Heterogeneous Households and Market Segmentation in a Hedonic Framework

By MARTIJN I. DRÖES*[a,b,c], STEVEN C. BOURASSA[d], MARTIN E. HOESLI[e,f,g]

This version: 12 December 2019

SUMMARY — This paper explores Rosen's (1974) suggestion that within the hedonic framework there are natural tendencies toward market segmentation. Taking housing markets as an example, we argue that in the presence of sorting of heterogeneous households, markets can become segmented. This results in the hedonic price line no longer being continuous or unique. We show that market segmentation can be estimated on the basis of an augmented hedonic model in which *both* marginal prices and housing attributes are separated by household characteristics into different classes. The classes can either be exogenously defined or endogenously determined based on an unsupervised machine learning algorithm or a latent class formulation. We illustrate the usefulness of these methods using American Housing Survey data for Louisville and show that there are distinct housing market segments within the Louisville metropolitan area based income, and family structure.

JEL codes — E02; R31; O18

Keywords — latent class; market segmentation; machine learning; hedonic model; heterogeneous households

I. Introduction

The price of a heterogeneous good is typically measured based on its characteristics. For example, a house with more square feet and a garden is worth more than a house that lacks

* Corresponding author, e-mail: m.i.droes@uva.nl, tel.: + 31 20 525 5414. We thank Marc Francke, John Clapp, Dorinth van Dijk, and seminar participants of the ERES conference 2019 in Cergy-Pontoise, the AREUEA intl. conference 2019 in Milan, and the AsRES conference 2019 in Shenzhen for useful comments.

[a] Amsterdam Business School, Faculty of Economics and Business, University of Amsterdam, Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands.
[b] Amsterdam School of Real Estate, Jollemanhof 5, 1019 GW Amsterdam, The Netherlands.
[c] Tinbergen Institute, Gustav Mahlerplein 117, 1082 MS Amsterdam, The Netherlands.
[d] School of Urban and Regional Planning and School of Public Administration, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA.
[e] Geneva Finance Research Institute and Swiss Finance Institute, University of Geneva, 40 boulevard du Pont-d'Arve, CH-1211 Geneva 4, Switzerland.
[f] University of Aberdeen Business School, University of Aberdeen, King's College, AB24 3FX, Aberdeen, Scotland.
[g] Kedge Business School, 680 Cours de la Libération, 33405 Talence, France.

those attributes. A limitation of focusing only on housing attributes, a purely hedonic approach, is that the characteristics of the owner of the good are typically ignored. That is, in the hedonic theory as formally laid down by Rosen (1974), the hedonic price function is a reduced form equation which only depends on housing characteristics. The classical hedonic model has been convenient as no information on household characteristics is necessary to estimate it. Since then, there have been ample of studies which have used household characteristics in combination with the hedonic model to, for example, analyze bargaining power (i.e., Harding et al., 2003), to capture unobserved amenities (among others, Bourassa et al., 1999), and to identify housing demand/individual preferences (e.g. Ekeland et al., 2004).

The aim of this paper is to use household information in combination with the hedonic model to delineate (housing) market segments. The main idea behind this traces back to a quote by Rosen (1974, p. 40) in which he notes that "… a clear consequence of the model is that there are natural tendencies towards market segmentation … segmented by distinct income and taste groups …". The implications, however, for the hedonic model were not explored in further detail. This paper aims to fill this gap.

Our contribution is twofold. First, using housing markets as an example, we redefine the hedonic price function to allow for secondhand markets in a novel way using an Edgeworth box. The standard hedonic approach is based on trade between consumers (households) and profit-maximizing firms. However, in many cases goods are sold via secondhand markets.[1] Although the Edgeworth box is a standard tool in consumer theory, and the role of bargaining in secondhand markets has been explored in earlier work by Harding et al. (2003), the Edgeworth box has, to the best of our knowledge, not been applied to the hedonic model. The model we propose has several unique features in comparison to previous hedonic and housing market models. The hedonic model based on the Edgeworth box is characterized not by bid (buyer) and offer (seller) curves but by household endowment (wealth) and marginal willingness to pay curves: *a consumer can be a buyer of some housing attributes, but a seller of others*. In addition, the Edgeworth box allows for outcomes as a result of perfect competition as well as those based on bargaining. We thus consider our model to be a generalization of Harding et al. (2003). As the attributes of a heterogeneous good are typically not perfectly separable (i.e., they are sold in bundles) and it is, therefore, unlikely that markets will clear, we define equilibrium between multiple consumers by connecting several Edgeworth boxes, creating a *trade chain*, and allowing households to buy the configuration of housing attributes they want using money (excess cash holdings) as an intermediary good.

[1] Rosen (1974, p. 37) states that the '… possibilities for resale of used items in secondhand markets are ignored … by assuming that secondhand markets do not exist …'. Alternatively, Rosen (1974) argues that goods can be assumed to be pure consumption goods.

The second contribution relates to the role heterogeneous households (preferences) play in determining market segmentation. In particular, assuming that different types of households sort themselves into different types of houses, we would expect marginal prices and quantities to be clustered by means of household characteristics. Conveniently, the fact that we center our attention on secondhand markets allows us to focus on the characteristics of the current owner of the house (household heterogeneity) only. We particularly show that in the presence of market segmentation the hedonic price function is *no longer continuous or unique*. There can either be gaps or overlaps in the hedonic price line. We propose three empirical approaches that incorporate both information on household and housing characteristics to estimate the clustering in marginal prices across groups: (1) an exogenous class model based on simple interaction terms between household and housing characteristics; (2) a two-step hybrid model that uses an unsupervised machine learning algorithm (k-means clustering) to cluster the data (backend) and the standard hedonic model to estimate marginal prices (frontend); and (3) a fully statistical latent class model that jointly determines marginal attribute prices and the probability that a house belongs to a particular class.

Since our theoretical framework suggests that differences in marginal prices are in themselves not sufficient to identify market segments as differences in marginal prices can occur at different parts of the same hedonic price line, we need to additionally identify whether houses with the same attributes are traded for different marginal prices (overlap hedonic price lines) or there is a gap in the hedonic price line for at least some of the attributes.[2] Because similarity is a matter of degree, we examine the overlap in the distributions of the housing variables across classes. We propose using the Bhattacharyya Coefficient (Bhattacharyya, 1943) – a measure not often used in economics but popular in pattern recognition – which estimates the exact degree of overlap between the distributions even if those distributions are not continuous, as is often the case in empirical work.

We estimate the models using the American Housing Survey (AHS) metropolitan public use file for the Louisville, KY-IN MSA for 2013. We use these particular data because we wanted to show that it is already possible to estimate these models using a single wave of a single MSA (i.e., with a moderate amount of observations). We propose using household income and family structure (the presence of children), two key determinants of housing choice, as clustering variables.

The results show that with all of the three empirical approaches there is strong evidence that the marginal prices are separated in distinct market segments. In terms of model fit, each consecutive model improves upon the standard hedonic model. The naïve approach of creating exogenous classes, where classes are defined based on high or low income and

[2] For example, the fact that a house with more square feet might be traded at a lower price per square foot is not sufficient to claim that this house belongs to a different market segment.

having children or not, already marginally increases the fit. The two-step hybrid model, which first clusters the sample based on housing and household characteristics, does not necessarily perform much better than the exogenous class model. However, the hybrid approach more soundly rejects the equality of marginal prices across clusters/classes and, thus, does a better job in finding those differences. Our preferred, three-class, latent class model shows an increase in the R-squared from 0.637 (standard hedonic model) to 0.782 and the Akaike's Information Criterion (AIC) is reduced by half. This supports our claim that household information can substantially improve the performance of the hedonic model.

Furthermore, the latent class estimates show that household income and having children determine their own separate and distinct classes. These results seem to support the hedonic model with exogenous separate classes instead of the hybrid model with joint classes. In terms of marginal prices, one particularly noteworthy difference is that those households with children seem to negatively value (-15.7 percent per floor level) living on a higher floor (proxy for living in a condominium) of a building while high income households seem to positively value (15.9 percent per floor level) the same attribute. Overall, the classical hedonic approach does not show a statistically significant effect for this variable, suggesting that the effect is averaged out. In addition, the results show that although the means of the housing attributes across classes are significantly different, there is a high degree of overlap in distributions (i.e., high Bhattacharyya Coefficient), which supports our claim that we identify separate and distinct market segments.

This paper relates to several strands of literature. The hedonic model as laid down by Rosen (1974) is, admittedly, a relatively old model, but its importance is still acknowledged today. A search on Google Scholar shows that there were 16,400 hedonic studies in February 2017, most of them on housing and most of them empirical, and 17,300 a year later – a growth of about 5 percent. Given the still growing literature on this topic and the usefulness of the hedonic model for a variety of purposes, such as estimating the willingness to pay for schools (Black, 1999), the external effects of wind turbines (Gibbons, 2015) and power plants (Davis, 2011), and the fact that this study innovates on both a theoretical and empirical front, again emphasizes our particular contribution.

That housing markets are segmented is long known. Schnare and Struyk (1976), for example, show that there are considerable differences in attribute prices in Boston using a very similar approach to the exogenous class method presented in this paper. Using hierarchical models, Goodman and Thibodeau (1998) find that the housing market in Dallas is segmented by the quality of public education. Alternatively, one particular recent and popular approach has been to define market segments using quantile regressions (e.g., McMillen, 2008). The quantile regression approach is a relatively agnostic approach since it creates market segments based on prices and hedonic characteristics alone but gives no guidance as to what is causing those differences in prices. Instead, using a seemingly unrelated regression methodology combined with principal component analysis and an

iterative hedonic model, Lipscomb and Farmer (2005) find evidence of market segmentation by showing that several distinct household types are living within the same neighborhoods. As mentioned, our paper adds to this literature by comparing several straightforward empirical approaches to link housing characteristics, household characteristics, and marginal prices.

The paper further connects to literature on residential segregation and house prices, including some early hedonic work (e.g., Kain and Quigley, 1970) and a number of papers that use AHS data to explore black-white differences in marginal prices (for an overview, see Zabel, 2008). In addition, although in our opinion heavily underused, the latent class/finite mixture modelling has been applied to other problems, such as distinguishing between different classes of expenditures and use of health care (Deb and Holmes, 2000) or to separate wine prices into distinct classes based on variables like the wine's score and years of aging (Caudill and Mixon, 2016). A closely related paper by Belasco et al. (2012) applies the latent class model to housing in Atlanta and shows that markets are separated by variables like education and age. Our paper provides theoretical justification for the use of the latent class model to price heterogeneous goods and thus further highlights its empirical potential.

The remainder of this paper is organized as follows. Section II presents the hedonic theory and the resulting empirical methodology. Section III discusses the data used in this study. In Section IV, we present the results and Section V shows some limitations and provides suggestions for future research. Section VI concludes.

II. Rosen's model, heterogeneous preferences, and market segmentation

This section first discusses the standard model of Rosen (1974). Subsequently, the equilibrium is redefined based on secondhand markets and a generalization towards multiple consumers is explored. Next, the role of market segmentation, including how to measure it empirically, is examined in more detail.[3]

A. Rosen's model

As in Rosen (1974), assume that the household's consumer choice can be described by the household maximizing utility $U(z, x; \propto_j)$ subject to the budget constraint $m_j = x + P(z)$, with x being the composite, non-housing, numeraire good, m_j being income for household type j, and z being a vector of k housing characteristics $\{z_1, \dots, z_k\}$. The parameter vector $\propto_j$ (heterogeneous preferences) governs the shape of the utility function and is specific to household type j. The function $P(.)$ is the total price for housing with characteristics z and

[3] The theory in this paper has benefitted much from the work of Harding et al. (2003), Malpezzi (2003), Day (2001), Sheppard (1999), and Orford (1999), among others.

its shape is the result of $p(z)$, which is a vector of k marginal attribute price functions $\{p_1(z), \dots, p_k(z)\}$ and $\frac{\partial P(z)}{\partial z_k} = p_k(z)$.

The indifference curves underlying this consumption choice are given by $x(z; u_j, \propto_j)$, where u_j is a specific value of utility, and the slope of the indifference curves of a particular house characteristic is the marginal rate of substitution, $MRS_{z_k} = -U_{z_k}/U_x$. As is standard, utility is maximized when this rate equals $-p_k(z)$ for each z_k. Defining a total bid $\theta = m_j - x(z; u_j, \propto_j)$ for a house with characteristics z as the remaining money a household can spend on housing, keeping utility and income constant, we get to the bid function $\theta(z; u_j, m_j, \propto_j)$. The bid function is a monetary and *inverted version of the indifference curve* as defined above and captures the maximum willingness to pay for a house with characteristics z. The slope for a particular house characteristic, $\theta_{z_1}(z; u_j, \propto_j)$, is thus U_{z_k}/U_x and the utility-maximizing values, x^* and z^*, occur when the total price $P(z^*) = \theta(z^*; u_j^*, m_j, \propto_j)$, which is when $\theta_{z_k}(z^*; u_j^*, \propto_j) = p_k(z^*)$ for all characteristics k.[4] This is equivalent to the statement made before regarding utility maximization and the indifference curves. The benefit of using the bid function is that it intuitively characterizes the economic choice in terms of bids and prices.

Figure 1 combines all of these elements in a single picture. We follow Rosen (1974) in showing a non-linear price function because it allows us to also depict the tangency conditions as stated above and it is also economically plausible (see Ekeland et al., 2004).[5] There are two (types of) consumers who consume different levels of z_1 due to differences in income and preferences. So far, the formulation of the consumer's choice is the same as in Rosen (1974) except for the fact that we immediately specify that utility, the bid function, and hence housing choice, are household specific (through $\propto_j$ and m_j). Consequently, if $\propto_j$ and m_j are clustered, (marginal) bids (and hence attribute prices) are also clustered. This

[4] In contrast to the total bid function, the marginal willingness to pay and hence the marginal attribute prices do not depend on income in this particular setup. This is because the bid function, for simplicity, is assumed to be linear in income (preferences are homothetic). This is not necessary. At a bare minimum the bid function (utility) needs to be defined in such a way (quasi-concave) that the utility maximization results in an optimum. The simplifying assumption we make implies that clustering of marginal prices is based only on preferences, not income, a distinction we will ignore in the rest of the paper.

[5] In line with Rosen (1974), if the housing good is divisible in its characteristics and the conditions are such that arbitrage takes place, the price function would be linear. That is, if we could combine two houses with two rooms into one house with four rooms or, alternatively, separate a house with four rooms into two houses with two rooms, arbitrage would ensure that the price of four rooms is double that of two rooms (ignoring potential conversion costs). The reverse is not necessarily true: a linear price function does not imply that housing is divisible and arbitrage takes place. A non-linear price function can occur due to, for example, economies of scale (see Harding et al., 2003). Also consistent with Rosen (1974), each additional unit of z_1 has an increasingly higher price. It may well be the other way around (i.e., decreasing marginal attribute prices). Ultimately, the exact shape of the hedonic function is determined by the distribution of income and preferences and is mainly an empirical question.

will be convenient later on when we examine market segmentation and heterogeneity in household preferences in more detail.

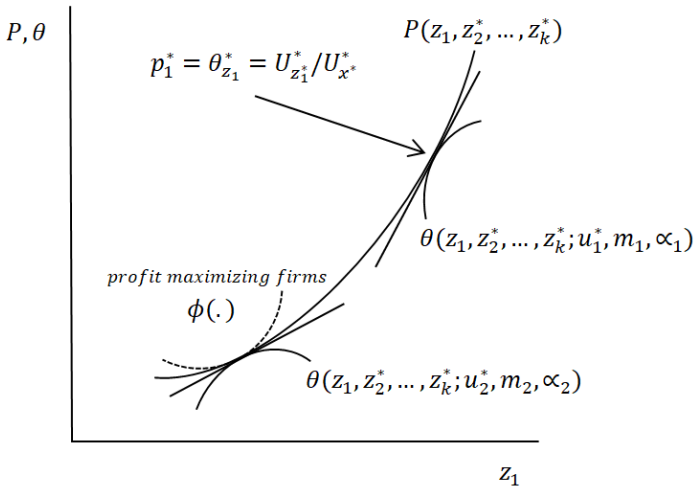

FIGURE 1—THE HEDONIC PRICE FUNCTION, BID FUNCTIONS, AND MARGINAL ATTRIBUTE PRICES

Note: This figure shows the standard utility maximizing solutions underlying the hedonic price function. Emphasis is placed on heterogeneous preferences. In Rosen (1974) profit maximizing firms determine the supply of housing.

B. A secondhand market

The hedonic model as defined by Rosen (1974) is based on the idea that profit-maximizing agents (whether they be firms or landlords) produce and sell heterogeneous goods to consumers. They do so based on their offer functions $\phi(.)$. Houses are sold at the points where bid and offer functions touch (see the dashed line in Figure 1). As is standard practice, general equilibrium – the vector of marginal attribute prices, optimal consumption, and thus the hedonic price line – occurs when aggregate demand equals aggregate supply.

One particular problem, however, is that many heterogeneous goods, like housing or cars, are predominately sold in a secondhand market. Although Harding et al. (2003) examine this issue in great detail, we take a different approach and directly highlight the duality of the problem by representing buyers and sellers through an Edgeworth box (Figure 2). The Edgeworth box is a standard and intuitive tool for examining trade between consumers, but it has not yet been applied to the hedonic framework. In Figure 2, the choice of an additional consumer is depicted by a mirror opposite of Figure 1. To accommodate this, it is necessary to add a secondary y and x axis as it is not possible to directly replace the offer functions of firms with the bid functions of consumers. Again, as the bid functions are

inverted utility functions, not adding a secondary axis would be inconsistent with utility maximization.

The seller in Figure 2 is not a profit-maximizing firm or landlord, but simply another consumer. For our purposes, this is convenient as we are no longer required to refer to both buyers and sellers (firms), but rather *consumers that are buyers with respect to some aspects of z and are sellers with respect to others*. This is a distinct feature of the model that deviates from other hedonic and housing market models. In particular, in the model there is assumed to be a single willingness to pay function $\theta(z; u_j, w_j, \propto_j)$ for each consumer, which, depending on the original endowments, w_j, can *either* be a bid function or an offer function. The main empirical benefit of this approach is that we can, thus, focus on household characteristics and wealth/endowments (i.e., ignoring firm characteristics) to delineate market segments.

To further elaborate, let $w_j = \{\theta_j^w, z_{1,j}^w, \dots, z_{k,j}^w\}$ be the initial endowment for consumer *j*. The endowment does not have to lie on the current hedonic price line as market conditions may well have changed over time. In Figure 2, the consumption choices of two consumers are depicted. The consumers start out at point A. The two indifference curves of consumers 1 and 2 crossing at point A create an efficient lens in which both consumers can achieve a Pareto improvement by trading. That is, equilibrium is now defined not by production and consumption but by trade between consumers. Such trade only occurs if preferences have shifted away from current endowments. The contract curve characterizes all of the mutually beneficial trading possibilities at which the marginal rate of substitution, $MRS_j = -U_{z_1}/U_x$, between consumers is equal and given the appropriate feasibility constraints (i.e., total consumption equals total endowment of a characteristic). In a competitive equilibrium, the marginal price p_1^* of z_1 as well as the final consumption choice (i.e., point B) for consumer *j*, $z_j^* = \{z_{1,j}^*, \dots, z_{k,j}^*\}$ and θ_j^*, will be directly determined by the initial allocation of endowments. In a non-competitive market and in line with Harding et al. (2003), the outcome is determined by bargaining (power). Hence, although the efficient lens in Figure 2 resembles the figure depicted in Harding et al. (2003) it is a generalization in the sense that it comprises both competitive and non-competitive outcomes, it depicts initial endowments, and it contains a secondary y and x axis.

In what follows, we will focus on the perfect competition case. In the example in Figure 2, consumer 1 decides to reduce z_1 and does so by selling his house to consumer 2 for the price θ_1^w and buying the house of consumer 2 for θ_1^* which results in a net transfer of money $C_1 = \theta_1^w - \theta_1^*$. Consumer 1 could for example be an elderly person who wants a house with fewer square feet of living space. Consumer 2, on the other hand, wants a bigger house and

pays $C_2 = (\theta_2^w - \theta_2^*)$ to consumer 1. In equilibrium, $C_1 = -C_2$ and consumers settle on a single marginal price p_1^* for z_1.[6]

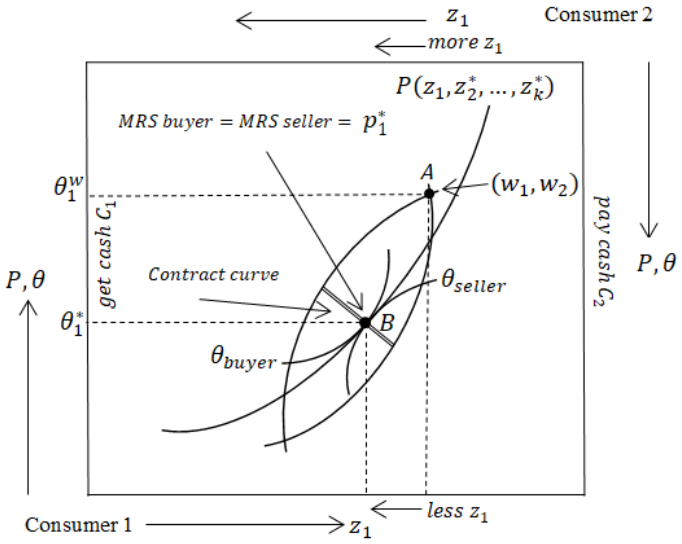

FIGURE 2—BUYERS AND SELLERS IN A SECONDHAND MARKET

Note: This figure shows the hedonic price function in an Edgeworth box as a result of matching between buyers and sellers in a secondhand market.

C. Multiple consumers, multilateral trade, and trade chains

In the case of multiple consumers, general equilibrium prices occur when aggregate demand equals aggregate endowment for each of the housing attributes $\sum_j z_{k,j}^w = \sum_j z_{k,j}^*, \forall k$, and there is no excess housing value in the economy $\sum_j \theta_j^w = \sum_j \theta_j^*$, which is equivalent to saying that there are no excess cash holdings in the economy, $\sum_j C_j = 0$.[7] One particular issue is that the housing attributes are bundled and cannot be separated. That is, consumer 1 needs to find someone (i.e., consumer 2) who would like to buy consumer 1's house – a configuration of housing attributes consumer 2 might not entirely like – and trade houses at the prevailing market price. This makes it unlikely for markets to clear. One particular

[6] This framework does not say anything about how households get matched or how long it takes (search time from the buyer's point of view and time on the market from the seller's point of view) to get matched. For more elaborate search and matching models see the classical model of Wheaton (1990). The model is also agnostic about the actual bidding process; there may be multiple bidders, but the actual trade occurs between one buyer and seller at the point where the willingness-to-pay curves touch. For a classical example including multiple bidders and the role of list prices, see Horowitz (1992).

[7] Interesting extensions would be that a consumer could consider to just own money (or borrow against the house) or to rent rather than owning a house, while possibly using the money toward a pension. This would imply the introduction of corner solutions and require modeling of tenure choice. We leave such considerations for future research.

solution is to allow for multilateral trade of bundles of housing characteristics and have households use money as an intermediary good to obtain the exact (configuration of) house they want. Figure 3 shows an example by connecting the consumption plots (i.e., Figure 1) of three consumers, in a different way than the Edgeworth box, creating a trade chain. To conserve space, the figure is rotated by 90 degrees.

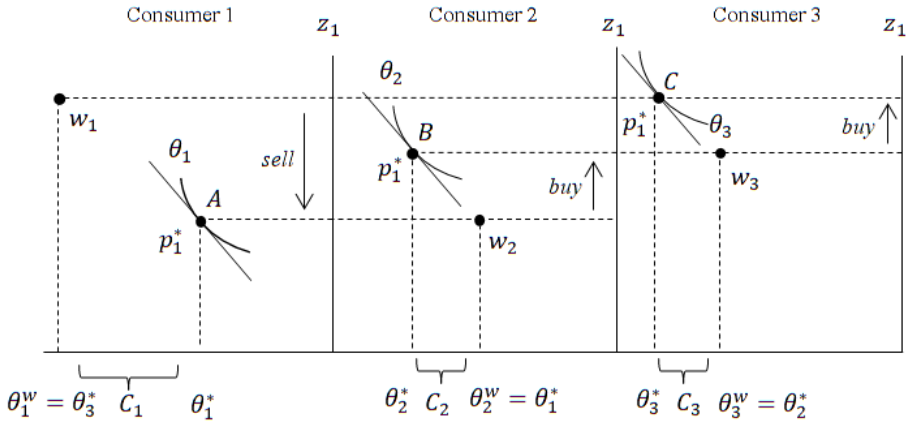

FIGURE 3—MULTIPLE CONSUMERS: A TRADE CHAIN

Note: This figure shows trade among three consumers in which each consumer moves from endowment point *w* to point A, B, or C, respectively. The horizontal axis contains the price and net cash that is paid or received by consumers. The vertical axis depicts a particular characteristic on which trade occurs. Some consumers are buyers, some are sellers. The figures are linked through the optimal point of a consumer (A, B, or C) and the endowment of the next consumer. In this example, consumer 1 buys the house of consumer 2, consumer 2 of consumer 3, and consumer 3 of consumer 1, creating a closed trade chain.

Let z_1 be the size of the house. Consumer 1 starts at the endowment point w_1 and, due to a shift in preferences, decides to sell the house for θ_1^w and to buy a smaller house for θ_1^*. This new optimum is depicted as point A in Figure 3. The net cash C_1 received by consumer 1 is the difference between the selling price of the old house and the purchase price of the new one. The new house of consumer 1 is the old house of consumer 2, which is depicted by endowment point w_2. The value of the house of consumer 2 at this particular endowment point θ_2^w is equivalent to the price θ_1^* received from consumer 1. Consumer 2 wants a somewhat larger house, but not as large as the one that consumer 1 owns. Consumer 2 buys a new house at point B. Note that consumer 2's net cash holdings/wealth (C_2) is negative. This does not necessarily mean consumer 2 has to borrow money (i.e., there is no financial market in this framework), but that there is less to spend on other consumer goods x. Consumer 2 purchases a house from consumer 3. Consumer 3 also wants a bigger house and ends up buying the house of consumer 1. Consumer 3 ends up at point C. Conveniently, in

this example markets are cleared at a single marginal price p_1^* for housing attribute z_1. A more realistic assumption is that due to differences in the types of houses traded and heterogeneous preferences there are multiple market segments (prices) for z_1, an extension we will discuss in the next section.

Trade in Figure 3 occurs only on the dimension of z_1. This is not necessary. Consumers can trade on multiple dimensions as long as the first order conditions stated before hold and markets for each of the attributes clear. An open question is how long the above-mentioned trade chains are and whether they are closed. Alternatively, open trade chains imply that markets do not clear and are in constant turmoil. This might be an alternative explanation, which we leave for future research, for multiple prices for the same housing attribute at the same point in time. Also, the degree of heterogeneity of the good may determine the number of consumers necessary for markets to clear. In thinly traded markets (i.e., at particular locations or particular points in time), some attributes (like square footage) might have a clearly defined price, while other attributes might not.

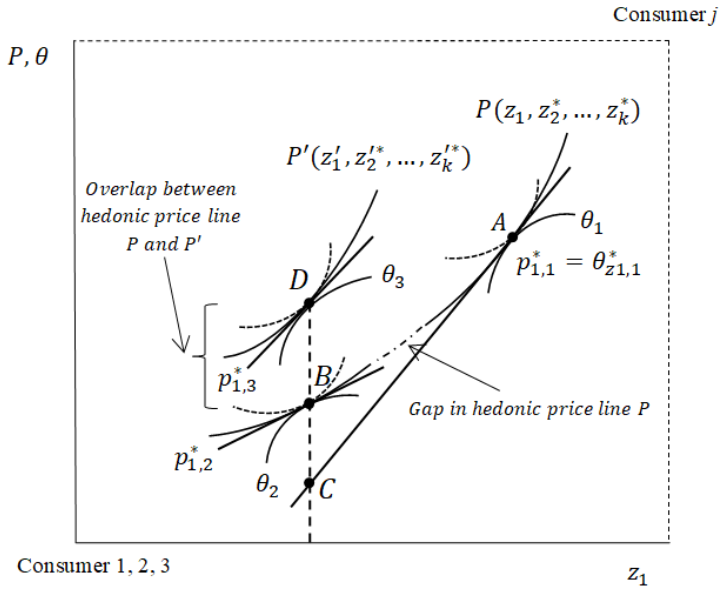

FIGURE 4—MARKET SEGMENTATION: THREE TYPES OF CONSUMERS

Note: This figure shows three types of consumers (1,2,3) trading at points A, B, and D with other consumers *j* (dashed line willingness to pay curves). The consumers at points B and D are clearly trading in two segmented markets. The marginal prices at the same level of z_1 are different. The hedonic price lines *P* and *P'* are overlapping. The consumers at point A are also paying a different marginal price (point C) from the consumers at point B, but may well be trading at a different part of the same hedonic price line. This may still reflect different market segments if the hedonic price line is interrupted in between (gap).

D. Market segmentation

The fact that due to heterogeneity in preferences ($\propto_j$) and income/wealth (w_j) different households buy different houses is not particularly insightful. However, it becomes interesting, as mentioned by Rosen (1974), when housing markets are naturally segmented by income and taste groups. Such markets can exist when goods are not perfect substitutes (no perfect arbitrage), which is likely the case for heterogeneous goods, and there is demand and supply for the good in each segment. Figure 4 shows an example of segmented markets. We distinguish between two different cases: a case where the hedonic price lines are overlapping (point B versus D) and one where there is a gap in the hedonic price line (point A versus B). Both cases represent what we define to be market segmentation.

To be specific, three types of consumers (1, 2, 3) are depicted in Figure 4. Assume that they sort themselves based on their preferences and wealth into particular types of houses and, as such, trade at equilibrium points A, B, and D. They have a different marginal willingness to pay $\theta^*_{z_1,j}$ and, consequently, are trading at different marginal prices $p^*_{1,j}$. Note that we are now explicit about the fact that there are different types of consumers by adding the subscript *j*. The consumers at point B are trading z_1 for marginal price $p^*_{1,2}$, which is at the optimum exactly what they are willing to pay ($\theta^*_{z_1,2}$). The total amount they will pay is again given by the bid function θ_2, which at point B has the value θ^*_2. The hedonic price line is again defined as $P(z_1, z^*_2 \dots, z^*_k)$. The consumers at point D, however, are trading at a different hedonic price line $P'(z_1, z^*_2 \dots, z^*_k)$. Markets are segmented. They pay a different marginal price than the consumers at point B *even at the same level* of z_1. In particular, the marginal price at point D is higher than at point B. Consumers at point D could for example be households with children who value an increase in square footage more than households without children who trade at point B. The third type of consumer trades at point A and *possibly* at a different marginal price than consumers at point B, even if evaluated at the same level of z_1 (point C). However, the consumers at point A may well be trading on the same hedonic price line as consumers at point B albeit at a different part of that line. That is, differences in marginal prices are not a necessary or sufficient condition to signal market segmentation. For that it is also important to look at which part of the hedonic price line (the level of z_1) trade occurs. If there is a gap in the hedonic price line in between points A and B, even if marginal prices are the same at those points, we also argue that markets are segmented.

Practically speaking, market segmentation is a matter of degree. To be concrete, let $S_j(z^*_1, \dots, z^*_k)$ define the set of optimally chosen house characteristics for household type *j*. If consumers are trading at different marginal prices, there is evidence of market segmentation if at least some of the characteristics of the houses they buy overlap. That is, $S_j \cup S_i$ is non-empty, where $j \neq i$. In addition, instead of trading at a particular point, as in Figure 4, consumer class *j* may be trading along a section of the hedonic price line. This is

particularly relevant for continuous characteristics, like square footage. In this case, S_j captures multiple points for each characteristic k and the question is whether the distribution of trades of a particular characteristic overlap for consumer class j relative to consumer class i. That is, are houses with, for example, a *similar* size traded for different marginal prices? Alternatively, if the distributions do not, or to a low degree, overlap this is in line with the second case of market segmentation as defined above. Market segmentation in the end is thus a reflection of both differences in (marginal) prices and quantities, resulting in a situation in which the hedonic price line is no long continuous and/or unique.[8]

The situation depicted in Figure 4 is an example of market segmentation but it is not specific about *why* such market segmentation exists. Market segmentation is the result of imperfect arbitrage which can have several different sources. Two well-known sources are search (Kim, 1992) and transaction costs (Lundborg and Skedinger, 1999). Therefore, the variation in marginal prices should be less in liquid markets or in an environment with better, more symmetric, and frictionless information.[9] Another source has to do with indivisibility of the housing good and product differentiation (for example in relation to zoning, see Henderson, 1985). Housing may be provided only in particular configurations, which impedes arbitrage and strengthens sorting outcomes. Therefore, we would expect that, for more homogeneous goods, like apartments, there should be less variation in marginal prices. Another important element of differentiation is housing quality. For example, a square foot in one house is not the same as a square foot in another, typically reflecting unobserved differences in housing quality (Epple et al., 2013). In what follows we will be rather agnostic about the underlying reasons for market segmentation. We will again touch upon this issue in the limitations and future research section of this paper.

E. Application to the hedonic regression model

The previous section suggests that marginal attribute prices $p^*_{k,j}$ may be separated into j different household classes and that those classes might be consuming different amounts of housing attributes. We show several models in which classes are either exogenously (via interaction terms) or endogenously (unsupervised machine learning and latent class analysis) determined. Each method has its pros and cons. The interaction model is easy to apply by including interaction terms between house and household characteristics. The

[8] A property can be traded for the same overall price but still belong to a different market segment because the underlying combinations of housing characteristics and marginal prices are, at least to some degree, different. Also, two houses can have the same characteristics but different marginal prices, and thus belong to different market segments, or houses can have the same marginal prices but different quantities that are being traded.

[9] For the bias in real estate valuation methods as a result of illiquidity, see Lin and Vandell (2007). For an example of the role of asymmetric information in commercial real estate, see Garmaise and Moskowitz (2004). Anenberg (2016) discusses the impact of information frictions on housing market dynamics.

classes are, however, exogenously determined by the researcher. Clear theoretical guidance is necessary to determine which interaction effects to look at.

The hybrid machine learning (clustering) hedonic model determines the clusters directly, but why these clusters are formed is less clear and open to interpretation. House and household characteristics are combined to create clusters (and there is not necessarily a distinction between the two) and formally testing which household characteristics determine the different clusters is not possible in this approach. It is a mechanical approach that can easily incorporate many different house and household characteristics but gives no guidance as to what variables should be included. In addition, this method is also not particularly suited for dummy variables and the results depend on starting values, which implies the model needs to be run several times to show robustness.

In contrast, the latent class model has an underlying statistical model determining class assignment based on household characteristics; it jointly estimates hedonic models for each class. The latent class model also allows us to formally test whether a particular class characteristic has a statistically significant effect on the probability of belonging to a class. The method itself is, however, rather complicated (joint estimation hedonic and multinomial logit model via maximum likelihood) and does not scale well with the number of parameters and size of the dataset (it may take a long time to estimate or the algorithm might not converge at all). The following paragraphs discuss the different methods in more detail.

Let the heterogeneous preferences/income that separate households into J different classes be measured by household characteristics h and a class j be defined by a particular combination of those characteristics. Then, an augmented hedonic price function would be:

$$\log(P_j) = \sum_j \sum_k \beta_{k,j} z_{k,j} + \varepsilon_j, \tag{1}$$

where P_j is the transaction price paid by a consumer in class j, $\beta_{k,j} = p^*_{k,j}$ and captures the (average) marginal attribute prices for class j, and ε_j is the error term. Note that this equation is stacked over all observations/housing transactions i, with $i = 1,\ldots,n$. In comparison to the standard hedonic model, equation (1) has j specific marginal prices for each house characteristic z_k. Naturally, each class j can also choose different amounts of the house characteristics (*e.g.* house size, number of rooms), $z_{k,j}$. When $\beta_{k,j} = \beta_{k,l} = \beta_k$, something we will empirically test, equation (1) reduces to the standard hedonic model. Hence, we will start by estimating the standard hedonic model for benchmarking purposes and subsequently examine the class specific estimates. Using the standard hedonic model when equation (1) applies is not necessarily incorrect, but the coefficient estimates would just be measuring the average of the marginal attribute prices across classes. In this article, we are particularly interested in the variation around the average.

Equation (1) can be implemented by either estimating hedonic price functions for each of the *j* classes or by pooling across classes and adding *j* binary class indicators and interaction terms between the house characteristics and binary class dummies.[10] We will start with the latter approach. In particular, we begin with a model with exogenous classes (dummy variables) based on household income and family structure (i.e., having children) and add interaction terms between the hedonic characteristics and the class indicators.[11] Standard F-tests on the interaction terms and comparing the adjusted R-squared with the standard hedonic model will help in determining whether this hedonic 'interaction effect' model is useful.

The unsupervised machine learning approach that is applied is k-means cluster analysis. While supervised machine learning is based on the idea that the clusters are a priori labelled and the model is trained based on a training dataset such that it can deal with new data, the k-means approach just clusters the (unlabeled) data into different sets. In this paper, we create clusters combining information on house prices, house characteristics, and household characteristics (household income, having children). Let these variables (vector for each observation *i)* be represented by $\boldsymbol{d} = \boldsymbol{d}_1, \dots, \boldsymbol{d}_n$. The observations are clustered into sets, $\boldsymbol{C} = C_1, \dots, C_J$, by calculating the centroid $\boldsymbol{\mu}_j$ of each cluster (we start with random clusters) and iteratively minimizing the within-cluster sum of squares (WSS), the Euclidean distance to the centroids, according to

$$\arg\min_{\boldsymbol{C}} \sum_{j=1}^{J} \sum_{\boldsymbol{d} \in C_j} \left\| \boldsymbol{d} - \boldsymbol{\mu}_j \right\|^2. \tag{2}$$

The number of clusters itself is again exogenously fixed. However, there are several goodness of fit measures that can be used to determine the optimal number of clusters. In particular, we can look at the WSS, the log of this measure, the $\eta^2 = 1 - WSS(j)/WSS(1)$ which is a measure very similar to the R-squared, and the proportional reduction in error (PRE) (see Makles, 2012). Since the clustering outcome depends on the initialization, we run 100 replications and calculate the average of the goodness of fit measures. Those measures are ordered by the number of classes and plotted in so-called scree plots. After having defined the optimal number of clusters, we use the result (clustering) with the highest η^2 within the 100 replications to estimate a hedonic regression for each class. Although less

[10] It is also possible to create continuous classes by interacting house characteristics with household characteristics directly. This would imply that there is a smooth transition in the marginal effects between classes. Although this might in some cases be true, the latent class model presented and estimated later typically shows that there is evidence for a few distinct classes.

[11] Our theoretical model suggests that we should use a measure of household wealth, which is not reported in the AHS. Instead, to simplify matters, we will use household income which is well known to be (highly) correlated with wealth.

efficient, this avoids a large number of interaction terms. While it is possible to estimate the equations separately, we added the parameters into a single vector and estimated the joint variance-covariance matrix (seemingly unrelated estimation, sandwich estimator). The coefficient estimates are the same as OLS, but the benefit of this approach is that it becomes easy to do cross-equation tests. In addition, although we will use distinct classes and the same hedonic functional form for each class, the variance-covariance matrix is robust to classes that are not strictly separate and even robust to different hedonic functional forms across classes. In comparison to a full-fledged (deterministic) machine learning model, the two-step hybrid approach still allows us to do standard hypothesis testing.

In contrast to the two-step method, a more elegant approach is to use a latent class formulation.[12] The general idea is that each observation is drawn from a population of j classes. Instead of the deterministic k-means approach, there is a probability π_j that an observation i belongs to a class j. The distribution of $\log(P_i)$, $g(.)$, is a mixture of the latent class distribution $f_j(.)$ according to:

$$g(\log(P_i)\,|\boldsymbol{\mu},\boldsymbol{\pi}) = \sum_j \pi_j{}^{d_{ij}} f_j\big(\log(P_i)\,\big|\mu_j\big)^{d_{ij}}, \tag{3}$$

where $\mu_j = \mu\big(z_{k,j},\beta_j\big)$ and d_{ij} is a binary variable indicating if $\log(P_i)$ belongs to class j. Assuming that π_j is distributed according to the multinomial logistic distribution, equation (3) can be estimated using maximum likelihood based on the Expectation Maximization (EM) algorithm.[13] The output is a jointly estimated multinomial logit model (that determines the probability that an observation i belongs to class j) and a hedonic model for each class j.[14] In our case, the probabilities are modeled as a function of household characteristics h. We use log income and a dummy for having children. The number of classes is exogenously fixed, but it is possible to compare models with different classes and potentially different covariates using the standard Akaike's Information Criterion (AIC). Goodness of fit can, among other things, be determined by examining the estimated posterior probabilities that an observation belongs to a class. In particular, the average should be close to one. Alternatively, a measure of entropy E_j determines the distinctiveness of the latent classes:

$$\mathrm{E}_j = 1 - \frac{\sum_i \sum_j -\hat{\pi}_{ij} ln \hat{\pi}_{ij}}{n\, ln\, J}, \tag{4}$$

[12] A good reference is Cameron and Trivedi (2005). For an empirical application see Belasco et al. (2012).
[13] We will choose starting values in an informed manner using factor analysis.
[14] An equivalent two-step approach would be to determine the probabilities using the multinomial logit model and then to estimate hedonic regressions for each class with weights equal to the probabilities that an observation belongs to the relevant class.

where there are in total *J* classes, *n* observations, and $\hat{\pi}_{ij}$ now specifically recognizes that the posterior probabilities are specific to observation *i*. The entropy measure ranges between zero and one. A higher entropy measure indicates that the observations are better classified into the latent classes.

Finally, of particular interest is the degree of overlap in the distribution of the housing characteristics across classes to measure market segmentation. A simple way is to examine the means and standard deviations of the house characteristics across classes. A paired t-test (two cluster comparison) or F-test using ANOVA (Bonferroni multiple-comparison test) gives an indication of whether the means are statistically significantly different from each other (the distributions are distinct), while the standard deviation gives an indication of whether the distributions are still to a degree overlapping. The problem with this measure of overlap is that it is reliable only when the variables (for each of the classes) are normally distributed, which is rarely the case. An alternative is to look at the overlap in the min-max spread across classes. This, however, does not measure the actual area of the distributions that are overlapping. That is, if the tails of two distributions are overlapping, it does not necessarily imply that both distributions are overlapping as this depends on the joint number of observations that are overlapping. Measuring the actual number of observations that overlap is easy to do in case of a continuous distribution but variables are typically not continuously distributed. A measure that solves all of these aforementioned issues is the Bhattacharyya Coefficient (BC) which was developed by Bhattacharyya (1943):

$$BC = \sum_m q_m l_m, \tag{5}$$

where the sample is split into m partitions and q_m and l_m are the proportion of members of each distribution that are part of the partition. A BC of one signals a perfect overlap in the distribution. A value of zero indicates that the distributions are disjoint. We would expect that, although there are distinct (statistically significant) differences in means, there is still a considerable degree of overlap between the distributions. Although the BC is an old measure and is particularly popular in pattern (image/speaker) recognition; it is not often used in economics. When there are more than two distributions, the average of the BC across all pairs of distributions can be used. The BC measure can be calculated per house characteristic. An overall measure is the average across house characteristics.

III. Data

The data are from the American Housing Survey (AHS) metropolitan public use file for the Louisville, KY-IN MSA for 2013. The MSA definition used for the 2013 survey includes four counties in Kentucky and two in Indiana. Information about location within the MSA in the public use file is limited to an indicator of whether the dwelling is within the central city

(Louisville). The survey defines the central city using its former boundaries, which were superseded when the City of Louisville merged with Jefferson County in 2003.

The AHS gives detailed information for a sample of dwellings, including the owner's estimate of the value of the house, which we will use as our main dependent variable, and characteristics of the structure and lot (if any).[15] It also provides information about the occupants of the house, such as household income and characteristics of household members. Hence the data are particularly useful for the present study because they allow for both hedonic price analysis and for the definition of classes based on household attributes. We selected owner-occupied dwellings including both single-family houses and condominiums. Observations were deleted from the sample if the estimated house value was implausible or if values for key variables were missing. The final sample size is 1,636 observations.

TABLE 1—SUMMARY STATISTICS: HOUSE PRICES, HOUSE CHARACTERISTICS AND HOUSEHOLD CHARACTERISTICS, LOUISVILLE (2013)

Variables	Mean	Std. Dev.	Min.	Max.
Housing variables				
Sale price (expected, $)	196,125	147,843	10,000	1,120,000
House size (sq. ft.)	2,212	1,334	99	7,235
Lot size (sq. ft.)	72,678	182,894	1	956,923
Age of structure (years)	40	24	0	94
Number of bathrooms	2.30	1.02	1	8
Number of rooms	6.64	1.76	2	13
Garage	0.79	0.40	0	1
Dishwasher	0.83	0.38	0	1
Fireplace	0.51	0.50	0	1
Floor	0.02	0.22	0	3
Louisville (former city)	0.17	0.38	0	1
Clustering variables				
Children	0.31	0.46	0	1
Household income ($)	80,319	62,546	1	456,869
Number of observations		1,636		

Note: Based on the AHS Louisville KY-IN metropolitan area public use file for 2013. Floor is the number of floors from the building main entrance to the unit, which is defined as zero for single-family houses and condominiums on the same floor as the main entrance. Children is a dummy variable for the presence of children under 18 in the household.

[15] We are well aware that the expected selling price is at best a crude proxy for transaction prices as it may reflect inaccuracy and strategic misreporting (for a discussion, see Choi and Painter, 2018).

Table 1 gives descriptive statistics for the AHS variables that we use for our empirical analyses. The average (expected) sale price is $196,000 against an average house size of 2,200 square feet. About 17 percent of the observations are within the former City of Louisville boundaries. To illustrate the workings of the three different empirical approaches (see the previous section), the variables used to create classes/clusters are children and (log) household income (as a proxy for wealth). We will discuss several other potential class determinants in the limitations and future research section (Section V.). About 31 percent of the households in the sample have a child (under 18) and the average household income is $80,000.

IV. Results

This section discusses the results of the hedonic models extended based on household characteristics (children and income), per equation (1). We start by discussing the standard hedonic model and then we add exogenous classes (interaction effects). Subsequently, we show the results using endogenous clustering based on equation (2). We highlight the statistically significant differences in marginal prices, including the descriptive statistics (means) for the classes, leaving the detailed description of the class characteristics (mean comparison test, Bhattacharyya coefficient) for the final latent class specification as in equation (3).

A. Standard hedonic model and exogenous classes

Table 2, specification (1), shows the results of a standard hedonic model for the Louisville MSA using the 2013 data. The results are straightforward. An increase of 1 percent in house size increases house prices by 0.3 percent. An increase in 1 percent in lot size increases house prices by 1.9 percent. Building age has a negative effect on house prices, but at a diminishing rate. Each bathroom adds 16.7 percent to house prices and each room 4.1 percent. A garage, dishwasher, and fireplace increase house prices by 13.1 percent, 27.8 percent, and 12.3 percent, respectively. The large effect of a dishwasher on house prices suggests that, at least in this particular hedonic model, this variable might be proxying for unobserved housing quality. In the case of condominiums, the floor level is not statistically significant. The same applies to the Louisville city dummy. Apparently, there are not enough observations to accurately estimate the difference in house prices between the former Louisville city area and the rest of the MSA (the coefficient is, however, also particularly small). The adjusted R-squared indicates that this fairly simple model explains 63.7 percent of the variation in house prices.

Specification (2) contains the estimates for equation (1) using interaction terms with the children (under 18) dummy variable and, to keep in line with the idea of distinct separate classes, a dummy variable for above or below the median sample income of $61,000. The joint significance tests of the interaction effects with the children dummy and the high-income dummy (F-statistics of 1.82 and 5.12, respectively) show that both sets of

interaction effects are jointly statistically significant. The marginal effects for the high-income group and the group with children are also jointly statistically significantly different (F-statistic of 2.09). The F-statistics are, however, not particularly high, suggesting that we might not have chosen classes particularly well.

TABLE 2—HEDONIC MODEL AND EXOGENOUS CLASSES, LOUISVILLE (2013)
(Dependent variable: log sale price)

	(1)	(2)			
	Hedonic	Exogenous classes			F-stat.
		Reference category	Interaction children	Interaction high income	Ref. + child = Ref. + income
House size (log)	0.309***	0.251***	-0.113*	0.184***	10.72***
	(0.0383)	(0.0601)	(0.0616)	(0.0664)	
Lot size (log)	0.0185***	0.0192***	0.00506	-0.000620	
	(0.00423)	(0.00643)	(0.00921)	(0.00796)	
Age of structure	-0.00675***	-0.00557**	0.00169	-0.00481	
	(0.00156)	(0.00273)	(0.00291)	(0.00310)	
Age of structure sq.	5.63e-05***	1.93e-05	-9.65e-06	9.69e-05***	4.04**
	(1.77e-05)	(2.84e-05)	(3.37e-05)	(3.53e-05)	
Number of bathrooms	0.167***	0.138***	0.0429	0.0194	
	(0.0154)	(0.0284)	(0.0294)	(0.0308)	
Number of rooms	0.0414***	0.0457***	0.0120	-0.0225	
	(0.00815)	(0.0153)	(0.0152)	(0.0172)	
Garage	0.131***	0.148***	-0.0359	-0.0111	
	(0.0258)	(0.0420)	(0.0461)	(0.0489)	
Dishwasher	0.278***	0.303***	-0.0825	-0.0829	
	(0.0319)	(0.0444)	(0.0597)	(0.0608)	
Fireplace	0.122***	0.114***	0.0840**	-0.0185	
	(0.0222)	(0.0354)	(0.0413)	(0.0438)	
Floor	0.0347	-0.0448	0.00148	0.296**	
	(0.0695)	(0.0692)	(0.123)	(0.123)	
Louisville (former city)	0.0330	0.00749	0.163**	-0.0367	
	(0.0377)	(0.0492)	(0.0813)	(0.0817)	
Joint sig. (F-stat.)			1.82**	5.12***	1.95**
Adj. R-squared	0.637		0.652		
Observations	1,636		1,636		

Note: Robust standard errors in parentheses. High income is defined as income above the sample median of $61,000. The exogenous class model also includes children and high income as separate variables. *, **, *** indicate 10%, 5%, 1% significance, respectively.

Although the model shows quite a few economically sizable differences in marginal prices, only a few of those differences are actually individually statistically significant. Focusing on those differences, high income households pay 0.18 percent more per square foot (relative to the reference category). In contrast, households with children pay 0.11 percent less. Households with children seem to particularly value a working fireplace. Being in a condominium (higher floor) is preferred by higher income households and households with children seem to pay a premium for housing within the former Louisville city limits. The adjusted R-squared of 65.2 percent suggests a fairly modest improvement in terms of fit relative to the standard hedonic model.

B. Classes based on a clustering algorithm

Figure 5 depicts the average goodness of fit measures (scree plots) for different numbers of clusters (or classes) based on 100 replications. We are looking for a kink in the scree plots as a signal to determine the number of clusters. Although not very distinct, there seems to be a kink at two classes. The PRE suggests that there is a 20 percent reduction in the WSS going from class 1 to class 2. In total there is a 30 percent reduction (η^2) with three classes. Even though there is some indication that the model improves using four or even five clusters, we will show the results for the two-cluster and three-cluster models as these lead to the largest reduction in the WSS. Out of the 100 replications we took the two-cluster and three-cluster models with the highest η^2. Now we also report the descriptive statistics for the different classes, in Table 3. Table 4 contains the equation-by-equation regression results for these two different clustering models.

Regarding the two-cluster model, Table 3 indicates that the second cluster, in comparison to cluster one, has relatively many households with children (42 percent), with relatively high average income ($111,076), living in relatively new and large houses (and not condominiums) with a relatively high price ($288,872) outside the former Louisville city limits. In comparison to the exogenous class model, the clustering model seems to suggest that income and children are correlated and jointly determine the classes. The regression estimates in Table 4, specification (3), shows that the equality of the hedonic coefficients between the two classes is now soundly rejected (χ^2 of 99.48) even though again not many of the coefficients are individually significantly different from each other. For example, an increase of one percent in house size increases house value by 0.35 percent for cluster two but only by 0.23 percent for cluster one. In other words, cluster two households are buying houses which are larger and they are also, ceteris paribus, paying more per square foot. This difference in marginal prices is, although sizeable, not statistically significantly different, which most likely reflects the low number of observations used in this study. Statistically significant differences are that cluster two households pay less for living in condominiums (proxied by the floor variable) and pay more for living outside the former Louisville city area. Although most of the second cluster households do not live within the old Louisville city limits they do seem to be willing to pay an additional price of 18.1 percent when living in the city in comparison to households in cluster one. The age of the structure decreases prices in both clusters but less so at higher ages for cluster two. The overall adjusted R-squared for this model is 0.65 which is actually very similar to the model with exogenous classes. This suggests that having the machine sort matters out does not, at least in this particular example and using this particular clustering method, lead to much gain.

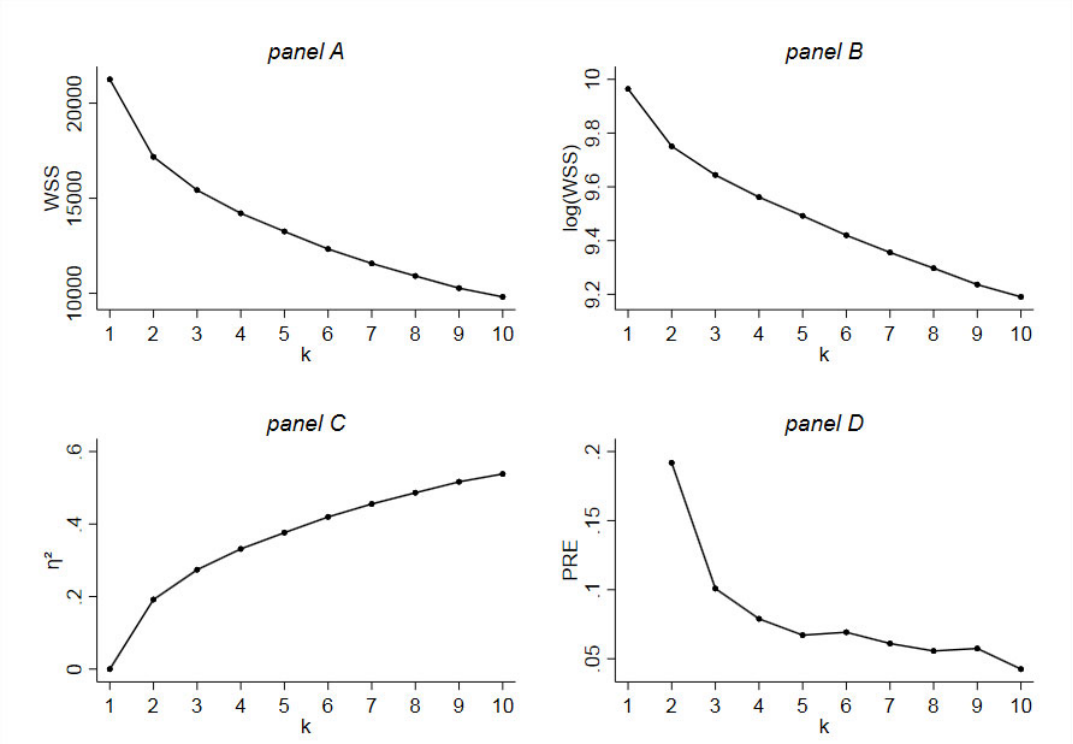

FIGURE 5—SCREE PLOTS CLUSTERING MODELS

Note: Based on 100 replications, this figure shows average goodness of fit measures for k number of classes. Panels A to D show the within sum of squares, the log of this measure, the η^2 (R-squared type of measure), and the proportional reduction of error, respectively.

TABLE 3—SUMMARY STATISTICS: MEANS PER CLUSTER, CLUSTERING MODEL, LOUISVILLE (2013)

	(3)		(4)		
	Cluster 1	Cluster 2	Cluster1	Cluster 2	Cluster 3
Sale price	125,189	288,872	112,063	181,041	413,745
House size	1,535	3,098	1,448	2,089	4,148
Lot size	52,074	99,617	33,306	89,207	101,612
Age of structure	48	29	59	31	29
Number of bathrooms	1.71	3.07	1.49	2.38	3.69
Number of rooms	5.69	7.87	5.59	6.53	9.11
Garage	0.66	0.96	0.52	0.91	0.98
Dishwasher	0.71	0.98	0.52	0.98	0.99
Fireplace	0.24	0.87	0.15	0.62	0.93
Floor	0.04	0.001	0.06	0.01	0.00
Louisville (former city)	0.24	0.08	0.39	0.04	0.13
Children	0.23	0.42	0.20	0.32	0.51
Household income	56,795	111,076	55,716	71,289	158,290
Observations	927	709	543	826	267

Note: The averages are based on the clusters underlying the cluster models of specifications (3) and (4).

TABLE 4 —HEDONIC MODEL, CLASSES BASED ON CLUSTERING ALGORITHM, LOUISVILLE (2013)

(Dependent variable: log sale price)

	(3)			(4)						
	Two-cluster model		F-stat.	Three-cluster model			F-stat.			
	Cluster 1	Cluster 2	1 = 2	Cluster1	Cluster 2	Cluster 3	1 = 2	1 = 3	2 = 3	1=2=3
House size (log)	0.234***	0.345***		0.203***	0.267***	0.272***				8.73***
	(0.058)	(0.046)		(0.076)	(0.040)	(0.074)				
Lot size (log)	0.0179***	0.0167***		0.0188*	0.0189***	0.00897				
	(0.0055)	(0.0062)		(0.010)	(0.0042)	(0.011)				
Age of structure	-0.00635***	-0.00985***		-0.00286	-0.0102***	0.00106			7.64***	
	(0.0022)	(0.0022)		(0.0048)	(0.0020)	(0.0035)				
Age of structure sq.	0.0000513**	0.000115***	2.86*	0.0000250	0.0000884***	-0.00000628			3.50*	
	(0.000023)	(0.000029)		(0.000041)	(0.000030)	(0.000041)				
Number of bathrooms	0.124***	0.165***		0.148***	0.116***	0.194***			6.17**	10.03***
	(0.025)	(0.018)		(0.041)	(0.018)	(0.026)				
Number of rooms	0.0352**	0.0311***		0.0612***	0.0238***	-0.00326				
	(0.016)	(0.0092)		(0.022)	(0.0091)	(0.017)				
Garage	0.100***	0.236**		0.107**	0.172***	-0.0321				
	(0.027)	(0.092)		(0.042)	(0.030)	(0.27)				
Dishwasher	0.275***	0.413***		0.314***	0.231**	0.0488				
	(0.034)	(0.12)		(0.042)	(0.11)	(0.044)				
Fireplace	0.0746**	0.117***		0.117*	0.120***	0.0257				
	(0.033)	(0.036)		(0.066)	(0.022)	(0.073)				
Floor	0.0502	-0.706***	33.31***	0.0623	-0.0617	-				
	(0.073)	(0.11)		(0.091)	(0.099)					
Louisville (former city)	-0.0236	0.182***	7.04***	-0.0726	0.112*	0.204**	5.74**	7.03***		10.29***
	(0.043)	(0.064)		(0.046)	(0.062)	(0.093)				
Equality coef. (χ^2)	99.48***			64.79***						
Adj. R-squared (per eq.)	0.298	0.568		0.252	0.343	0.341				
Adj. R-squared (overall)	0.650			0.662						
Observations	927	709		542	826	267				

Note: Robust standard errors in parentheses. The clusters are based on the k-means clustering method on house prices, house characteristics, household income, and whether there are children under 18 in the household. The hedonic coefficient estimates are the same as separate OLS estimates for the individual clusters, but the (co)variance matrix is different (jointly estimated, sandwich estimator). The adjusted R-squared is based on the regular OLS estimates. *, **, *** indicate 10%, 5%, 1% significance, respectively.

For the three-cluster model, cluster one has the lowest household incomes and house prices, while cluster three has the highest incomes and prices. Cluster three also has a high percentage of households with children (51 percent). A similar type of ordering as before applies with regard to many of the house characteristics with, for example, higher house and lot sizes in cluster three. However, cluster one now has the highest percentage of households living in Louisville city (39 percent). Instead, cluster two households are by all accounts an intermediate group with, for example, household income and the presence of children lying in between group one and three.

Table 4, specification (4), again indicates some differences in the marginal attribute prices across clusters. However, the fact that the joint test of equality of the coefficients has a lower χ^2 suggests that this model might be less appealing than the two-cluster model. Also, several of the individual coefficients are no longer statistically significant. Nevertheless, the adjusted R-squared statistic increases to 0.662 suggesting a better (linear) fit of the model. Some of the results are in line with the two-cluster model, but there are also some noticeable differences. In particular, the three-cluster model now suggests that the differences in the marginal prices of house size are jointly statistically significant with cluster three having the highest marginal price. In addition, a difference in the effect of age now occurs between clusters two and three. There also seem to be differences in the price paid per bathroom. Also, households in cluster three do not live in condominiums (i.e., this variable is omitted from the hedonic estimates). Finally, cluster two is already willing to pay a 11.2 percent premium for living in Louisville while cluster three households pay a 20.4 percent premium.

The results in this subsection again emphasize that separating hedonic models using housing and household characteristics can lead to a statistically significant improvement relative to the standard hedonic model. Relative to the exogenous class model, it also seems to better separate the data into different classes with the χ^2 suggesting highly statistically significant differences, although the model does not necessarily perform better in terms of goodness of fit.

One further issue with this type of hybrid hedonic, machine-learning model is that it is difficult, from an economic point of view, to explain the differences between the groups based on the household characteristics. The fact that households (with high income and children) in cluster three are willing to pay both for housing in the city *and* a large house seems to be a bit in contradiction to each other. Households with children should be willing to pay less for living in the city, while high income households are expected to pay more.[16] Also, we would expect some economically meaningful differences between groups regarding living in a condominium (floor variable). The weakness of the machine learning (clustering)

[16] This argument might not necessarily apply to Louisville, which has attractive older neighborhoods within the former city limits that contain large, relatively expensive houses.

model in terms of economic interpretation stems from the fact that we can choose which variables to include in the clustering algorithm but we do not know how they contribute to the formation of classes (i.e., there is no underlying regression or classification model that we could test). This is exactly what the latent class model does allow for.

C. A latent class model

Table 5 reports the descriptive statistics for the different classes. However, in comparison to Table 3, we now also add standard deviations and t-tests to compare group means. In addition, we report the Bhattacharya Coefficient to compare the overlap in distributions across classes as in equation (5). Table 6 contains the estimates for several latent class models in line with equation (3). In particular, we estimated a latent class model with two classes and one with three classes. The upper part of the table contains the hedonic estimate, while the lower part also shows the simultaneously estimated multinomial logit model with children and the logarithm of income as the independent variables.

Specification (5) is the two-class latent model. The equality of the coefficients across classes is rejected (χ^2 of 187) even more strongly than with the cluster model. The AIC also suggests that this model is a substantial improvement over the single-class (standard hedonic) model.[17] The decrease in AIC from the standard hedonic model to the two-class latent class model is 442. If anything, this suggests that the latent class model better fits the data than the standard hedonic model. This overall increase in fit is also evident by the increase in the adjusted R-squared to 0.735 which is higher than the standard hedonic model, exogenous class model, and cluster model.[18]

The two-class model suggests that having children has a positive and statistically significant effect on the probability of belonging to class two (relative to class one). In particular, having children increases the log-odds ratio by 0.682. This is sizable against the average log odds of 1.749 (i.e., $\log(0.851/0.148)$). The coefficient on income is positive but statistically insignificant. Based on the most likely class assignment, about 85 percent of the observations belong to class two. This suggests that the model performs well in assigning observations to classes. The entropy measure of 0.46, however, implies that the distinctiveness of the classes is not very high. In addition, Table 5 shows that, although many of the means of the house and household characteristics are statistically significantly different between classes, the means of two important variables, house size and price, are not significantly different. This is an indication that the two-class model might not be optimally assigning observations into different classes.

---

[17] To give an indication of the order of magnitude, a decrease in the AIC of 50 already suggests that the model with the lower AIC is $\exp(50/2) = 72$ billion times as likely to minimize the information loss in comparison to the baseline model. The information loss measures the divergence in the probability distribution of the actual data generating process f and the model g we use to estimate it.

[18] We calculated the adjusted R-squared using the most likely latent class assignment and using a similar (interaction effect) methodology as the exogenous classes model.

TABLE 5 —SUMMARY STATISTICS: MEAN, STANDARD DEVIATION, AND OVERLAP PER CLUSTER, LATENT CLASS MODEL, LOUISVILLE (2013)

	(5)				(6)				
	Class 1	Class 2			Class1	Class 2	Class 3		
	Mean (sd)	Mean (sd)	T-test	Overlap	Mean (sd)	Mean (sd)	Mean (sd)	F-test	Overlap
Sale price ($)	205,761	194,444	0.66	0.905	221,468	138,744	234,090	*87.25*	0.870
	(262,474)	(116,949)			(292,793)	(61,608	(152,277)		
House size (sq. ft.)	2,297	2,198	0.90	0.989	2,060	2,036	2,357	*11.71*	0.979
	(1,627)	(1,276)			(1429)	(1,209)	(1,389)		
Lot size (sq. ft.)	138,155	61,256	*4.21*	0.980	158,910	52,240	76,745	*16.62*	0.981
	(276,627)	(158,452)			(302,670)	(150,122)	(181,770)		
Age of structure	54	38	*8.78*	0.945	52	41	38	*17.53*	0.964
	(27)	(23)			(26)	(23)	(24)		
Number of bathrooms	2.14	2.33	*-2.51*	0.986	2.16	2.12	2.44	*19.60*	0.983
	(1.11)	(1.00)			(1.08)	(0.93)	(1.05)		
Number of rooms	6.53	6.66	-1.00	0.983	6.20	6.44	6.84	*13.33*	0.968
	(1.90)	(1.74)			(1.73)	(1.64)	(1.82)		
Garage	0.74	0.80	*-2.25*	0.997	0.75	0.76	0.83	*6.74*	0.997
	(0.44)	(0.40)			(0.43)	(0.43)	(0.38)		
Dishwasher	0.65	0.86	*-6.32*	0.971	0.64	0.78	0.88	*29.67*	0.978
	(0.48)	(0.35)			(0.48)	(0.42)	(0.32)		
Fireplace	0.44	0.52	*-2.32*	0.998	0.45	0.45	0.57	*11.30*	0.995
	(0.50)	(0.50)			(0.50)	(0.50)	(0.50)		
Floor	0.07	0.02	*1.93*	0.996	0.01	0.03	0.02	0.99	0.997
	(0.39)	(0.17)			(0.10)	(0.26)	(0.20)		
Louisville (former city)	0.35	0.14	*6.32*	0.971	0.35	0.17	0.16	*12.82*	0.984
	(0.48)	(0.35)			(0.48)	(0.37)	(0.36)		
Children	0.16	0.34	*-6.89*	0.977	0.11	0.45	0.24	*51.06*	0.961
	(0.36)	(0.47)			(0.31)	(0.50)	(0.43)		
Household income ($)	54,066	84,898	*-6.97*	0.929	48,766	54,143	103,287	*157.85*	0.880
	(64,025)	(61,163)			(54,780)	(42,179)	(66,271)		
Overlap, house char., average	-	-	-	0.975	-	-	-	-	0.965
Number of observations	243	1,393	-	-	109	637	890	-	-

Note: The averages are based on the most likely assignment to a latent class of specification (5) and (6); see Table 6. The overlap is based on the Bhattacharyya Coefficient. For the three-cluster case the average of the coefficient of all pairs of distributions is shown based on separating the sample into ten bins. The t-test column shows t-values based on a paired t-test of the means allowing for unequal variances. In the three-cluster model F-values are reported based on an ANOVA (Bonferroni multiple-comparison) test. The values are in italics if the differences are statistically significant at the 10 percent level or better.

TABLE 6 —LATENT CLASS HEDONIC MODEL, LOUISVILLE (2013)

(Dependent variable: log sale price)

	(5)			(6)						
	Two-class model		χ^2	Three-class model			χ^2			
Hedonic variables	Class 1	Class 2	1 = 2	Class1	Class 2	Class 3	1 = 2	1 = 3	2 = 3	1=2=3
House size (log)	0.271**	0.345***		0.335	0.169***	0.411***			7.42***	8.38**
	(0.106)	(0.0375)		(0.220)	(0.0653)	(0.0486)				
Lot size (log)	0.0456**	0.00967**		0.0649**	0.00702	0.0148*				
	(0.0232)	(0.00481)		(0.0316)	(0.0131)	(0.00842)				
Age of structure	-0.00254	-0.0119***		-0.00714	-0.00436	-0.00995***				
	(0.00501)	(0.00158)		(0.0116)	(0.00313)	(0.00215)				
Age of structure sq.	-2.89e-05	0.000136***	7.83***	7.82e-06	-3.52e-06	0.000122***			9.31***	12.29***
	(4.78e-05)	(2.32e-05)		(0.000114)	(4.61e-05)	(2.51e-05)				
Number of bathrooms	0.237***	0.138***	2.97*	0.280***	0.123***	0.154***				
	(0.0488)	(0.0179)		(0.0911)	(0.0305)	(0.0201)				
Number of rooms	0.0394	0.0408***		0.0188	0.0460***	0.0362***				
	(0.0291)	(0.00830)		(0.0511)	(0.0106)	(0.0114)				
Garage	0.115	0.133***		0.121	0.0936**	0.132***				
	(0.0758)	(0.0239)		(0.169)	(0.0374)	(0.0329)				
Dishwasher	0.368***	0.138***	4.10**	0.510***	0.200***	0.0999**		5.06***		7.07***
	(0.106)	(0.0301)		(0.184)	(0.0629)	(0.0432)				
Fireplace	0.0506	0.158***		-0.0126	0.136***	0.153***				
	(0.0703)	(0.0200)		(0.137)	(0.0407)	(0.0298)				
Floor	-0.0274	0.0928		0.0934	-0.157***	0.159**			10.45***	10.73***
	(0.0908)	(0.110)		(0.174)	(0.0572)	(0.0690)				
Louisville (former city)	-0.140	0.121***	5.07**	-0.109	-0.0596	0.131***			9.85***	9.91***
	(0.101)	(0.0393)		(0.201)	(0.0383)	(0.0466)				
Multinomial logit variables										
Children		0.682**			1.353**	0.476				
		(0.284)			(0.573)	(0.621)				
Household income (log)		0.493			0.125	1.183***				
		(0.401)			(0.113)	(0.301)				
Log pseudo likelihood		-452.63			-364.32					
AIC (single class = 1,405)		963			819					
Adj. R-squared		0.735			0.782					
Average posterior prob.		0.849			0.716					
Entropy		0.455			0.428					
Equality coef. (χ^2)		186.56***			211.31***					
Frequency, most likely class	243	1,393		109	637	890				
	(14.8%)	(85.1%)		(6.7%)	(38.9%)	(54.4%)				
Observations		1,636			1,636					

Note: Robust standard errors in parentheses. The clusters are based on the latent class (multinomial logit) model and simultaneously estimated with the hedonic models using the EM algorithm. The adjusted R-squared is based on the model estimated using the most likely class assignment as dummy variables using the same interaction effect methodology as in the exogenous classes model. *, **, *** indicate 10%, 5%, 1% significance, respectively.

Nevertheless, there are quite a few differences in the marginal attribute prices of the hedonic model across the two classes. Focusing on those that are statistically significant, living within the former Louisville city boundaries adds a positive premium to house prices for class two and the building age-price profile is more curved for this class. This is similar to the previous cluster model. Class two households also seem to pay less for bathrooms and having a dishwasher. Class two contains households that have higher average income ($85,000) and are more likely to have kids (34 percent), although these distinctions are less pronounced than in the two-cluster model. Again, from an economic point of view, it is difficult to distinguish whether the regression results are due to having children, high income, or a combination of both.

Conceptually, the two-class model is similar to the two-cluster model since in both cases we use the children and household income variables jointly instead of separately (as in the exogenous class model). The benefit of the latent class model is that we can formally test whether the variables jointly or separately determine the classes. To do so, we estimate a three-class model. Table 6, Specification (6), shows the latent class estimates based on three classes where each class is determined by the children and household income variables. The results suggest that income and children are important determinants of the class probabilities, but separately (not jointly). In particular, having children has a positive effect on belonging to class two, while household income is an important determinant of the probability of being in class three.[19] Having a child increases the log-odds ratio of belonging to class two by 1.353 against an average of 1.758, while each percent increase in household income increases the log-odds ratio of belonging to class three by 1.183, which is again sizable against the average log-odds ratio of 2.094. Indeed, Table 5 suggests that class two has a relatively high share of households with children (about 45 percent) and class three is dominated by households with high incomes (on average $103,000). Class one can be interpreted as a reference group.

The regression estimates show that there are quite a few differences in the marginal prices across classes. These differences are again highly statistically significant (χ^2 of 221). In comparison to the two-class model, the average posterior probabilities and entropy measure are lower. This is not surprising as it indicates that it is simply more difficult to separate the observations into three classes instead of two classes. The AIC of 818, however, suggests that this model is again an improvement over the two-class model as is also evident by the higher adjusted R-squared of 0.782. The high-income group (class three) is now represented by only 54 percent of the observations in comparison to 85 percent in the

[19] An additional benefit of the latent class model is that it is possible to exclude the household income variable in the multinomial logit model of class two and the children variable in the model of class three (that is, to have separate determinants for each class). This leads to very similar hedonic latent class estimates (not reported).

two-class model. This is because the class with a high percentage of households with children (class two) is now no longer included in the same class.

According to the results in Table 6, class two households pay less per square foot in percentage terms (marginal price of 0.169), while class three household pay more (marginal price of 0.411). Again, this might reflect differences in quality and location of the house, but the fact remains that those differences are there. Interestingly, we find that high income households (class three) value the age of the building more as it is a highly statistically significant determinant while it does not affect prices in class one and two. There is also a strong exponential curvature in this effect. Each year of age decreases house prices by about 1 percent and this effect becomes 0.1 percentage point smaller for each additional year. Note that this particular class (class three) also buys on average newer houses (see Table 5). Furthermore, Table 6, specification 6, suggests that marginal prices for dishwashers are also clustered per class with the highest marginal price for class one and the lowest for class three.

One of the most convincing regression results is the effect of the floor level (proxy for living in a condominium). The high-income group (class three) values living on a high floor. For each unit increase in floor level, this group pays a 15.6 percent higher house price. In contrast, for households with children it is very inconvenient to live on a high floor. For this group every unit increase in floor level is associated with a discount of 15.7 percent. In the standard hedonic model this variable had no statistically significant effect (see Table 2), as only the average effect of this variable was measured. Similarly, high income households seem to value the city more (with a premium of 13.3 percent), while the effect is now negative, albeit not statistically significant, for families with children.

Finally, we take a closer look at the descriptive statistics for the classes underlying the three-class model, see Table 5. The means of the housing characteristics are statistically significantly different across classes. This suggests that the classes are really separate classes and, in comparison to the two-class model, this now also applies to house prices and house size. The standard deviation and overlap measure between classes (Bhattacharyya Coefficient, see equation (5)), however, imply that the distributions across classes are to a substantial degree overlapping. For example, the average age of the property differs substantially, ranging from 52 years in class one to 38 years in class three. However, both lie within one standard deviation from each other and the overlap coefficient of 0.97 implies a high degree of overlap. This is suggestive of market segmentation: the marginal prices across classes are different, but the underlying house characteristics at which houses are traded are to a degree similar.

Overall, these results suggest that: (1) the use of household information to separate the hedonic model into separate classes can increase the fit of the model substantially; (2) these classes reflect different market segments; and (3) the differences in marginal prices across classes have an economically meaningful interpretation. The latent (three-)class model

seems to be most useful in this regard and supports the idea of classes determined separately by either income or having children (relative to a benchmark class), which is in line with the exogenous (but naively specified) class model and in contrast to the hybrid model based on machine learning.

V. Limitations and future research

This section provides several directions for future research. First, the hedonic model is a rather static framework. It does, for example, not directly allow for capital gains or life cycle considerations. A model that does incorporate such considerations is presented by Ortalo-Magné and Rady (1999; 2006). They show that an increase in house prices increases demand and creates upward pressure on house prices higher up the property ladder. It would be particularly interesting to examine how such dynamic effects would change the hedonic equation. For example, McMillen (2008) shows that house price appreciation in Chicago for high-priced homes is not explained by structural characteristics or the location of the house (i.e., types of houses sold), but mainly as a result of changes in the hedonic (quantile) coefficients. How do such changes in coefficients relate to changes in market segments?

Second, although there are many possible extensions to the modeling framework in this paper, an obvious one is to allow for mismatch between housing and household attributes (see Glaeser and Luttmer, 2003). To examine this in more detail it would be necessary to use the panel data structure of the AHS separating the marginal prices over time. This would for example also allow for a longitudinal analysis of mismatch that could relate such mismatch to residential mobility. In particular, we would expect that an increase in mismatch increases the probability that a household moves. From a broader perspective, this is related to the concept of trade chains as discussed in the theory part of this paper. How do these trade chains extend over space and time?

Third, although we focused on two particular household characteristics, income and having children, there are potentially more variables that affect the class assignment of households. We started this study with a grid search to determine which variables might be important. The results showed that many of them – ranging from ethnicity to age, unemployment, and marital status – might separate marginal prices into distinct classes.[20] This relates to the broader question about the role of sorting within the hedonic model (see Yinger, 2015).[21] Particularly, what are the different mechanisms behind sorting and how

[20] Some of the individual characteristics used in this study may also proxy for bargaining power (see Harding et al., 2003). The point is that even in the absence of bargaining (i.e., in the perfect competition case) individual characteristics are expected to separate marginal prices into different classes. Further research should focus on how bargaining and market segmentation are interrelated.

[21] Yinger (2015) estimates amenity demand elasticities focusing on school quality and neighborhood ethnic composition and shows, within the hedonic framework, that these are the result of sorting. Although the framework provided in our paper is based on the idea that specific types of households

does this affect the hedonic equation? This requires the use of multiple types of data, not just the housing characteristics that usually accompany transactions data. How do we incorporate such multiple types of data within a single framework? Although it turns out that in our study the latent class model performs particularly well, more so than the hybrid machine learning hedonic model, it is evident that as datasets grow bigger new techniques are required. Machine learning might turn out to be particularly up to the task, although more benchmarking measures are still necessary. Machine learning is, to a degree, still seen as a black box approach (for a discussion see Hutson, 2018). In this study a simple clustering algorithm is used, but there are already more advanced methods available. It might for example be possible to use random forest techniques to better determine (optimize) the hedonic variable selection (Antipov and Pokryshevskaya, 2012; Yoo et al., 2012).

Fourth, more research is necessary about the role location plays in the hedonic model. On the one hand, location dummies are typically used to measure the effect of unobserved amenities since data to capture all different amenities in a neighborhood may be scarce. One location is not the same as another and this might be a broader reflection of the quality of the buildings and location. As such, there is a tendency in the hedonic literature to include ever more detailed location fixed effects. On the other hand, too detailed location fixed effects would account for normal variations within a market, and those variations are what we are particularly interested in. One empirical extension of the models estimated above, that mitigates this trade-off, is to add detailed neighborhood information (i.e., census data) to the hedonic model to better explain some of the variations we currently find in marginal prices within the market (Louisville MSA) that is investigated in this study. However, the broader question remains how exactly to define markets.[22] There is a long-standing tradition of using a spatial definition of (housing) markets (see Muth, 1961; Bourassa et al., 2003; Tu et al., 2007, amongst others). Reestimating the models for other MSAs would thus be useful. But do distinctly separate (sub)markets even exist, or are they all interconnected, and is it only physical distance which defines their boundaries?

Finally, there is a long-standing literature on the use of household information to estimate individual housing demand curves and the identification problems associated with such estimation (Brown and Rosen, 1982; Ekeland et al., 2004). We find a strong association between average marginal implicit prices and household characteristics. From a causal perspective, however, implicit prices do not necessarily say something about the underlying willingness to pay functions of households as implicit prices are an equilibrium outcome. It

---

sort themselves into specific types of houses, and we show a consistent way to estimate the resulting clustering in marginal prices, the model itself is agnostic about the exact process of sorting.

[22] In part, the answer to this question relates to the reasons why there are differences in marginal prices as discussed in the theory section of this paper. In particular, we argued that this is related to imperfect arbitrage as a result of financial frictions, indivisibility of the housing good, and incomplete information. Regarding the latter, we would for example expect that there is an impact of benchmarking platforms such as Zillow on the (marginal) price dispersion in the housing market.

is well known that other methods and/or additional information is necessary to estimate the household's demand curve (individual preferences). Information on similar households across markets (Epple, 1987), information over time within a market (Arguea and Hsiao, 1993), functional form restrictions (Quigley, 1982) or, alternatively, the better use of the non-linear functional form of the hedonic price function (Ekeland et al., 2004) have all been proposed to deal with the above identification problem. Although our results suggest that household information might not only have an indirect role in determining marginal attribute prices but may also be used directly in classifying different market segments, further research should focus on the clustering of preferences and its underlying causes.

VI. Conclusion and discussion

This paper has examined the role of heterogeneous households and market segmentation in a hedonic framework. Using housing as an example, we redefined the hedonic equilibrium allowing for a secondhand market using an Edgeworth box and with trades occurring based on a trade chain between consumers. Households exhibiting heterogeneous preferences sort themselves into particular types of houses. This leads marginal attribute prices to be clustered into separate classes. We demonstrated three empirical methods for using household information to estimate those classes in a hedonic model: an exogenous (interaction effects) class model, a hybrid machine learning hedonic model, and a latent class model. We estimated those models using the American Housing Survey (AHS) metropolitan public use file for the Louisville, KY-IN MSA for 2013.

The estimation results indicate that each of the three models is an improvement over the standard hedonic model in terms of model fit. Our final preferred specification of the latent class model consists of three classes, with having children and household income determining the class probabilities. We find that having children increases the log-odds ratio of belonging to class two by 1.353 (relative to the low income, low presence of children, reference class) against an average of 1.758. Similarly, an increase of one percent in household income increases the log-odds ratio to be part of class three by 1.183, which is again sizable relative to the average log-odds ratio of 2.094. In comparison with the standard (one class) hedonic model, we observe an increase in the adjusted R-squared from 0.637 to 0.782 and a reduction by almost half of the AIC.

We find that classes with higher income pay a premium for living in Louisville city and on higher floors (i.e., condominiums), and they pay more per square foot. In contrast, households with children pay a negative premium for living on a higher floor or in Louisville city and are willing to pay less per square foot. Although the average house characteristics are statistically significantly different across classes, and households within those classes seem to trade at a distinct set of marginal prices, the distributions are according to the Bhattacharyya Coefficient to a high degree overlapping. This is highly suggestive of market

segmentation in which (in some aspects) similar houses are traded for different marginal prices.

These results indicate that household information should have a direct place in the hedonic equation. Household information can help to determine market segments and better predict house prices. One could, however, question the usefulness of our approach in case the standard hedonic model already explains 80, 90, or even more percent of the variation in house prices, as is not uncommon. The framework we propose is particularly useful when the focus is not on predicting house prices, per se, but in better understanding why marginal prices are different across submarkets. Our approach is also more in line with current professional practice in which (heterogeneous) goods are typically marketed (to target audiences) based on their physical attributes. A car may, for example, be classified as a typical family car because it has a large backseat area, considerable trunk, and the configuration of both areas can easily be amended to accommodate future changes in family composition. These cars are thus mainly bought by families and are traded at a distinct, but bounded, set of marginal prices.

References

Anenberg, E., 2016. Information Frictions and Housing Market Dynamics, *International Economic Review*, 57(4), pp. 1149-1479.

Antipov, E.A., Pokryshevskaya, E.B., 2012. Mass Appraisal of Residential Apartments: An application of Random Forest for Valuation and a CART-Based Approach for Model Diagnostics, *Expert Systems with Applications*, 38(2), pp. 1772-1778.

Arguea, N.M., Hsiao, C., 1993. Econometric Issues of Estimating Hedonic Price Functions: With an Application to the U.S. Market for Automobiles, *Journal of Econometrics,* 56(1-2), pp. 243-267.

Belasco, E., Farmer, M.C., Lipscomb, C.A., 2012. Using a Finite Mixture Model of Heterogeneous Households to Delineate Housing Submarkets, *Journal of Real Estate Research* 34(4), pp. 577-594.

Bhattacharyya, A., 1943. On a Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions, *Bulletin of the Calcutta Mathematical Society*, 35, pp. 99-109.

Black, S., 1999. Do Better Schools Matter? Parental Valuation of Elementary Education, *Quarterly Journal of Economics* 114(2), pp. 577-599.

Bourassa, S.C., Hoesli, M.. Peng, V.S., 2003. Do Housing Submarkets Really Matter? *Journal of Housing Economics*, 12(1), pp. 12-28.

Brown, J.N., Rosen, H.S., 1982. On the Estimation of Structural Hedonic Price Models, *Econometrica*, 50(3), pp. 765-768.

Cameron, A.C., Trivedi, P.K., 2005. Microeconometrics: Methods and Applications. Cambridge University Press: New York, USA.

Caudill, S.B., Mixon Jr., F.G., 2016. Estimating Class-Specific Parametric Models Using Finite Mixtures: An Application to a Hedonic Model of Wine, *Journal of Applied Statistics*, 43(7), pp. 1253-1261.

Choi, J.H., Painter, G., 2018. Self-Reported vs. Market Estimated House Values: Are Homeowners Misinformed or are They Purposely Misreporting? *Real Estate Economics*, 46(2), pp. 487-520.

Davis, L., 2011. The Effect of Power Plants on Local Housing Values and Rents, *Review of Economics and Statistics* 93(4), pp. 1391-1402.

Day, B.H., 2001. The Theory of Hedonic Markets: Obtaining Welfare Measures for Changes in Environmental Quality Using Hedonic Market Data. CSERGE Publications, Centre for Social and Economic Research on the Global Environment (CSERGE): London, UK.

Deb, P., Holmes, A.M., 2000. Estimates of Use and Costs of Behavioral Health Care: A Comparison of Standard and Finite Mixture Models, *Health Economics*, 9(6), pp. 475-489.

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

Epple, D., 1987. Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products, *Journal of Political Economy,* 95(1), pp. 59-80.

Epple, D., Quintero, L., Sieg, H., 2013. Estimating Hedonic Price Functions when Housing is Latent, GSAI working paper, No 2013-E26.

Garmaise, M.J., Moskowitz, T.J., 2004. Confronting Information Asymmetries: Evidence from Real Estate Markets, *Review of Financial Studies*, 17(2), pp. 405-437.

Gibbons, S., 2015. Gone With the Wind: Valuing the Local Impacts of Wind Turbines Through House Prices, *Journal of Environmental Economics and Management* 72, pp. 177-196.

Glaeser, E.L., Luttmer, E.F.P., 2003. The Misallocation of Housing under Rent Control, *American Economic Review*, 93(4), pp. 1027-1046.

Goodman, A.C., Thibodeau, T.G., 1998. Housing Market Segmentation, *Journal of Housing Economics,* 7(2), pp. 121-143.

Harding, J.P., Rosenthal, S.S., Sirmans, C.F., 2003. Estimating Bargaining Power in the Market for Existing Homes, *Review of Economics and Statistics*, 85(1), pp. 178-188.

Henderson, J.V., 1985. The Impact of Zoning Policies Which Regulate Housing Quality, *Journal of Urban Economics*, 18(3), pp. 302-312.

Horowitz, J.L., 1992. The Role of the List Price in Housing Markets: Theory and an Econometric Model, *Journal of Applied Econometrics*, 7(2), pp. 115-129.

Hutson, M., 2018. Has Artificial Intelligence Become Alchemy? *Science*, 360(6388), p. 478.

Kain, J., Quigley, J., 1970. Measuring the Value of Housing Quality, *Journal of the American Statistical Association*, 65(330), pp. 532-548.

Kim, S., 1992. Search, Hedonic Prices and Housing Demand, *Review of Economics and Statistics*, 74(3), pp. 503-508.

Lin, Z., Vandell, K.D., 2007. Illiquidity and Pricing Biases in the Real Estate Market, *Real Estate Economics*, 35(3), pp. 291-330.

Lipscomb, C.A., Farmer, M.C., 2005. Household Diversity and Market Segmentation within a Single Neighborhood, *Annals of Regional Science*, 39(4), pp. 791-810.

Lundborg, P., Skedinger, P., 1999. Transaction Taxes in a Search Model of the Housing Market, *Journal of Urban Economics*, 45(2), pp. 385-399.

Makles, A., 2012. Stata Tip 110: How to Get the Optimal K-Means Cluster Solution, *The Stata Journal*, 12(2), pp. 347-351.

Malpezzi, S., 2003. Hedonic Pricing Models: A Selective and Applied Review. In *Housing Economics and Public Policy*, O'Sullivan, T., Gibb, K. (Editors), Blackwell Publishing: Oxford, UK, pp. 67-89.

McMillen, D.P., 2008. Changes in the Distribution of House Prices over Time: Structural Characteristics, Neighborhood, or Coefficients?, *Journal of Urban Economics,* 64(3), pp. 573-589.

Muth, R.F., 1961. The Spatial Structure of the Housing Market, *Papers in Regional Science,* 7(1), pp. 207-220.

Orford, S., 1999. The Hedonic House Price Function. In *Valuing the Built Environment: GIS and House Price Analysis*, Routledge: Abingdon, UK, pp. 14-43.

Ortalo-Magné, F., Rady, S., 1999. Boom In, Bust Out: Young Households and the Housing Price Cycle, *European Economic Review*, 43(4-6), pp. 755-766.

Ortalo-Magné, F., Rady, S., 2006. Housing Market Dynamics: On the Contribution of Income Shocks and Credit Constraints, *Review of Economic Studies*, 73(2), pp. 459-485.

Quigley, J.M., 1982. Nonlinear Budget Constraints and Consumer Demand: An Application to Public Programs for Residential Housing, *Journal of Urban Economics*, 12(2), pp. 177-201.

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

Schnare, A.B., Struyk, R.J., 1976. Segmentation in Urban Housing Markets*, Journal of Urban Economics,* 3(2), pp. 146-166.

Sheppard, S., 1999. Hedonic Analysis of Housing Markets. In *Handbook of Regional and Urban Economics*, Cheshire, P., Mills, E.S. (Editors), Volume 3, Chapter 41, Elsevier: Amsterdam, The Netherlands, pp. 1596-1635.

Tu, Y., Sun, H., Yu, S.M., 2007. Spatial Autocorrelations and Urban Housing Market Segmentation, *Journal of Real Estate Finance and Economics,* 34(3), pp. 385-406.

Wheaton, W.C., 1990. Vacancy, Search, and Prices in a Housing Market Matching Model, *Journal of Political Economy*, 98(6), pp. 1270-1292.

Yinger, J., 2015. Hedonic Markets and Sorting Equilibria: Bid-Function Envelopes for Public Services and Neighborhood Amenities, *Journal of Urban Economics*, 86(C), pp. 9-25.

Yoo, S., Im, J., Wagner, J.E., 2012. Variable Selection for Hedonic Modeling Using Machine Learning Approaches: A Case Study in Onondaga County, NY, *Landscape and Urban Planning*, 107(3), pp. 293-306.

Zabel, J.E., 2008. Using Hedonic Models to Measure Racial Discrimination and Prejudice in the U.S. Housing Market. In *Hedonic Methods in Housing Markets: Pricing Environmental Amenities and Segregation*, Baranzini, A., Ramirez, J., Schaerer, C., Thalmann, P. (Editors), Springer: New York, USA, pp. 177-201.