Analysis of port choice:
A methodological proposal adjusted with public data

Abstract

The availability of public databases provides a great amount of data for research, but their use can involve a lack of detailed information about the decision-makers. This fact prevents the analysis of systematic variations in their preferences. In addition, the influence of the explanatory variables included in the utility function of discrete choice models is not necessarily linear. The analysis of the port choice process through this kind of models has usually been addressed neglecting both the existence of heterogeneity in the preferences and the non-linearity in the variables. In this paper, a Box-Cox Mixed Logit Model is proposed to overcome both constraints, and it is introduced through a case study.

In the aim to highlight the interest of the proposal made, the obtained results were compared with those of the more traditional formulations. The conclusion is that this model provides more accurate results. Therefore, it can better help policy-makers when assessing hypothetical scenarios to define their competitive strategies because conclusions can vary significantly, as can be seen from the prognosis carried out for the Spanish case.

Keywords:

Port choice; Box-Cox Mixed Logit Model; Heterogeneity; Non-linearity; Spanish case.

1. Introduction

From a broad review of the literature on the Port Economics field, Pallis, Vitsounis and De Langen (2010) found seven main research areas: terminal studies, ports in transport and supply chains, port governance, port planning and development, port policy and regulation, spatial analysis of seaports and, finally, port competition and competitiveness. The port choice analysis is included in this last category, and is one of the topics attracting most interest from researchers (Mennis, Platis, Lagoudis & Nikitakos, 2008).

Since the 80's, Discrete Choice Models (DCM) have been increasingly used to analyze how and why the customers of the transport sector make their choices. Specifically, passenger transport demand modelling has accumulated long experience. However, the greater complexity of the freight transport systems makes it hard to take advantage of that knowledge (Román, Arencibia & Feo-Valero, 2017). This is particularly true when analyzing the port choice because there are many different agents involved in the process. Indeed, the choice does not result from a single agent but from multiple decisions taken by different stakeholders, all of them engaged in the supply chain. Additionally, each component of that chain has its particular characteristics and its own commercial and logistics requirements and objectives (see, for instance, Meersman, Van De Voorde and Vanelslander, 2010; Sanchez, Ng and Garcia-Alonso, 2011 or Nugroho, Whiteing and de Jong, 2016).

Paixão, Carvalho and Oliveira (2010) identified 56 articles published between 1981- 2009, and found that discrete choice models (e.g., Ben-Akiva & Lerman, 1985) is the methodological approach proposed in 20% of the papers. More recently, Martínez Moya and Feo Valero (2017) highlighted that most of the articles analyzing the port choice topic from the Discrete Choice Theory perspective propose a multinomial logit model (MNL). The pioneers were Malchow and Kanafani (2001). They proposed a disaggregate MNL, which is the most widely proposed specification because of its simplicity. However, that advantage could turn into a disadvantage if it leads to misleading conclusions. Attempting to provide more accurate results, Anderson, Opaluch and Grigalunas (2009), S. Veldman, Garcia-Alonso and Vallejo-Pinto (2013) and Vega, Cantillo and Arellana (2019) proposed to address the port choice problem with a nested logit model (NL). With this approach, these authors first tested the coastline choice and then the port selection (i.e., ports were previously grouped by coastlines). Those three articles improved the results obtained in previous analysis respectively made about the port choice process in USA, Spain and Colombia (all countries with two clearly differentiated coastlines). Certainly, there are more advanced models in the DCM literature (see e.g. Garrow, 2010, for an overview), but they have not usually been applied to the port choice problem. A synthesis of the main articles on DCM applied to port choice can be seen in Tables 1 and 2, where case studies using aggregate and disaggregate data are shown, respectively.

Reference	DCM Model	Geog. area	Data	Obs.*	Variables ‡
Blonigen and Wilson (2006)	MNL (OLS)	USA	US Customs and Border	95680 (M)	Maritime and Inland Costs. Port Efficiency
Garcia-Alonso and Sanchez-Soriano (2009)	MNL (MLE)	Spanish peninsula	Spanish Customs Statistics	-	Inland distance.
S. Veldman and Gopkalo (2011)	MNL (OLS)	Russia	Russian containerized	241 + 287 (M+X)	Inland time and cost. Maritime time and cost. South Basis routing.
S. Veldman, Garcia-Alonso and Vallejo-Pinto (2011)	MNL (OLS)	Spanish peninsula	Spanish Customs Statistics	1984 + 2211 (M+X)	Inland cost. Maritime cost. Total cost. Port quality service (Mohoring effect).
S. Veldman, Garcia-Alonso and Vallejo-Pinto (2013)	NL (OLS)	Spanish peninsula	Spanish Customs Statistics	1984 + 2211 (M+X)	Coastal side. Inland and maritime cost. Back-haul effect. Mohring effect
S. Veldman, Garcia-Alonso and Liu (2015)	MNL (OLS)	Spanish peninsula	Spanish Customs Statistics	1984 + 2211 (M+X)	Inland time and cost. Maritime cost. Non-monetary cost at sea and in port. Feeder port. Mohring effect

Table 1: Classification of references using DCM to study Port Choice (aggregate data).

‡ All variables listed here are significant (non-significant variables are not included in references).

* The traffic flow direction in parentheses is indicated: imports (M) and exports (X).

Reference	DCM Model	Geog. area	Data	Obs.*	Variables ‡
Malchow and Kanafani (2001)	MNL (MLE)	USA	PIERS	4842 (X)	Maritime and inland distance. *Frequency. Vessel capacity*
Malchow and Kanafani (2004)	MNL Cham-berlain	USA	PIERS	4842 (X)	Maritime and inland distance. Sailing headway. Recurring port. *Vessel capacity*
Nir, Lin and Liang (2003)	MNL (MLE)	Taiwan	Survey (shippers)	309 (M+X)	Highway travel time and cost. Number of routes. Recurring user. *Frequency. Closeness to port chosen*
Tongzon and Sawant (2007)	MNL[§]	Singapore and Malaysia	Survey (shipping line)	31	Infrastructure. Port charges. Port services. *Vessel size. Connectivity. Efficiency. Deep-water port*
Anderson et al. (2009)	NL (MLE)	USA	PIERS	470766 (M)	Inland distance. Sea Time. Freight charge. Truck trip one way. Full truck trip. Port

					reliability. Coastal variables: coastal side, same coastal source, same side of major market.
Steven and Corsi (2012)	MNL (OLS)	Pittsburgh metropolitan area (USA)	PIERS	19556 (M)	Maritime and Inland transit time. Crane productivity. Port Congestion. Frequency. Form of port governance. Ocean freight charges. Port size. Shipper size *Cargo size*#
Vega et al. (2019)	MNL, NL, EC (MLE)	Colombia	DIAN	20000 (M+X)	Oceanic cost and time. Inland Cost. Frequency. Coastal side. Containerized cargo
Yang, Wang and Li (2016)	MNL	Bohai Bay (China)	China Ports	1721 (M+X)	Destination trade markets. Rapid boutique lines. Hinterland's GDP. Port capacity. Port capacity. *Foreign investment in Port. Highway mileage.*
Kashiha et al. (2016)	MNL (MLE)	Austria	PIERS	2084857 (X)	Inland and maritime distance. Border count. Route infrastructure. Efficiency. Connectivity. Mediterranean hub. Geographical circumstances (coastal, landlocked, pseudo-landlocked). Shipper size, shipment volume and value
Martínez-Pardo et al. (2018)	MNL (MLE)	Spanish peninsula	Spanish Customs Statistics	5432556 (X)	Inland Distance. Oceanic Distance. Cranes. Degree of use of port facilities.

Table 2: Classification of references using DCM to study Port Choice (disaggregate data).

‡ Non-significant variables (or with an unexpected sign) are in italics.

* The traffic flow direction in parentheses is indicated: imports (M) and exports (X).

§ Binary logit model ($j \in \{0,1\}$). This model can be viewed as a special case of the MNL, or the MNL a direct extension of the Binary logit model.

Some combinations of variables (interaction terms) are also non-significant.

Usually, DCM in transport are based on survey data (either revealed or stated preferences). As pointed out by Ben-Akiva et al. (2008), this kind of data is expensive and usually proprietary. Fortunately, nowadays there are a lot of sources of public data that reveal choices. The advantages of using public data are price, replicability and continuity in time. However, this kind of database does not usually include detailed information about the decision makers. In this case, when preference heterogeneity cannot be addressed in a systematic way, the use of the random coefficients approach allows to take heterogeneity into account and to obtain accurate models in this context (Orro, Novales and Benitez, 2007).

One of the aims of this paper is to consider the existence of possible differences in the stakeholders' choice behavior. These can be introduced in the model in a systematic way if the variation can be estimated according to measured attributes of the decision makers or of the shipment. These attributes can be related with type of route, type of traffic, type of decision-maker, etc. Remaining heterogeneity can be taken into account using random coefficients specification of the Mixed Logit model (ML; see, McFadden and Train, 2000 and Train, 2009) which has become very popular in the last two decades, with many contributions in different fields.

Ben-Akiva, Bolduc and Park (2008) used a random coefficients model to predict shippers' choice of mode but, to the best of the authors' knowledge, only two studies controlled for preference heterogeneity in the port choice field, both of them by including the size of port users as a continuous variable in the form of an interaction term: Steven & Corsi (2012) and Kashiha, Thill & Depken (2016). Their hypothesis is the same: large shippers give greater control to carriers. According to this hypothesis, factors that benefit the carriers should be statistically significant for large shippers, while factors that directly benefit the shippers should be statistically significant for the small shippers. Both of them find significance heterogeneity in preferences between larger and smaller shippers. Addtionally, Kashiha et al. (2016) also studies heterogeneity through shipment characteristics (volume and value per unit) and geographical circumstances. More recently, Vega et al. (2019) proposed an ML with an error components specification (EC) to address the correlation between port alternatives, but not the random taste heterogeneity. An alternative approach to the ML used in this research could be Latent Class Models (see Greene and Hensher, 2003).

The second aim of this paper is to face an additional source of problems in the study of port choice: the misspecification of the influence of each attribute. Most authors consider linear influence in the utility function, and this may not reflect the reality of port choice. For example, a saving of ten kilometers in the origin distance from the port would mean the same change of utility for a 60-kilometer trip as it would for a 200-kilometer trip. There are several ways to introduce non-linearity in the specification of DCM. The analyst can do it using, among others, logarithms, powers or interaction between several attributes. It can be introduced in the utility function and tested in the model estimation. Alternatively, the functional form of the relationship can be estimated in a parametric way, avoiding to set it *a priori*. This can be done with a Box-Cox transformation (BC) (Box & Cox, 1964). In the DCM field, its first application was the Box-Cox Logit (BCL) (Gaudry and Wills, 1978). Although less popular than ML, BC transformations have been used in discrete choice models in different applications since then, but no study with BC transformation applied to port choice has been found in the review done.

In short, this paper contributes to the previous literature in port choice by proposing a Box-Cox Mixed Logit model (BCML), which combines the advantages of the use of random coefficients and the estimation of non-linear relationships (Orro, 2005). That is, it allows to take into account decision makers' heterogeneity when using public databases and the use of the Box-Cox transformation to estimate non-linearity in variables. In the context of labour supply, a Mixed Logit model with estimated Box-Cox transformation was presented in Razzolini (2010). Lapparent, Frei and Axhausen (2009) presented a model with random parameters and BC transformation applied to a long distance travel problem.

In order to highlight the interest of this proposal, the consequences of ignoring both circumstances are exposed through the case study addressed: the container port choice in Spain made from the inland side. The data are approximately 5.5 million observations from the years 2004 to 2012, obtained from the Foreign Trade Statistics of the Customs and Excise Duties Department of the Spanish Tax Agency (SCS).

The remainder of the paper is structured as follows. The next section presents the methodological background of the research. In Section 3 the case study is presented, with the data and the model specification to test. In Section 4 the estimated models and goodness-of-fit measures are provided and discussed, together with validation with data not used for calibration. The final part of the study, Section 5, is concerned with the impact of considering or not random heterogeneity and non-linearity in different situations. Section 6 summarizes the conclusions drawn.

2. Background and methodological proposal

The aim of the model is to analyze the choice of port for each of the shipments. In the context of DCM, port j will be chosen to channel shipment n when it provides the highest utility (U_{nj}) to the decision maker. In the random utility maximization framework, the utility can be expressed as:

$$U_{nj} = V_{nj}\left(\beta_j, x_{nj}\right) + \varepsilon_{nj} \tag{1}$$

where V_{nj} is the observed part and ε_{nj} is the unobserved part. V is a representative utility function that depends on a vector of observed variables (x_{nj}) and a vector of parameters (β_j) to be estimated.

The multinomial logit model (MNL, McFadden, 1973) assumes that ε_{nj} are independently and identically distributed (IID) Gumbel or extreme value type I. Among other considerations, it supposes identical tastes for all the individuals who share the same observed variables.

For V_{nj}, the hypothesis of linear in parameters specification with attributes in natural form is the most common. Considering k explicative attributes and alternative-specific constants (ASC_j), the representative utility is:

$$V_{nj} = ASC_j + \sum_{k=1}^{K} \beta_{kj} x_{knj} \tag{2}$$

As has been shown in the literature review, most of the models applied to port choice use MNL, and employ linear in parameters specification. The pillars of the present methodological proposal are:

- To take into account the heterogeneity between decision makers without the knowledge of their characteristics, due to the use of public databases of shipments, and
- To estimate the presence of non-linear influence of the attributes without previous specification of the functional form of the relationship.

In order to reach these objectives, a Box-Cox Mixed Logit (BCML) approach is proposed. It allows simultaneous specification of random heterogeneity among decision makers and estimated non-linearity in the influence of the attributes. It will be compared with MNL, ML and BCL specifications, which can be seen as restricted cases of the BCML. The utility function for the BCML with non-random ASC can be formulated according to Equations 3 and 4.

$$U_{nj} = ASC_j + \sum_{k=1}^{K} \beta_{knj} x_{knj}^{(\lambda_k)} + \varepsilon_{nj} \tag{3}$$

$$x^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln x \text{ if } \lambda = 0 \end{cases} \tag{4}$$

The preferences of port decision-makers that choose n are represented by β_{knj}. These preferences vary randomly within the population according to a density function $f(\beta|\theta)$ dependent on certain underlying parameters θ (e.g. mean and standard deviation). $x^{(\lambda k)}_{knj}$ are the observed variables relating to port j for the shipment n, and the Box-Cox transformation of parameter λ_k (Eq. 4) is applied to some or all of them. Finally, the ε_{nj} are the IID Gumbel error term. The estimation of the model consists in obtaining the parameters which will be those underlying the coefficient distribution and the exponents of the BC transformations. The process is carried out by maximizing the simulated log-likelihood of the sample.

3. Case study

3.1. Ports and database

In order to illustrate the methodological proposal, a study of the choice between the top Spanish peninsular container ports is presented. These ports are Algeciras, Barcelona, Bilbao and Valencia (see Figure 1). They have jointly managed around 90% of the annual container traffic over the period studied and, as stated in García-Alonso et al (2019) and García-Alonso and Márquez (2017), they are the ports which compete most fiercely for Spanish export flows, which is the traffic analyzed here.

Figure 1. Location of the ports studied

The data at shipment level (n) used in this analysis are revealed preference data; that is, they reflect the choices actually made by the decision-makers. We use the database of the Foreign Trade Statistics of the Customs and Excise Duties

Department of the Spanish Tax Agency (SCS)[1], particularly suitable for the analysis of the inter-port distribution of the Spanish extra-Community maritime flows[2]. It is freely available and provides information about the composition of the Spanish foreign trade by province and mode of transport (in Euros and tons), taking into account the country of origin/destination. The port chosen for channeling each flow is assumed to be that located in the province where the flow is managed, as in previous studies (see, for instance, Veldman et al. 2011, 2013 or 2015).

The data of the four ports from 2004 to 2012 is screened from SCS to form the dataset, only considering export container flows from peninsular provinces to non-European countries. The wide range of years allows to consider the influence on choices of the changes in port facilities and other attributes. A summary of the dataset, after eliminating incomplete records, is given in Table 3.

Container export shipments to non-European countries								
Algeciras		Barcelona		Bilbao		Valencia		Total (4 ports)
316136	(5.82%)	2815136	(51.82%)	1647	(0.03%)	2299637	(42.33%)	5432556
Container traffic (thousand Ton)								
Algeciras		Barcelona		Bilbao		Valencia		Port system
367199	(35%)	185489	(18%)	51232	(5%)	346648	(33%)	1058528

Table 3: Summary of traffic (accumulated total between 2004 and 2012).

Source: Authors' own elaboration from data provided by the SCS (2014) and Spanish Ports Authority (2018).

The use of historical series, official and publicly accessible data has several advantages: a high number of data points, economy, replicability or the capacity to analyze the influence of variables that vary over time. The main disadvantage is the lack of knowledge of the specific data of the companies that carry out the shipment, due to confidentiality issues. It prevents the use of panel data and specifying systematic taste heterogeneity with company attributes. The methodological approach with the use of random taste variation seeks to compensate the absence of information about the port choice decision-maker.

3.2. Attributes and base utility specification

To better appreciate the interest of the methodological proposal made, the starting point is a previous model developed for the same case study in Martínez-Pardo, Garcia-Alonso, and Orro (2018). Therefore, the explanatory

[1] The complete relation of the variables included in the database can be found (in Spanish) in https://www.agenciatributaria.es/static_files/AEAT/Aduanas/Contenidos_Privados/Estadisticas_Comercio_Exterior/comercio_exterior/datos_mensuales_maxima_desagregacion/diseno226.pdf

[2] See Escamilla-Navarro et al. (2010). These authors provide a comprehensive analysis of this database, and conclude that this is the only data source available in Spain to identify the provincial origin-destination of the maritime flows.

variables selected are the distance and the degree of use of port facilities, as shown in Table 4.

Variable	Definition
DO_{nj}	Distance by road between port j and province of origin of shipment n
DD_{nj}	Length of maritime route between port j and country of destination for shipment n
$TC_j(t-1)$	Degree of use of port facilities indicator, at port j for each year $t-1$

Table 4: Description of variables

On the one hand, inland and maritime distance have been used in many previous studies (see Tables 1 and 2). On the other, the degree of use of port facilities aims to synthesize information about the level of service offered by a port, which has also been considered in several analysis through alternative indicators as capacity, productivity, congestion or efficiency. In this particular case, the degree of use of port facilities are considered following Martínez-Pardo, Garcia-Alonso, and Orro (2018)[3].

The distance in origin is computed measuring the kilometers by road between the port and province of origin of trade. The length of the sea route to the destination is computed by measuring the kilometers needed to reach the main port of the destination country. This maritime distance reflects the vessel's actual routes, which might consist of a sequence of scheduled ports of call until reaching the country of destination and can be seen as a proxy of the availability of direct routes between ports. The data used is the mean of the most frequent routes according to SeaRates (2015).

The degree of use of port facilities indicator is obtained at port j as the container traffic in TEUs divided by the number of ship-to-shore (STS) gantry cranes. In this paper, it is lagged by one year to avoid endogeneity problems ($TC_j\ (t-1)$). For the Spanish case, the STS gantry cranes and TEUs moved per port can be found in Traffic Statistics or the Statistical Yearbooks of Spanish ports (Spanish Port Authority, 2016).

This attribute is studied in categories; that is, it is divided into Q equal sized segments and each category is introduced in the model as a dummy variable. This way of specifying TC allows a non-linear influence of the degree of use without the need to impose it a priori. A detailed description of how the ranges of TC were specified, including issues such as the number of intervals or their values, can be found in Martínez-Pardo, Garcia-Alonso, and Orro (2018). In this study it was seen that there is a point of saturation of the port facilities from which an increase in the variable will have a negative effect; so, it is specified here in the same way.

Of course, there are other variables that can influence the choice of port, but they are not available with the kind of data used or they appear as non-significant in the models estimated according to statistical tests in the case study. SCS variables tested and excluded from the model are weight/value, type

[3] As pointed out in Martínez-Pardo et al. (2018, p. 516): "…the more traffic a port has, the more attractive it becomes because of the economies of agglomeration, scale and network effects but only up to a certain point when the port starts to be saturated". They found that the attractiveness of the port is conditioned by the degree of use of its port facilities, whose influence is not linear.

of international commercial transaction (INCOTERM) and type of good (CN code). The number of cranes appears as significant in previous versions of the model, but this variable has been removed from the final specification due to potential endogeneity issues. Variables related to frequency, cost or specific route characteristics are outside the scope of this research, because it focuses on total shipments from all origins to all destinations over a period of several years. Variables such as recurring user or recurring port, as well as shipper characteristics, cannot be determined from the SCS data. In the framework of the Spanish port market, the possibility of competing in port charges is very limited, so it has not been considered.

Therefore, the base MNL model used as comparison reference in this paper to estimate the utility of port *j* for shipment *n* is:

$$U_{nj} = ASC_j + \beta_{DO} \cdot DO_{nj} + \beta_{DD} \cdot DD_{nj} + \sum_{q=1}^{Q} \beta_{TC}^{q} \cdot TC_j^q \left(t-1\right) + \varepsilon_{nj} \tag{5}$$

where ASC_j are the alternative-specific constants and β_{DO}, β_{DD} and $\beta^q{}_{TC}$ are the generic coefficients.

3.3. Proposed models

As mentioned previously, the objective of this work is to introduce heterogeneity in preferences and estimated non-linearity in variables through Box-Cox transformations starting with the previous MNL specification. We propose three different models: a Mixed Logit Model (ML), a Box-Cox Logit Model (BLC) and a Box-Cox Mixed Logit Model (BCML).

Firstly, only heterogeneity is introduced in the preferences of port choice decision-makers: to that end, the coefficients of DO_{nj} and DD_{nj} are assumed to be random coefficients that represent different values for each shipment, with a distribution to be estimated in the population (random coefficients specification). The choice of the appropriate distribution for the random parameters is not a trivial matter because, in general, information of the actual form of that distribution is not available (for a general discussion on the distributions used in random coefficients see Hensher and Greene (2003), Train and Sonnier (2005) and Hess, Bierlaire and Polak (2005)). Equation (6) presents the specification for the ML model. The coefficients are assumed to be generic coefficients and follow normal distributions with mean μ_k and variance σ_k. ξ_{kn} is a random variable with a standardized distribution (zero mean and unit variance) analogous to that of the corresponding coefficient. Log-normal and triangular distributions were also tested; although the random coefficients can be better delimited with either of them, there was no improvement in final log-likelihood and their use led to computational complexity and slow convergence.

$$\begin{aligned} U_{nj} = {} & ASC_j + \left(\mu_{DO} + \sigma_{DO} \cdot \xi_{DOn}\right) \cdot DO_{nj} + \left(\mu_{DD} + \sigma_{DD} \cdot \xi_{DDn}\right) \cdot DD_{nj} + \\ & + \sum_{q=1}^{Q} \beta_{TC}^{q} \cdot TC_j^q \left(t-1\right) + \varepsilon_{nj} \end{aligned} \tag{6}$$

Secondly, a model with Box-Cox transformation on the attributes is specified. Equation (7) presents the specification for the BCL. The Box-Cox transformation (see Eq. 4) is applied to DO_{nj} and DD_{nj} with a unique exponent for all the alternatives: λ_{DO}, λ_{DD}.

$$U_{nj} = ASC_j + \beta_{DO} \cdot DO_{nj}^{(\lambda_{DO})} + \beta_{DD} \cdot DD_{nj}^{(\lambda_{DD})} + \sum_{q=1}^{Q} \beta_{TC}^{q} \cdot TC_j^q (t-1) + \varepsilon_{nj} \tag{7}$$

Finally, heterogeneity in preferences and Box-Cox transformation in variables are jointly introduced. The detailed specification for BCML can be expressed as:

$$\begin{aligned} U_{nj} = {} & ASC_j + \left(\mu_{DO} + \sigma_{DO} \cdot \xi_{DOn}\right) \cdot DO_{nj}^{(\lambda_{DO})} + \left(\mu_{DD} + \sigma_{DD} \cdot \xi_{DDn}\right) \cdot DD_{nj}^{(\lambda_{DD})} + \\ & + \sum_{q=1}^{Q} \beta_{TC}^{q} \cdot TC_j^q (t-1) + \varepsilon_{nj} \end{aligned} \tag{8}$$

4. Estimation results and discussion

The aim of this section is to present the results of the four models estimated, to determine whether non-linearity and random heterogeneity are present and to analyze which is the best specification according to usual statistical tests.

4.1. Methods and general results

The model estimation was carried out in BIOGEME (Bierlaire, 2016). The maximization is performed using the CFSQP algorithm (Lawrence, Zhou & Tits, 1997) using a Sequential Quadratic Programming method. For ML and BCML simulated likelihood estimation with Modified Latin Hypercube Sampling draws (MLHS) are used. A discussion can be found in Hess, Train and Polak (2006) comparing the use of MLHS versus Halton draws.

Given the complexity of the models to be estimated, the use of the complete database of 5.5 million observations is not computationally adequate. A random sample was extracted from the revealed preference base defined in Section 3. To ensure consistency of the results, different sizes of samples, with a different number of MLHS extractions, were estimated. Since only slight differences were found, the model is considered stable for the sample of 1 million observations and 250 MLHS extractions.

The estimation results of the models are presented in Table 5. The alternative-specific constant of Barcelona and category B for the degree of use of port facilities are set to zero (ASC_{Bar} = 0 and β^{B}_{TC} = 0) to identify the models. To analyze whether the parameters are statistically significant or not, note that the null hypothesis is that the parameter is equal to zero for all parameters except the exponents of the BC transformation, that is one. In all cases there is a low p-value (the highest is $p < 0.04$ for β^{A}_{TC} in MLN), so the null hypothesis with 96% of confidence can be rejected.

In models with Box-Cox transformation, the t-test values estimated implicitly depend on the unit of measurement of the Box-Cox transformed variables (see Spitzer, 1984 or Dagenais and Dufour, 1994 for the case of regression models). After checking this situation, it has been found that the t-tests for μ_{DO}, σ_{DO}, μ_{DD} and σ_{DD} (but not for λ_{DO} and λ_{DD}) vary with the scale of DD and DO. An additional estimation with estimated λ_{DO} and λ_{DD} fixed is included in Table 5, its parameters do not depend on scale.

	MNL				ML				BC							BCML					
Parameter (units)	value	*RSE*	*r t-test*		value	*RSE*	*r t-test*		value	*RSE*	*r t-test*	value	*RSE*	*r t-test*		value	*RSE*	*r t-test*	value	*RSE*	*r t-test*
ASC_{Alg}	-1.49	0.01	-138.40		-3.21	0.03	-120.83		-1.64	0.01	-143.7	-1.64	0.01	-148.67		-2.42	0.02	-142.38	-2.42	0.02	-152.41
ASC_{Bar}	*0.00*	*[fixed]*			*0.00*	*[fixed]*			*0.00*	*[fixed]*		*0.00*	*[fixed]*			*0.00*	*[fixed]*		*0.00*	*[fixed]*	
ASC_{Bil}	-8.03	0.07	-120.41		-22.40	0.12	-181.12		-7.93	0.07	-118.48	-7.93	0.07	-120.36		-14.70	0.09	-157.77	-14.70	0.09	-165.74
ASC_{Val}	-0.53	0.00	-112.29		-0.91	0.01	-132.64		-0.36	0.00	-72.92	-0.36	0.00	-76.82		-0.70	0.01	-110.31	-0.70	0.01	-113.53
β_{DO}	-6.86	0.01	-844.76	μ_{DO}	-13.00	0.05	-260.79	β_{DO}	-3.70	0.01	-267.21	-3.70	0.01	-734.35	μ_{DO}	-5.11	0.03	-166.17	-5.11	0.02	-273.39
				σ_{DO}	6.00	0.03	176.88								σ_{DO}	2.41	0.02	154.21	2.41	0.01	171.81
								λ_{DO}	0.58	0.00	290.23	0.58	*[fixed]*		λ_{DO}	0.46	0.00	290.23	0.46	*[fixed]*	
β_{DD}	-7.98	0.09	-86.98	μ_{DD}	-10.40	0.16	-65.58	β_{DD}	-3.64	0.02	-42.92	-3.64	0.02	-225.13	μ_{DD}	-5.61	0.02	-42.92	-5.61	0.07	-86.16
$\beta_{TC}{}^{A}$				σ_{DD}	38.00	0.31	124.29								σ_{DD}	6.67	0.00	106.90	6.67	0.10	67.30
								λ_{DD}	0.25	0.01	39.49	0.25	*[fixed]*		λ_{DD}	0.41	72.40	39.88	0.41	*[fixed]*	
$\beta_{TC}{}^{A}$	-0.28	0.13	-2.09		-0.51	0.20	-2.55	$\beta_{TC}{}^{A}$	-0.33	0.13	-2.49	-0.33	0.13	-2.49	$\beta_{TC}{}^{A}$	-0.80	0.13	-2.49	-0.80	0.14	-5.79
$\beta_{TC}{}^{B}$	*0.00*	*[fixed]*			*0.00*	*[fixed]*		$\beta_{TC}{}^{B}$	*0.00*	*[fixed]*		*0.00*	*[fixed]*		$\beta_{TC}{}^{B}$	*0.00*	*[fixed]*		*0.00*	*[fixed]*	
$\beta_{TC}{}^{C}$	0.21	0.01	32.98		0.35	0.01	40.00	$\beta_{TC}{}^{C}$	0.20	0.01	30.84	0.20	0.01	30.84	$\beta_{TC}{}^{C}$	0.31	0.01	30.84	0.31	0.01	39.05
$\beta_{TC}{}^{D}$	-0.61	0.01	-45.22		-0.68	0.03	-25.95	$\beta_{TC}{}^{D}$	-0.33	0.01	-23.94	-0.33	0.01	-23.94	$\beta_{TC}{}^{D}$	-0.23	0.01	-23.94	-0.23	0.02	-11.80
Number of estimate parameters			8				10							10							12
Final LL/(S)LL			-404618.33				-372633.15							-368663.49							-353425.29
ρ^2			0.71				0.73							0.73							0.75
$\bar{\rho}^2$			0.54				0.57							0.58							0.59

Number of observations	1000000	
Algeciras	57972	(5.80%)
Barcelona	518019	(51.80%)
Bilbao	306	(0.03%)
Valencia	423703	(42.37%)

Table 5: Estimated parameters

Note: *RSE* is the robust standard error and *r t-stat* is the robust t-statistics.

Overall, it can be said for the four models estimated (MNL, ML, BCL and BCML) that all of the estimated parameters are statistically significant and have the expected sign. The results show that increases in the values of DO and DD reduce the utility of the port under study and, therefore, the probability that it will be chosen. Analyzing the influence of each attribute on the utility, according to the variation of the values in the sample, DO and DD have higher influence on the utility than TC.

The influence of the degree of use of port facilities has been specified as non-linear by means of coefficients for each category. The results are similar for all the models, BCML values are shown in Figure 2. They embody the conflict between economies of agglomeration, scale and network effects with the level of use of the facilities. It can be said that the attractiveness of the port is conditioned by the degree of use of its facilities and that a threshold can be identified beyond which the attractiveness of the port diminishes. For a discussion on this topic see Martínez-Pardo et al. (2018).

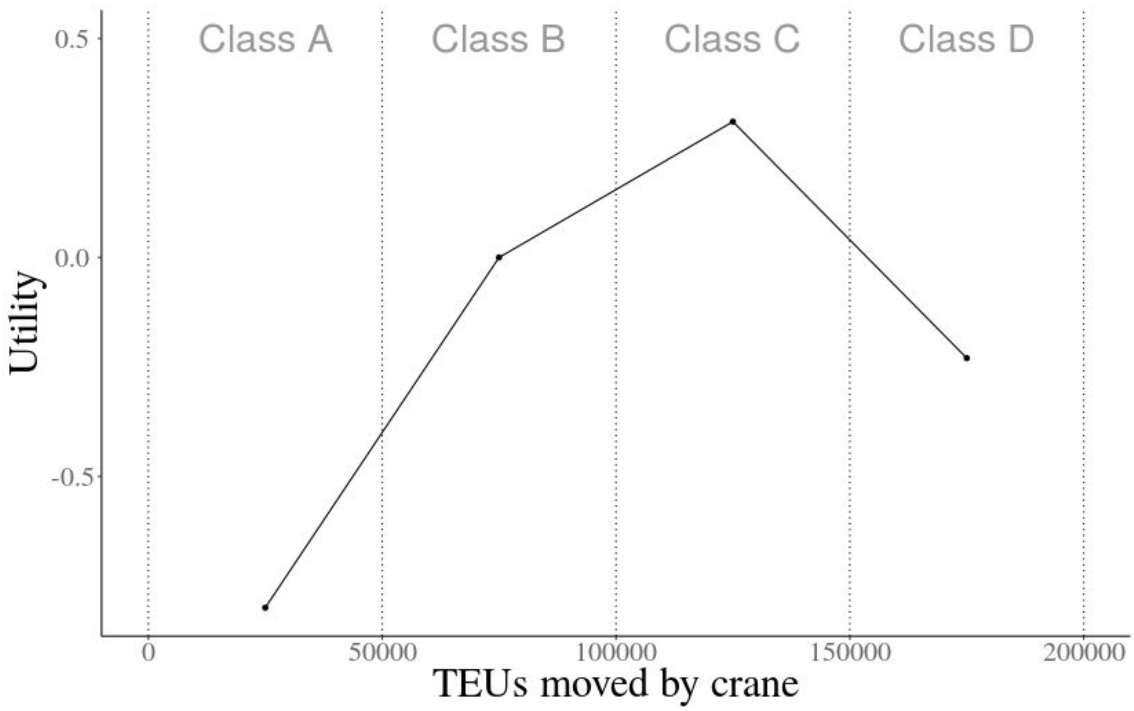

Figure 2. Level of use of port facilities vs utility of the port (BCML).

4.2. Estimated random heterogeneity and non-linearity

The ML model shows that the influence of the DO and DD variables has a significant heterogeneity in the sample. σ_{DO} and σ_{DD} are significantly different from 0 and have values in the same order of magnitude as the means of the distributions.

Nevertheless, it is possible that a confounding effect between non-linearity influence of the attributes and heterogeneity in the preferences of the decision-makers is present.

In fact, if the real influence of a variable on the utility function is non-linear and a linear model with random coefficients is specified, this coefficient will appear as randomly distributed in the port-choice population. However, this randomness responds to an incorrect specification; that coefficient should really be interpreted as a systematic variation among situations of choice based on the value of the variable (non-linearity) and not as a variation distributed in the population (random coefficients). That is, if reality were non-linear with constant coefficients, all individuals would have the same coefficient for the same value of the variable; in other words, they would value that attribute equally. However, the same individual would give different importance to a unit of that attribute depending on the total value. Analysis of this problem can be found in Orro, Novales and Benitez (2005, 2010).

To check if the values of σ_{DO} and σ_{DD} obtained in the ML model are due to heterogeneity in tastes and not due to the non-linearity of a variable, the results obtained by the BCL model should be observed. The transformation exponents (λ_{DO} and λ_{DO}) are significantly different from the unit. Therefore it seems clear that non-linear influence cannot be disregarded. With respect to BCML, it can be seen that all of them are strongly significant, and this confirms the presence of both. So, the coefficients of *DO* and *DD* are random parameters distributed across choice-makers and those variables have a non-linear influence on the utility of the port alternatives.

4.3 Models goodness of fit

Table 6 summarizes model fitting information for the four models estimated. Note that the log-likelihoods are simulated in ML and BCML. Two ratio likelihood indexes are used: ρ^2 and adjusted $\bar{\rho}^2$ (see Ortúzar and Willumsen (2011)). In this case, it is valid to say that the model with the higher ρ^2 or $\bar{\rho}^2$ fits the data better. The values of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are also reported. The most appropriate model is that which provides the minimum value of these measures.

Model	LL/(S)LL	ρ^2	$\overline{\rho}^2$	AIC	BIC	Parameters	LR	DF	χ^2_{99}	Models compared
null	-1386294.361	-	-	-	-	-	-	-	-	-
cte	-872135.382	0.371	-	-		3	-1028317.957	3	11.340	cte - null
MNL	-404618.334	0.708	0.536	809228.668	404673.596	8	-935034.097	5	16.810	MNL - cte
ML	-372633.149	0.731	0.573	745256.298	372702.227	10	-63970.370	2	9.210	ML - MNL
BC	-368663.485	0.734	0.577	737316.970	368732.563	10	-71909.698	2	9.210	BC - MNL
BCML	-353425.287	0.745	0.595	706838.574	353508.180	12	-102386.094	4	13.280	BCML - MNL
							-38415.724	2	9.210	BCML - ML
							-30476.396	2	9.210	BCML - BC

Table 6: Models goodness of fit comparison

At convergence, the lowest log-likelihood and the best fit in terms of ρ^2, $\bar{\rho}^2$, AIC and BIC is obtained by BCML. This shows that taking into account both the heterogeneity in preferences and the non-linear influence of the attributes, it supposes a significantly better adjustment to the data used for the calibration. The likelihood ratio test (LR, Ben-Akiva and Lerman, 1985) comparison also shows that models which include non-linearity and/or random heterogeneity strongly surpass their counterparts without any of these characteristics.

4.4. Model validation

The validation procedure is in essence a process to ensure that the choices observed in a sample (other than that those on which the model was estimated) are consistent with the probabilities predicted by the fitted or calibrated model. In this paper, 16 validation samples were checked. They are random sub-samples of 25000 observations not used in the estimation. The choice probabilities obtained in validation samples were compared with the results of the estimation sample by using the "average probability of correct prediction" (APCP) as defined in Başar and Bhat (2004). The APCP is the average of the probability given by the model for the chosen alternative across the sample. APCP allows to check if the models have been over-fitted to the estimation data (e.g. Hess, 2005: 181). The results shown in table 8 support that none of the models are over-adjusted. Values of APCP are quite similar in all cases, but the BCML is always the model that gives highest probabilities in average to the alternative really chosen for the shipments, so it fits better from this disaggregated point of view. In all the samples, BCML is followed by ML, BCL and MNL, so it can be observed that random heterogeneity increases APCP more than BC transformation for this case, although both of them improve the fit.

Sample Type	MNL	ML	BCL	BCML
Estimation	0,7811	0,8021	0,7973	0,8100
Validation 1	0,7807	0,8017	0,7970	0,8093
Validation 2	0,7833	0,8045	0,7995	0,8125
Validation 3	0,7830	0,8037	0,7992	0,8114
Validation 4	0,7825	0,8034	0,7990	0,8113
Validation 5	0,7830	0,8035	0,7990	0,8115
Validation 6	0,7833	0,8044	0,7996	0,8123
Validation 7	0,7839	0,8049	0,7995	0,8125
Validation 8	0,7826	0,8038	0,7991	0,8119
Validation 9	0,7817	0,8030	0,7977	0,8107
Validation 10	0,7816	0,8022	0,7972	0,8095
Validation 11	0,7830	0,8041	0,7982	0,8112
Validation 12	0,7804	0,8015	0,7962	0,8090
Validation 13	0,7792	0,8001	0,7957	0,8082
Validation 14	0,7823	0,8032	0,7988	0,8114
Validation 15	0,7815	0,8024	0,7982	0,8107
Validation 16	0,7838	0,8047	0,8004	0,8129

Table 7: Average probability of correct prediction for estimation and validation.

A Chi-Squared goodness of fit test was also used to compare the observed number of individuals choosing each alternative for all years and the estimated number according to the fitted model. The results are shown in table 8, with detail for the first validation sample and only the Chi-Squared value for the others. The critical value at 95% for 3 degrees of freedom is 7.82, so all models yield predictions that are consistent with the data in all validation samples.

			Predicted values validation sample 1							
Port Alternative	Observed values		MNL		ML		BCL		BCML	
Algeciras	1406,00	5,62%	1422,24	5,69%	1481,36	5,93%	1431,90	5,73%	1446,48	5,79%
Barcelona	13074,00	52,30%	13033,20	52,13%	13013,40	52,05%	13038,70	52,15%	13076,60	52,31%
Bilbao	12,00	0,05%	7,69	0,03%	13,65	0,05%	7,60	0,03%	14,36	0,06%
Valencia	10508,00	42,03%	10536,80	42,15%	10491,60	41,97%	10521,80	42,09%	10462,60	41,85%
Total	25000,00		24999,93		25000,01		25000,00		25000,04	
		χ^2								
		Sample 1	2,81		4,34		3,13		1,72	
		Sample 2	1,81		2,14		2,66		2,82	
		Sample 3	3,21		3,30		3,40		1,77	
		Sample 4	0,55		7,13		0,53		6,02	
		Sample 5	5,22		0,87		5,50		0,50	
		Sample 6	0,15		3,48		0,11		2,69	
		Sample 7	1,16		2,42		1,01		1,47	
		Sample 8	0,83		6,54		0,83		4,57	
		Sample 9	7,05		1,34		7,05		0,48	
		Sample 10	3,57		0,60		4,52		2,38	
		Sample 11	1,21		5,60		1,03		4,54	
		Sample 12	5,30		7,49		4,95		2,71	
		Sample 13	2,94		0,86		2,62		1,28	
		Sample 14	1,65		3,03		1,98		0,76	
		Sample 15	6,86		6,21		7,35		2,41	
		Sample 16	2,56		1,23		2,84		1,36	

Table 8: Comparison of the predicted and the observed values for the validation sample 1 and Chi-Squared test for all samples

It is well known that even a MNL model with only constants can reproduce market shares in the estimation sample. As random samples for all years have similar quotas, table 8 cannot be used for selection of the best approach.

Ten sets of new random validation samples were taken with data not used for estimation. Each sample has 25000 observations for each of the years from 2004 to 2012. The objective is to test the capability for prognosis of the changes of port quota over the years of the different models. For example, Barcelona share ranges from 44% to 57% and Valencia from 37% to 50% along the years and the samples. The annual base is used because it is when the port variable included in the model changes. In reality, response to changes in the degree of use of port facilities will be gradual. For each set, port, model and year the difference between the number of shipments forecasted and the actual number of shipments is calculated. The Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE) across ports and years, for each set and model are reported in table 9. In the case of RMSE, ML clearly performs better than MNL in all sets, and both BCL and BCML are clearly better, but similar between themselves. When analyzing MAE, the order is the same, although difference between ML and MNL is smaller. However, the difference between BCL and BCML is a little bigger and BCML performs better in seven out of ten cases. Hence, it could be concluded that non-linearity seems to be more important than random heterogeneity when market shares are predicted in this case.

	Validation set										
	1	2	3	4	5	6	7	8	9	10	Mean
RMSE MNL	205.0	218.8	212.4	209.5	217.0	207.2	202.7	206.6	215.4	222.3	211.7
RMSE ML	191.8	208.4	202.0	198.3	203.9	194.5	194.3	194.6	201.5	208.2	199.8
RMSE BCL	152.8	165.3	162.3	158.3	160.0	155.9	151.8	154.4	166.5	168.9	159.6
RMSE BCML	153.4	165.9	160.7	160.3	158.7	156.4	154.0	154.1	165.2	163.9	159.3
MAE MNL	138.0	145.3	142.8	146.3	143.3	143.4	143.4	140.4	148.0	153.7	144.5
MAE ML	137.1	148.1	142.8	146.9	143.1	139.5	140.9	140.3	146.1	151.8	143.6
MAE BCL	105.3	114.1	114.1	117.5	113.5	113.2	112.0	110.7	118.1	124.1	114.3
MAE BCML	106.8	114.2	111.8	114.6	115.9	112.7	108.4	108.6	115.9	116.6	112.5

Table 9: Number of shipments for port and year error measures

5. Implications

Once the presence of non-linearity and random heterogeneity in this case study is confirmed, two key questions arise: what are the consequences of ignoring both circumstances when applying the model to new situations? Are there significant differences in the results obtained?

Although differences among the models in terms of fitting are not great in general terms (see Table 6), the probability of choice of each alternative would be slightly underestimated or overestimated depending on whether the prediction is made or not with the BCML model versus other models.

A hypothetical situations analysis is performed to quantify the differences of using BCML, ML, BCL or MNL for the prognosis at shipment level. The port probability of choice of six hypothetical shipments is analyzed. Firstly, the baseline situation is defined: two provinces of origin and three different countries of destination are selected. Secondly, there are two situations to see how changing the degree of use of port facilities affects the probabilities of choice of each port. Finally, the results of the four models are compared for each of the situations.

The selected provinces of origin are Madrid and Zaragoza, which are the main inland provinces of origin of the database and have similar road distances to more than one of the ports considered. The countries of destination chosen are United States, Brazil and China. These countries are among the ten main countries of destination of the case study and have a sufficiently different geographical position to be able to perceive clearly if there is a different behavior in the shipments. All the variables are calculated for each one of the shipments.

The results for all situations are shown in Table 10: by rows, the shipments; by columns, the probabilities of choice of each port, for each of the four calibrated models. In the initial situation, for the four models, it can be seen that while for shipments made from Madrid (I, II and III) there is a strong preference for Valencia, if the shipment is made from Zaragoza (IV, V and VI) the preference is for Barcelona. Considering the country of destination, for the MNL, BCL and BCML, the probability of Barcelona increases for shipments to China and Algeciras, Bilbao and Valencia for shipments made to the United States or Brazil. For ML, the exceptions are Valencia and Bilbao, whose probability of choice increases for shipments to China made from Madrid or Zaragoza.

The situation (A) presents, ceteris paribus, an increase in traffic in the Port of Valencia that would increase the degree of use of its port facilities (TC moves from class C to class D) above the congestion threshold shown in Figure 2. The results confirm that the probability of Valencia decreases in favor of the other alternatives for next year's shipments. In this case, the use of the BCML model in comparison with the rest shows differences in percentage close to 10 %.

In the second proposed situation (B), there is a decrease in traffic in the port of Valencia that would lead to a reduction in the degree of use of its port facilities (TC moves from class C to class A). Less traffic implies lower impact of the economies of agglomeration and scale or network effects, so the probability of choosing the port of Valencia for future shipments would decrease and, in this case, the decrease would be more marked than in situation (A). For this situation, the difference in the forecast of the probability of choice thrown by the BCML model with respect to the other models reaches absolute differences of 18 % for the most unfavorable case; ML for the probability of choice Barcelona or Valencia for Madrid – China. In Barcelona, there is a change from 20.39% to

38.54%, which is an 89% increase in relative terms. With respect to the usually employed MNL, the absolute difference of taking into account non-linearity and random heterogeneity reaches 14% in several cases.

Initial situation

		Prob Algeciras				Prob Barcelona				Prob Bilbao				Prob Valencia			
		MNL	ML	BCL	BCML	MNL	ML	BCL	BCML	MNL	ML	BCL	BCML	MNL	ML	BCL	BCML
I	Madrid - USA	6.40%	6.25%	5.35%	2.95%	13.91%	9.96%	20.66%	17.16%	0.04%	0.00%	0.03%	0.00%	79.65%	83.79%	73.97%	79.89%
II	Madrid - China	2.41%	1.51%	3.82%	1.60%	19.18%	12.33%	22.92%	20.05%	0.01%	0.15%	0.01%	0.00%	78.41%	86.02%	73.24%	78.35%
III	Madrid - Brasil	6.41%	5.71%	5.86%	3.40%	13.91%	10.40%	20.03%	16.68%	0.02%	0.00%	0.02%	0.00%	79.66%	83.89%	74.09%	79.91%
IV	Zaragoza - USA	0.31%	1.64%	0.72%	0.88%	57.41%	62.57%	56.10%	62.96%	0.03%	0.00%	0.02%	0.00%	42.26%	35.79%	43.15%	36.16%
V	Zaragoza - China	0.09%	0.98%	0.49%	0.47%	65.49%	73.17%	59.00%	67.43%	0.00%	0.04%	0.01%	0.00%	34.41%	25.80%	40.50%	32.10%
VI	Zaragoza - Brasil	0.31%	1.37%	0.81%	1.35%	57.41%	62.44%	55.27%	61.67%	0.02%	0.00%	0.02%	0.00%	42.27%	36.19%	43.91%	36.97%
		MNL -BCML	ML -BCML	BC -BCML		MNL -BCML	ML -BCML	BC -BCML		MNL -BCML	ML -BCML	BC -BCML		MNL -BCML	ML -BCML	BC -BCML	
I	Madrid - USA	3%	3%	2%		-3%	-7%	3%		0%	0%	0%		0%	4%	-6%	
II	Madrid - China	1%	0%	2%		-1%	-8%	3%		0%	0%	0%		0%	8%	-5%	
III	Madrid - Brasil	3%	2%	2%		-3%	-6%	3%		0%	0%	0%		0%	4%	-6%	
IV	Zaragoza - USA	-1%	1%	0%		-6%	0%	-7%		0%	0%	0%		6%	0%	7%	
V	Zaragoza - China	0%	1%	0%		-2%	6%	-8%		0%	0%	0%		2%	-6%	8%	
VI	Zaragoza - Brasil	-1%	0%	-1%		-4%	1%	-6%		0%	0%	0%		5%	-1%	7%	

Situation (A): increase in the degree of use of the facilities of the port of Valencia

		Prob Algeciras				Prob Barcelona				Prob Bilbao				Prob Valencia			
		MNL	ML	BCL	BCML	MNL	ML	BCL	BCML	MNL	ML	BCL	BCML	MNL	ML	BCL	BCML
I	Madrid - USA	11.56%	9.28%	7.69%	4.22%	25.10%	19.28%	29.66%	24.22%	0.07%	0.00%	0.04%	0.00%	63.28%	71.44%	62.61%	71.55%
II	Madrid - China	4.29%	2.43%	5.47%	2.12%	34.19%	22.98%	32.77%	28.15%	0.01%	0.10%	0.02%	0.00%	61.51%	74.49%	61.74%	69.74%
III	Madrid - Brasil	11.56%	10.09%	8.42%	4.89%	25.11%	19.20%	28.79%	23.65%	0.04%	0.00%	0.03%	0.00%	63.29%	70.71%	62.76%	71.45%
IV	Zaragoza - USA	0.40%	3.09%	0.88%	1.18%	75.20%	78.53%	68.18%	73.79%	0.03%	0.00%	0.03%	0.00%	24.37%	18.39%	30.91%	25.03%
V	Zaragoza - China	0.12%	0.52%	0.59%	0.53%	81.12%	87.62%	70.77%	77.69%	0.00%	0.15%	0.01%	0.00%	18.76%	11.71%	28.63%	21.78%
VI	Zaragoza - Brasil	0.40%	2.79%	0.98%	1.30%	75.21%	78.50%	67.42%	72.87%	0.02%	0.00%	0.02%	0.00%	24.37%	18.70%	31.58%	25.83%
		MNL -BCML	ML -BCML	BC -BCML		MNL -BCML	ML -BCML	BC -BCML		MNL -BCML	ML -BCML	BC -BCML		MNL -BCML	ML -BCML	BC -BCML	
I	Madrid - USA	7%	5%	3%		1%	-5%	5%		0%	0%	0%		-8%	0%	-9%	
II	Madrid - China	2%	0%	3%		6%	-5%	5%		0%	0%	0%		-8%	5%	-8%	
III	Madrid - Brasil	7%	5%	4%		1%	-4%	5%		0%	0%	0%		-8%	-1%	-9%	
IV	Zaragoza - USA	-1%	2%	0%		1%	5%	-6%		0%	0%	0%		-1%	-7%	6%	
V	Zaragoza - China	0%	0%	0%		3%	10%	-7%		0%	0%	0%		-3%	-10%	7%	
VI	Zaragoza - Brasil	-1%	1%	0%		2%	6%	-5%		0%	0%	0%		-1%	-7%	6%	

Situation (B): decrease of use of the facilities of the port of Valencia

		Prob Algeciras				Prob Barcelona				Prob Bilbao				Prob Valencia			
		MNL	ML	BCL	BCML	MNL	ML	BCL	BCML	MNL	ML	BCL	BCML	MNL	ML	BCL	BCML
I	Madrid - USA	9.23%	8.84%	7.68%	5.69%	20.05%	17.32%	29.65%	33.84%	0.05%	0.00%	0.04%	0.00%	70.67%	73.84%	62.62%	60.48%
II	Madrid - China	3.44%	2.67%	5.47%	3.04%	27.46%	20.49%	32.77%	38.54%	0.01%	0.04%	0.02%	0.00%	69.09%	76.81%	61.75%	58.42%
III	Madrid - Brasil	9.23%	9.31%	8.42%	6.72%	20.05%	16.31%	28.78%	32.59%	0.03%	0.00%	0.03%	0.00%	70.69%	74.38%	62.77%	60.69%
IV	Zaragoza - USA	0.37%	2.91%	0.88%	1.18%	68.54%	76.40%	68.17%	82.97%	0.03%	0.00%	0.03%	0.00%	31.06%	20.69%	30.92%	15.85%
V	Zaragoza - China	0.11%	0.83%	0.59%	0.84%	75.48%	85.74%	70.76%	85.61%	0.00%	0.09%	0.01%	0.00%	24.41%	13.34%	28.64%	13.54%
VI	Zaragoza - Brasil	0.37%	1.96%	0.98%	1.47%	68.55%	77.00%	67.41%	82.18%	0.02%	0.00%	0.02%	0.00%	31.07%	21.03%	31.59%	16.35%
		MNL -BCML	ML -BCML	BC -BCML		MNL -BCML	ML -BCML	BC -BCML		MNL -BCML	ML -BCML	BC -BCML		MNL -BCML	ML -BCML	BC -BCML	
I	Madrid - USA	4%	3%	2%		-14%	-17%	-4%		0%	0%	0%		10%	13%	2%	
II	Madrid - China	0%	0%	2%		-11%	-18%	-6%		0%	0%	0%		11%	18%	3%	
III	Madrid - Brasil	3%	3%	2%		-13%	-16%	-4%		0%	0%	0%		10%	14%	2%	
IV	Zaragoza - USA	-1%	2%	0%		-14%	-7%	-15%		0%	0%	0%		15%	5%	15%	
V	Zaragoza - China	-1%	0%	0%		-10%	0%	-15%		0%	0%	0%		11%	0%	15%	
VI	Zaragoza - Brasil	-1%	0%	0%		-14%	-5%	-15%		0%	0%	0%		15%	5%	15%	

Table 10: Hypothetical situations analysis

6. Conclusions

The use of current available public records for discrete choice models allows a great quantity of data, with advantages of continuity and replicability. Nevertheless, this kind of data presents a lack of information of decision-makers. It does not allow using systematic variations among its preferences. In that context, the use of random heterogeneity through random coefficient approaches can be an adequate way to take into consideration these differences.

For the case study of port choice in Spain, the most important variables obtained in the choice of port for container exports are distance by road (DO) and length of maritime route (DD). The degree of use of port facilities (TC) has also appeared as significant. The presence of non-linearity in the influence of variables like DO, DD and TC has been shown. As it has been previously proved that random heterogeneity can be confused with non-linearity, it is recommended that both of the characteristics be taken into account simultaneously. The proposed Box-Cox Mixed Logit (BCML) approach allows it for estimated non-linearity. In that case, there is heterogeneity in preferences in the population in the importance of DD and DO. The BCML model presents a significant better fit in comparison with Box-Cox Logit, Mixed Logit and Multinomial Logit. The BCML model was also validated and it was concluded that the model is not over-adjusted and that its predictions are consistent with the data. Joint specification of preference heterogeneity and non-linearity allows obtaining better fit to individual choices and market shares in the validation samples. Preference heterogeneity seems more important at shipment level and non-linearity at market share level.

The choice of model is relevant for policy-makers. It has been shown with several hypothetical situations that the use of BCML can suppose considerable differences in the estimated probability of choice of a port for a shipment, as great an increase, in relative terms, of more than 3 quarters in the chance of Barcelona for a Madrid – China shipment in relation with the estimation with the ML model.

This paper has established that the consideration of the presence of non-linearity and preference heterogeneity could be important for port choice models. A methodology for simultaneous estimation of both characteristics is presented and applied for a case study in Spain. These results are potentially useful for future research and policy decisions. It would be interesting to carry out similar studies in other regions and contexts. Alternative approaches for taking into account individual heterogeneity, like Latent Class Models or even Latent Class Mixed Logit (Greene and Hensher, 2012) can also be applied to this problem.

Acknowledgements

Orro and Garcia-Alonso acknowledge, respectively, the financial support provided by the Government of Spain (MCIU/AEI/FEDER, UE) under the projects RTI2018-097924-B-I00 and PGC2018-097965-B-I00. The authors wish to thank four anonymous reviewers for their helpful comments and suggestions.

References

Anderson, C. M., Opaluch, J. J. & Grigalunas, T. A. (2009). The Demand for Import Services at US Container Ports. Maritime Economics & Logistics, 11(2), 156–185. doi: 10.1057/mel.2009.4

Başar, G. & Bhat, C. (2004) A parameterized consideration set model for airport choice: an application to the San Francisco Bay Area. Transportation Research Part B 38, 889–904

Ben-Akiva, M., Bolduc ,D. & Park, J.Q. (2008), Discrete Choice Analysis of Shippers' Preferences, in Ben-Akiva, M., Meersman, H. & van de Voorde, E. (ed.) Recent Developments in Transport Modelling, pp.135 - 155

Ben-Akiva,M. & Lerman, S. S. (1985). Discrete Choice Analysis: Theory and Application to Travel Demand. Cambridge, Massachusetts: MIT Press.

Bierlaire. (2016). Estimating choice models with latent variables with Python Biogeme (Tech. Rep. No. TRANSP-OR 160628). Transport and Mobility Laboratory, ENAC, EPFL.

Blonigen, B. A. & Wilson, W. W. (2006). International trade, transportation networks, and port choice. In American economic association annual meeting. Retrieved from http://www.corpsnets.us/docs/PortDevInternalTransport/PortChoice114.pdf

Box, G. E. P. & Cox, D. R. (1964). An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological), 26(2), 211–252.

Cantillo, J., Cantillo, V. and Arellana, J. (2018). Modelling with joint choice of ports and countries of origin and destination: application to Colombian ports. Maritime Policy and Management, 45(6), 720-738.

China Ports. (2014). China Ports Yearbook (In Chinese). Shanghai, China: Chinese Port Magazine Publisher.

Dagenais, M.G., Dufour, J.-M. (1994). Pitfalls of Rescaling Regression Models with Box-Cox Transformations. The Review of Economics and Statistics 76, 3, 571-75, August.

DIAN. (2012). Registro de importación y exportación de la Dirección Nacional de Impuestos y Aduanas Nacionales. Retrieved from [2016-03-01]http://www.dian.gov.co/

Escamilla-Navarro, L., García-Menéndez, L., Pérez-García, E. (2010). Integration of foreign trade and maritime transport statistics in Spain. Maritime Policy & Management 37(4), 347–375. doi:10.1080/03088839.2010.486641

Garcia-Alonso, L. & Sanchez-Soriano, J. (2009). Port selection from a hinterland perspective. Maritime Economics & Logistics, 11(3), 260–269. doi: 10.1057/mel.2009.9

García-Alonso, L., Monios, J. & Vallejo-Pinto, J.A. (2019). Port competition through hinterland accessibility: the case of Spain. Maritime Economics & Logistics, 21(2), 258-277.

García-Alonso, L. & Marquez; M.A. (2017). The role of the market potential in the port choice process: A case study. International Journal of Transport Economics, 44(3), 381-397. Doi: 10.19272/201706703002.

Garrow, L. A. (2010). Discrete Choice Modelling and Air Travel Demand: Theory and Applications. Ashgate Publishing, Ltd.

Gaudry,M. J. & Wills,M. J. (1978). Estimating the functional form of travel demand models. Transportation Research, 12(4), 257–289.

Greene,W. H. & Hensher, D. A. (2003) A latent class model for discrete choice analysis: contrasts with mixed logit. Transportation Research Part B 37, 681–698 DOI: 10.1016/S0191-2615(02)00046-2

Greene,W. H. & Hensher, D. A. (2013) Revealing additional dimensions of preference heterogeneity in a latent class mixed multinomial logit model, Applied Economics, 45:14, 1897-1902, DOI: 10.1080/00036846.2011.650325

Hensher, D. A. & Greene,W. H. (2003). The Mixed Logit model: The state of practice. Transportation, 30(2), 133–176.

Hess, S. (2005). Advanced discrete choice models with applications to transport demand (Doctoral dissertation). Retrieved from http://www.cts.cv.ic.ac.uk/documents/theses/HessPhD.pdf

Hess, S., Bierlaire, M. & Polak, J. W. (2005). Estimation of value of travel-time savings using mixed logit models. Transportation Research Part A: Policy and Practice, 39(2-3), 221–236. doi: 10.1016/j.tra.2004.09.007

Hess, S., Train, K. E. & Polak, J. W. (2006). On the use of a Modified Latin Hypercube Sampling (MLHS) method in the estimation of a Mixed Logit Model for vehicle choice. Transportation Research Part B: Methodological, 40(2), 147–163. doi: 10.1016/j.trb.2004.10.005

Kashiha, M., Thill, J.-C. & Depken, C. A. (2016). Shipping route choice across geographies: Coastal vs. landlocked countries. Transportation Research Part E: Logistics and Transportation Review, 91, 1–14. doi: 10.1016/j.tre.2016.03.012

Lapparent, M., Frei, A. and Axhausen . K.W. (2009) Choice of mode for long distance travel. Current SP-based models from three European countries. Arbeitsberichte Verkehrs- und Raumplanung Volume 579. European Transport Conference 2009 (ETC 2009), Noordwijkerhout, The Netherlands, October 5-7, 2009. https://doi.org/10.3929/ethz-a-005888713

Lawrence, C., Zhou, J. & Tits, A. (1997). User's guide for CFSQP version 2.5: A C code for solving (large scale) constrained nonlinear (minimax) optimization problems, generating iterates satisfying all inequality. Technical Report TR-94-16r constraints (Tech. Rep.). College Park: Institute for Systems Research, University of Maryland.

Malchow, M. B. & Kanafani, A. (2001). A disaggregate analysis of factors influencing port selection. Maritime Policy & Management, 28(3), 265–277.

Malchow, M. B. & Kanafani, A. (2004). A disaggregate analysis of port selection. Transportation Research. Part E, 40(4), 317–337.

Martínez Moya, J. & Feo Valero, M. (2016). Port choice in container market: a literature review. Transport Reviews, 1–22.

Martínez-Pardo, A., Garcia-Alonso, L. & Orro, A. (2018). The role of the degree of use of the facilities in the port choice process: the Spanish dockside cranes case. International Journal of Shipping and Transport Logistics 10 (5/6), 514–532. DOI: 10.1504/IJSTL.2018.10016513

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In Frontiers in econometrics (pp. 105–142).

McFadden, D. & Train, K. (2000). Mixed MNL models for discrete response. Journal of Applied Econometrics, 15(5), 447 – 470.

Meersman, H., Van De Voorde, E. & Vanelslander, T. (2010). Port Competition Revisited. Review of Business and Economic Literature, 55(2), 210–233. Retrieved from http://econpapers.repec.org/RePEc:sen:rebelj:v:55:i:2:y:2010:p:210-233

Mennis, E., Platis, A., Lagoudis, I. & Nikitakos, N. (2008). Improving port container terminal efficiency with the use of Markov theory. Maritime Economics and Logistics, 10(3), 243–257.

Nir, A.-S., Lin, K. & Liang, G.-S. (2003). Port choice behaviour–from the perspective of the shipper. Maritime Policy & Management, 30(2), 165–173. doi: 10.1080/0308883032000069262

Nugroho, M. T., Whiteing, A. & de Jong, G. (2016). Port and inland mode choice from the exporters' and forwarders' perspectives: Case study — Java, Indonesia. Research in Transportation Business &Management. doi: 10.1016/j.rtbm.2016.03.010

Orro, A. (2005). Modelos de elección discreta en transportes con coeficientes aleatorios. (Doctoral dissertation, Universidad de A Coruña). Retrieved from http://caminos.udc.es/ferrotrans/orro/documentos/tesis_orro_final.pdf

Orro, A., Novales, M. and Benítez, F.G. (2007) Systematic and random variations of the value of travel time savings in discrete choice models. 11th World Conference on Transport Research, Berkeley CA, United States, 2007-6-24 to 2007-6-28

Orro, A., Novales, M. & Benitez, F. G. (2005). Nonlinearity and Taste Heterogeneity Influence on Discrete Choice Models Forecasts. In European transport conference, 2005.

Orro, A., Novales, M. & Benitez, F. G. (2010). "Box-Cox Mixed Logit Model dor Travel Behaviour Analysis". Orro, A., Novales, M. y Benítez, AIP Conf. Proc. 1281, 679 (2010); DOI: 10.1063/1.3498568

Ortúzar, J. d. D. & Willumsen, L. G. (2011). Modelling Transport (Fourth ed.). Chichester, UK: JohnWiley & Sons, Ltd.

Paixão, A., Carvalho, S. & Oliveira, M. (2010). Port choice in the European short sea shipping market: the view of point of port authorities. In Iame 2010 annual conference proceedings.

Pallis, A. A., Vitsounis, T. K. & De Langen, P.W. (2010). Port Economics, Policy and Management: Review of an Emerging Research Field. Transport Reviews, 30(1), 115–161. doi: 10.1080/01441640902843208

Razzolinim, T. (2010) Study on labour supply when tax evasion is an option with Box–Cox functional forms and random parameters, Applied Economics, 42:28, 3669-3684, DOI: 10.1080/00036840802243797

Román, C., Arencibia, A. I. & Feo-Valero, M. (2017). A latent class model with attribute cut-offs to analyze modal choice for freight transport. Transportation Research Part A: Policy and Practice, 102, 212–227. doi: 10.1016/j.tra.2016.10.020

Sanchez, R. J., Ng, A. K. Y. & Garcia-Alonso, L. (2011). Port Selection Factors and Attractiveness: The Service Provider's Perspective. Transportation Journal, 50(2). Retrieved from http://trid.trb.org/view.aspx?id=1106509

SCS. (2014). Spanish Customs Statistics. Retrieved from [2016-03-01]http://www.agenciatributaria.es

SeaRates. (2015). Freight Exchange platform. Retrieved from [2016-03-01]www.searates.com

Spanish Port Authority. (2018). Traffic Statistics and Statistical yearbooks of the Spanish Port Authority. Retrieved from [2016-03-01]http://www.puertos.es

Spitzer, J.J. (1984). Variance Estimates in Models with the Box-Cox Transformation: Implications for Estimation and Hypothesis Testing. The Review of Economics and Statistics 66, 645-652.

Steven, A. B. & Corsi, T.M. (2012). Choosing a port: An analysis of containerized imports into the US. Transportation Research Part E: Logistics and Transportation Review, 48(4), 881–895.

Tavasszy, L., Minderhoud, M., Perrin, J.-F. & Notteboom, T. (2011). A strategic network choice model for global container flows: specification, estimation and application. Journal of Transport Geography, 19(6), 1163–1172. doi: 10.1016/j.jtrangeo.2011.05.005

Tiwari, P., Itoh, H. & Doi, M. (2003). Shippers' Port and Carrier Selection Behaviour in China: A Discrete Choice Analysis. Maritime Economics & Logistics, 5(1), 23–39. doi: 10.1057/palgrave.mel.9100062

Tongzon, J. L. & Sawant, L. (2007). Port choice in a competitive environment: from the shipping lines' perspective. Applied Economics, 39(4), 477–492. doi: 10.1080/00036840500438871

Train, K. (2009). Discrete choice methods with simulation (2nd ed.). Cambridge University Press.

Train K. & Sonnier G. (2005) Mixed Logit with Bounded Distributions of Correlated Partworths. In: Scarpa R., Alberini A. (eds) Applications of Simulation Methods in Environmental and Resource Economics. The Economics of Non-Market Goods and Resources, vol 6. Springer, Dordrecht)

Vega, L. A., Cantillo, V. M. & Arellana, J. A. (2019). Assessing the impact of major infrastructure projects on port choice decision: The Colombian case. Transportation Research Part A: Policy and Practice (120), 132-148. doi: 10.1016/j.tra.2018.12.021

Veldman, S., Garcia-Alonso, L. & Liu, M. (2015). Testing port choice models using physical and monetary data: a comparative case study for the Spanish container trades. Maritime Policy &Management, 1–14.

Veldman, S., Garcia-Alonso, L. & Vallejo-Pinto, J. A. (2011). Determinants of container port choice in Spain. Maritime Policy & Management, 38(5), 509–522. doi: 10.1080/03088839.2011.597450

Veldman, S., Garcia-Alonso, L. & Vallejo-Pinto, J. A. (2013). A port choice model with logit models: a case study for the Spanish container trade. International Journal of Shipping and Transport Logistics, 5(4/5), 373.

Veldman, S. & Gopkalo, O. (2011). A Model on Port Basin Choice with Russian Container Trade. Proceedings of the International Forum on Shipping, Ports and Airports (IFSPA) 2010.

Veldman, S. J. & Bückmann, E.H. (2003). A Model on Container Port Competition: An Application for the West European Container Hub-Ports. Maritime Economics & Logistics, 5(1), 3–22.

Veldman, S. J., Bückmann, E. H. & Saitua, R. N. (2005). River Depth and Container Port Market Shares: The Impact of Deepening the Scheldt River on the West European Container Hub-Port Market Shares. Maritime Economics & Logistics, 7(4), 336–355. doi: 10.1057/palgrave.mel.9100142

Yang, J., Wang, G.W. & Li, K. X. (2016). Port choice strategies for container carriers in China: a case study of the Bohai Bay Rim port cluster. International Journal of Shipping and Transport Logistics, 8(2), 129.