School of Economic Sciences

Working Paper Series
WP 2019-5

Assessing the Importance of an Attribute in a Demand System
Structural Model versus Machine Learning

Syed Badruddoza, Modhurima Dey Amin,
Jill J. McCluskey

December 2019

WASHINGTON STATE UNIVERSITY
World Class. Face to Face.

Assessing the Importance of an Attribute in a Demand System
Structural Model versus Machine Learning

Syed Badruddoza

Modhurima Dey Amin

Jill J. McCluskey*

Abstract

Firms can prioritize among the product attributes based on consumer valuations using market-level data. However, a structural estimation of market demand is challenging, especially when the data are updating in real-time and instrumental variables are scarce. We find evidence that Random Forests (RF)—a machine-learning algorithm—can detect consumers' sensitivity to product attributes similar to the structural framework of Berry-Levinsohn-Pakes (BLP). Sensitivity to an attribute is measured by the absolute value of its coefficient. We check the RF's capacity to rank the attributes when prices are endogenous, coefficients are random, and instrumental or demographic variables are unavailable. In our simulations, the BLP estimates correlate with the RF importance factor in ranking (68%) and magnitude (79%), and the rates increase with the sample size. Consumer sensitivity to endogenous variables (price) and variables with random coefficients are overestimated by the RF approach, but ranking of variables with non-random coefficients match with BLP's coefficients in 96% cases. These estimates are pessimistically derived by RF without parameter-tuning. We conclude that machine-learning does not replace the structural framework, but provides firms with a sensible idea of consumers' ranking of product attributes.

Key words: Machine-Learning, Random Forests, Demand Estimation, BLP, Discrete Choice.

JEL Codes: C55, D11, Q11.

* Syed Badruddoza and Modhurima Amin are doctoral students and Jill McCluskey is a Regents Professor in the School of Economic Sciences at Washington State University. We thank AAEA annual meeting attendees for their constructive suggestions and helpful comments. Any errors are our own. Correspondence: mccluskey@wsu.edu.

ASSESSING THE IMPORTANCE OF AN ATTRIBUTE IN A DEMAND SYSTEM: STRUCTURAL MODEL VERSUS MACHINE LEARNING

Products are bundles of characteristics. Incorporation of consumer preferences over the characteristics space resulted in a stream of demand theory and econometric models (e.g., Lancaster 1966; McFadden 1974; Nevo 2011). Finding the prominence of a product attribute can be necessary for many reasons; for instance, producers or retailers may want to learn how consumers value a particular feature of the product, e.g., shelf-life of a cereal, flight schedule of an airline (Gramming, Hujer, and Scheidler 2005; Ortega et al. 2011). A firm would like to know consumers' sensitivity to an existing feature, or if adding a particular feature to the product is worth to consumers' valuation, e.g., organic certification (Meas et al. 2014). Aggregate valuation of a product characteristic defines its market demand. The literature on location models indicates that failure to match consumer tastes on characteristics space may adversely affect firms' market share and profitability (e.g., Irmen and Thisse 1998).

In empirical industrial organization, estimation of demand systems in both characteristics space and product space remains a popular exercise. Academicians, businesses, and policymakers frequently conduct demand analysis to gain insights about elasticities, market structure, policy interventions, and the importance of a product attribute. Many models and econometric tools have been developed over the years to estimate the relevant parameters (see Nevo 2011 for a discussion).

However, the costs of construction and computation of a model should be compared to the added value the complexity offers to the researcher. There is a clear trade-off between data dimensionality and model rigor (Yu et al. 2006). Availability of continuously updating large datasets has generated interest in machine-learning models that are faster to apply and generate out-of-sample predictions better than structural demand models (Bajari et al. 2015). Structural models have

clear interpretability, whereas the machine-learning model provides reliable predictions (Wang and Zhao 2018a). Economists tend to emphasize statistical inference and the identification of causal effects; but machine-learning models concentrate on predictive fit, handling model uncertainty, and dimension reduction. A common approach in economics is to use observed product attributes and market share to estimate consumer valuation of attributes. Use of attributes instead of product indicator variables is rationalized by dimension reductions (Nevo 2001). However, large market-level data sets often have high dimensionality—e.g., product attributes outnumber the observed products—which further limits the use of structural models.

The current article compares the performance of machine-learning with structural demand model estimation in estimating the importance of a product attribute to the consumers. Attribute importance is measured by the magnitude of parameter estimates in the structural model, and the predictability of market outcomes in machine-learning. To the best of our knowledge, this is the first study to compare the strength of the two methods in identifying the significance of a product characteristic. Random coefficient discrete choice, popularly known as the BLP model for Berry, Levinsohn, and Pakes (1995), represents the structural approach, and Random Forests (RF) is used as a machine-learning approach. The former model originates from McFadden's (1974) work and augmented by Nevo (2000), and many others. The latter is entirely data-driven, non-parametric model (Breiman, 2001, 2002). Both models appeal to researchers for their estimation capacity from the aggregate market level data, and an inexpensive alternative to the household survey. In the BLP approach, observed product attributes explain the market share, and the coefficient of each attribute shows its valuation to consumers. Contrarily, the RF approach selects the attributes based on their importance in predicting the market outcomes. We argue that if both approaches similarly rank product attributes, researchers can adopt the RF approach to save time and resources.

In this study, we simulate a dataset following BLP's structure, with endogenous prices and random coefficients, and estimate a BLP model using a set of instruments. We then check the RF's capacity to rank attributes given price is endogenous, coefficients are random, and instrumental or demographic variables are unavailable. Our simulations provide evidence that the RF approach fairly captures (70%) consumer sensitivity to a product attribute, without parameter tuning, demographic or instrumental variables. Sensitivity to an attribute is measured by the absolute value of its coefficient. RF overestimates the importance of endogenous variables, e.g., price, and the variables with random coefficients in BLP model. However, it is able to rank the importance of product attributes, especially non-random attributes with considerable accuracy (96%) and consistency. We conclude that machine-learning does not replace the structural framework, but provides firms with a sensible idea of consumers' ranking of product attributes.

Modelling Consumer's Valuation of an Attribute

Economists often derive consumer valuation for product attributes using experimental data, either in a lab-setting or in the market (Melton et al. 1996, Lusk et al. 2001). Obtaining household valuation of product attributes via survey is expensive, and household-level large datasets are often complicated by zero values in the dependent variables because the household may not purchase the product (Yen, Biing-Hwang and Smallwood 2003). Although market-level data is easier to obtain, generating consumer valuation is difficult to estimate from observational and aggregated data. The simplest form is to estimate demand with time-series data, which is vulnerable to cross price collinearity and endogeneity. Another approach is Rosen (1974)'s hedonic pricing model, which can generate sensible results when the model is flexible and the market is homogeneous (e.g., Ready and Abdalla 2005; Kuminoff, Parmeter, and Pope 2010). However, the hedonic model is criticized in the literature for misspecification and endogeneity (e.g., Ekeland, Heckman, and Nesheim 2004).

A more refined approach is to find own- and cross product elasticities through budget shares, which is often used for policy analysis regarding consumer welfare (Lewbel and Pendakur 2009; Unnevehr et al. 2010). Demand systems in product space build on the theoretical relationship between quantity and price and use constraints, aggregation, symmetry or separability to reduce the data dimension. Products can be grouped and their elasticities can be estimated within and between groups. Multistage demand systems work better when there are a small number of products that are relatively constant across large number of markets. Analyzing consumer sensitivity to a product feature from aggregate market-level data has some major problems (as pointed out by Nevo 2011). For example, a linear demand system with J products requires estimation of J^2 parameters, and allowing for a different price coefficient for each product is not feasible. The requirement can be somewhat reduced by simplifying assumptions (e.g., symmetry of the Slutsky matrix) but the number of parameters to be estimated is still proportional to J^2. Most importantly, the product space approach does not provide insights into individual consumer behavior and the distribution of taste heterogeneity. Products are often vertically differentiated, so capturing quality differences and predicting demand is difficult for new goods with unprecedented characteristics. In general, product space models are aggregate and rarely host explicit parametrization of consumer tastes, e.g., storage decisions, and bandwagon effects. Customizing consumer utility functions and then deriving aggregate demand is more convenient rather than going the other way. From an empirical perspective, it is almost impossible to find exogenous instruments for prices of a large set of narrowly defined, highly collinear products. Zhen et al. (2013) use the Exact Affine Stone Index incomplete demand system with instrumental variables, and observe that neglecting price endogeneity or estimating a conditional demand model significantly overestimates the absolute value of elasticity.

On the contrary, demand systems in characteristics space solve the dimensionality problem. Such systems mainly include logit type models that capture consumer valuation of a product characteristic. Huang, Rojas and Bass (2008) conducted a Monte Carlo experiment to find that the logit model yields unbiased estimates for a certain size of the assumed market potential. Nevertheless, consumers are heterogeneous; so the most popular approach to derive consumer valuation from market data is random coefficient discrete choice (Nevo 2011). Discrete choice models start with consumer indirect utility. Hence they are customizable, accommodate unobserved product characteristics, and assume that the choice process is itself probabilistic (Tversky 1972). The parameters may depend on demographic features, and have respective fixed and random components. Discrete choice models can also be used to measure consumer welfare or mergers or introduction of a product (Nevo 2000; Petrin 2002; Nevo 2003).

Empirical application of discrete choice models may face issues such as continuity of choices, weak instruments, and errors in measuring price and market share (Nevo 2011). The problem worsens when business decision requires automated real-time analysis of large data set containing numerous products and characteristics. Therefore, firms in the retail, health care, and Internet industries often employ machine-learning methods for demand estimation. Bajari et al. (2015) compare between structural models and commonly used machine-learning algorithms for sales prediction, and observe better out-of-sample predictive power of the latter. An optimal solution would be to utilize the wealth of data efficiently without compromising economic intuition. Several studies attempt using economic models in big data and machine-learning. For example, Athey et al. (2018) utilize anonymous locational information to model consumers' choices of restaurants. Wang and Zhao (2018a) show that the maximum likelihood estimation of discrete choice models is a special case of deep neural networks, and find a tradeoff between interpretability and predictability. In an empirical follow-up, the authors generate the probabilities of travel choices

using both methods, but observe large estimation errors and irregularity of the probability space of neural network (Wang and Zhao 2018b). A few other studies check prediction accuracy of vehicle ownership using Random Forests and multinomial logit and find mixed results (Paredes et al. 2017; Lee et al 2019).

Discrete Choice versus Random Forests

Studies discussed above compare between the two methods in terms of predictive accuracy. Our goal is to apply machine learning to derive some usable information regarding consumer taste, without violating economic logic. Below we point out the links between a discrete choice model and a machine learning algorithm, and check their association using simulated data in later section.

Structural Approach: BLP Model

BLP (1995) provide a commonly used structural approach in demand analysis. The indirect utility of consumer $i = 1, \dots, N$ with income y from brand $j = 1, \dots, J$ in market $m = 1, \dots, M$ is,

$$u_{ijm} = \alpha_i(y_i - p_{jm}) + \mathrm{x}_{jm}\beta_i + \xi_{jm} + \epsilon_{ijm} \quad (1)$$

where, x_{jm} and ξ_{jm} are vectors of observed and unobserved product attributes, respectively; p is price, ϵ is mean-zero stochastic term, and $\{\alpha, \beta\}$ is a set of parameter vectors to be estimated. For instance, if j is car brands, X_1 is rear cargo room, $\beta_{X1} > 0$ indicates higher utility from an additional unit of cargo space. Similarly, a higher α_i means the consumer's marginal utility from a dollar is higher. Assume another attribute X_2 is the charging time for electric cars, and a hypothetical consumer receiving disutility from a longer charging time would be indicated by $\beta_{X2} < 0$. Thus, the absolute values of parameters $|\beta_{x1}|, |\beta_{x2}|$ indicate consumer sensitivity to a particular product attribute. If we standardize the Xs for comparison, then $|\beta_{x1}| < |\beta_{x2}|$ means that the consumer is

more sensitive to charging time than rear cargo room. Then a profit-seeking car manufacturer should focus more on reducing charging time than adding more cargo space. We use this logic to construct the machine-learning algorithm in the next subsection.

Some parameters are heterogeneous across consumers, specified as random coefficients. Nevo (2001) defined two sources of variation in the parameters—observed variation that comes from demographic features of the market, and unobserved variation that follows a parametric distribution e.g., $\beta_i = \mathbf{\Omega}(demographics_i) + error_i$, where Ω is a matrix of how parameters depend on consumer observables. Demographic observables and errors are independent by assumption. Notice that we do not include demographic features in a simulation study. Because randomly drawn normal demographic features added with normal errors produce another random normal variable and provide the same results as randomly drawn β_i. We obtain the popular logit specification if parameters are not heterogeneous (e.g., Alfnes et al. 2006)

The probability that one brand j will be preferred to another brand k is

$$(2)\quad P(u_{ijm} > u_{ikm}) = P(\epsilon_{ikm} - \epsilon_{ijm} < \alpha_i p_{jm} - \alpha_i p_{km} + \mathrm{x}_{km}\beta_i - \mathrm{x}_{jm}\beta_i + \xi_{km} - \xi_{jm})$$
$$= \int_{\epsilon} \mathbb{I}(\alpha_i p_{jm} - \alpha_i p_{km} + \mathrm{x}_{km}\beta_i - \mathrm{x}_{jm}\beta_i + \xi_{km} - \xi_{jm})\, dG(\epsilon_{ikm} - \epsilon_{ijm})$$

where, $\mathbb{I}$ is an indicator function representing the inequality, and G is the distribution of difference in error terms. BLP assume that ϵ has Type-I extreme value distribution, so the difference has a Logit form (Train 2003). Following Nevo (2001), we assume an outside brand that is not included in the analysis and provides zero utility to consumer. In addition, ties between two brands occur with zero probability. An individual consumer's probability of purchasing brand j is,

$$(3)\quad s_{ijm} = \frac{\exp(-\alpha_i p_{jm} + \mathrm{x}_{jm}\beta_i + \xi_{jm})}{1 + \sum_{k=1}^{J} \exp(-\alpha_i p_{km} + \mathrm{x}_{km}\beta_i + \xi_{km})}$$

and the probability of choosing brand j over the outside brand is, $\ln(s_{ijm}) = -\alpha_i p_{jm} + \mathrm{x}_{jm}\beta_i + \xi_{jm}$. Nevo (2001) approximates brand j's share in market m as an aggregation of individual preferences,

$$s_{jm}(p, \mathrm{x}, \xi; \alpha, \beta) = \frac{1}{N_m}\sum_{i=1}^{N_m} s_{ijm} \tag{4}$$

where, N_m is the number of consumers randomly drawn from the market m.

There are two practices for aggregating individual probabilities. First, dropping the assumption of heterogeneous consumer and assuming ϵ_{ijt} to be Type-I extreme value independently and identically distributed (iid) gives the simple logit specification for the market share, s_{jm}. Second, the random coefficients logit assumes ϵ_{ijt} iid with Type-I extreme value distribution, but the consumer characteristics affect the calculation of elasticities, hence they are less likely to impose restrictions on cross-price elasticities compared to the Logit model (Nevo 2011). Prices are endogenous to market share, so the BLP estimation requires instrumental variables (IV) in a Generalized Method of Moments (GMM) framework. Endogeneity arises from correlation between price and the unobserved product characteristics, ξ_{jt}.

Estimation issues of BLP are well-discussed in the literature (e.g., Nevo 2000; Dube, Fox, and Su 2009) and beyond the scope of the paper. The idea is to minimize the distance between the log of observed market share and the log of theoretical market share, $\ln(S_{.m}) - \ln[s_{.m}(p, \mathrm{x}, \xi; \alpha, \beta)]$. In the next subsection, we exploit this relationship to predict $\ln(\mathrm{S})$ by $p, \mathrm{x}, \mathrm{j}$ using RF without using any demographic or instrumental variables.

Machine Learning Approach: Random Forests

The RF approach refers to a supervised machine-learning algorithm widely used for classification and regression. The algorithm constructs a multitude of decision trees at training time with

bootstrap sampling and randomly chosen predictors, and then aggregates them for prediction (Breiman 2001). RF is a powerful out-of-sample predictor because the randomness of choosing predictors and resampling corrects for decision trees' tendency to overfit their training set. It accounts for correlations between sub-samples, need not rely on parametric assumptions, and is robust to outliers (Biau and Scornet 2016). It generates variable importance and improves on unstable estimates, especially for big and high-dimensional data, where finding a structural model is impossible because the scale and complexity of the problem (Kleiner et al. 2014).

We chose RF to test the model discussed above for two reasons. First, RF has a comprehensive, tree-based nature compared to other convoluted machine-learning models, such as an artificial neural network, and is rapidly gaining popularity in economic applications (e.g., Davis and Heller 2017). Second, RF generates variable importance, which we plan to use to rank product characteristics. Unlike other importance generators, such as Gradient Booster, RF is capable of making out-of-sample predictions without requiring tuning of hyper-parameters.

The structural framework above can be linked to RF. Assume there are $q = 1, \dots, Q$ observed determinants (e.g., price, attributes, product identity) $\chi \subset \mathbb{R}^Q$, and the goal is to predict log of observed market share vector $\ln(S)$. Define the nonparametric regression function as $\Phi(\chi) = \mathbb{E}[\ln(S)|\chi]$, the sample as $\mathcal{D}$, and a typical sub-sample $\mathcal{D}_w = \{\ln(S), p, \mathrm{x}, \mathrm{j}\}_w$ randomly drawn with replacement prior to tree construction. A random forest is a predictor consisting of a collection of T randomized regression trees. The t-th tree estimation takes the form

$$\Phi_w(\chi; \lambda_t, \mathcal{D}) = \sum_{\mathcal{D}_{t,w}} \frac{\mathbb{I}_{\chi \in A_w(\chi; \lambda_t, \mathcal{D})} \ln(S)}{N_w(\chi; \lambda_t, \mathcal{D})} \tag{5}$$

Where the random variable λ is used to resample and successive directions for splitting, $\mathbb{I}$ is an indicator function, $A(\chi; \lambda_t, \mathcal{D}_w)$ is the cell containing χ, and $N_w(\chi; \lambda_t, \mathcal{D})$ is the number of

preselected points contained in $A_w(\chi; \lambda_t, \mathcal{D})$ (Ishwaran 2007; Biau and Scornet 2016). The overall model prediction is the unweighted average of trees,

$$\bar{\Phi}(\chi; \lambda, \mathcal{D}) = T^{-1} \sum_1^T \Phi_w(\chi; \lambda_t, \mathcal{D}). \quad (6)$$

Ideally, the prediction approaches the log of observed market share, $\ln(S)$ as $T \to \infty$ (Louppe 2014). RF regression has a quadratic loss function $L(\ln(S), \bar{\Phi})$ that measures the discrepancy between its two arguments.

We define the importance of q^{th} determinant by the mean increase in the root mean square error (RMSE) of prediction when the determinant is randomly removed from the model—also known as permutation importance. A greater increase in RMSE indicates additional importance of the determinant. Following Breiman (2001, 2002), the importance factor of a determinant $\chi_q \in \chi$ is,

$$\pi_q = \mathbb{E}_{\mathcal{D}_v}\left[\frac{1}{MJ}\sum_{\mathcal{D}_v} L\left(\ln(S), T^{-1}\sum \Phi_v(\chi; \lambda_t, \mathcal{D})\right)\right] - \frac{1}{MJ}\sum_{\mathcal{D}_w} L\left(\ln(S), T^{-1}\sum \Phi_w(\chi; \lambda_t, \mathcal{D})\right) \quad (7)$$

where $\mathbb{E}$ refers to expected value, $\mathcal{D}_v$ denotes a subsample in which the values of χ have been randomly permuted, so the first-loss function involves trees that have been built from bootstrap samples that did not include $\mathcal{D}_w$—the so-called "out-of-bag features" (Louppe 2014). Both loss functions are averaged across the full sample, which includes multiple of markets and brands. The rationale is that randomly omitting a determinant χ_q should substantially increase errors of prediction, if the determinant is associated with $\ln(S)$.

We use the package "ranger" in R software for fast implementation of RF (Wright, Wager, and Probst 2019). The replication appendix contains a typical RF algorithm.

Valuation of an Attribute in BLP and its Importance in Machine-Learning

Now we find a relationship between the BLP-estimated consumer sensitivity to the product attribute, $|\beta_q|$, and its importance predicted by machine-learning, π_q. We propose that if the

structural model is true, then $\mathbb{E}(\pi_q|\beta_q|) \neq 0 \; \forall \; |\beta_q| > 0$. The proof is straightforward. If the consumer is sensitive to a particular product attribute, then the associated coefficient is nonzero, $\partial u/\partial X_q = \beta_q \neq 0$. That attribute predicts some of the changes in the market share, so $\partial \ln(S) /\partial X_q \neq 0$. Then the attribute should affect the loss functions when arbitrarily removed, thus affecting the variable importance. Assume $\partial \ln(S)/\partial X_q \neq 0 \Rightarrow |\beta_q| > 0$. Following the previous notations where $X_q \in \mathcal{D}_w$ and $X_q \notin \mathcal{D}_v$ we have,

$$\mathbb{E}[L(\ln(S), \Phi_v)] - L(\ln(S), \Phi_w) > 0 \; \forall \; t \in T \Rightarrow \pi_q > 0. \tag{8}$$

That is, importance must be positive if permuting the predictor increases error for all trees. Hence, π_q is not independent of $|\beta_q|$, which implies their covariance, $\mathbb{E}(\pi_q|\beta_q|) - \mathbb{E}(\pi_q)\mathbb{E}(|\beta_q|) \neq 0$. That means their correlation coefficient is nonzero. However, the direction of association between variable importance and that coefficient depends on too many factors to derive a closed-form expression. We test the association under different simulated scenario in the following section.

Data Generation and Simulation

To check the similarity between structural model and machine learning, we avoid observational data so that the true values of the parameters can be predefined for the comparison. Moreover, the results from the observational data is more likely to depend on the chosen sample. Figure 1 provides an overview of the data generation and simulation process. The process assumes that the data follow BLP structure: the market shares are aggregations of consumer choices that depends on product attributes. The objective is to test how machine-learning performs when the BLP model is true. This specification gives us a solid benchmark for comparison and minimizes the noise compared to BLP estimation using observational data.

We use an R-package, "BLPEstimatoR," for data generation and BLP estimation (Brunner, Weiser, and Romahn 2019). The package allows for creating a BLP data set with varied numbers of parameters and samples. The replication codes for data generation, BLP estimation, RF regression, and correlations are provided in the appendix.

The price variable is created endogenously to the market share, and 20 instrumental variables are created for GMM estimation of BLP. There are ten exogenous attribute variables, $X_1, \dots, X_{10}$ of which the first two have random coefficients. Attribute variables are standardized for comparison. Inclusion of more attribute variables (15, 20) generates similar results.

One-hundred combinations of brands and markets are presumed at the beginning of the simulation, such that, $j = \{10, 20, \dots, 100\}$ and $m = \{10, 20, \dots, 100\}$. We randomly assign values to α, β—that give consumers' marginal utility of attributes. Then we estimate the BLP parameters and rank their absolute values to compare with variable importance factors generated by RF. Greater absolute values of estimated parameters indicate higher consumer sensitivity to the attribute. We then estimate the RF regression of $\ln(S)$ on price, other attributes and market indicators, without any demographic or instrumental variables, and derive the importance of each attribute in predicting $\ln(s)$. Finally, we compare the importance with the absolute values of BLP estimates using two measures. Pearson's correlation coefficient indicates if the estimated magnitude of consumer sensitivity and RF variable importance factors are correlated,

$$\rho_{Pearson} = \frac{Cov(\pi, |\beta|)}{\sqrt{Var(\pi) Var(|\beta|)}}. \tag{9}$$

However, our main interest is Spearman's rank correlation coefficient that measures if the rank of product attribute by consumer correlates with that of RF importance factors,

$$\rho_{Spearman} = 1 - \frac{6 \sum d_i^2}{Q(Q^2 - 1)}. \tag{10}$$

Where d represents the difference in two ranks and Q is the number of coefficients with attributes and price ($Q = 10 + 1 = 11$ in our model).

Notice that greater heterogeneity in consumer tastes (higher $Var(|\beta|)$) lowers the correlation. That is, correlations with RF importance might be different for random and non-random coefficients β. We check the correlations for all coefficients first, and then separately for the coefficients with non-random attributes (with price and random parameters present in the model). The reason for the latter exercise is to check how consumer heterogeneity affects the results and to compare the RF importance with structural model under usual logit specification where the parameters are non-random.

Results

Table 1 presents a sample simulation result out of 100 simulations. The BLP model performs well in estimating the parameters, mainly because the data are constructed under structural framework. The RF importance factors are generated using equation (7). BLP rank provides the order of BLP estimates of their absolute values—the greater the absolute value of a BLP estimate, the higher the rank. RF ranks are similarly generated from RF importance factors. Pearson correlation coefficients show the correlation between the absolute values of BLP estimates and RF importance. Spearman's rank correlation coefficients measure the association between BLP and RF ranks.

Some patterns can be observed in Table 1. If consumers care little about an attribute, its importance predicted by RF is usually the lowest, e.g., X9. If consumers are sensitive to an attribute, that is the absolute value of the parameter estimate is high, its importance is predicted high by RF, e.g., X6. The correlation between the BLP estimates in absolute values and RF importance often coincide—as shown by Pearson's correlation (0.721). Spearman's rank correlation indicates that there is a significant positive association (0.691) between RF importance and the absolute value of

the BLP estimates. The closer the consumer sensitivity of two attributes, the closer their importance, e.g., X5 and X7.

The order of the estimated marginal utilities may not coincide with the order of importance for endogenous (EN) and random coefficient (RC) attributes. Random coefficients are often miss-ranked by RF due to their random nature. Price, being an endogenous attribute, appears to be more important in RF prediction (4^{th}) than it actually is in consumers' utility (9^{th}). The endogenous variable is most likely to create noise and change orders frequently, which reduces the correlation.
Table 2 provides correlation coefficients from 100 simulations with varying markets and brands. The mean and median of the association is around 70% when endogenous and random coefficients are present. The association grows to about 96% if we only check the rank correlation for non-random attributes.

Figures 2A presents the association between RF predicted importance and the absolute value of BLP estimates by brands and markets. The size of the circles represents Spearman's rank correlation coefficients, whereas the shades of the circles indicate Pearson's correlation coefficients. Spearman's correlation varies between 0.5 and 0.8, and Pearson's correlation ranges from 0.4 to 0.9 for all coefficients (Figure 2A). Circles imply that there is a nonnegative correlation between RF importance and consumers' sensitivity of attributes in BLP model. It is possible to achieve high correlations using small sample sizes e.g., 10 brands and 10 markets in Figure 2A. However, the correlation increases in general as we include more data in the sample, especially more brands. The association between the two methods weakens due to the volatile nature of endogenous and random coefficient variables. Figure 2B plots the correlations for non-random coefficients only. The correlation between the structural model and machine-learning is higher for non-random coefficients; the Pearson statistic ranges from 0.91 to 0.97 whereas the Spearman statistic varies between 0.8 and 1.

Figure 2C plots correlations from individual simulations against sample size. A greater sample size improves both Pearson and Spearman correlations. The association revolves around 71% in presence of endogenous and random coefficient variables. However, the correlation approaches one with narrow variations for non-random attributes only.

Figure 2D shows the relationship between RF importance factors and BLP estimates by coefficient type. RF importance factors and BLP estimates have a quadratic association for non-random coefficients—created by a quadratic-loss function of importance. Random coefficients exhibit positive associations with large confidence intervals. Importance of endogenous and random coefficients (with the price variable) is highly overestimated by RF—suggesting an exponential relationship.

In short, the magnitude and order of $|\beta|$ estimated by the BLP model positively correlates with the feature importance predicted by RF. This is particularly true for non-random attributes (96%), and moderately works for endogenous and random coefficient attributes. We further extended the analysis with 15 and 20 product attributes and obtain similar results. Different initial values of parameters were also employed without obtaining any new insights.

The results provide the retail market researchers with an instrument to rank consumer valuation of different attributes under the constraints of time, demographic or instrumental variables, and data dimensionality. Before designing a structural model, researchers may consider a machine-learning algorithm to gain primary insights about the market, particularly before adding or removing a product characteristic.

Concluding Remarks

We performed a simulation study to evaluate whether the importance of an attribute predicted by machine-learning matches the ordering of the absolute value of coefficients estimated by a structural

model of demand estimation. We find a positive association between the two, particularly when the coefficients are non-random. The machine-learning model does not replace structural model because it cannot provide us with the actual valuation of a product attribute by consumers. Neither does it generate the direction of consumer valuation for an attribute, i.e., whether the consumer likes the attribute or dislikes it. However, it provides a reasonable approximation of the importance of product attributes, especially under time, computation, and data-related constraints. The direction of consumer valuation can then be obtained empirically using a simple linear regression. Firms can focus on product attributes that are more important to consumers, thus saving valuable resources and receiving better appreciation of their products.

The costs of construction and computation of a model should be compared to the added value the complexity offers to the researcher. This study is a small step towards constructing computationally efficient models for retail market analysis with economic intuition. Future research will examine further at the relationship between RF importance factors and BLP estimates in order to generate an index or multiplier for endogenous, non-random, and random coefficients, so that structural parameter estimates can be retrieved from machine-learning. Another way to capture more variations in random coefficients and improve the correlation would be to use observed demographic variables, appropriate instruments and parameter tuning in RF.

References

Alfnes, F., A.G. Guttormsen, G. Steine, and K. Kolstad. 2006. "Consumers' Willingness to Pay for the Color of Salmon: A Choice Experiment with Real Economic Incentives." *American Journal of Agricultural Economics* 88 (4): 1050-1061.

Athey, S., D. Blei, R. Donnelly, F. Ruiz, and T. Schmidt. 2018. "Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data." *AEA Papers and Proceedings* 108: 64-67.

Bajari, P., D. Nekipelov, S.P. Ryan, and M. Yang. 2015. "Machine Learning Methods for Demand Estimation." *American Economic Review* 105 (5): 481-485.

Berry, S., J. Levinsohn, and A. Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63 (4): 841-890.

Biau, G., and E. Scornet. 2016. "A Random Forest Guided Tour." *Test* 25 (2): 197-227.

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

Breiman, L. 2002. *Manual on Setting Up, Using, and Understanding Random Forests.* Report v3, Berkely, CA: University of California Berkeley Statistics Department.

Brunner, D., C. Weiser, and A. Romahn. 2019. "BLPestimatoR: Performs a BLP Demand Estimation." *CRAN*, May 14, Version 0.2.9. Available https://cran.r-project.org/web/packages/BLPestimatoR/BLPestimatoR.pdf.

Davis, J., and S.B. Heller. 2017. "Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs." *American Economic Review* 107(5):546-50

Dube, J.P., J.T. Fox, and C.L. Su. 2012. "Improving the Numerical Performance of Static and Dynamic Aggregate Discrete Choice Random Coefficients Demand Estimation." *Econometrica* 80 (5): 2231-2267.

Ekeland, I., J.J. Heckman, and L. Nesheim. 2004. "Identification and Estimation of Hedonic Models." *Journal of Political Economy* 112 (S1): S60-S109.

Gramming, J., R. Hujer, and M. Scheidler. 2005. "Discrete Choice Modelling in Airline Network Management." *Journal of Applied Econometrics* 20(4): 467-486.

Huang, D., C. Rojas, and F. Bass. 2008. "What Happens When Demand is Estimated with a Misspecified Model?" *The Journal of Industrial Economics* 56 (4): 809-839.

Irmen, A., and J.F. Thisse. 1998. "Competition in Multi-Characteristics Spaces: Hotelling was Almost Right." *Journal of Economic Theory* 78 (1): 76-102.

Ishwaran, H. 2007. "Variable Importance in Binary Regression Trees and Forests." *Electronic Journal of Statistics* 1 (2007): 519-537.

Kleiner, A., A. Talwalkar, P. Sarkar, and M.I. Jordan. 2014. "A Scalable Bootstrap for Massive Data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (4): 795-816.

Kuminoff, N.V., C.F. Parmeter, and J.C. Pope. 2010. "Which Hedonic Models Can we Trust to Recover the Marginal Willingness to Pay for Environmental Amenities?" *Journal of Environmental Economics and Management* 60 (3): 145-160.

Lancaster, K.J. 1966. "A New Approach to Consumer Theory." *Journal of Political Economy* 742: 132-157.

Lee, D., J. Mulrow, C.J. Haboucha, S. Derrible, and Y. Shiftan. 2019. "Attitudes on Autonomous Vehicle Adoption Using Interpretable Gradient Boosting Machine." *Transportation Research Record* 1-14. doi:10.1177/0361198119857953.

Lewbel, A., and K. Pendakur. 2009. "Tricks with Hicks: The EASI Demand System." *American Economic Review* 99 (3): 827-863.

Louppe, G. 2014. *Understanding random forests: From theory to practice.* PhD dissertation, Liege, Belgium: University of Liege Department of Electrical Engineering and Computer Science.

Lusk, J.L., J.A. Fox, T.C. Schroeder, J. Mintert, and M. Koohmaraie. 2001. "In-store Valuation of Steak Tenderness." *American Journal of Agricultural Economics* 83 (3): 539-550.

McFadden, D. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics*, edited by Paul Zarembka, 105-142. New York City, NY: Academic Press.

Meas, T., W. Hu, M.T. Batte, T.A. Woods, and S. Ernst. 2014. "Substitutes or Complements? Consumer Preference for Local and Organic Food Attributes." *American Journal of Agricultural Economics* 97 (4): 1044-1071.

Melton, B.E., W.E. Huffman, J.F. Shogren, and J.A. Fox. 1996. "Consumer Preferences for Fresh Food Items With Multiple Quality Attributes: Evidence from an Experimental Auction of Pork Chops." *American Journal of Agricultural Economics* 78 (4): 916-923.

Nevo, A. 2000. "A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand." *Journal of Economics and Management Strategy* 9 (4): 513-548.

Nevo, A. 2001. "Measuring Market Power in the Ready-to-Eat Cereal Industry." *Econometrica* 69 (2): 307-342.

Nevo, A. 2003. "New Products, Quality Changes and Welfare Measures Computed from Estimated Demand Systems." *The Review of Economics and Statistics* 85(2): 266-275.

Nevo, A.. 2011. "Empirical Models of Consumer Behavior." *Annual Review of Economics* 3 (1): 51-75.

Ortega, D.L., H. Wang, N.J. Olynk, L. Wu, and J. Bai. 2011. "Chinese Consumers' Demand for Food Safety Attributes: A Push for Government and Industry Regulations." *American Journal of Agricultural Economics* 94 (2): 489-495.

Paredes, M., E. Hemberg, O.M. O'Reilly, and C. Zegras. 2017. "Machine Learning or Discrete Choice Models for Car Ownership Demand Estimation and Prediction?" *IEEE International Conference on Models and Technologies for Intelligent Transportation Systems* (MT-ITS): 780-785.

Petrin, A. 2002. "Quantifying the Benefits of New Products: The Case of the Minivan." *Journal of Political Economy,* 705-29.

Ready, R.C., and C.W. Abdalla. 2005. "The Amenity and Disamenity Impacts of Agriculture: Estimates from a Hedonic Pricing Model." *American Journal of Agricultural Economics* 87 (2): 314-326.

Rosen, S. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy* 82 (1): 34-55.

Train, K. 2003. *Discrete Choice Methods with Simulation.* Cambridge, UK: Cambridge University Press.

Tverskey, A. 1972. "Elimination by Aspects: A Theory of Choice." *Psychological Review,* 79:281-299.

Unnevehr, L., J. Eales, H. Jensen, J. Lusk, J.J. McCluskey, and J. Kinsey. 2010. "Food and Consumer Economics." *American Journal of Agricultural Economics* 92 (2): 506-521.

Wang, S., and J. Zhao. 2018a. "Framing Discrete Choice Model as Deep Neural Network with Utility Interpretation." *arXiv* 1810.10465 (Oct): 1-16.

Wang, S., and J. Zhao. 2018b. "Using Deep Neural Network to Analyze Travel Model Choice with Interpretable Economic Information: An Empirical Example." *arXiv* 1812.04528 (Dec): 1-16.

Wright, M.N., S. Wager, P. Probst. 2019. "ranger: A Fast Implementation of Random Forests" *CRAN,* Version 0.11.2. Available https://cran.r-project.org/web/packages/ranger/ranger.pdf.

Yen, S.T., L. Biing-Hwang, and D.M. Smallwood. 2003. "Quasi- and Simulated-Likelihood Approaches to Censored Demand Systems: Food Consumption by Food Stamp Recipients in the United States." *American Journal of Agricultural Economics* 85 (2): 458-478.

Yu, L., K.K. Lai, S. Wang, and W. Huang. 2006. "A Bias-Variance-Complexity Trade-Off Framework for Complex System Modeling," in *Computational Science and its Applications*, O. Gervasi, V. Kumar, C. J. Tan, D. Taniar, A. Lagana, Y. Mun and H. Choo, eds. 518-527. Berlin: Springer.

Zhen, C., E.A. Finkelstein, J.M. Nonnemaker, S.A. Karns, and J.E. Todd. 2013. "Predicting the Effects of Sugar-Sweetened Beverage Taxes on Food and Beverage Demand in a Large Demand System." *American Journal of Agricultural Economics* 96 (1): 1-25.

Table 1. Sample Simulation Result

Variable	Type	Parameter true values	BLP estimates	BLP p-values	RF importance	BLP rank	RF rank
Price	RC, EN	-0.2	-0.197	0.000	1.036	9	4
X1	RC, EX	0.1	-0.104	0.863	0.243	10	7
X2	RC, EX	-2	-2.906	0.053	0.314	2	6
X3	NR, EX	-0.32	-0.299	0.000	0.060	8	9
X4	NR, EX	1	0.956	0.000	0.325	5	5
X5	NR, EX	2.2	2.261	0.000	2.484	3	3
X6	NR, EX	3	2.977	0.000	4.481	1	1
X7	NR, EX	-2.2	-2.233	0.000	2.855	4	2
X8	NR, EX	0.5	0.473	0.000	0.044	7	10
X9	NR, EX	0	-0.016	0.530	0.001	11	11
X10	NR, EX	0.7	0.692	0.000	0.117	6	8

Note: RC=has random coefficient, NC=has non-random coefficient, EN=endogenous, EX=exogenous. This iteration uses 100 brands and 100 markets. $\rho_{Pearson} = 0.721, \rho_{Spearman} = 0.691$.

Table 2. Overall Simulation Results (100 brand-market combinations)

Measure of association	Mean	SD	Median
Pearson correlation coefficient	0.795	0.134	0.825
Pearson correlation coefficient for non-random attributes only	0.962	0.009	0.964
Spearman's rank correlation coefficient	0.682	0.079	0.682
Spearman's correlation for non-random attributes only	0.954	0.039	0.976

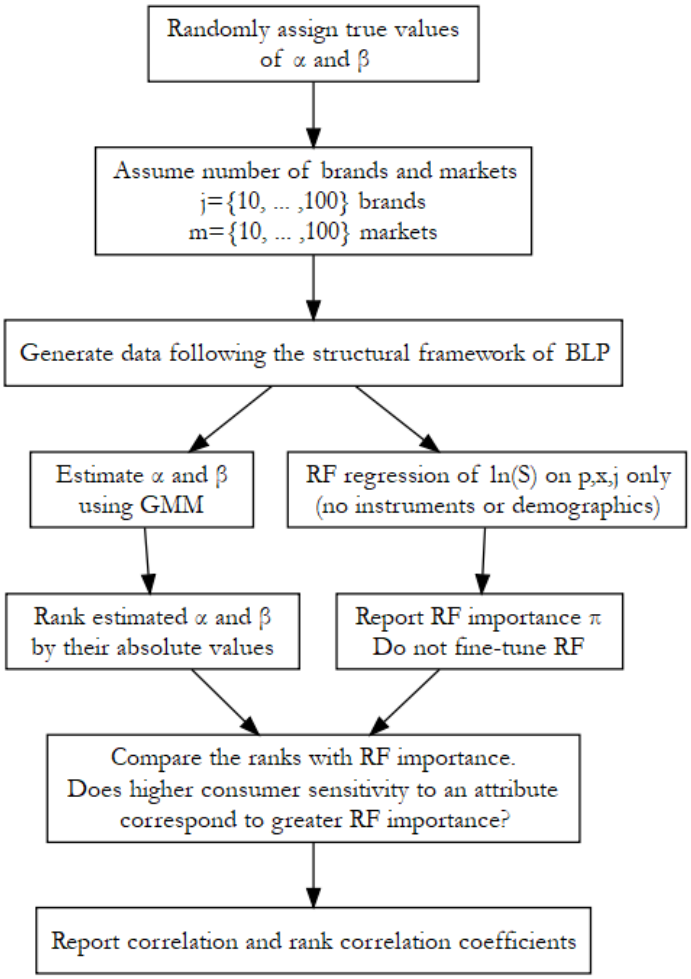

Figure 1. Simulation Flow Chart

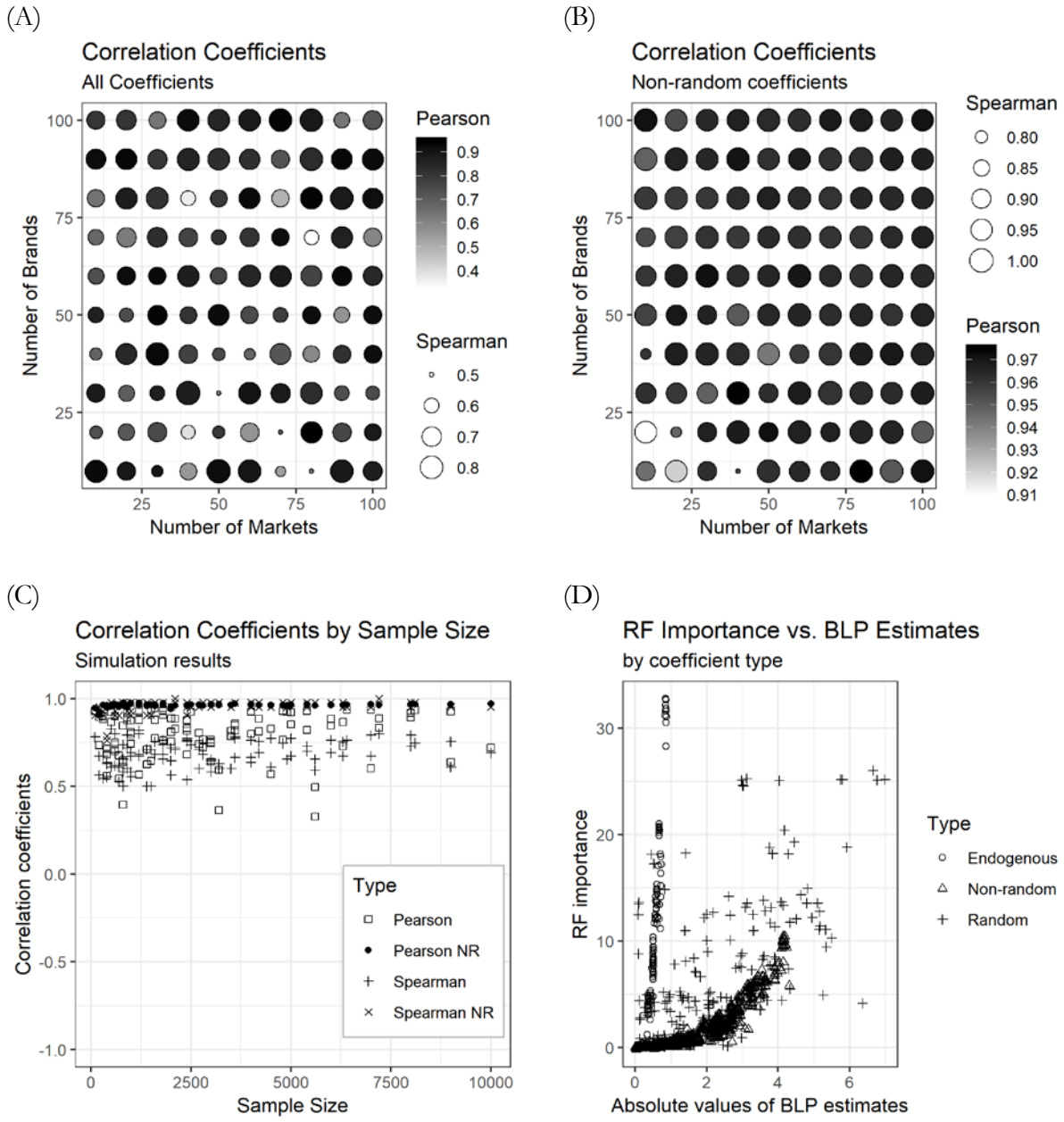

Figure 2. Simulation Results

Replication Codes

The following codes can be used in R software to replicate the table and figures. Package BLPestimatoR, ranger, and ggplot2 must be installed in the system before running the code. Please change the working directory to an address where you want the outputs. The code in its current format runs for assumed true values of the BLP parameters. Please remove the hashtag sign before random true values in Step 1 for generating random true values for each iteration.

Note:

A typical RF algorithm has the following steps.

1. Randomly sample from the training dataset with replacement
2. Fully grow a decision tree for each sample drawn in (1), i.e., no further splits are possible
3. At each node of the tree, select the best split among randomly selected predictors where number of predictors in the subset is chosen as square-root of total predictors.
4. Repeat until T trees are grown.

Replication File: Badruddoza_Amin_McCluskey_2019.R