NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Evgeniy M. Ozhegov, Alina Ozhegova

REGRESSION TREE MODEL FOR ANALYSIS OF DEMAND WITH HETEROGENEITY AND CENSORSHIP

BASIC RESEARCH PROGRAM
WORKING PAPERS

SERIES: ECONOMICS
WP BRP 174/EC/2017

Regression tree model for analysis of demand with heterogeneity and censorship

Evgeniy M. Ozhegov [†], Alina Ozhegova [††]

Abstract

In this research we analyze new approach for prediction of demand. In the studied market of performing arts the observed demand is limited by capacity of the house. Then one needs to account for demand censorhip to obtain unbiased estimates of demand funnction parameters. The presence of consumer segments with different purposes of going to the theatre and willingness-to-pay for performance and ticket characteristics causes a heterogeneity in theatre demand. We propose an estimator for prediction of demand that accounts for both demand censorhip and preferences heterogeneity. The estimator is based on the idea of classification and regression trees and bagging prediction aggregation extended for prediction of censored data. Our algorithm predicts and combines predictions for both discrete and continuous parts of censored data. We show that our estimator performs better in terms of prediction accuracy compared with estimators which accounts either for censorship, or heterogeneity only. The proposed approach is helpful for finding product segments and optimal price setting.

Keywords: demand, performing arts, machine learning, regression tree, censored data, pricing.

JEL codes: Z11, C53, D12.

[†]National Research University Higher School of Economics (Perm, Russia). Research fellow, Group for Applied Markets and Enterprises Studies. E-mail: tos600@gmail.com

[††]National Research University Higher School of Economics (Perm, Russia). Young research fellow, Group for Applied Markets and Enterprises Studies. E-mail: arbuzanakova@gmail.com

1 Introduction

Currently, firms, households and society as a whole generate, collect and store enormous volume of data starting from level of individuals to level of countries. Availability of large data sets in turn opens the way for modelling consumer behavior for retail, banking and other industries to improve their business decisions. Analysis based on big data also matters to marketing, since customers provide lots of information that may be potentially useful for marketing decisions. Up to date analysis problems have been restricted by standard econometric models. Development of computer science techniques and large volumes of data result in the implementation of machine learning techniques in solving of current business and marketing problems.

In terms of marketing and business decisions pricing is a crucial factor for company on the way to increasing profits. Effective pricing strategy helps the company to determine the price point at which it maximizes profits on sales of products. When the product is differentiated, that is possesses the set of attributes, the seller should understand customer's preferences according to each of attribute. Researchers of demand analyze customer's preferences using willingness-to-pay, that demonstrate added value of each attribute. Pricing strategy development also supposes customer sesitivity to a price change. Knowledge about how sensitive are customers to price change contributes to a better understanding of customer behavior.

Availability of large data sets about customers' purchases in various industries permits to estimate demand function. The results of estimation allow to provide recommendations about significance of particular product attributes, especially price. Preceding papers study demand primarily on aggregated data (Schimmelpfennig, 1997; Levy-Garboua and Montmarquette, 1996; Lange and Luksetich, 1984; Pommerehne and Kirchgassner, 1987; Throsby, 1994). Models on aggregated data allow to draw an inference about averaged consumer. These conclusions may be useful for general recommendation. In the case of heterogeneous consumers, detailed results would be more valuable. Most recent studies employ disaggregated data about customer behavior. Detailed consumer choice data allow to account for heterogeneity of consumers. This in turn leads to more concrete recommendations and more precise forecasting of demand using results for product or consumer segments.

For proper modelling of demand function there is an issue of demand censorship that should be discussed in details. Whereas there are various measures of demand, such as revenue, quantity of product sold or percent of product sold, the value of demand is at least limited by zero on the left side. In the case of limited capacity when the seller cannot supply a product over particular amount per unit time, the demand is limited by maximum stock level on the right side. The majority of early studies do not take into account the censored character of demand, in many cases the authors ignore the fact of censorship. Dropping the limited nature of demand leads to biased estimate and incorrect inferences about demand effects.

Majority of studied dedicated to demand estimation employ the models of econometrics. Econometric methods have gained popularity among researchers of demand for a number of reasons. Indeed, econometric methods prove themselves in estimation of models on a few number of observations. When a large amount of observations is hard to obtain, advanced methods are powerless, while econometric methods allow to estimate elementary models on small datasets. Econometric techniques are also realized in various statistical packages, that make them accessible for broad application. Significant advantage of econometric estimation is interpretability of results. The estimates of effects in regression equations as a rule may be red without additional transformations. Concerning the censorship econometric techniques have significantly progressed in consistent estimation of regressions on limited data. There are traditional methods dealing with censorship of dependent variable (Tobin, 1958; Heckman, 1977) as well as modern semiparametric or nonparametric approaches.

Econometric models possess some drawbacks that restrict their use in some cases. In most cases techniques of econometricians assume homogeneity of studied objects. This leads to averaged estimates for a whole population and inability to divide the population into segments according to different economic effects. Still some econometric models account for heterogeneity of effect, but they require to set the source of heterogeneity in advance. Parametric econometric methods also need distributional assumptions on dependent variable or error term that restrict their applicability. Over the past decades, there has been a high level of interest among researchers of demand in practical using of machine learning (ML) methods. Generally, machine learning methods demonstrate high quality in forecasting of demand where the number of observations allows consistent estimation of model. Furthermore, machine learning methods account for objects heterogeneity without *a priori* assumptions on the sources. That is, the model is absolved from incorrect hypotheses about the sources of heterogeneity. At the same time, these methods possess higher rate of convergence compared to nonparametric econometric models that make them highly sought when data sets contain large number of predictors.

While the methods of machine learning cope with heterogeneity of consumers, they do not account for censorship of demand. Since the censored data is very common in tasks of demand estimation, there is a need to develop an algorithm that would use the principles of machine learning methods and take into account censorship of data. In this research, we adapt ML prediction methods to censored demand data. The features of developed algorithm for demand estimation are absence of *a priori* distributional assumptions as well as assumptions on sources of heterogeneity and account for data censorship.

We develop the method and apply it to a censored demand estimation problem in a market of performing arts. We use data on tickets sales that are taken from the Perm Opera and Ballet Theatre, which is considered as one the best regional opera theatre in Russia. The data cover all performances for four seasons between August 2011 and July 2015 and include information on ticket purchase and performance characteristics. Structure

of data disaggregated to the level of particular pricing area in a house allows to control on quality of seat as well. We use performance (production type, composer, band director etc.), play (month, day of a week, time of a day, premiere play) and seat (seating area dummies) characteristics to predict attendance rate and study variables importance. The number of observations is 2682, the unit of observation presents demand for a particular seating area in a house on a particular performance. Since the prediction of demand presents an important problem for theatre management, the results of this research allow to propose recommendations upon price differentiation over seats and performances.

We find relatively good performance of our algorithm in terms of predictive power compared with parametric methods (OLS and median regression), methods (Tobit model and censored quantile regression) which account for data censorship but not for heterogeneity and method (quantile regression tree) which account for heterogeneity only. ML methods allow to reveal the most relevant variables, that explain essential part of heterogeneity in demand. We find that the most relevant variables are type of production (ballet or opera), seating area, nationality of composer (Russian or foreign), world fame of production and band director. We estimate the distribution of price effect comparing the predictions of attendance with current prices and prices increased by 10%. The price elasticity varies from 0 to -0.30 with a median equal to -0.07 that indicates on weakly elastic demand. We find that price elasticity of demand varies substantially with less elastic demand for ballets, Russian ballets among ballets and foreign operas among operas, and seats in the center of stalls.

2 Literature review

2.1 Approaches for demand prediction

Machine learning methods divide the forecasting problem into problems of variables and models selection and identification of similar objects groups. There are two common types of variables to be predicted. Categorical variables express belonging of an object to a certain class from discrete set. Continuous variables reflect a quantitative measure of object state. Meanwhile, models of social and economic predictions require statistical methods dealing with discrete-continuous variables such as censored or limited dependent variables. Censored data often arise in the models of individual consumption, where consumers either do not demand the good (zero consumption) or demonstrate positive amount of consumption. Since the consumption is left censored by zero, the data of this kind consist of discrete (choice of zero or non-zero consumption) and continuous (choice of amount when consumption is non-zero) components. Models of product demand with limited capacity are also suffered from the problem of censored data. Since the seller cannot supply a good over particular amount per unit time, the demand is right-censored by maximum stock level. In such a case, the potential demand, the amount of good that consumers are willing to purchase, may exceed

the observed demand, the amount of good that the consumers are willing to purchase and the seller is capable to supply. In the former case ignoring censored nature of data results in biased prediction of consumption. Model calibration on uncensored observations only allows to correctly estimate the change in consumption but fails to predict transition to the group of consumers that demonstrate zero consumption. In relation to the latter, traditional methods ignoring the fact of censorship lead to underestimated effects and biased prediction of consumption. Inaccurate estimation causes non optimal pricing policy and loss of expected gain.

Within the context of demand estimation, it is crucial to account for demand heterogeneity, that arises from differentiated goods with a variety of characteristics and consumers with different preferences. Model of demand that does not take into account customer and product heterogeneity tends to estimate the effects and predict the consumption for an averaged good. Modelling the heterogeneity allows to detect the differences in customer preferences towards good characteristics, to reveal willingness-to-pay for different goods and to adapt pricing policy to certain product and consumer segments.

Econometric methods for applied demand research progress in consistent estimation of regressions on censored data. Traditional methods of limited dependent variable (LDV) estimation (Tobin, 1958; Heckman, 1977) are based on distributional assumptions of dependent variable or error term. This approach is sensitive to the choice of distributional assumption. At the same time, the lack of tests on assumption validity limits the accuracy of results. Modern nonparametric extensions of LDV models (Das, Newey & Vella, 2003; Matzkin, 2012) relax distributional assumptions. However, nonparametric estimation with several independent variables leads to computational burden and slow rate of convergence that result in practical limitation on the number of explanatory variables and partial linearizing of a model. Semiparametric approach of censored quantile regression (Chernozhukov, Hong, 2002; Chernozhukov, Fernandez-Val & Kowalski, 2015) also allows to model demand on censored data without distributional assumption. Model estimation on different levels of quantile is a convenient way to account for heterogeneity of effects. Meanwhile, this approach is not suitable for prediction goals, since it requires the value of quantile for estimation of effects, that is unobservable in out-of-sample data.

Over the past decade, econometricians have demonstrated an interest in using methods of machine learning in the field of demand estimation. This interest has been partially provoked by the necessity to deal with large consumer datasets and requirement in more precise estimation. There is a number of studies where the authors bridge the gap between standard deterministic approach and methods of machine learning (Belloni, Chernozhukov & Hansen, 2014; Bajari, Nekipelov, Ryan & Yang, 2015a, 2015b; Athey, Imbens, 2016). In the paper (Bajari, Nekipelov, Ryan & Yang, 2015a) the authors demonstrate several popular machine learning methods to the problem of demand estimation and compare them with standard approaches. They show that the methods from machine learning possess superior

predictive power compared to standard approaches. Our motivation in using machine learning approach is encouraged by the possibility to consider the heterogeneity of effects without ad-hoc assumptions about sources of preferences differences. One of the ML techniques based on the principle is classification and regression trees (CART) (Breiman *et al.*, 1984). CART and its extensions (Breiman, 1996, 2001) are gradually spread among econometricians and even now are widely used in prediction models of heterogeneous demand (Bajari *et al.*, 2015). The core of subject consists of partitioning the characteristic space into a series of hyper-cubes and model calibration for each of those partitions. Thus, the algorithms of considered methods allow to account for heterogeneity in effects of demand without underlying assumptions about its sources. In this project we develop the CART approach by accounting for censored nature of demand also.

2.2 Review of performing arts market studies

A growing body of empirical studies is devoted to the estimation of demand for performing arts. The majority of research is concerned with the issue of price elasticity, since there is still a question, whether art is a luxury good. This raises the topics of audience sensitivity to price change and an absence of close substitutes to the art goods. Estimation of demand function is complicated by the specificity of good. The art is generally described as a merit good, that is confirmed by public policy directed to maintenance of its production and consumption (Towse, 2011). Experience nature of art implies that the taste is acquired over time with exposure. The theory of learning by consuming states that consumer possesses the taste for arts, that she reveals during the process of consumption (Levy-Garboua & Montmarquette, 1996). The study of demand becomes more involved due to heterogeneous tastes of consumers.

There has been an extensive work concerning empirical estimation of demand function (Seaman, 2006). We give an overview of those papers, that are most closely related to our research. Earlier papers model the demand mainly as function of price (Moore, 1966). Particular discussion in the literature is dedicated to the issue of quality assessment in the demand model (Throsby, 1990; Abbe-Decarroux, 1994; Withers, 1980). In earlier papers quality is accounted as constant in the model of demand (Hansmann, 1981). Following Lancaster theory (Lancaster, 1966) Throsby *et al.* (1983) propose variables such as repertory classification, standard of performance and production, the author of production and design to account for quality. Abbe-Decarroux (1994) extends the set of quality measures including reputation of play through reviews, reputation of director and playwright. The authors conclude that the perception of quality ex-ante is an important determinant for consumer that seeking information before ticket purchase.

Since pricing is one of the most important issues in marketing, most of empirical studies consider the issue of price elasticity, that is motivated by managerial concern about appropriate

price levels. Earlier performing arts research based on aggregated data show that demand is generally inelastic by price (Moore, 1966; Houthakker & Taylor, 1970; Touchstone, 1980; Gapinski, 1984; Bonato, Gagliardi & Gorelli, 1990). Price inelasticity is also more prominent for researches on aggregated data. Consumer theory suggests that the heterogeneity of consumer tastes should also be considered when modelling a price elasticity. More sophisticated studies based on disaggregated data demonstrate different elasticity indicators for subgroups of the population (Levy-Garboua & Montmarquette, 1996; Lange & Luksetich, 1984). In the paper (Pommerehne & Kirchgassner, 1987) the authors reveal lower price elasticity for consumers with higher income. Price elasticity also may vary for different seats in a theater (Schimmelpfennig, 1997). Demand for seats in the stalls, the circle and the back-end of the tiered stalls is elastic, whereas the demand is inelastic in the central part of the tiered stalls. Throsby (1994) found demand for higher arts to be less elastic relative to immediately accessible. Given the importance of price elasticity modelling on disaggregated data, it is worthwhile to notice that it expects prior assumptions about sources of heterogeneity. This challenge in the case of high-dimensional data with huge number of variables, when looping through an array, becomes intricate. Manual search also may lead to inaccurate selection of variables that cause the heterogeneity in effects.

In the context of demand modelling, we should discuss the problem of censorship. The demand equation is a relation between the volume of tickets purchased and tickets prices and performance characteristics. Demand can be measured by the number of tickets sold per performance, per unit of time or by the percent of theatre house occupancy. The majority of early studies based on aggregated data are not taking into account the censored character of demand. In this case, the number of tickets sold for the performance is the only observed demand, while potential demand may exceed the capacity of a house. Dropping the distinction between potential and observed demand may affect the estimates of parameters and lead to estimates bias. In early papers authors employ the simplest approaches dealing with censored data: to ignore the fact of data censorship or to exclude the censored observations from the sample. Apparently, neither of these approaches are effective. The former approach leads to biased estimates of parameters and, consequently, inaccurate inferences. The latter lose information about the observations beyond the censoring bound. Some previous papers include inflexible capacity in a model of demand as explanatory variable. More sophisticated studies employ various statistical techniques, that are generally known as demand untruncation or uncensoring methods. Tobin (1958) became one of the earliest who proposed a method dealing with censored data problem. The author developed a model of relationship between a non-negative dependent variable and vector of independent variables that is currently known as Tobit model. There have been several studies that extended the model for different cases of dependent variable censorship. All proposed variations of Tobit model require an assumption on distribution of dependent latent variable (often use normal distribution), that clearly may be violated. Heckman model that in a manner became an extension of Tobit allows to deal

with censoring data using two-step model. On the first step one models the probability to be censored for each observation. The second step allows to estimate the latent dependent variable on a data on uncensored observations. The model again requires an assumption about joint distribution of error terms from both steps. In the paper McGill (1995) the censorship of dependent variable is handled by an adaptation of expectation-maximization method (EM). In general, EM method is computationally demanding. In addition, the method assumes the underlying demand distribution that may apply a restriction on its appropriateness. A number of studies propose various deterministic approaches to inferring latent demand. Lee (1990) proposes to employ a model that assume a Poisson distribution of demand. Modelling an airline demand Wickham (1995) employs a method of booking curve detruncation. In the paper (Liu, Smith, Orkin & Carey, 2002) the authors develop parametric regression models to estimate censored demand using Poisson, Weibull, exponential and normal distributions. Considered approaches develop new techniques and propell forward the issue of modelling censored data. At the same time, these methods have underlying assumptions about the distribution of latent demand, that may diverge from the true data generating process and lead to inconsistent estimation of effects. In addition to the distributional assumption, modelling of heterogeneity requires prior assumptions about sources of observations' heterogeneity. Censored quantile regression (CQR) solves the problem of data censorship in the modelling of demand and assumes heterogeneity in effects at various levels of demand quantiles. However, this gives a small notion on preferences heterogeneity that may be useful for marketing and pricing purposes.

3 Data

The data for research are taken from the Perm Opera and Ballet Theatre, which is considered as one the best regional opera theatre in Russia. It is famous for its modern musical productions, nonstandard classical performances, and unconventional festival projects. It is also a major Russian center for opera and ballet, where the quality of the musical performance is paramount. Every year the theatre performs forty regular productions and three to five new productions.

The Perm Opera and Ballet Theatre is a non-commercial organization and as such is loss-making. Its main source of funding is a Perm state budget. As a non-commercial venture the goal of the theatre is to make ballet and symphonic art available for Perm residents. The theatre does have to, at least partially, recoup the expenses with production revenue in order to produce new ones. Consequently, the theatre constantly tries to balance between being affordable and covering costs using pricing mechanism and charging different prices for different performances and seats.

The data collected cover all performances for four seasons between August 2011 and July 2015. There were 298 performances out of 36 repertoire productions at the main venue. The

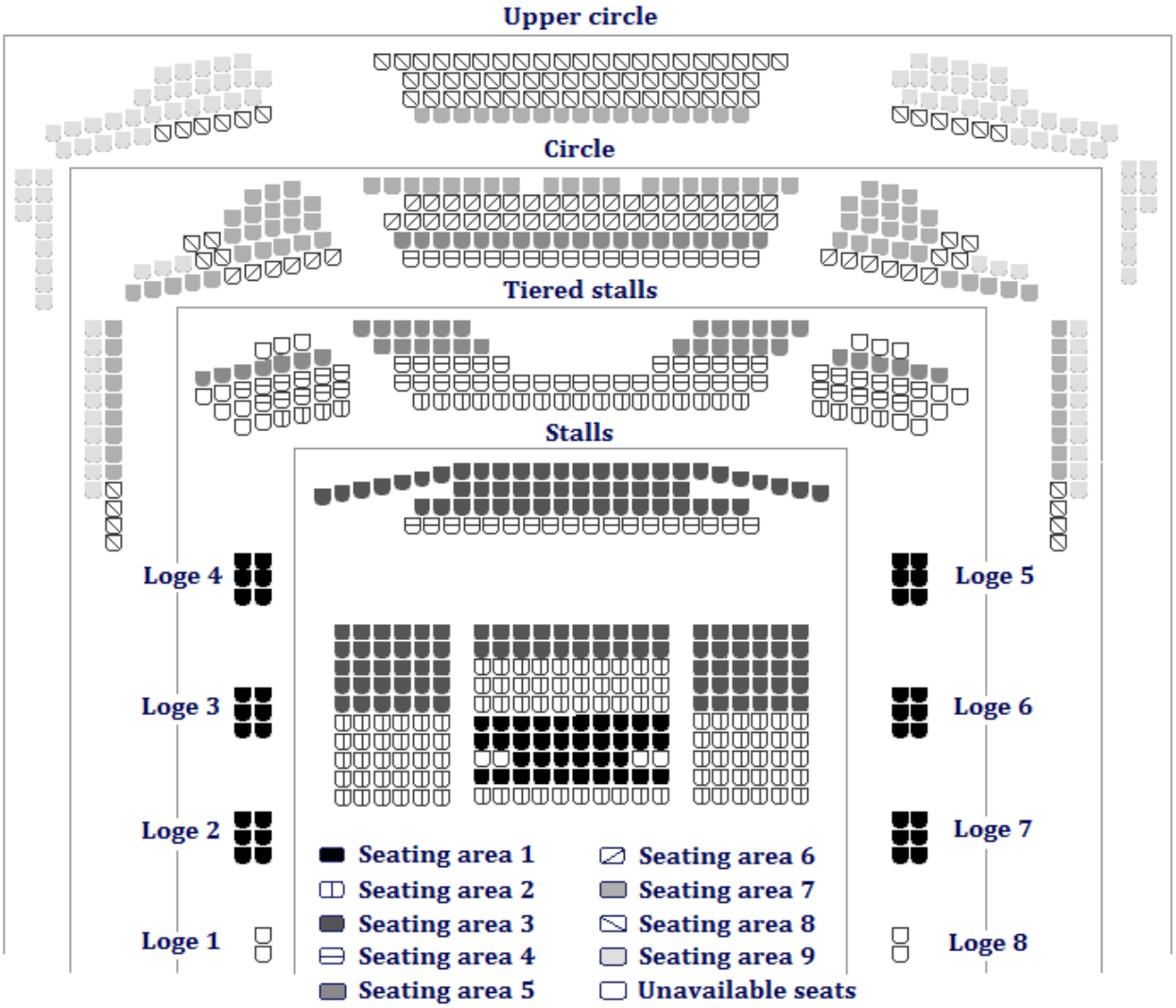

Fig. 1. The scheme of the house

data include information on the name of production, the date and time of play (season, year, month, the day of week and time of day), the price of a ticket, time and date of ticket purchase and the location of a seat in a house. The house of the theatre is divided into sectors: loges, the stalls, tiered stalls, the circle and the upper circle. In the sectors, the seats are identified by row and place. Further, the house is divided into nine seating areas according to the distance from the stage (Figure 1).

The seats in different seating areas vary by the quality of view and sound, prestige and price. Whereas the seats located in one seating area are considered as homogeneous in terms of price and quality. The theatre also has a system of discounts for special segments of the population (students, students of the ballet school, retired people). Thus, for every ticket purchased we have information on the basic price charged by the theatre and on the actual price of a sale with discount. We use only the basic price of the ticket as a measure of the price considering that the administration of the theatre manages the basic price and remains the system of discounts.

In addition to the information provided by the theatre, we collected information on performance characteristics which explains the demand according to previous research (Corning and Levy, 2002; Seaman, 2006). We classify productions into operas and ballets, into classical (written before 1900) and modern (written after 1900) ones. We collect information on the composer and construct dummy responsible for the nationality of the composer (Russian/foreign) and the dummy on whether the production is a premiere one. We classified performances according to the age recommended for attendance: children (without restriction), family (12+) and adult (16+). Information on conductors allows estimating the contribution of a particular person. Among conductors, we identified two persons that are especially successful and in-demand. Perm Opera and Ballet Theatre has been regularly nominated for the prestigious Russian theatre award "Golden Mask". For each production, we collect information on the number of nominations and awards won. In order to measure the world popularity of musical composition, we collect the data on various ratings. We use data from the worldwide rating of operas and their composers (operabase.com) and of ballets (listverse.com). Descriptive statistics of performances characteristics are presented in Table 1.

To estimate the model of demand, we aggregate data on sales and prices by seating areas. For each seating area we calculate the attendance rate as a number of sold tickets to the total number of seats in the area and assign the basic price in accordance with one of 8 theatre pricing schemes. The pricing scheme is the set of prices for 9 seating areas. Prices for the most expensive tickets (the first seating area) vary from 300 to 2000 rubles while the cheapest tickets (the ninth seating area) are always sold for 100 rubles.

Apart from the seats in the house, the productions may also be heterogeneous. Figure 2 shows that half of the observations are filled over 80%. The remaining seating areas show lower demand which tells us about the heterogeneity of productions.

One more issue to be discussed is a potentially different quality of seats for different types of productions. Seats are heterogeneous in terms of view and sound quality which are not ordered strictly according to seating area number (and price of a ticket). Thus, seats closer to stage are not the best to watch a ballet since the level of stalls is lower than a level of the stage. Theatre experts' opinion is that the best seats for watching a ballet are located in the center of circle which corresponds to fourth to sixth seating areas. This is supported by the data on attendance of performances and seats disaggregated by production type (Table 2). The most filled areas at ballets are areas 4-7 while for operas the most filled areas are 2-4. This corresponds to a higher quality of sound in this areas and higher importance of sound quality in operas compared to ballets. The quality of seat in terms of the view and sound quality should be also taken into account in a model of demand with an attention to potentially different seats quality estimate for various production types. It also may result in different estimates of price elasticity over types of production and seats since willingness-to-pay for a particular seat associated with its quality may vary over operas and ballets.

Tab. 1. Descriptive statistics

Variable	Total	Share
Day of week	2682	
Working days	1440	46.3
Weekend	1242	53.7
Time of day	2682	
Before 2 am	342	12.8
After 2 am	2340	87.2
Type of performance	2682	
Ballet	954	35.6
Opera	1728	64.4
World rating of performance	2682	
Rated	1017	37.9
Not rated	1665	62.1
Language of opera	2682	
Foreign	378	14.1
Russian	2304	85.9
Recommended age	2682	
Without restrictions	1107	41.3
From 12 y.o.	1170	43.6
From 16 y.o.	405	15.1
Awards	2682	
Presence	144	5.4
Absence	2538	94.6
The nationality of composer	2682	
Russian	1521	56.7
Foreign	1161	43.3
Band director	2682	
Valeriy Platonov	1494	55.7
Teodor Currentzis	279	10.4
Others	909	33.9

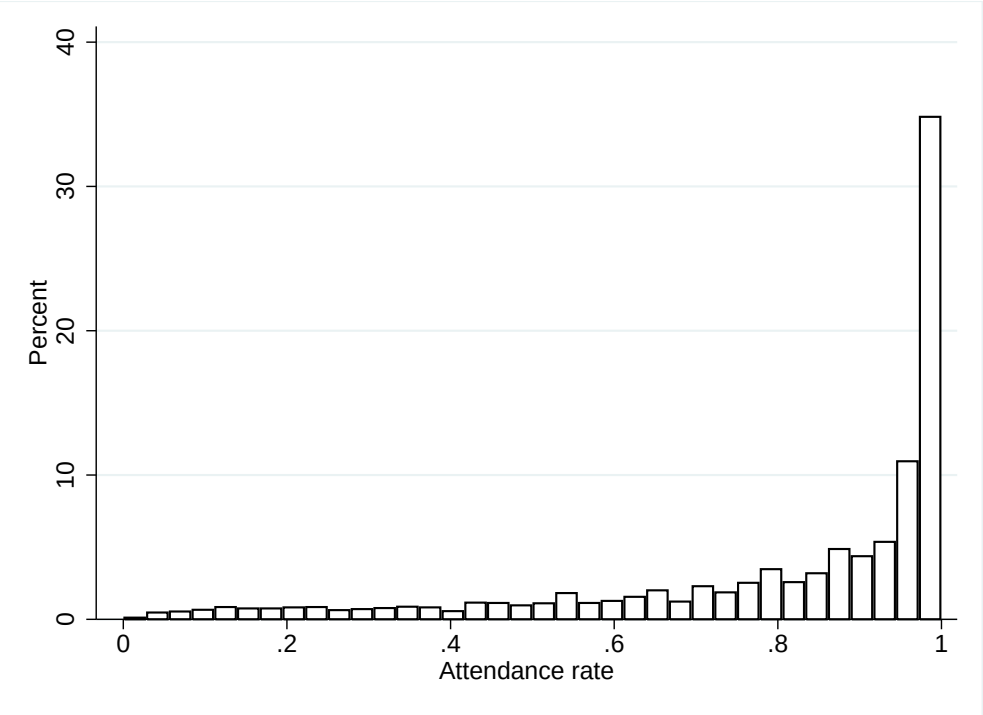

Fig. 2. The distribution of attendance

Tab. 2. Descriptive statistics for attendance rate

Variable	Operas	Ballets	Total
Attendance rate	0.72	0.96	0.80
Attendance (area 1)	0.83	0.92	0.85
Attendance (area 2)	0.87	0.96	0.89
Attendance (area 3)	0.85	0.96	0.89
Attendance (area 4)	0.86	0.97	0.90
Attendance (area 5)	0.76	0.98	0.84
Attendance (area 6)	0.68	0.98	0.80
Attendance (area 7)	0.52	0.97	0.70
Attendance (area 8)	0.46	0.95	0.65
Attendance (area 9)	0.64	0.87	0.72

4 Methodology

A regression tree is a collection of rules that determine the parameters values of a regression function. Tree-based methods partition the characteristic space into a series of hyper-cubes, and fits effects to each partition depend on the value of right-hand side variables, X. Trees are characterized by a hierarchical series of nodes, with a decision rule associated at each node. Following (Bajari et al., 2015), define a pair of half-planes:

$$R_1(j;s) = X|X_j \leq s \tag{1}$$
$$R_2(j;s) = X|X_j > s$$

where j indexing a splitting variable and s is a split point (threshold). Starting with the base node at the top of the tree, the rule for that node is formed by the following optimization problem:

$$\min_{j,s}[\min_{\theta_1} \sum_{i:x_i \in R_1(j,s)} L(y_i - f(x_i|\theta_1)) + \min_{\theta_2} \sum_{i:x_i \in R_2(j,s)} L(y_i - f(x_i|\theta_2))] \tag{2}$$

The inner optimization is solved by setting θ optimal according to prespecified loss function L and regression function f. For ordinary regression tree L is a squared function and f is a linear function of x with parameters θ. We employ a tree of median regressions setting L as an absolute deviation function to control for influential observations. In a classification problem, classification tree is built based on binary choice (probit) function f with loss L associated with errors in classification. We use ordinary classification accuracy measure, a number of misclassified observations, since relative importance of type I and II errors is not defined.

The outer optimization problem is a problem of finding an optimal splitting point s for each possible splitting variable and then choosing a variable x to split by. Once the splitting variable and point are found, the same procedure is then performed on each resulting partitioning, finally giving a partition of characteristics space.

In the limit, each value of $x \in X$ is assigned to value of $y = f(x)$, which is a perfect reconstruction of the underlying function f for in-sample prediction. In practice, we are interested in out-of-sample prediction. Therefore, the tree is expanded until a value of loss function for out-of-sample data falls. Often, tree is grown until a specific number of splits or a minimal number of observations in subsamples is achieved.

The literature has proposed several variations on the regression tree estimators to obtain "honest" prediction and predicted values robust to influential observations. One is bagging (Breiman, 1996), which uses resampling and model averaging to obtain a predictor. The idea is to sample the data with replacement B times, train a regression tree on each resampled set of data, and then predict the outcome at each x through a simple average of the predictions under each of the B trees. We use the same idea of resampling, taking each time random

subset of observations to train the model and predict the values of y for remain observations. Calibration of 200 regression trees with 75% of random observations for model training at each resample expectedly gives 50 out-of-sample predictions for each observation which we average to obtain "honest" prediction.

Since we have a problem of censored data prediction, we construct an algorithm for prediction of both discrete and continuous components of dependent variable x and combination of these predictions. An algorithm has following steps:

1. Construct dummy for observation censorship $d := I\{y = 1\}$.
2. Classify observations into censored and uncensored ($\hat{d} \in \{0; 1\}$) based on bagging prediction from classification (probit) trees and predict $\hat{p} = E[d|X]$.
3. Predict and trim y by censoring bound using median regression tree trained on classified as uncensored ($\hat{d} = 0$) data with $\hat{p}$ as predictor:

$$\hat{y} = \min\{Q_{y|X,\hat{p}}(0.5); 1\} \tag{3}$$

4. Combine the predictions of discrete ($\hat{d}$) and continuous ($\hat{y}$) components:

$$\hat{\hat{y}} = \begin{cases} 1, \hat{d} = 1 \\ \hat{y}, \hat{d} = 0 \end{cases} \tag{4}$$

5 Results

5.1 Models comparison

Firstly, we compare the predictive accuracy of the proposed estimator compared with parametric ones and one nonparametric estimator that accounts for heterogeneity only. We perform 4 parametric estimators and construct bagging predictions (average out-of-sample prediction from 200 models trained on 75% random subset of observations) similar to the proposed tree-based algorithm. Two parametric estimators (OLS and quantile regression) do not account for censorship and heterogeneity while two more (Tobit model and censored quantile regression) account for censorship only. We also construct quantlile regression tree algorithm that predicts attendance on the sample of both censored and uncensored data to control for heterogeneity but not for censorship. Our estimator (tree of censored quantile regressions) outperforms parametric estimators in terms of explained variance and prediction error. Quantile regression tree prediction has higher variance but lose to regression tree that account for censorship in terms of predictive accuracy (RMSE). Results presented in Table 3 show that given the data on theatre demand it is neccessary to account for demand censorship and heterogeneity.

Tab. 3. Prediction accuracy

	Mean	SD	R^2	RMSE	Min	Max
y (Attendance rate)	0.803	0.263			0.009	1
Model for $\hat{y}$:						
CQR Tree	0.813	0.201	0.588	0.052	0.089	1
CQR	0.823	0.171	0.422	0.083	0.209	1
Tobit	0.823	0.183	0.488	0.080	0.188	1
QR Tree	0.804	0.206	0.618	0.059	0.043	1
QR	0.842	0.121	0.212	0.098	0.459	1
OLS	0.793	0.159	0.370	0.110	0.310	1
Number of observations	2682					
Number of predictors	36					
Number of replications	200					

5.2 Findings for Perm Opera and Ballet Theatre

Next, we analyze the variables importance for trees grown calculating the share of partitions by a certain variable among all partitions in estimated trees. Importance of variables for data partition shows the main sources of heterogeneity in effects which matters for demand prediction. We separately calculate importance for growing trees for prediction of discrete ($\hat{d}$) and continuous ($\hat{y}$) parts of the data. Results for variables importance are presented in Table 4. Results vary for prediction of $\hat{d}$ and $\hat{y}$ since they are estimated on different subsamples of the data. A model for $\hat{d}$ is calibrated on the whole sample while a model for $\hat{y}$ is calibrated only on the expectedly uncensored observations. Results show that main sources of observations heterogeneity are those related to prestige of tickets (seating area, dummy for premeire play and dummy for play of "Golden Mask" nominee), content of performance (type of performance, world rating of the production, nationality of composer, language of opera singing and recommended age) and band director while time and day of performance give only small explanation of overall demand heterogeneity.

Among all of the heterogeneous effects on demand, we are mainly interested in studying the heterogeneity in price elasticity of demand. Price elasticity matters when we calculate willingness-to-pay for a certain ticket or performance characteristic as well as for a particular seating area in house. Price elasiticity estimation is also crucial when one need to find optimal price distribution among house and/or performances. To calculate price elasticity we construct two predictions of attendance, the first one is a prediction with current prices and the second one is a prediction with prices increased by 10% from the current level. We find the total elasticity range from -0.3 to 0, that corresponds to weak elasticity of demand (See Figure 3). Zero elasticity for a substantial share of observations indicates that for fully

Tab. 4. Variables importance

Variable	Share of splits $\hat{d}$	Share of splits $\hat{y}$	Total
Seating area	0.142	0.032	0.105
Premiere	0.080	0.105	0.088
Laureat of GM	0.244	0.050	0.180
Ballet	0.043	0.077	0.055
Rated opera	0.048	0.048	0.048
Rated ballet	0.095	0.153	0.114
Russian composer	0.085	0.052	0.074
Foreign language	0.043	0.087	0.058
Band director: Platonov	0.083	0.164	0.110
Band director: Currentzis	0.007	0.009	0.008
12+	0.082	0.098	0.087
16+	0.013	0.032	0.019
Evening	0.010	0.025	0.015
Friday	0.000	0.007	0.002
Saturday	0.002	0.025	0.010
Sunday	0.023	0.036	0.027
Number of trees	200	200	200
Number of splits	1204	1253	2457
Mean number of splits in a tree	6.0	6.1	6.0

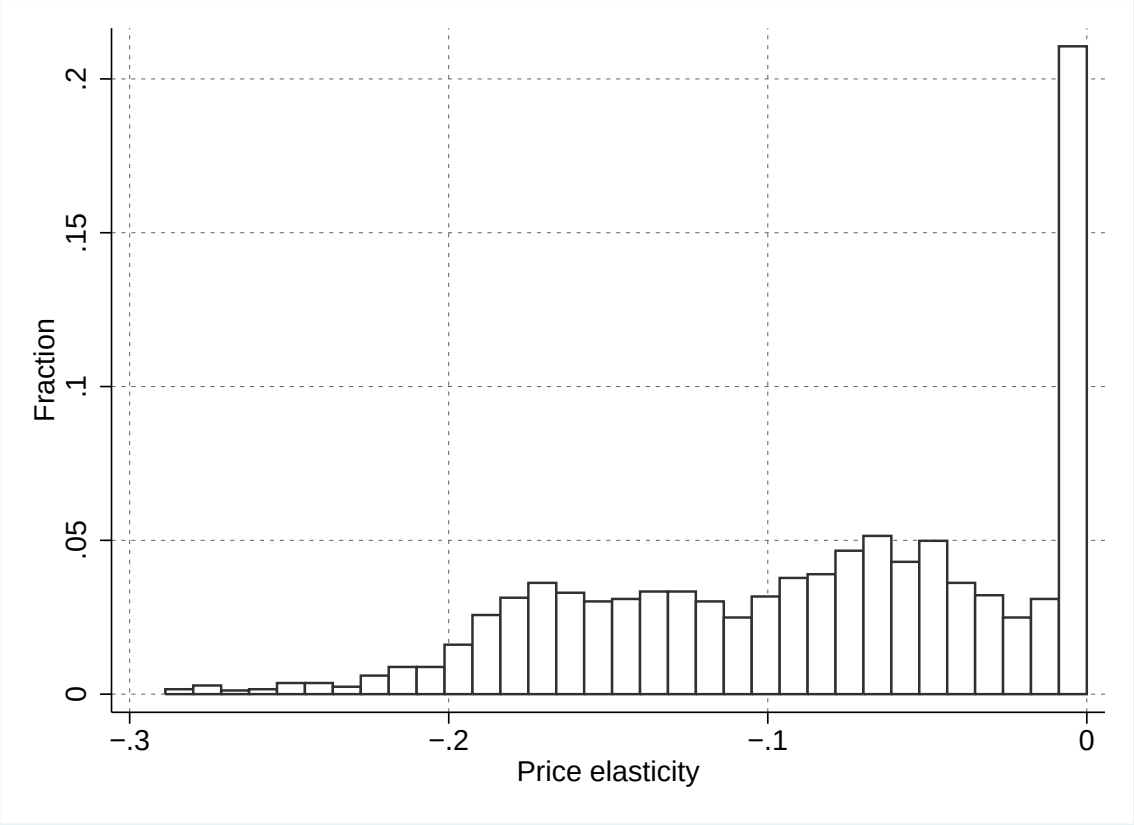

Fig. 3. The distribution of price elasticity

occupied seating areas the potential demand is significantly higher than capacity. Increase of the price for 10% will decrease the potential demand but not below the capacity level. Then the observed demand will remain on the same level of area capacity. The gap between a potential demand and area capacity sheds light for points of price optimization.

Given the relative importance of variables in explanation of overall demand heterogeneity, we aggregate the estimates of price elasticity over subsamples of the data (Table 5). Price elasticity substantially vary over types of production with less elastic demand for ballets. Among the seats in a house, the less elastic demand is in the seating areas with the highest quality of sound and prestige (2-4 seating areas). By the content of the performance we reveal less elastic demand for world famous productions, russian ballets and foreign operas performed on foreign language. Demand elasticity also varies by band directors with less elastic demand for performances conducted by the Theatre art director Theodor Currentzis, especially on weekend evenings. This results correspond to the exceptional status of performances conducted by Currentzis and higher willingness-to-pay for such theatre tickets.

5.3 Managerial insights

Our approach provides sellers with insights that can not be obtained from sales reports and market share data. Marketers may use results from regression tree to see how product segments are organized. Studying regression trees structure allows to find product subgroups

Tab. 5. Price elasticity estimates by subsamples

Variable	Operas	Ballets	Total
All seating areas	-0.107	-0.043	-0.086
Seating area 1	-0.100	-0.059	-0.086
Seating area 2	-0.097	-0.030	-0.074
Seating area 3	-0.098	-0.028	-0.074
Seating area 4	-0.100	-0.032	-0.076
Seating area 5	-0.109	-0.049	-0.088
Seating area 6	-0.116	-0.038	-0.089
Seating area 7	-0.123	-0.048	-0.101
Seating area 8	-0.111	-0.050	-0.092
Seating area 9	-0.113	-0.056	-0.096
Rated	-0.096	-0.033	
Non-rated	-0.112	-0.054	
Russian composer	-0.130	-0.038	-0.103
Foreign composer	-0.073	-0.048	-0.063
Russian language	-0.086		
Foreign language	-0.062		
Band director: Others	-0.111	-0.044	-0.087
Band director: Currentzis	-0.085	-0.016	-0.075
Other days	-0.087	-0.017	
Friday	-0.038		
Saturday		-0.004	

with similar predicted sales count, consumer sensitivity to price change and willingness-to-pay for certain product charactersitics. Subgroups may be found depending on the value of observed product characteristics. For example, in this research we apply our technique for finding theatre tickets subgroups with the different sensitivity of consumers to tickets price. Subgroups are found depending on algoritmically chosen important product characteristics. This approach is helpful when a marketer is not assuming any *a priori* criteria to segment products or when several criteria jointly form product segments.

Sellers may also use our approach of censored quantile regression trees to make better pricing decisions. Specifically, observing predicted volume of sales equal to the stock level, one may increase the price without loss of observed demand. In our example, the potential demand for tickets on a performance in a certain seating area may exceed a capacity of a seating area. Then we observe a demand equal to an area capacity. Price manager may increase a price but need to know an exact level of price increase to remain potential demand not less than a stock level. Our model of demand prediction distincts between observed and potential demand when makes a prediction about the level of observed demand. Having this distinction and unbiased estimates of price sensitivity allow to perform an analysis of seller's pricing strategy optimality.

6 Conclusion

In this research we analyze the demand for performing arts on the ticket sales data obtained from Perm Opera and Ballet Theatre. Data contain information on the attendance of seating areas for 298 performances played in 2011-2015. Since the observed demand is limited by capacity of the house and the third of seating areas are fully occupied, one needs to account for demand censorhip. The presence of consumer segments with different purposes of going to the theatre and willingness-to-pay for performance and ticket characteristics causes a heterogeneity in theatre demand.

We propose an estimator for prediction of demand that accounts for both demand censorhip and preferences heterogeneity. The estimator is based on the idea of classification and regression trees (CART) and bagging prediction aggregation. We extend CART for the problem censored dependent variable prediction. The algorithm consists of three steps: 1) Bagging prediction of dummy whether the dependent variable is on the censoring bound using classification trees; 2) Bagging prediction of dependent variable for observations classified as uncensored using median regression trees; 3) Trimming of second-step prediction and its combination with first-step prediction of censored observations.

We find relatively good performance of our algorithm in terms of predictive power compared with parametric methods (OLS and median regression), with parametric methods (Tobit model and censored quantile regression) which account for data censorship but not for heterogeneity as well with method (quantile regression tree) which accounts for heterogeneity

only but not for data censorship. We study the importance of variables for explanation of demand heterogeneity. The most frequent variables for splitting the sample on subsamples with different effects on demand are type of production (ballet or opera), seating area, nationality of composer (russian or foreign) and band director. We estimate the distribution of price effect comparing the predictions of attendance with current prices and prices increased by 10%. The price elasticity varies form -0.30 to 0 with a median equal to -0.07 that indicates on weakly elastic demand. We find that price elasticity of demand varies substantially with less elastic demand for ballets, Russian ballets among ballets and foreign operas among operas, and seats in the center of stalls. This result is useful for price optimization and product subgroups among different seats in a house and performances with different characteristics.

References

Abbe-Decarroux, F. (1994). The perception of quality and the demand for services: Empirical application to the performing arts. *Journal of Economic Behavior & Organization*, 23(1), 99-107.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353-7360.

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015a). Machine learning methods for demand estimation. *The American Economic Review*, 105(5), 481-485.

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015b). *Demand estimation with machine learning and model combination*. National Bureau of Economic Research (No. w20955).

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 608-650.

Bonato, L., Gagliardi, F. & Gorelli, S. (1990). The demand for live performing arts in Italy. *Journal of Cultural Economics*, 14(2), 41-52.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2), 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Chernozhukov, V., & Hong, H. (2002). Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association*, 97(459), 872-882.

Chernozhukov, V., Fernandez-Val, I., & Kowalski, A. E. (2015). Quantile regression with censoring and endogeneity. *Journal of Econometrics*, 186(1), 201-221.

Corning, J., & Levy, A. (2002). Demand for live theater with market segmentation and seasonality. *Journal of Cultural Economics*, 26(3), 217-235.

Das, M., Newey, W. K., & Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 70(1), 33-58.

Gapinski, J. H. (1984). The economics of performing Shakespeare. *The American Economic*

Review, 74(3), 458-466.

Hansmann, H. (1981). Nonprofit enterprise in the performing arts. *The Bell Journal of Economics*, 341-361.

Heckman, J. (1977). Sample selection bias as a specification error. *Econometrica*.

Houthakker, H. S. & Taylor, L. D. (1970). *Consumer Demand in the United States*. Harvard University Press, Cambridge.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy*, 74(2), 132-157.

Lange, M. D. & Luksetich, W. A. (1984). Demand elasticities for symphony orchestras. *Journal of Cultural Economics*, 8(1), 29-47.

Lee, A. (1990). *Airline Reservations Forecasting*. Erewhon, NC: Prentice-Hall.

Levy-Garboua, L. & Montmarquette, C. (1996). A microeconometric study of theatre demand. *Journal of Cultural Economics*, 20(1), 25-50.

Liu, P. H., Smith, S., Orkin, E. B., & Carey, G. (2002). Estimating unconstrained hotel demand based on censored booking data. *Journal of Revenue and Pricing Management*, 1(2), 121-138.

Matzkin, R. L. (2012). Identification in nonparametric limited dependent variable models with simultaneity and unobserved heterogeneity. *Journal of Econometrics*, 166(1), 106-115.

McGill, J. I. (1995). Censored regression analysis of multiclass passenger demand data subject to joint capacity constraints. *Annals of Operations Research*, 60(1), 209-240.

Moore, T. G. (1966). The demand for broadway theater tickets. *Review of Economics and Statistics*, 48(1), 79-87.

Seaman, B. A. (2006). Empirical studies of demand for the performing arts. *Handbook of the Economics of Art and Culture*, 1, 415-472.

Schimmelpfennig, J. (1997). Demand for ballet: a nonparametric analysis of the 1995 royal ballet summer season. *Journal of Cultural Economics*, 21(2), 119-127.

Throsby, D. (1990). Perception of quality in demand for the theatre. *Journal of Cultural Economics*, 14(1), 65-82.

Throsby, D. (1994). The production and consumption of the arts: A view of cultural economics. *Journal of Economic Literature*, 32(1), 1-29.

Throsby, D., Withers, G. A., Shanahan, J. L., Hendon, W. S., Hilhorst, I. T. H. & van Straalen, J. (1983). Measuring the demand for the arts as a public good. In *Economic support for the arts*. [Volume 3, Proceedings of the Second International Conference on Cultural Economics and Planning, May 26-28, 1982, Netherlands]. (pp. 37-52). Association for Cultural Economics, University of Akron.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society*, 24-36.

Touchstone, S. K. (1980). The effects of contributions on price and attendance in the lively arts. *Journal of Cultural Economics*, 4(1), 33-46.

Towse, R. (2011). Opera and ballet. In: Towse, R. (Ed.). *A Handbook of Cultural Economics*. Edward Elgar Publishing.
Wickham, R.R. (1995) Evaluation of forecasting techniques for short-term demand of air transportation. PhD thesis, *MIT, Flight Transformation Lab*, Cambridge, MA.
Withers, G. (1980). Unbalanced growth and the demand for the performing arts: An econometric analysis. *Southern Economic Journal*, 46, 735-742.

Authors:

Evgeniy M. Ozhegov
National Research University Higher School of Economics (Perm, Russia).
Research Group for Applied Markets and Enterprises Studies. Research Fellow;
E-mail: tos600@gmail.com

Alina Ozhegova
National Research University Higher School of Economics (Perm, Russia).
Research Group for Applied Markets and Enterprises Studies. Young Research Fellow;
E-mail: arbuzanakova@hse.ru