in MAPE(see Fig. 8 where the MAPE for RC-LSVR is 10.23% and for MO-LSVR is 10.97%), but the RC-XGBoost (MAPE=10.14%) is 3.82% worse than the MO-XGBoost (MAPE=9.75%) in MAPE. According to the RMSE, there is no significant trend between the MO and the RC structure. The RC-XGBoost (RMSE=0.00255) is 2.35% worse than the MO-XGBoost (RMSE=0.00249)in RMSE, but the RC-LSVR (RMSE=0.00268) is 4.28% better than the MO-LSVR (RMSE=0.0028) in RMSE.

The results of all performance evaluations seem to indicate the performance of the MO and RC framework is very similar, while in this case MO-XGBoost is the best model according to both RMSE and MAPE. The reason why the MO structure is slightly better than the RC structure is the accumulation of errors from previous regression results in the training of a new regression model.

5.5 Comparison between the MAPE and the RMSE under different numbers of prediction targets

Figure 10 shows the detailed RMSE vs. MAPE in predicting all combinations not including "Train" or "Bus", and the points are colored by the number of prediction targets. In this figure, each point presents the outcome of a certain

regressor (such as RF, LR, MO-LSVR, and so on) when predicting a combination of predicted targets/modes (such as [Car as a Driver, Walk only] and [Work at home, Walk only, Car as passenger]). As we can see, when predicting 1 mode, both RMSE and MAPE are very small which is the best performance because the prediction tasks are the simplest. When predicting 2,3,4 modes at a time, there are two split bands of points and there is a trend that RMSE will decrease by increasing the prediction length while the MAPE will increase by increasing the number of prediction targets.

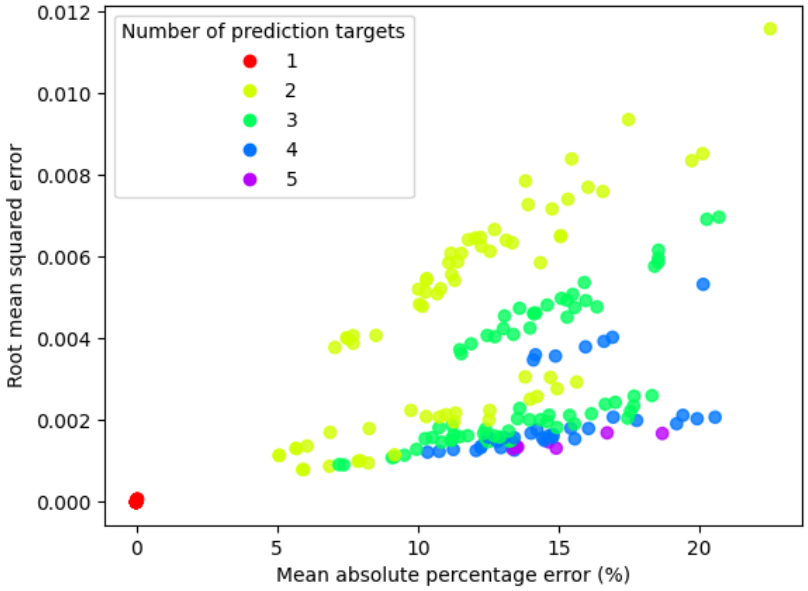

Fig. 10 Detailed RMSE and MAPE in predicting combinations not including "Train" or "Bus", colored by the prediction feature length.

6 Conclusions

Contribution summary: In this paper, we proposed a new multi-chain and multi-output machine learning predicting framework, to predict the total number of travellers choosing any transport modes to commute to work on a daily basis. The application domain is the New South Wales, Australia, leveraged by 10 years worth of data regarding traveller behaviour in the state. We proposed several regressors that originally have the capability of performing multi-output regression, as well as embed those regressor that are not compatible with multi-output regression in both multi-output regression strategy and the chain regression mechanism. The main outcomes are excellent RMSE and

MAPE results (almost close to zero) for either single modes and multi-modal prediction problems. The most problematic transport modes to be predicted are trains and buses due to the lower range of travellers choosing these modes in Australia, which is mostly a car-focused country. However, our proposed modelling has achieve a MAPE of 9.7% and RMSE of 0.0024% which are usually very good results in transportation modelling. These results also outperform significantly the results of baseline ML models such as: Linear Regression, K.NN, DTs and RF regressors.

Benefits of this work: By knowing the predicted number of travellers on all modes in the city, traffic managers can make better decisions of line closures, incident management and disruption planning such as in the COVID-19 case, where entire LGAs had to be cut-off from their daily servicing due to high numbers of cases. Patronage prediction is also a very important topic for management centres which can help them to better plan and manage resources.

Limitation and future direction: Given the limited data set for NSW among all transport modes, a future direction is to conduct the analysis for more concentrated areas in large metropolitan areas where bus and train patronage is much higher than in rural NSW, for example. This can help the ML framework to learn better the patterns of travellers between any areas and provide better estimates of public transport patronage.

7 Acknowledgement

I'd like to express my deepest thanks to Dr. Seunghyeon Lee and Dr. Yuming Ou for your discussion, advice and suggestion during the experiment design as well as conduction. Thank you for sharing your experience and appreciate your support and help in reviewing and revising this paper. Especially, I would like to thank Dr. Adriana-Simona Mihăiţă for the guidance and suggestion through the Machine Learning experiments. Thank you for your great help for revising this paper.

References

[1] de Dios Ortúzar, J., Willumsen, L.G.: Modelling Transport. John wiley & sons, ??? (2011)

[2] Ben-Akiva, M.E., Lerman, S.R., Lerman, S.R.: Discrete Choice Analysis: Theory and Application to Travel Demand. Cambridge, MIT Press, ??? (1985)

[3] Hensher, D.A., Rose, J.M., Rose, J.M., Greene, W.H.: Applied Choice Analysis: a Primer. Cambridge university press, ??? (2005)

[4] Wen, C.-H., Koppelman, F.S.: The generalized nested logit model. Transportation Research Part B: Methodological **35**(7), 627–641 (2001)

[5] Hensher, D.A., Greene, W.H.: The mixed logit model: the state of practice. Transportation **30**(2), 133–176 (2003)

[6] Xie, C., Lu, J., Parkany, E.: Work travel mode choice modeling with data mining: decision trees and neural networks. Transportation Research Record **1854**(1), 50–61 (2003)

[7] Zhang, Y., Xie, Y.: Travel mode choice modeling with support vector machines. Transportation Research Record **2076**(1), 141–150 (2008)

[8] Omrani, H.: Predicting travel mode of individuals by machine learning. Transportation Research Procedia **10**, 840–849 (2015). https://doi.org/10.1016/j.trpro.2015.09.037. 18th Euro Working Group on Transportation, EWGT 2015, 14-16 July 2015, Delft, The Netherlands

[9] Hagenauer, J., Helbich, M.: A comparative study of machine learning classifiers for modeling travel mode choice. Expert Systems with Applications **78**, 273–282 (2017)

[10] Golshani, N., Shabanpour, R., Mahmoudifard, S.M., Derrible, S., Mohammadian, A.: Modeling travel mode and timing decisions: Comparison of artificial neural networks and copula-based joint model. Travel Behaviour and Society **10**, 21–32 (2018)

[11] Wang, F., Ross, C.L.: Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model. Transportation Research Record **2672**(47), 35–45 (2018)

[12] Wong, M., Farooq, B., Bilodeau, G.-A.: Discriminative conditional restricted boltzmann machine for discrete choice and latent variable modelling. Journal of choice modelling **29**, 152–168 (2018)

[13] Lhéritier, A., Bocamazo, M., Delahaye, T., Acuna-Agost, R.: Airline itinerary choice modeling using machine learning. Journal of choice modelling **31**, 198–209 (2019)

[14] Rasouli, S., Timmermans, H.J.: Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. European Journal of Transport and Infrastructure Research **14**(4) (2014)

[15] Richards, M., Zill, J.: Modelling mode choice with machine learning algorithms. 41st Australasian Transport Research Forum (ATRF) (2019)

[16] Wang, S., Mo, B., Zhao, J.: Predicting travel mode choice with 86 machine learning 1 classifiers : An empirical benchmark study. (2019)

[17] Assi, K., Shafiullah, M., Nahiduzzaman, K.M., Mansoor, U.: Travel-to-school mode choice modelling employing artificial intelligence techniques: A comparative study. Sustainability **11** (2019). https://doi.org/10.3390/su11164484

[18] Kim, E.-J.: Analysis of travel mode choice in seoul using an interpretable machine learning approach. Journal of Advanced Transportation **2021**, 1–13 (2021). https://doi.org/10.1155/2021/6685004

[19] Buijs, R., Koch, T., Dugundji, E.: Using neural nets to predict transportation mode choice: Amsterdam network change analysis. J Ambient Intell Human Comput **12**, 121–135 (2021). https://doi.org/10.1007/s12652-020-02855-6

[20] Džeroski, S., Demšar, D., Grbović, J.: Predicting chemical parameters of river water quality from bioindicator data. Applied Intelligence **13**(1), 7–17 (2000)

[21] Kocev, D., Džeroski, S., White, M.D., Newell, G.R., Griffioen, P.: Using single-and multi-target regression trees and ensembles to model a compound index of vegetation condition. Ecological Modelling **220**(8), 1159–1168 (2009)

[22] Borchani, H., Varando, G., Bielza, C., Larranaga, P.: A survey on multi-output regression. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **5**(5), 216–233 (2015)

[23] Australia, P.: Sydney Population 2021 (2021). https://www.population.net.au/sydney-population/

[24] Aalen, O.O.: A linear regression model for the analysis of life times. Statistics in medicine **8**(8), 907–925 (1989)

[25] Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to Linear Regression Analysis. John Wiley & Sons, ??? (2021)

[26] Mao, T., Mihăită, A.-S., Chen, F., Vu, H.L.: Boosted genetic algorithm using machine learning for traffic control optimization. IEEE Transactions on Intelligent Transportation Systems (2021)

[27] Goldberger, A.S.: Best linear unbiased prediction in the generalized linear regression model. Journal of the American Statistical Association **57**(298), 369–375 (1962)

[28] Devroye, L.: The uniform convergence of nearest neighbor regression function estimators and their application in optimization. IEEE Transactions on Information Theory **24**(2), 142–151 (1978)

[29] Devroye, L., Gyorfi, L., Krzyzak, A., Lugosi, G.: On the strong universal consistency of nearest neighbor regression function estimates. The Annals of Statistics, 1371–1385 (1994)

[30] Maltamo, M., Kangas, A.: Methods based on k-nearest neighbor regression in the prediction of basal area diameter distribution. Canadian Journal of Forest Research **28**(8), 1107–1115 (1998)

[31] Song, Y., Liang, J., Lu, J., Zhao, X.: An efficient instance selection algorithm for k nearest neighbor regression. Neurocomputing **251**, 26–34 (2017)

[32] Czajkowski, M., Kretowski, M.: The role of decision tree representation in regression problems – an evolutionary perspective. Applied Soft Computing **48**, 458–475 (2016). https://doi.org/10.1016/j.asoc.2016.07.007

[33] Loh, W.-Y.: Fifty years of classification and regression trees. International Statistical Review **82**(3), 329–348 (2014)

[34] Ortuño, F.M., Valenzuela, O., Prieto, B., Saez-Lara, M.J., Torres, C., Pomares, H., Rojas, I.: Comparing different machine learning and mathematical regression models to evaluate multiple sequence alignments. Neurocomputing **164**, 123–136 (2015)

[35] Ou, Y., Mihaita, A.-S., Chen, F.: Big data processing and analysis on the impact of covid-19 on public transport delay. Data Science for COVID-19: Volume 2: Societal and Medical Perspectives, 257 (2021)

[36] Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P.: Random forest: a classification and regression tool for compound classification and qsar modeling. Journal of chemical information and computer sciences **43**(6), 1947–1958 (2003)

[37] Segal, M.R.: Machine learning benchmarks and random forest regression (2004)

[38] Cootes, T.F., Ionita, M.C., Lindner, C., Sauer, P.: Robust and accurate shape model fitting using random forest regression voting. In: European Conference on Computer Vision, pp. 278–291 (2012). Springer

[39] Grömping, U.: Variable importance assessment in regression: linear regression versus random forest. The American Statistician **63**(4), 308–319 (2009)

[40] Shafiei, S., Mihăiţă, A.-S., Nguyen, H., Cai, C.: Integrating data-driven and simulation models to predict traffic state affected by road incidents.

Transportation Letters, 1–11 (2021)

[41] Trafalis, T.B., Ince, H.: Support vector machine for regression and applications to financial forecasting. In: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, vol. 6, pp. 348–353 (2000). IEEE

[42] Meyer, D., Leisch, F., Hornik, K.: The support vector machine under test. Neurocomputing **55**(1-2), 169–186 (2003)

[43] Zhang, L., Zhou, W., Jiao, L.: Wavelet support vector machine. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **34**(1), 34–39 (2004)

[44] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., *et al.*: Xgboost: extreme gradient boosting. R package version 0.4-2 **1**(4), 1–4 (2015)

[45] Shaffiei, S., Mihăiţă, A., Cai, C.: Demand estimation and prediction for short-term traffic forecasting in existence of non-recurrent incidents. In: ITS World Congress 2019 (ITSWC2019), Singapore (2019)

[46] Sheridan, R.P., Wang, W.M., Liaw, A., Ma, J., Gifford, E.M.: Extreme gradient boosting as a method for quantitative structure–activity relationships. Journal of chemical information and modeling **56**(12), 2353–2360 (2016)

[47] Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W., Vlahavas, I.: Multi-target regression via input space expansion: treating targets as inputs. Machine Learning **104**(1), 55–98 (2016)

[48] Wen, T., Mihăiţă, A.-S., Nguyen, H., Cai, C., Chen, F.: Integrated incident decision-support using traffic simulation and data-driven models. Transportation research record **2672**(42), 247–256 (2018)

[49] Mihaita, A.-S., Li, H., He, Z., Rizoiu, M.-A.: Motorway traffic flow prediction using advanced deep learning. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 1683–1690 (2019). IEEE

[50] Shafiei, S., Mihaita, A., Nguyen, H., Bentley, C., Cai, C.: Short-term traffic prediction under non-recurrent incident conditions integrating data-driven models and traffic simulation. In: Transportation Research Board 99th Annual Meeting (2020)

[51] Mihaita, A.-S., Li, H., Rizoiu, M.-A.: Traffic congestion anomaly detection and prediction using deep learning. arXiv preprint arXiv:2006.13215

(2020)