TRANSPORTATION LETTERS

Revisiting Kernel Logistic Regression under the Random Utility Models perspective. An Interpretable Machine Learning Approach

José Ángel Martín-Baos[a,b] , Ricardo García-Ródenas[a,b] and Luis Rodriguez-Benitez[c]

[a]Departamento de Matemáticas, Escuela Superior de Informática. University of Castilla-La Mancha, Spain.
[b]Instituto de Matemática Aplicada a la Ciencia y la Ingeniería (IMACI), University of Castilla-La Mancha.
[c]Departamento de Tecnologías y Sistemas Informáticos, Escuela Superior de Informática. University of Castilla-La Mancha, Spain.

ABSTRACT

The success of machine-learning methods is spreading their use to many different fields. This paper analyses one of these methods, the Kernel Logistic Regression (KLR), from the point of view of Random Utility Model (RUM) and proposes the use of the KLR to specify the utilities in RUM, freeing the modeller from the need to postulate a functional relation between the features. A Monte Carlo simulation study is conducted to empirically compare KLR with the Multinomial Logit (MNL) method, the Support Vector Machine (SVM) and the Random Forests (RF). We have shown that, using simulated data, KLR is the only method that achieves maximum accuracy and leads to an unbiased willingness-to-pay estimator for non-linear phenomena. In a real travel mode choice problem, RF achieved the highest predictive accuracy, followed by KLR. However, KLR allows for the calculation of indicators such as the value of time, which is of great importance in the context of transportation.

KEYWORDS
Random Utility Models; Kernel Logistic Regression; Machine Learning; Willingness to Pay; Value of Time

1. Introduction

Nowadays, Artificial Intelligence (AI) has gained great popularity due to its success in applications such as autonomous vehicles, intelligent robots, image and voice recognition, automatic translation, etc. The construction of these intelligent machines is mainly based on Machine Learning (ML) methods, which has led to an increased use of these methods and a growing interest in expanding the domain of applications where ML methods are applied, such as in the field of transport modelling.

Corresponding author: José Ángel Martín-Baos (JoseAngel.Martin@uclm.es)

Traditionally, travel behaviour research has been primarily supported by discrete choice models, which describe how a rational decision-maker chooses an alternative from among a set of choices depending on the attributes of each one of the alternatives and the characteristics of the individual (Ben-Akiva and Bierlaire, 1999a; McFadden, 1978; Train, 2009). Random Utility Model (RUM) assume that in the decision process there are some latent (unobservable) functions, called *utility functions*, which measure the interest of each alternative to an individual. The decision-maker is assumed to be rational, i.e. they choose the alternative that maximises their utility.

In RUM the utility of each alternative is the sum of two terms, one deterministic and one stochastic. With respect to the stochastic part, the probability distribution determines the resulting model. The most widespread example is the Multinomial Logit (MNL) model, which is obtained when a independent and identically distributed (i.i.d.) Gumbel distribution is assumed for the stochastic term. These models are estimated using a maximum likelihood estimation methodology, which allows an asymptotic distribution of the estimators to be determined and makes it possible to test hypotheses with respect to the values of the parameters. Linear and non-linear functions in the parameters can be used to define the deterministic part. Many authors have already pointed out that non-linear utility functions are more suitable for some applications, such as departure time choice. However, there are two main limitations when it comes to non-linear utility functions: firstly, choosing the right function; innumerable functions have been proposed and each of them gives a very different result; and secondly, calibrating the parameters is also problematic. As a result, linear functions are often used by both researchers and practitioners in preference to non-linear ones.

Nowadays, the scientific community is evaluating the classical RUM with the new proposals based on ML. Preliminary results have revealed a significantly higher predictive capacity of ML methods, such as Hagenauer and Helbich (2017); Wang et al. (2019b); Cheng et al. (2019), which encourages further investigation. However, there are also studies with apparently contradictory results that can be explained by the differences in the adjustments of the hyperparameters, or in the differences between the problems themselves. As regards the disadvantages of introducing these new methods, the difficulty or impossibility of obtaining econometric information with these new proposals is highlighted. These two points motivated us to analyse the Kernel Logistic Regression (KLR) technique, whose application in the field of transport is innovative. This method can be considered as the introduction of choice probabilities into the Support Vector Machine (SVM), a method that shares a prominent place in the ML community along with Deep Neural Networks and ensemble methods. In this paper the KLR technique is reinterpreted within the RUM framework, in order to guarantee its interpretability and the calculation of econometric indices. Then, a simulation study is conducted, which allows the evaluation of the predictive capacity of the most promising ML proposals with respect to the maximum predictive capacity that can be obtained from the simulated data and the effect of compliance with the assumptions used in deriving RUM on the performance of ML methods.

This paper is organised as follows. First, Section 2 describes the increasing interest in finding alternative methods to RUM in modelling individual behaviour, focusing our attention on ML techniques and the interpretability of those methods. Section 3 introduces RUM and KLR methodologies in detail. Next, Section 4 sets out how to compute interpretable economic information on both RUM and KLR models. Then, Section 5 presents two computational experiments: a Monte Carlo simulation study and a real-world travel mode choice problem. Finally, Section 6 presents the main conclusions.

2. Literature Review

Discrete choice methods based on RUM (McFadden, 1978; Ben-Akiva and Lerman, 1985; Ben-Akiva and Bierlaire, 1999a; Train, 2009) have been developed over the last four decades, and they have now acquired a high degree of sophistication. These models have dominated the analysis of travel behaviour since their formulation. Rasouli and Timmermans (2012) review some prior works on RUM applied to travel demand forecasting and identify gaps in the literature and future research.

Nowadays, there is increasing interest in alternative methods to RUM with the aim of making them more flexible and adaptable to different applications. Discrete choice modelling can also be considered as a classification problem, in the sense that the output is a categorical variable, i.e. the choice; therefore, ML methods can provide an alternative to the traditional RUM. Over the last few years, ML methods have been successfully applied to a wide variety of fields. Fernández-Delgado et al. (2014) conducted an extensive study with 179 ML classifiers over a total of 121 datasets from various domains and found Random Forests (RF) to be the best classifiers. Nonetheless, Wainberg et al. (2016) nuance their results due to certain methodological issues, and find that RF does not have significantly higher accuracy than SVM and neural networks.

In the transport field, a large number of recent studies suggest that individual travel behaviour can be accurately estimated using ML classifiers. In addition, these studies generally affirm that ML classifiers outperform the traditional RUM, such as the MNL. The work of Hagenauer and Helbich (2017) assesses seven automatic-learning methods applied to a transport-choice study on a sample of daily trips. These authors found that ML methods achieve better results than the MNL model, and that the RF technique could be the most promising method. Recent studies of Cheng et al. (2019) and Lhéritier et al. (2019) found that the RF method outperforms the standard and the latent class MNL model in terms of accuracy and computational time when they are applied to a travel mode choice and an airline itinerary choice problem, respectively. Wang and Ross (2018) reported that extreme gradient boosting models have higher prediction accuracy than the MNL. Lindner et al. (2017) show that artificial neural networks and classification trees provides good estimations for travel mode choice problems. Moreover, a recent work of Cheng et al. (2019) applies an ensemble-based MNL model obtaining better prediction results than MNL and the possibility to deal with high dimensional data efficiently.

One of the first studies to combine machine-learning techniques and MNL in modelling transport was Celikoglu (2006), which specifies the utilities of the MNL using neural networks that were based on radial basis functions (RBFNN) and generalised regression (GRNN). The author argues that this hybrid model improves the performance of a basic MNL model, whereas a basic artificial neural network does not exceed the performance of the MNL.

RUM allow a set of measures such as Willingness To Pay (WTP), Value of Time (VOT), elasticities, market shares, etc. to be obtained, which allow the result of any intervention in the transport system to be assessed. Several proposals have been presented in the literature on how to calculate these indices using ML methods. They are based on the numerical approximation of probability derivatives with respect to the feature vector. Zhao et al. (2020) use this technique to estimate marginal effects and arc elasticities. Wang et al. (2019a) use the knowledge of RUM to design a particular deep neural network architecture with alternative-specific utility functions. This model exhibits some computational difficulties in calculating the probabilities of the alternatives and, therefore, in calculating the indicators as well. These problems stem from the large variability of individual models due to optimisation difficulties in the minimisation of the non-convex risk function of deep neural networks.

In this paper we are interested in ML methods with some mechanisms that can be analogous to the utility functions defined in RUM. The combination of MNL with radial basis functions is known in the ML community as Kernel Logistic Regression (KLR) (Zhu and Hastie, 2005; Cawley and Talbot, 2005; Maalouf and Trafalis, 2011; Liu et al., 2016; Ouyed and Allili, 2018; Martín-Baos et al., 2020). The motivation underlying the use of this method, as opposed to other prominent ML methods such as RF, is that it allows for the identification of utility functions. In KLR the parameter estimation problem is based on a penalised maximum likelihood estimation in which the goodness of fit criterion weighs the empirical risk and its complexity. To operate with KLR models it is sufficient to choose one of the so-called *kernel functions*. These models have also been derived from Gaussian Process (GP) and, in certain areas, they are known as kernel-based MNL approaches, which allude to a common way to define the *non-parametric* utilities based on kernels. The non-linear MNL models require the modeller to specify a functional expression based on the attribute set and a parameter vector (parametric utilities), while the kernel-based MNL addresses non-linearity by using semi- or non-parametric utility specifications that do not require a priori assumptions on the functional form of the logit link.

Few papers have considered the use of non-parametric utilities. One of the studies that can be considered seminal is Abe (1999), which introduced a spline-based utility specification in a semi-parametric MNL. Other types of radial basis functions have been used, such as penalised B-spline functions (Kneib et al., 2007) or cubic spline functions (Fukuda and Yai, 2010). Espinosa-Aranda et al. (2018) propose a Nested Logit (NL) model with restrictions in which utilities are specified by radial basis functions. This paper generalises the KLR method to a constrained NL model in which the constraints reflect the exogenous or endogenous factors affecting the decision process.

The kernel-based MNL has seldom been used in empirical applications. Langrock et al. (2014) apply this approach to analyse party preferences based on the characteristics of the individuals. Espinosa-Aranda et al. (2015) and García-Ródenas and López-García (2015) use the kernel-based NL model in a passenger-centred train-timetabling problem. Recently, Bansal et al. (2019) uses this methodology to analyse the factors associated with institutional births (as opposed to home births) in India.

The literature review shows that the use of non-parametric utilities has not been extensively applied in the calculation of certain econometric indicators. This paper analyzes from a computational viewpoint the use of KLR to obtain these indices.

2.1. *Major contributions*

A preliminary work from the authors (Martín-Baos et al., 2020) suggests the use of the KLR method as a promising tool in modelling individual behaviour. KLR has potential use in a constellation of disciplines such as marketing research, social sciences, health economics, among others. However, that work focuses only on comparing the goodness of fit between the MNL and the KLR.

In this paper, transport research has been taken as the reference point. The contributions of this work can be summarised as follows:

- It presents a review of KLR methods from a RUM perspective, showing that these approaches provide a way to specify non-parametric utilities, avoiding the requirement for the modeller to explicitly state a functional expression for these utilities.

- It enriches the comparison of ML methods with RUM found in the literature, which are mainly limited to analysing which methods have the highest predictive capabilities in discrete choice problems, and rarely examine the behavioural outputs that can be derived from ML models and compares the results with those obtained for MNL.
- It performs a controlled computational experiment using Monte Carlo simulation methods, to motivate the use of KLR when dealing with non-linear phenomena.
- It has been shown that KLR is capable of obtaining unbiased estimates of WTP for non-linear utilities without the need to know their functional expression.
- Given that the application of KLR is problematic due to the resource-intensive nature of KLR estimation process, it proposes the use of the L-BFGS-B algorithm to estimate the KLR model. This algorithm reduces model estimation time by a factor of 8–15, compared to the BFGS or Newton's method.
- It includes the development of a Python package for the estimation of KLR method, which is called PyKernelLogit. This package has been used in the numerical section of the paper.

3. RUM and KLR methodologies

In this section the RUM and KLR methods are reviewed in order to introduce a common notation to observe the KLR methods from a RUM perspective, enhancing its use in the transport research community.

3.1. *Random utility models*

As described in Ben-Akiva and Bierlaire (1999b), utility theory assumes that the decision-maker's preference for an alternative can be captured by a value, which is called *utility*, and the decision-maker selects the alternative with the highest associated utility from their choice set. This approach has limitations in practical applications because the underlying assumptions of this concept are often violated. Utility theory assumes that the decision-maker has perfect discriminatory capacity, but the analyst has incomplete information and, therefore, uncertainty has to be taken into account in the specification of the utilities.

In RUM the utility defined for a decision-maker n when choosing an alternative i from the choice set $C = \{1, \ldots, I\}$ is given by

$$U_{in} = V_{in} + \varepsilon_{in}, \tag{1}$$

where V_{in} is the deterministic (also called systematic) component of the utility, and ε_{in} is the unobserved component, which is a random term used to include the impact of all the unobserved variables on the utility function. Hence, the probability that a decision-maker n chooses an alternative i from the choice set C is

$$P_{in} = \mathbb{P}\left(U_{in} \geq U_{jn} \quad \forall j \in C\right) = \mathbb{P}\left(U_{in} = \max_{j \in C} U_{jn}\right). \tag{2}$$

Some assumptions are necessary to make the random utility model operational. The hypothesis about the error distribution ε_{in} determines the probability of choosing each alternative. The MNL models assume that ε_{in} are independently Gumbel distributed with variance $\sigma^2 = \pi^2/(6\mu^2)$, where μ is the Gumbel scale parameter and the location parameter is reset such that the expected value of the Gumbel error term is zero. Therefore, in this case the probability of each alternative is given by the expression:

$$P_{in} = \frac{\exp(V_{in})}{\sum_{j=1}^{I} \exp(V_{jn})}, \tag{3}$$

where I is the total number of alternatives.

These models have been generalised in multiple directions. For instance, more general distributions have been assumed, such as Generalised Extreme Value (GEV), Generalised Nested Logit (GNL) or Cross-Nested logit (CNL) models. All these models have closed forms for the calculation of the probabilities. Moreover, errors have also been modelled assuming a multivariate normal distribution leading to the Multinomial Probit (MNP) model. Other models assume that the systematic part varies from one decision-maker to another, assuming the parameters are a normal multivariate variable. These models are known as Mixed Logit (MXL) models. MNP and MXL models do not have closed forms for calculating probabilities and require the approximation of Gaussian integrals.

The RUM framework is completely specified once the functional form of the deterministic utility and the distribution of the error term are specified. The utility functions depend on a vector of parameters which needs to be estimated. We denote this functional relationship by $V_{in} = V_i(\mathbf{x}_{in}|\boldsymbol{\beta})$. For this purpose, it is assumed that a sample $\mathbf{X}_n = \{\mathbf{x}_{in}\}_{i=1}^{I}$ of features for each decision-maker $n = 1, \cdots, N$ has been observed. In addition, the decisions made by each of the decision-makers have been collected and stored in the matrix $\mathbf{y}$, where $y_{in} = 1$ if decision-maker n chooses alternative i or $y_{in} = 0$, otherwise. The likelihood of the sample $\mathbf{y}$ is

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{n=1}^{N} \prod_{i=1}^{I} \mathbb{P}(i|\mathbf{X}_n, \boldsymbol{\beta})^{y_{in}}. \tag{4}$$

The estimate of the $\boldsymbol{\beta}$ parameters of the utilities is obtained from Penalised Maximum Likelihood Estimation (PMLE), by solving

$$\underset{\boldsymbol{\beta}}{\text{Maximise}} \ \log \mathcal{L}(\boldsymbol{\beta}) - \lambda \mathcal{P}(\boldsymbol{\beta}), \tag{5}$$

where λ is a penalisation parameter, which controls the trade-off between goodness of fit and complexity of the model, and the penalisation term $\mathcal{P}$ is defined by a convex function over the parameters. The canonical method for estimating $\boldsymbol{\beta}$ is the Maximum Likelihood Estimation (MLE), obtained by setting $\lambda = 0$. A fundamental property of the MLE is that the estimate has an asymptotic multivariate normal distribution. The Ridge and Lasso estimations are among the simple techniques used to overcome the *over-generalisation* or *over-fitting* problem. The Lasso method uses the penalisation $\mathcal{P}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ and the Ridge method uses $\mathcal{P}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$.

Once the main concepts related to the RUM model have been established, the next section introduces KLR.

3.2. *Multinomial kernel logistic regression*

Many ML methods approach the problem of classification from a non-statistical point of view. These procedures do not intend to explain the process of choosing for a specific user but to develop procedures with the smallest classification error. KLR is considered a variant of SVM (Cortes and Vapnik, 1995), which not only predicts the classification of an object (an individual's choice), but also estimates the probability of belonging to each category.

KLR builds several *latent functions*, $V_i(\mathbf{x})$ for all $i \in C$, which are equivalent to the systematic utility functions of RUM and, therefore, they are denoted in the same way. Nevertheless, KLR operates with this latent functions as black boxes where the relationship between the feature vector and the utility is not explicitly stated. These latent functions $V_i : \mathcal{X} \mapsto \mathbb{R}$ for $i = 1, \cdots, I$ are searched within function spaces named *Reproducing Kernel Hilbert Spaces* (RKHS). The RKHS space is a vector space which is univocally generated by the so-called *kernel function* $k(\mathbf{x}, \mathbf{x}')$, and its associated RKHS space is denoted by $\mathcal{H}_k$. The family of functions $\{k(\mathbf{x}, \mathbf{x}')\}_{x' \in \mathcal{X}}$ constitutes a basis of the vector space. Any element from $\mathcal{H}_k$ can be represented as a linear combination of basis elements, in particular for $V_i(\mathbf{x}) \in \mathcal{H}_k$. The expression of the latent functions, which from now on will be referred to as utilities, is given by:

$$V_i(\mathbf{x}|\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n k(\mathbf{x}_{in}, \mathbf{x}). \tag{6}$$

The main difference between the approach followed by the KLR models, which is referred to as *non-parametric*, and that followed by RUM, which is denominated as *parametric*, is how they specify the utility function. In the parametric approach, denoted by $V(\mathbf{x}_{in}|\boldsymbol{\beta})$, it is necessary to define a functional expression in advance, establishing from the beginning the effect of each attribute against the others. However, the non-parametric approach, denoted by $V(\mathbf{x}_{in}|\boldsymbol{\alpha})$, does not take a predetermined form because it is constructed according to the information derived from the data. The choice of the kernel function, $k(\mathbf{x}, \mathbf{x}')$, determines the RKHS $\mathcal{H}_k$ where the utilities $V(\mathbf{x}_{in}|\boldsymbol{\alpha}) \in \mathcal{H}_k$ are searched.

The expression (6) shows that obtaining the $V_i(\mathbf{x}|\boldsymbol{\alpha})$ functions requires estimating the parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)^\top$. Hastie et al. (2001); Zhu and Hastie (2005); Ouyed and Allili (2018) propose a regularized function estimation in RKHS for the estimation of $\boldsymbol{\alpha}$. This method suggests estimating the parameters by solving the following optimisation problem

$$\underset{\boldsymbol{\alpha}}{\text{Minimise}} \sum_{n=1}^{N} \sum_{i=1}^{I} L(y_{in}, V_i(\mathbf{x}_{in}|\boldsymbol{\alpha})) + \frac{\lambda}{2} \sum_{i=1}^{I} \|V_i(\mathbf{x}|\boldsymbol{\alpha})\|^2_{\mathcal{H}_k}, \tag{7}$$

where $L(\cdot)$ is a loss function that measures discrepancies between predicted and observed classifications, λ is a regularization parameter that controls the trade-off between goodness of fit and complexity of the model and the norm of the utility functions is computed in the space RKHS. Eq. (7) has the form of *loss* + *penalty*. The loss function $L(\cdot)$ allows many different ways of measuring the model adjustment. KLR uses the negative value of the log-likelihood function as a loss function. This procedure is called minimising Regularised Negative Log-Likelihood (RNLL) or, equivalently, maximising PMLE.

Now, we will derive the objective function for RNLL. But first, it is necessary to introduce the Gram matrices $\mathbf{K}_i$, which are defined as

$$[\mathbf{K}_i]_{n,n'} = k(\mathbf{x}_{in}, \mathbf{x}_{in'}) \text{ for } n, n' = 1, \ldots, N. \tag{8}$$

The Gram matrices are symmetric and positive definite because $k(\cdot, \cdot)$ is a kernel function. All computations are referred to these matrices, and it is necessary to introduce the following notation. We denote by $\mathbf{K}_i^{(n)}$ the $n-$th column of $\mathbf{K}_i$ and by $\mathbf{K}_n = \{\mathbf{K}_i^{(n)}\}_{i=1}^I$ the set of transformed attribute vectors for each decision-maker $n = 1, \cdots, N$.

Similar to what happens with the utility functions in RUM, the latent functions $V_i(\mathbf{x})$ are over-specified. Therefore, without prejudicing the explanatory capacity of the model, it can be assumed that $V_I(\mathbf{x}) = 0$. KLR uses the softmax function to provide estimates of the posterior probability of the alternatives given $\mathbf{K}_n$, as follows:

$$\begin{aligned}
\mathbb{P}(i|\mathbf{K}_n, \boldsymbol{\alpha}) &= \tfrac{\exp(V_i(\mathbf{x}_{in}|\boldsymbol{\alpha}))}{1+\sum_{j=1}^{I-1}\exp(V_j(\mathbf{x}_{jn}|\boldsymbol{\alpha}))} &&= \frac{\exp\left(\mathbf{K}_i^{(n)\top}\boldsymbol{\alpha}\right)}{1+\sum_{j=1}^{I-1}\exp\left(\mathbf{K}_j^{(n)\top}\boldsymbol{\alpha}\right)}, i \in \{1, \ldots, I-1\} \\
\mathbb{P}(I|\mathbf{K}_n, \boldsymbol{\alpha}) &= 1 - \textstyle\sum_{i=1}^{I-1}\mathbb{P}(i|\mathbf{K}_n, \boldsymbol{\alpha}) &&= \frac{1}{1+\sum_{j=1}^{I-1}\exp\left(\mathbf{K}_j^{(n)\top}\boldsymbol{\alpha}\right)}.
\end{aligned} \tag{9}$$

We can derive the log-likelihood function $\log \mathcal{L}(\boldsymbol{\alpha})$ by combining Eq. (9) with Eq. (4). On the other side, the norm of the utilities can be expressed as $\|V_i(\mathbf{x}, \boldsymbol{\alpha})\|_{\mathcal{H}_k}^2 = \boldsymbol{\alpha}^\top \mathbf{K}_i \boldsymbol{\alpha}$ and $\|V_I(\mathbf{x})\|_{\mathcal{H}_k}^2 = \|\mathbf{0}\|_{\mathcal{H}_k}^2 = 0$. Therefore, RNLL can be formulated as follows:

$$\underset{\boldsymbol{\alpha}}{\text{Minimise}} \sum_{j=1}^{I-1} -\mathbf{y}^{(j)\top}\mathbf{K}_j\boldsymbol{\alpha} + \mathbf{1}^\top\boldsymbol{\Delta}(\boldsymbol{\alpha}) + \frac{\lambda}{2}\sum_{j=1}^{I-1}\boldsymbol{\alpha}^\top\mathbf{K}_j\boldsymbol{\alpha}, \tag{10}$$

where $\mathbf{y}^{(j)} = (y_{j1}, \ldots, y_{jN})^\top$, $\mathbf{1} = (1, \ldots, 1)^\top$ is an $N-$dimensional vector of ones and

$$\boldsymbol{\Delta}(\boldsymbol{\alpha}) = \begin{bmatrix} \log\left(1+\sum_{j=1}^{I-1}\mathbf{K}_j^{(1)}\boldsymbol{\alpha}\right) \\ \log\left(1+\sum_{j=1}^{I-1}\mathbf{K}_j^{(2)}\boldsymbol{\alpha}\right) \\ \cdots \\ \log\left(1+\sum_{j=1}^{I-1}\mathbf{K}_j^{(N)}\boldsymbol{\alpha}\right) \end{bmatrix}$$

4. Computing interpretable economic information

The purpose of this section is to introduce a numerical procedure that allows to work with the non-parametric utilities in order to calculate economic measures derived from KLR.

The results of the RUM can be interpreted simply and intuitively. Similar to any other statistical model, analysts can easily understand how the estimated RUM model works by studying the sign, magnitude, and statistical significance of the model's coefficients. These results can also be applied to conduct further analysis on travel behaviour, obtaining indicators such as the marginal effect, elasticities, the Willingness To Pay (WTP) and Value of Time (VOT) among other indicators. These applications can be validated through explicit mathematical formulations and derivations, allowing analysts to clearly understand what is happening.

In the parametric scheme, the linear utility functions are the most widely used. These utility functions are stated as follows

$$V_{in} = V_i(\mathbf{x}_{in}|\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \mathbf{x}_{in} = \sum_{k=1}^{K} \beta_k x_{ink}, \tag{11}$$

where the feature vector belongs to $\mathbb{R}^K$. The non-parametric approach presents the fundamental advantage that its utility specification, stated in Eq. (6), allows very diverse linear and non-linear phenomena to be approximated by the same kernel function, without the need to have prior knowledge of the phenomenon.

Note that the parametric and non-parametric utilities, Eq. (11) and Eq. (6), are linear in the parameters but they use different feature vectors. In addition, the dimensionality of these vectors fulfills $K << N$. In the context of a discrete choice model the $\boldsymbol{\beta}$ parameters (11) are understandable while the $\boldsymbol{\alpha}$ parameters are not. Table 1 shows some of the most widely used economic measures and how they are calculated by linear MNL models. These formulae show how the estimated parameters $\boldsymbol{\beta}$ are involved in their calculation. One relevant index used in the field of transport is the VOT which is defined as the amount of money that an individual is willing to pay to save one unit of time. In this case the formula for the linear case adopts a simple expression as a function of the estimated parameters β_t (coefficient associated with the time attribute) and β_c (parameter associated with the cost attribute).

However, despite this fact, both approaches allow the calculation of marginal effects, elasticities, WTP and VOT among other indicators. This is the fact that is going to be highlighted in this section. The key fact is that the definition of these economic measures is independent of the functional expression of the utility and is therefore defined for non-parametric utilities. In the linear case these expressions have a simple form in terms of the parameters $\boldsymbol{\beta}$. In the case of using non-parametric utilities $V(\mathbf{x}_{in}|\boldsymbol{\alpha})$ the partial derivatives can be calculated using the chain rule, which obtains a closed-form expression, or using numerical differentiation.

To illustrate this numerical approach, we focus on the calculation of the WTP. Let $\mathbf{x}_{in}$ be the current value of the feature vector. Consider two scenarios where all covariates take the same value except for the covariate under study that takes the values $x^+_{ink} = x_{ink} + h$ and $x^-_{ink} = x_{ink} - h$. We denote these new attribute vectors as $\mathbf{x}^+_{in}$ and $\mathbf{x}^-_{in}$, respectively. The Center Divided Difference Method approximates the $k-$th partial derivative as follows:

$$\frac{\partial V_i}{\partial \mathbf{x}_{ik}}(\mathbf{x}_{in}|\boldsymbol{\alpha}) = \frac{V_i(\mathbf{x}^+_{in}|\boldsymbol{\alpha}) - V_i(\mathbf{x}^-_{in}|\boldsymbol{\alpha})}{2h} + O(h^2) \tag{12}$$

In the same way, the partial derivative is calculated with respect to the cost attribute c_i. The calculation of WTP$_{ik}$ applies twice the formula (12) and therefore requires four evaluations of the $V_i(\mathbf{x}|\boldsymbol{\alpha})$.

Table 1. Definition of economic indicators and computation using linear MNL models

Description	Definition	Linear MNL
Marginal effects of feature x_{ik} for alternative j	$M_{j,ik} = \dfrac{\partial P_j}{\partial x_{ik}}$	$= \begin{cases} P_i\,(1-P_i)\,\beta_k & \text{if } j = i \\ -P_j P_i \beta_k & \text{if } j \neq i \end{cases}$
Arc elasticity of feature x_{ik} for alternative j	$E_{j,ik} = \dfrac{\partial P_j\,/\partial x_{ik}}{P_j}$	$= \begin{cases} (1-P_i)\,\beta_k & \text{if } j = i \\ -P_i \beta_k & \text{if } j \neq i \end{cases}$
Willingness-to-pay for the attribute k for alternative i	$WTP_{ik} = -\dfrac{\partial V_i\,/\partial x_{ik}}{\partial V_i\,/\partial c_i}$	$= -\dfrac{\beta_k}{\beta_c}$
Value of time for alternative i	$VOT_i = \dfrac{\partial V_i\,/\partial t_i}{\partial V_i\,/\partial c_i}$	$= \dfrac{\beta_t}{\beta_c}$

As an alternative to the numerical methods, the chain rule can be applied for the kernel function $k(\cdot,\cdot)$ under consideration in the case study. As an example, for the isotropic Gaussian kernel used in the numerical tests, the following closed formula is obtained:

$$\frac{\partial V_i}{\partial \mathbf{x}_{ik}}(\mathbf{x}_{in}|\boldsymbol{\alpha}) = -\theta \sum_{n'=1}^{N} \alpha_{n'}(x_{ink} - x_{in'k}) \exp\left(-\theta \|\mathbf{x}_{in} - \mathbf{x}_{in'}\|_2^2\right). \tag{13}$$

Numerical differentiation methods prevent the development of specific formulae for each of the chosen kernels $k(\cdot,\cdot)$. In the computational experiments we have made use of these numerical techniques.

5. Numerical results

In this section some numerical experiments have been conducted to evaluate and motivate the use of KLR against other alternatives. One of these alternatives is the previously discussed linear MNL model. Moreover, since ML methods are gaining lot of importance in the transport field, it has been decided to also include in this experiment the SVM and the RF methods. The reason behind the selection of these ML methods is the fact that they have achieved better results than the traditional MNL methods on numerous recent studies, such as in Ballings et al. (2015); Hagenauer and Helbich (2017); Zhao et al. (2020); Lhéritier et al. (2019); Wang and Ross (2018), where the SVM and RF methods have achieved the highest accuracy rates.

This section consists of two parts. In the first part the four approaches have been analysed and compared using synthetic data generated by Monte Carlo simulation techniques. Through the use of simulated data, the actual results are known in advance, and the outcome of the methods can be evaluated with respect to the expected results. The second part focuses on assessing the four previous methods on a travel mode choice problem using real data.

All the numerical tests have been coded using Python 3 programming language. The PyKernelLogit[1] package, which has been developed by the authors, has been used to estimate the MNL and KLR. This package extends the functionalities of a previous Python package called PyLogit (Brathwaite and Walker, 2018) and provides some extra functionalities that allow the estimation of discrete choice models based on KLR. For the SVM and RF, the Scikit-learn package was employed (Pedregosa et al., 2011).

5.1. *Monte Carlo simulation experiment*

Firstly, a Monte Carlo simulation study was designed, in order to control the error term and the utility specification. These experiments consider $I = 3$ alternatives and two explanatory variables. In this study several models have been generated in which the utility of the alternative i for the individual n is given by the expression:

$$U_{in} = V(x_{in1}, x_{in2}) + \varepsilon_{in};\ \text{with } i \in \{1, 2, 3\}, \tag{14}$$

where the error terms ε_{in} are i.i.d. random variables drawn from a Gumbel distribution with scale parameter μ and location parameter 0. These models are the ground-truth defined to compare the performance of the four approaches that are being analysed.

For the simulation experiment three systematic utilities have been supposed:

$$V(x_{in1}, x_{in2}) = \quad \beta_1 x_{in1} + \beta_2 x_{in2} \qquad \text{Linear} \tag{15}$$

$$V(x_{in1}, x_{in2}) = \quad {x_{in1}}^{\beta_1} {x_{in2}}^{\beta_2} \qquad \text{Cobb-Douglas (CD)} \tag{16}$$

$$V(x_{in1}, x_{in2}) = \quad \min\{\beta_1 x_{in1}, \beta_2 x_{in2}\} \qquad \text{Minimum} \tag{17}$$

For each of the previous models three pairs of parameters have been considered, being the parameter $\beta_1 = 1$ and $\beta_2 \in \{0.5, 1, 2\}$. Moreover, each of these nine models has been defined using two levels of uncertainty, by means of the scale parameter $\mu \in \{0.2, 0.02\}$. Therefore, 18 different models have been considered. For each model, a total of $N = 1000$ individuals were generated using a uniform distribution of $\mathbf{x}_{in}$ on the square $[0, 1] \times [0, 1]$ for $i = 1, 2, 3$. Finally, to obtain more accurate results, we have generated 200 samples of each model, 100 of them have been used to train the methods which have been compared in this experiment, whereas the other 100 have been used to test the performance of those methods.

In this numerical experiment four different approaches have been compared: MNL, KLR, SVM and RF. The MNL has been estimated using the linear utility specification defined on Eq. (18), where the vector of parameters $\boldsymbol{\beta}$ has been estimated using MLE. The intercept parameter of the first alternative is always fixed to 0, i.e. $\beta_0^1 = 0$.

$$V(\mathbf{x}|\boldsymbol{\beta}_i) = \beta_0^i + \beta_1^i x_1 + \beta_2^i x_2 \tag{18}$$

Concerning KLR, it has been estimated using the utility specification defined on Eq. (19), where the $\boldsymbol{\alpha}$ vector of parameters have also been estimated using MLE. This specification uses an isotropic Gaussian kernel with $\theta = 1$. An intercept parameter, β_0^i, has also been considered as in the previous method. Note that Eq. (6) considers different utility functions $V_i(\mathbf{x}|\boldsymbol{\alpha})$ for all $i \in C$. Nevertheless, on the numerical experiments we have assumed that all utility functions for the alternatives are identical

[1] PyKernelLogit is available on the repository: `https://github.com/JoseAngelMartinB/PyKernelLogit`

except for the interception term, i.e. $V_i(\mathbf{x}|\boldsymbol{\alpha}) = \beta_0^i + V(\mathbf{x}|\boldsymbol{\alpha})$ for all $i \in C$. To this end, the same points $\mathbf{x}_{in}$ should be taken in each alternative. These points are denoted by $\boldsymbol{x}_m$, and they are generated as a square grid of 30×30 elements taken uniformly from the square $[0,1] \times [0,1]$. Finally, the utility function is expressed as follows:

$$V_i(\mathbf{x}|\boldsymbol{\alpha}) = \beta_0^i + \sum_m \alpha_m \exp\left(-\theta \|\mathbf{x} - \mathbf{x}_m\|_2^2\right). \tag{19}$$

The SVM is a ML method for binary classification problems. The idea is to map the input vector into a very high-dimension feature space in which linear decision surfaces can be constructed (Cortes and Vapnik, 1995). Since they are binary classifiers, a one-over-rest (OvR) strategy has been used, which involves training a single classifier per alternative and selecting the alternative whose classifier reports the highest confidence score. Hence, a SVM classifier has been estimated using a Gaussian kernel. The regularisation hyperparameter has been adjusted using a random search with 1000 iterations where the possible values for the hyperparameter are sampled from a uniform distribution in the range $[0, 10]$. The kernel coefficient hyperparameter is set to $1/(\text{number of features} \cdot \text{variance of the feature vector})$.

Finally, RF (Breiman, 2001) is an ensemble method which combines a set of decision tree predictors that are trained in parallel using bootstrap samples. A subset of the variables in the model determines each split at the nodes. The prediction of the selected alternatives is determined by the majority voting among all the individual decision trees. Here, two hyperparameters have been adjusted by using a random search with 1000 iterations over a discrete uniform distribution: the first one is the number of estimators, which has been adjusted using the range $[1, 200]$; and the second one is the maximum number of features to consider when looking for the best split, which has been adjusted by using the range $[1, 6]$.

5.1.1. Model adjustment

All the previous methods have been used to estimate each one of the 100 train samples generated for each one of the 18 different models. For the MNL and KLR methods a goodness of fit indicator called McFadden R-squared value has been computed as $\rho^2 = 1 - \frac{LL(\widehat{\boldsymbol{\Theta}})}{LL(0)}$, where $\widehat{\boldsymbol{\Theta}}$ is the estimate of the vector of parameters. Tables 2 and 3 report the McFadden R-squared value for the MNL and KLR methods and the CPU time needed to estimate all the methods. For each measure, the mean value of the 100 train samples and the standard deviation is reported.

In Table 2 ($\mu = 0.02$) the effect of the error term is small, and therefore, the generated data contains more information about the phenomenon. As it can be observed by comparing the goodness of fit ρ^2 index, the KLR method outperforms the MNL method as it is able to generalise better and, consequently, it is possible to determine whether the model is non-linear and to adapt better to it. In this way, KLR gives better results for the Cobb-Douglas and minimum models. In case of the linear models, both methods obtain the same results. Table 3 ($\mu = 0.2$) shows similar results but the differences are less significant because the effect of the error term is bigger, and therefore, the uncertainty strongly influences the decision process.

One disadvantage that has been highlighted in the literature is the high computational cost of KLR (Ouyed and Allili, 2018; Zhu and Hastie, 2005), leading to the development of so-called sparse KLR. In this work, it has been tested the Newton's method and BFGS algorithms, which are the canonical methods for solving the MLE

Table 2. Estimation process assessment (Case $\mu = 0.02$). The mean value of the train samples and the standard deviation (between brackets) is reported.

Utility function	**Parameters**	**Goodness of fit (ρ^2)**		**CPU time (s)**			
		MNL	KLR	MNL	KLR	SVM	RF
CD	$\beta_1 = 1$	0.77	0.90	0.16	1.34	0.01	0.35
	$\beta_2 = 0.5$	(0.02)	(0.01)	(0.01)	(0.48)	(0.00)	(0.01)
	$\beta_1 = 1$	0.74	0.89	0.16	0.91	0.01	0.09
	$\beta_2 = 1$	(0.02)	(0.02)	(0.01)	(0.24)	(0.00)	(0.00)
	$\beta_1 = 1$	0.70	0.84	0.16	1.21	0.02	0.24
	$\beta_2 = 2$	(0.02)	(0.02)	(0.01)	(0.24)	(0.00)	(0.00)
Linear	$\beta_1 = 1$	0.93	0.93	0.18	1.24	0.01	0.36
	$\beta_2 = 0.5$	(0.01)	(0.01)	(0.01)	(0.36)	(0.00)	(0.01)
	$\beta_1 = 1$	0.94	0.94	0.18	1.47	0.02	0.19
	$\beta_2 = 1$	(0.01)	(0.01)	(0.01)	(0.37)	(0.00)	(0.00)
	$\beta_1 = 1$	0.97	0.97	0.19	1.79	0.02	0.19
	$\beta_2 = 2$	(0.01)	(0.01)	(0.01)	(0.66)	(0.00)	(0.00)
Minimum	$\beta_1 = 1$	0.55	0.81	0.16	1.90	0.02	0.36
	$\beta_2 = 0.5$	(0.02)	(0.02)	(0.01)	(0.50)	(0.00)	(0.01)
	$\beta_1 = 1$	0.56	0.88	0.16	1.49	0.02	0.42
	$\beta_2 = 1$	(0.02)	(0.02)	(0.01)	(0.45)	(0.00)	(0.01)
	$\beta_1 = 1$	0.57	0.89	0.16	2.19	0.01	0.35
	$\beta_2 = 2$	(0.02)	(0.01)	(0.01)	(0.36)	(0.00)	(0.01)

problem. However, by using the L-BFGS-B optimisation algorithm (Byrd et al., 1995), it has been achieved a limited memory usage and a lower computational time comparing to the BFGS or Newton's method. More concretely, it is reduced by a factor ranging from 8 to 15. For the sake of simplicity, these results are not reported in the paper and all the numerical results have been calculated applying the L-BFGS-B algorithm. This highlight allows the practical application of the KLR to the field of transport since the data is usually collected through surveys, whose size is moderate. It can also be noticed that the computation time for the KLR is considerably reduced when uncertainty increases, for example in the case $\mu = 0.2$.

5.1.2. Model assessment

Once the four methods have been adjusted, they have been assessed using the 100 test samples generated for each model. It has been decided to use accuracy to measure the classification performance of the different methods. The accuracy measures the average number of correctly classified observations. In Tables 4 and 5 the mean accuracy for each method over the 100 test samples is computed. Due to the fact that the test data has been generated by a Monte Carlo simulation, the *maximum accuracy* which can be achieved by any method on the test set is known beforehand. As far as we know, this is the first time that this index has been reported within comparative method studies. The maximum accuracy is defined as the percentage of observations in which users would make the same choice regardless of whether it is considered the whole utility function or only the systematic part of the utility function. That is, the random term does not change the decision. Since RUM and ML methods are only capable of learning the systematic part of the utility, then, it is impossible to determine the random part of that utility, which produces part of the misclassification error. This index has been numerically calculated from the 100 test samples generated by simula-

Table 3. Estimation process assessment (Case $\mu = 0.2$). The mean value of the train samples and the standard deviation (between brackets) is reported.

Utility function	**Parameters**	**Goodness of fit (ρ^2)**		**CPU time (s)**			
		MNL	KLR	MNL	KLR	SVM	RF
CD	$\beta_1 = 1$	0.26	0.28	0.15	0.73	0.03	0.36
	$\beta_2 = 0.5$	(0.02)	(0.02)	(0.01)	(0.05)	(0.00)	(0.01)
	$\beta_1 = 1$	0.23	0.26	0.15	0.71	0.03	0.45
	$\beta_2 = 1$	(0.02)	(0.02)	(0.01)	(0.06)	(0.00)	(0.01)
	$\beta_1 = 1$	0.18	0.22	0.15	0.79	0.03	0.36
	$\beta_2 = 2$	(0.02)	(0.02)	(0.01)	(0.11)	(0.00)	(0.01)
Linear	$\beta_1 = 1$	0.39	0.38	0.15	0.69	0.03	0.11
	$\beta_2 = 0.5$	(0.02)	(0.02)	(0.01)	(0.07)	(0.00)	(0.00)
	$\beta_1 = 1$	0.48	0.47	0.16	0.68	0.02	0.35
	$\beta_2 = 1$	(0.02)	(0.02)	(0.01)	(0.06)	(0.00)	(0.01)
	$\beta_1 = 1$	0.64	0.64	0.16	0.75	0.03	0.34
	$\beta_2 = 2$	(0.02)	(0.02)	(0.01)	(0.11)	(0.00)	(0.01)
Minimum	$\beta_1 = 1$	0.10	0.12	0.15	0.73	0.03	0.23
	$\beta_2 = 0.5$	(0.01)	(0.02)	(0.01)	(0.08)	(0.00)	(0.00)
	$\beta_1 = 1$	0.22	0.28	0.15	0.71	0.03	0.27
	$\beta_2 = 1$	(0.02)	(0.02)	(0.01)	(0.07)	(0.00)	(0.00)
	$\beta_1 = 1$	0.26	0.33	0.15	0.88	0.03	0.26
	$\beta_2 = 2$	(0.02)	(0.02)	(0.01)	(0.13)	(0.00)	(0.00)

tion. Therefore, a t-test was used to determine if the mean accuracy reported by each of the methods is significantly different from the maximum accuracy achievable on the test set. Several levels of statistical significance have been defined and are denoted by using stars (***<0.001, **<0.01, *<0.05). Finally, the 95% confidence interval of the accuracy has been computed using the simulation results for each method.

Analysing the results provided on Tables 4 and 5, several facts can be observed. Firstly, only the MNL and KLR methods are capable of achieving the maximum accuracy on some models, i.e. the applied t-test fails to reject the null hypothesis that the mean accuracy is equal to the mean maximum accuracy, mostly on the results in the Table 5. The maximum accuracy is obtained by the MNL method when the data meet all the hypotheses (linear utilities). The KLR obtains the maximum accuracy for both the linear case and the non-linear Cobb-Douglas utilities. Note that the specification used in KLR is unique for all models and it approximates both phenomenons. This highlight releases the modeller from having to specify the functional expression of the utility, since it is derived from the data.

MNL obtains better results than any other method for the linear models. MNL also obtains better results than RF for Cobb-Douglas models. These results contrasts with those shown in several works such as Ballings et al. (2015); Hagenauer and Helbich (2017); Zhao et al. (2020); Lhéritier et al. (2019); Wang and Ross (2018), where ML methods outperformed MNL. The reason for this is that the samples used in this experiment consists of synthetic data where the hypotheses of the MNL models are satisfied. Nevertheless, Wang et al. (2017) obtain very similar results to those presented in this experiment using two real datasets, as the MNL obtains a similar prediction accuracy to the tree-based ML algorithms but it gets a higher accuracy than the kernelized SVM. When there is a higher level of uncertainty as shown in Table 5, i.e. the data contain limited information about the underlying utility functions, then the linear approximations used in the MNL work properly and their results are closer to

Table 4. Assessment of the model's accuracy (Case $\mu = 0.02$). The mean value obtained on the test samples and the 95% confidence interval (between brackets) is reported.

Utility function	**Parameters**	**Accuracy (%)**				**Maximum accuracy (%)**
		MNL	KLR	SVM	RF	
CD	$\beta_1 = 1$	89.72 ***	95.02***	92.11***	88.75 ***	95.51
	$\beta_2 = 0.5$	(87.96, 91.48)	(93.70, 96.35)	(90.54, 93.68)	(86.35, 91.14)	(94.17, 96.85)
	$\beta_1 = 1$	88.94 ***	94.78	91.92***	87.43 ***	94.94
	$\beta_2 = 1$	(86.94, 90.93)	(93.44, 96.13)	(90.04, 93.81)	(85.36, 89.51)	(93.58, 96.30)
	$\beta_1 = 1$	88.07 ***	92.01**	89.23***	86.49 ***	92.35
	$\beta_2 = 2$	(86.23, 89.90)	(90.28, 93.74)	(87.16, 91.30)	(83.99, 88.98)	(90.67, 94.03)
Linear	$\beta_1 = 1$	96.39 *	96.36**	93.41***	88.86 ***	96.59
	$\beta_2 = 0.5$	(95.23, 97.55)	(95.20, 97.52)	(91.58, 95.24)	(86.84, 90.88)	(95.57, 97.61)
	$\beta_1 = 1$	97.04 **	97.04**	93.81***	88.37 ***	97.24
	$\beta_2 = 1$	(95.99, 98.09)	(96.01, 98.08)	(92.06, 95.57)	(85.95, 90.79)	(96.25, 98.22)
	$\beta_1 = 1$	98.15 **	98.04***	94.14***	89.16 ***	98.30
	$\beta_2 = 2$	(97.33, 98.97)	(97.22, 98.85)	(92.42, 95.86)	(87.22, 91.09)	(97.57, 99.02)
Minimum	$\beta_1 = 1$	78.80 ***	91.01***	87.21***	88.21 ***	92.43
	$\beta_2 = 0.5$	(76.48, 81.13)	(89.19, 92.82)	(85.45, 88.97)	(86.01, 90.41)	(91.04, 93.81)
	$\beta_1 = 1$	79.64 ***	93.97***	89.57***	89.78 ***	95.50
	$\beta_2 = 1$	(77.05, 82.24)	(92.36, 95.57)	(87.4, 91.74)	(87.43, 92.14)	(94.16, 96.84)
	$\beta_1 = 1$	79.83 ***	94.42***	90.11***	90.52 ***	96.32
	$\beta_2 = 2$	(77.53, 82.13)	(92.90, 95.93)	(88.07, 92.14)	(88.39, 92.65)	(95.08, 97.56)

Note: ***<0.001, **<0.01, *<0.05

those of the ML or KLR methods.

For all of the 18 models considered, KLR provides better results than the other ML methods analysed (SVM and RF), which are commonly the best methods in literature. Moreover, KLR also achieves better results than the MNL method for the non-linear models, i.e. the Cobb-Douglas and the minimum model. In the case of the linear models, the results obtained by KLR and MNL methods are statistically indistinguishable and the 95% confidence interval overlaps between both algorithms. This confirms the capability of the KLR models to adapt to non-linear phenomena, achieving accuracy results that are very close to the maximum accuracy of the simulation data.

5.1.3. Computing Willingness To Pay economic indicator

One important disadvantage of ML methods highlighted in the literature is their lack of interpretability. Some of these methods do not provide probabilities, such as SVM, and consequently cannot calculate some indicators like elasticities or marginal effects. In other methods, such as RF, utility functions are not used and therefore the WTP or VOT indicators cannot be estimated. KLR does not present those issues and, for this reason, in this section it will be assessed its advantages and disadvantages in the estimation of the WTP indicator.

The MNL and KLR methods have been evaluated in the computation of the WTP economic indicator. As it is shown in Table 1, WTP can be computed as the partial derivative of the utility function. In the case of the MNL methods the partial derivative of the utility functions results into a closed formula which is specified in the previous table. However, for the KLR it is necessary to use numerical numerical differentiation methods to compute the WTP, which allows any arbitrary kernel function $k(\mathbf{x}, \mathbf{x}')$ to be considered.

The main advantage of working with simulated data is that it allows to compute the real WTP value for each of the generated samples. By applying the WTP formula

Table 5. Assessment of the model's accuracy (Case $\mu = 0.2$). The mean value obtained on the test samples and the 95% confidence interval (between brackets) is reported.

Utility function	**Parameters**	**Accuracy (%)**				**Maximum accuracy (%)**
		MNL	KLR	SVM	RF	
CD	$\beta_1 = 1$	63.92 ***	65.12	63.83 ***	61.43 ***	65.28
	$\beta_2 = 0.5$	(61.02, 66.82)	(62.06, 68.19)	(60.69, 66.98)	(58.28, 64.59)	(62.32, 68.24)
	$\beta_1 = 1$	62.52 ***	63.41	59.28 ***	59.69 ***	63.65
	$\beta_2 = 1$	(59.40, 65.63)	(60.55, 66.27)	(59.28, 65.31)	(56.67, 62.71)	(60.76, 66.53)
	$\beta_1 = 1$	59.14 **	59.48	55.42 ***	56.2 ***	59.86
	$\beta_2 = 2$	(55.99, 62.29)	(56.72, 62.25)	(55.42, 62.03)	(52.79, 59.61)	(56.83, 62.9)
Linear	$\beta_1 = 1$	70.51	70.50	69.62 ***	66.85 ***	70.59
	$\beta_2 = 0.5$	(67.59, 73.43)	(67.87, 73.12)	(66.81, 72.43)	(64.09, 69.62)	(67.94, 73.24)
	$\beta_1 = 1$	75.03	75.01	73.83 ***	71.52 ***	75.25
	$\beta_2 = 1$	(72.32, 77.74)	(72.48, 77.54)	(70.85, 76.81)	(68.79, 74.25)	(72.59, 77.91)
	$\beta_1 = 1$	83.24	83.32	82.14 ***	79.58 ***	83.44
	$\beta_2 = 2$	(80.95, 85.54)	(81.01, 85.64)	(79.78, 84.49)	(77.21, 81.95)	(81.01, 85.86)
Minimum	$\beta_1 = 1$	50.93 ***	53.13 **	51.44 ***	49.21 ***	53.83
	$\beta_2 = 0.5$	(47.85, 54.01)	(49.78, 56.48)	(48.03, 54.85)	(46.00, 52.42)	(50.53, 57.12)
	$\beta_1 = 1$	61.16 ***	64.95	63.02 ***	61.07 ***	65.26
	$\beta_2 = 1$	(58.36, 63.96)	(62.12, 67.78)	(60.00, 66.03)	(57.73, 64.42)	(62.44, 68.08)
	$\beta_1 = 1$	63.24 ***	67.34 ***	65.3 ***	64.17 ***	68.16
	$\beta_2 = 2$	(60.37, 66.12)	(64.6, 70.07)	(62.50, 68.1)	(61.06, 67.27)	(65.02, 71.29)

Note: ***<0.001, **<0.01, *<0.05

to each model the following expressions can be obtained,

$$WTP_{\text{Cobb-Douglas}} = -\frac{\beta_1 x_{n2}}{\beta_2 x_{n1}}, \tag{20}$$

$$WTP_{\text{Linear}} = -\frac{\beta_1}{\beta_2}, \tag{21}$$

$$WTP_{\text{Minimum}} = \begin{cases} \nexists & \text{if } \beta_2 x_{n2} \geq \beta_1 x_{n1} \\ 0 & \text{if } \beta_2 x_{n2} < \beta_1 x_{n1} \end{cases}. \tag{22}$$

The utility of each alternative is different for each decision-maker depending on their feature vector. Hence, each decision-maker would be willing to pay a different amount for the same increase in an attribute for a given alternative. For this reason, this simulation study considers for the values of the feature vector the points $(0.25, 0.75)$, $(0.50, 0.50)$ and $(0.75, 0.25)$.

Tables 6 and 7 show the obtained WTP values for the first alternative. For each of the three points being considered, it is represented the mean and its standard deviation of the WTP value computed over the 100 train samples using the MNL and KLR methods. Finally, the *real WTP* column contains the theoretical WTP value for each point computed using expressions (20) to (22). Notice that some theoretical WTP values are not defined on some points for the minimum models because the denominator is zero in the corresponding WTP formula.

Analysing those tables, it can be observed that the WTP value computed using the MNL is independent of the vector x_{in}. Roughly speaking, linear MNL estimates an average WTP value over the domain of x_{in}. In this way, linear MNL allow the theoretical values of the WTP for non-linear utilities in the point $\bar{x}_{in}$ to be estimated reasonably well, where $\bar{x}_{in}$ is the mean value of the attribute vector x_{in} in the sample. This can be observed in the results obtained for the point $\bar{x}_{in} = (0.50, 0.50)$ using Cobb-Douglas and minimum utilities, where the theoretical values are approximated but the results are slightly worse than those obtained with KLR. It should be noticed that the MNL model is not capable of capturing the non-linearity of these models and

Table 6. Comparison of MNL and KLR models for obtaining WTP economic indicator (Case $\mu = 0.02$). The mean value obtained on the train samples and the standard deviation (between brackets) is reported.

Utility function	**Parameters**	**P1 (0.25, 0.75)**			**P2 (0.50, 0.50)**			**P3 (0.75, 0.25)**		
		MNL	KLR	Real WTP	MNL	KLR	Real WTP	MNL	KLR	Real WTP
CD	$\beta_1 = 1$ $\beta_2 = 0.5$	-1.60 (0.11)	-5.80 (2.21)	-6.00	-1.60 (0.11)	-2.03 (0.19)	-2.00	-1.60 (0.11)	-0.63 (0.05)	-0.67
	$\beta_1 = 1$ $\beta_2 = 1$	-1.00 (0.06)	-3.02 (0.37)	-3.00	-1.00 (0.06)	-1.00 (0.04)	-1.00	-1.00 (0.06)	-0.34 (0.04)	-0.33
	$\beta_1 = 1$ $\beta_2 = 2$	-0.65 (0.05)	-1.56 (0.10)	-1.50	-0.65 (0.05)	-0.50 (0.04)	-0.50	-0.65 (0.05)	-0.20 (0.06)	-0.17
Linear	$\beta_1 = 1$ $\beta_2 = 0.5$	-2.01 (0.07)	-2.05 (0.18)	-2.00	-2.01 (0.07)	-1.94 (0.09)	-2.00	-2.01 (0.07)	-2.04 (0.12)	-2.00
	$\beta_1 = 1$ $\beta_2 = 1$	-1.00 (0.02)	-1.00 (0.04)	-1.00	-1.0 (0.02)	-1.00 (0.03)	-1.00	-1.0 (0.02)	-1.01 (0.05)	-1.00
	$\beta_1 = 1$ $\beta_2 = 2$	-0.50 (0.01)	-0.50 (0.03)	-0.50	-0.50 (0.01)	-0.51 (0.02)	-0.50	-0.50 (0.01)	-0.50 (0.03)	-0.50
Minimum	$\beta_1 = 1$ $\beta_2 = 0.5$	-0.53 (0.05)	-7.73 (2.66)	$\nexists$	-0.53 (0.05)	-0.01 (0.11)	0.00	-0.53 (0.05)	0.00 (0.06)	0.00
	$\beta_1 = 1$ $\beta_2 = 1$	-1.00 (0.08)	8.02 (2.78)	$\nexists$	-1.00 (0.08)	-0.99 (0.08)	$\nexists$	-1.00 (0.08)	0.13 (0.04)	0.00
	$\beta_1 = 1$ $\beta_2 = 2$	-1.89 (0.17)	4.10 (119.48)	$\nexists$	-1.89 (0.17)	7.83 (193.73)	$\nexists$	-1.89 (0.17)	-0.12 (0.04)	0.00

produces biased estimates when estimating the WTP at points which are distant from x_{in}. Nevertheless, the linear MNL method is more efficient than KLR in the estimation of the WTP for the linear models, although the difference is not very noticeable. These estimates are unbiased in all cases with a standard error for the estimates of an order of magnitude 10^{-2} and 10^{-1} for the cases $\mu = 0.2$ and $\mu = 0.02$.

The advantage of KLR over linear MNL method appears when the decision process is driven by non-linear utilities and it is desired to calculate the value of the WTP at a point distant from $\bar{x}_{in}$. This statement can be observed for Cobb-Douglas and minimum utilities at the points $(0.25, 0.75)$ and $(0.75, 0.25)$. KLR is capable of adapting to linear and non-linear phenomena and produces (apparently) unbiased WTP estimates for both points in the three utility functions. A large error of the estimate for KLR is associated with the non-existence of WTP in the minimum function. This behaviour is also exhibited on the standard deviation values for the Cobb-Douglas model with $\beta_2 = 0.5$ at point P1 $(0.25, 0.75)$, which may be caused because the value of the denominator in Equation (20) is close to zero at point P1.

Next, with a new experiment an actual problem is evaluated to identify whether results with non-synthetic data still satisfy our theoretical approach.

5.2. *Models assessment on a travel mode choice problem*

In this section, a travel mode choice application is undertaken with the goal of evaluating the previous methods on real data. In this scenario, as opposed to the previous experiments, the actual values of the parameters to be estimated and the WTP are unknown, thus, it is difficult to know whether a model estimates better or worse the behaviour of the decision-makers. The aim is therefore to check whether new situations or issues arise that have not been covered using the synthetic data.

The data used for this evaluation come from a case study involving a travel mode choice problem using revealed preference data collected in Switzerland between 2009 and 2010 (Atasoy et al., 2013). The main goal of this survey was to collect data for analysing the travel behaviour of people in low-density areas. The respondents

Table 7. Comparison of MNL and KLR models for obtaining WTP economic indicator (Case $\mu = 0.2$). The mean value obtained on the train samples and the standard deviation (between brackets) is reported.

Utility function	**Parameters**	**P1 (0.25, 0.75)**			**P2 (0.50, 0.50)**			**P3 (0.75, 0.25)**		
		MNL	KLR	Real WTP	MNL	KLR	Real WTP	MNL	KLR	Real WTP
CD	$\beta_1 = 1$ $\beta_2 = 0.5$	-1.60 (0.25)	-5.94 (36.28)	-6.00	-1.60 (0.25)	-1.70 (0.16)	-2.00	-1.60 (0.25)	-0.64 (0.12)	-0.67
	$\beta_1 = 1$ $\beta_2 = 1$	-0.98 (0.14)	-3.34 (1.18)	-3.00	-0.98 (0.14)	-0.99 (0.10)	-1.00	-0.98 (0.14)	-0.32 (0.12)	-0.33
	$\beta_1 = 1$ $\beta_2 = 2$	-0.67 (0.11)	-1.49 (0.84)	-1.50	-0.67 (0.11)	-0.59 (0.14)	-0.50	-0.67 (0.11)	-0.20 (0.31)	-0.17
Linear	$\beta_1 = 1$ $\beta_2 = 0.5$	-2.05 (0.29)	-1.93 (0.47)	-2.00	-2.05 (0.29)	-2.05 (0.20)	-2.00	-2.05 (0.29)	-1.99 (0.41)	-2.00
	$\beta_1 = 1$ $\beta_2 = 1$	-1.01 (0.10)	-1.00 (0.12)	-1.00	-1.01 (0.10)	-1.01 (0.06)	-1.00	-1.01 (0.10)	-1.05 (0.13)	-1.00
	$\beta_1 = 1$ $\beta_2 = 2$	-0.50 (0.05)	-0.51 (0.06)	-0.50	-0.50 (0.05)	-0.50 (0.04)	-0.50	-0.50 (0.05)	-0.54 (0.08)	-0.50
Minimum	$\beta_1 = 1$ $\beta_2 = 0.5$	-0.50 (0.13)	-2.85 (16.09)	$\nexists$	-0.50 (0.13)	-0.44 (0.11)	0.00	-0.50 (0.13)	0.32 (0.19)	0.00
	$\beta_1 = 1$ $\beta_2 = 1$	-1.02 (0.15)	12.53 (61.71)	$\nexists$	-1.02 (0.15)	-1.00 (0.11)	$\nexists$	-1.02 (0.15)	0.06 (0.07)	0.00
	$\beta_1 = 1$ $\beta_2 = 2$	-1.94 (0.32)	-60.46 (940.25)	$\nexists$	-1.94 (0.32)	-3.46 (1.34)	$\nexists$	-1.94 (0.32)	-0.22 (0.01)	0.00

were asked to register all the trips performed during a specified day. The possible values for the choice variable are: public transport (train, bus, tram, etc.), private modes (car, motorbike, etc.) and soft modes (bike, walk, etc.). During a preprocessing step, observations with unknown selected alternative were excluded, obtaining a final dataset with 1880 observations.

In this second experiment it has been decided to use the the MNL utility specification proposed by Bierlaire (2018), where the utility functions are defined as:

$$\begin{aligned}
V_{PT} =& \beta_{\text{Time\_Fulltime}} * \text{TimePT} * \text{fulltime} + \\
& \beta_{\text{Time\_Other}} * \text{TimePT} * \text{notfulltime} + \\
& \beta_{\text{Cost}} * \text{MarginalCostPT} \\
V_{Car} =& \beta_{\text{ASC\_Car}} + \\
& \beta_{\text{Time\_Fulltime}} * \text{TimeCar} * \text{fulltime} + \\
& \beta_{\text{Time\_Other}} * \text{TimeCar} * \text{notfulltime} + \\
& \beta_{\text{Cost}} * \text{CostCarCHF} \\
V_{SM} =& \beta_{\text{ASC\_SM}} + \\
& \beta_{\text{Dist\_Male}} * \text{distance\_km} * \text{male} + \\
& \beta_{\text{Dist\_Female}} * \text{distance\_km} * \text{female} + \\
& \beta_{\text{Dist\_Unreported}} * \text{distance\_km} * \text{unreportedGender}
\end{aligned} \tag{23}$$

With respect to KLR, it has been estimated using the same utility specification as in the previous experiment, which is defined on Eq. (19). The feature vector $\mathbf{x}_{in}$ has been scaled to the range $[0, 1]$ because it has been used an isotropic Gaussian kernel. Finally, concerning the SVM and RF methods, their hyperparameters have been adjusted using a random search with 1000 iterations using the same configuration as in the previous experiment.

Table 8 shows the results that have been obtained in this experiment. Since only one sample is available, a resampling-based method such as cross validation (Kohavi, 1995) should be applied for assessing the effectiveness of the models. More concretely, it was decided to use a 5×2 cross validation procedure, as suggested by Dietterich (1998), which consists of executing 5 times a 2-fold cross validation technique. For each of the metrics in the table (goodness of fit, CPU time and accuracy value), it has been reported the mean value of the 5×2 cross validation procedure and the standard deviation.

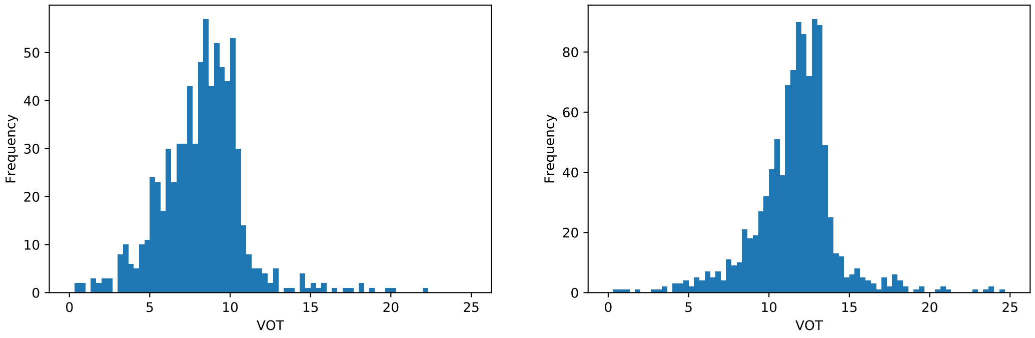
Table 8. Assessment of the models on a travel mode choice problem using real data. The mean value obtained on the 5×2 cross validation and the standard deviation (between brackets) is reported.

	MNL	**KLR**	**SVM**	**RF**
Goodness of fit (ρ^2)	0.41 (0.02)	0.44 (0.02)	-	-
CPU time (s)	0.18 (0.03)	9.13 (2.68)	0.02 (0.00)	0.17 (0.07)
Accuracy (%)	70.06 (0.87)	72.93 (0.90)	70.23 (1.08)	76.32 (1.14)

Analysing Table 8 some conclusions can be derived. Firstly, it can be observed that KLR has a higher computational cost compared to other methods, however, this cost is affordable. It should be noticed that all the ML methods have obtained higher accuracy results than the MNL method. In the case of the SVM the accuracy is similar than the MNL method. The RF has achieved the best accuracy score, as shown by numerous recent studies (Ballings et al., 2015; Hagenauer and Helbich, 2017; Zhao et al., 2020; Lhéritier et al., 2019; Wang and Ross, 2018). These results evidence that using real data the error terms not always follow the i.i.d. Gumbel distribution. Finally, the KLR method has achieved a better accuracy result than the MNL and SVM methods. It should be taken into account that the KLR method avoids the difficulty of establishing a functional expression for utilities beforehand, simplifying its use and enabling non-linear behaviour to be modelled. Note that if the dataset presents a significant imbalance between alternatives, then it is also helpful to calculate other measures, such as the sensitivity of the method, in order to evaluate properly its performance.

To conclude, the previous MNL and KLR methods have been used to estimate the VOT. It has been computed the VOT for full-time and non full-time employees that used the private transport alternative. Using the estimated MNL model, the VOT values obtained are 6.40 CHF (Swiss franc)/hour for full-time employees and 2.12 CHF/hour for non full-time employees. These values are similar to the ones reported in Bierlaire (2018) using a nested logit model. In this work, the author stated that the values obtained were too low and attributed this to a poor specification of the model's utilities. This highlights again the problem that KLR aims to avoid, the necessity of having to specify the functional expression of the utility functions.

Unlike the MNL method, KLR obtains a different value of the VOT for each decision-maker. This novel aspect can been incorporated to evaluate confidence intervals for the VOT, obtaining the histograms presented in Figure 1(a) and 1(b) for full-time and not full-time workers, respectively. The average values obtained for each alternative are 7.75 CHF/hour and 11.23 CHF/hour. These values are larger than those obtained using the MNL method and, therefore, more in line with our expectations.



(a) full-time workers

(b) Non full-time workers

Figure 1. Distribution of the value of time in the sample using KLR

6. Conclusions

In this paper, KLR method has been numerically evaluated against the MNL method and other prominent ML methods for discrete choice analysis, such as SVM and RF. Two important conclusions have been obtained: the first one is that KLR is able to reach the maximum accuracy for both linear models and non-linear models for the uncertainty level with parameter $\mu = 0.2$, whereas MNL only reaches this maximum accuracy on linear models. Hence, KLR is capable of modelling non-parametric specifications of the utility functions, relieving the analyst from specifying a functional relationship between the features in advance. It is noteworthy that with synthetic data the RF method performs worse than the KLR. The second highlight is that it appears that the KLR obtains unbiased WTP estimates for both linear and non-linear models, whilst the linear MNL only obtains unbiased WTP estimates for linear models.

This numerical study has been complemented with an application to a real travel mode choice problem. Three highlights have been obtained: i) The application of the L-BFGS-B optimisation algorithm produces a lower computational cost for the estimation of the KLR method and allows its application to real problems (a disadvantage that was foreseen a priori); ii) RF achieves the highest accuracy, followed by KLR. The fact that RF is one of the best methods of ML in various applications has been widely reported in the literature, however, RF does not employ utility functions and consequently it cannot calculate economic indices like the WTP; and iii) The KLR model allows to compute the VOT index for all the decision-makers in the sample.

The utility functions are a powerful tool for modelling user behaviour in the applications, such as combined traffic assignment models (Adnan et al., 2009; Cantelmo and Viti, 2019). Parametric utilities allow, through analytical manipulation, to understand how the different assumptions of the utility model affect the model's outcomes. This is a limitation of data-based models such as KLR. A second limitation of the KLR methodology is the choice of the kernel function, which might be a source of bias, affecting the results obtained. A plausible solution to this problem is to include the kernel function as an additional hyperparameter to be adjusted before the model is estimated.

Our numerical experiments provide strong evidence for practical effectiveness of non-parametric utilities. Some theoretical aspects should be addressed in depth in future work, such as the introduction of a general error term in the non-parametric utilities. One approach is to consider the systematic utility as a Gaussian Process, which leads to a special type of MXL model. A priori, this approach has the advantage that it does not require the specification of a distribution of varying tastes across individuals, which differs from the parametric MXL. Efficient maximum simulated likelihood estimation methods should be developed for this purpose. For this reason, it might result interesting to expand the numerical comparison presented with more sophisticated approaches, such as MXL.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research of Martín-Baos has been supported by the FPU Predoctoral Program of the Spanish Ministry of Science, Innovation and Universities (Grant Ref. FPU18/00802); and the research of García-Ródenas was supported by Project TRA2016-76914-C3-2-P of the Spanish Ministry of Science and Innovation, co-funded by the European Regional Development Fund.

References

Abe, M. (1999). A generalized additive model for discrete-choice data. *Journal of Business and Economic Statistics 17*(3), 271–284.

Adnan, M., D. Watling, and T. Fowkes (2009). Model for Integrating Home–Work Tour Scheduling with Time-Varying Network Congestion and Marginal Utility Profiles for Home and Work Activities. *Transportation Research Record 2134*(1), 21–30.

Atasoy, B., A. Glerum, and M. Bierlaire (2013). Attitudes towards mode choice in Switzerland. *disP - The Planning Review 49*(2), 101–117.

Ballings, M., D. Van Den Poel, N. Hespeels, and R. Gryp (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications 42*(20), 7046–7056.

Bansal, P., R. A. Daziano, and N. Sunder (2019, 6). Arriving at a decision: A semiparametric approach to institutional birth choice in India. *Journal of Choice Modelling 31*, 86–103.

Ben-Akiva, M. and M. Bierlaire (1999a). Discrete choice methods and their applications to short term travel decisions. In R. W. Hall (Ed.), *Handbook of transportation science*, pp. 5–33. Massachusetts: Kluwer Academic Publishers.

Ben-Akiva, M. and M. Bierlaire (1999b). Discrete Choice Methods and their Applications to Short Term Travel Decisions. In R. W. Hall (Ed.), *Handbook of Transportation Science*, pp. 5–33. Boston, MA: Springer US.

Ben-Akiva, M. E. and S. R. Lerman (1985). *Discrete choice analysis: Theory and Application to Travel Demand*, Volume 9. Cambridge, Massachusetts: MIT press.

Bierlaire, M. (2018). Calculating indicators with PandasBiogeme. Technical report, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne.

Brathwaite, T. and J. L. Walker (2018). Asymmetric, closed-form, finite-parameter models of multinomial choice. *Journal of Choice Modelling 29*, 78–112.

Breiman, L. (2001). Random Forests. *Machine Learning 45*(1), 5–32.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing 16*(5), 1190–1208.

Cantelmo, G. and F. Viti (2019). Incorporating activity duration and scheduling utility into equilibrium-based Dynamic Traffic Assignment. *Transportation Research Part B: Methodological 126*, 365–390.

Cawley, G. and N. Talbot (2005). The evidence framework applied to sparse kernel logistic regression. *Neurocomputing 64*(1-4 SPEC.), 119–135.

Celikoglu, H. B. (2006). Application of radial basis function and generalized regression neural networks in non-linear utility function specification for travel mode choice modelling. *Mathematical and Computer Modelling 44*(7-8), 640–658.

Cheng, L., X. Chen, J. De Vos, X. Lai, and F. Witlox (2019). Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society 14*, 1–10.

Cheng, L., X. Lai, X. Chen, S. Yang, J. De Vos, and F. Witlox (2019). Applying an ensemble-based model to travel choice behavior in travel demand forecasting under uncertainties. *Transportation Letters*.

Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine Learning 20*(3), 273–297.

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation 10*(7), 1895–1923.

Espinosa-Aranda, J., R. García-Ródenas, M. López-García, and E. Angulo (2018). Constrained nested logit model: formulation and estimation. *Transportation 45*(5), 1523–1557.

Espinosa-Aranda, J., R. García-Ródenas, M. Ramírez-Flores, M. López-García, and E. Angulo (2015). High-speed railway scheduling based on user preferences. *European Journal of Operational Research 246*(3), 772–786.

Fernández-Delgado, M., E. Cernadas, S. Barro, and D. Amorim (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research 15*, 3133–3181.

Fukuda, D. and T. Yai (2010). Semiparametric specification of the utility function in a travel mode choice model. *Transportation 37*(2), 221–238.

García-Ródenas, R. and M. López-García (2015). Utilization of reproducing kernel hilbert spaces in dynamic discrete choice models: An application to the high-speed railway timetabling problem. In *Transportation Research Procedia*, Volume 10, pp. 544–553.

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

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Kneib, T., B. Baumgartner, and W. J. Steiner (2007). Semiparametric multinomial logit models for analysing consumer choice behaviour. *AStA Advances in Statistical Analysis 91*(3), 225–244.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Articial Intelligence (IJCAI)*, Montreal, Canada, pp. 1137–1145.

Langrock, R., N. B. Heidenreich, and S. Sperlich (2014). Kernel-based semiparametric multinomial logit modelling of political party preferences. *Statistical Methods and Applications 23*(3), 435–449.

Lhéritier, A., M. Bocamazo, T. Delahaye, and R. Acuna-Agost (2019). Airline itinerary choice modeling using machine learning. *Journal of Choice Modelling 31*, 198–209.

Lindner, A., C. S. Pitombo, and A. L. Cunha (2017). Estimating motorized travel mode choice using classifiers: An application for high-dimensional multicollinear data. *Travel Behaviour and Society 6*, 100–109.

Liu, W., H. Liu, D. Tao, Y. Wang, and K. Lu (2016). Manifold regularized kernel logistic regression for web image annotation. *Neurocomputing 172*, 3–8.

Maalouf, M. and T. Trafalis (2011). Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics and Data Analysis 55*(1), 168–183.

Martín-Baos, J., R. García-Ródenas, M. López-García, and L. Rodriguez-Benitez (2020). Discrete choice modeling using Kernel Logistic Regression. In *Transportation Research Procedia*, Volume 47, pp. 457–464.

McFadden, D. L. (1978). Modelling the Choice of Residential Location. In A. K. et al. (Ed.), *Spatial Interaction Theory and Planning Models*, pp. 75–96. Amsterdam, The Netherlands: North Holland.

Ouyed, O. and M. S. Allili (2018). Feature weighting for multinomial kernel logistic regression and application to action recognition. *Neurocomputing 275*, 1752–1768.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Rasouli, S. and H. Timmermans (2012). Uncertainty in travel demand forecasting models: Literature review and research agenda. *Transportation Letters 4*(1), 55–73.

Train, K. (2009). *Discrete Choice Methods with Simulation.* Cambridge: Cambridge University Press.

Wainberg, M., B. Alipanahi, and B. J. Frey (2016). Are random forests truly the best classifiers? *Journal of Machine Learning Research 17*, 1–5.

Wang, F. and C. L. Ross (2018, 12). 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.

Wang, K., X. Ye, R. M. Pendyala, and Y. Zou (2017). On the development of a semi-nonparametric generalized multinomial logit model for travel-related choices. *PLOS ONE 12*(10), e0186689.

Wang, S., Q. Wang, and J. Zhao (2019a). Deep Neural Networks for Choice Analysis: Extracting Complete Economic Information for Interpretation. *arXiv preprint arXiv:1812.04528v2*.

Wang, S., Q. Wang, and J. Zhao (2019b). Multitask Learning Deep Neural Networks to Combine Revealed and Stated Preference Data.

Zhao, X., X. Yan, A. Yu, and P. Van Hentenryck (2020). Prediction and Behavioral Analysis of Travel Mode Choice: A Comparison of Machine Learning and Logit Models. *Travel Behaviour and Society 20*, 22–35.

Zhu, J. and T. Hastie (2005). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics 14*(1), 185–205.