Department of Economics
Johannes Kepler University of Linz

How to Measure Product Differentiation

by

Franz HACKL
Michael HÖLZL-LEITNER
Dieter PENNERSTORFER

Working Paper No. 2101
January 2021

Johannes Kepler University of Linz
Department of Economics
Altenberger Strasse 69
A-4040 Linz - Auhof, Austria
www.econ.jku.at

dieter.pennerstorfer@jku.at

How to Measure Product Differentiation*

Franz Hackl[a], Michael Hölzl-Leitner[a], Dieter Pennerstorfer[a]

[a] *Johannes Kepler University Linz*

January 1, 2021

Abstract

In this article, we provide a novel measure of product differentiation by observing consumer search behavior directly. We track individual consumers in a price search engine and generate a measure of distance in product space, based on goods surveyed conjointly within individual search episodes. This metric performs well in an application to digital cameras as an example of complex products. Regression results show that differences in product characteristics are correlated with our measure of distance to a surprisingly high degree, and that prices are significantly lower if products have to compete with a larger number of close substitutes.

Keywords: product differentiation, characteristic space, consumer search, price search engine, clickstream.

JEL Classification Numbers: D83, D43, L13, L63

*Corresponding author: Dieter Pennerstorfer, Johannes Kepler University of Linz, Department of Economics, Altenberger Straße 69, 4040 Linz, Email: dieter.pennerstorfer@jku.at. Financial support from funds of the Oesterreichische Nationalbank (Oesterreichische Nationalbank, Anniversary Fund, project number: 16255) is gratefully acknowledged.

1 Introduction

In many markets firms sell complex products, characterized by a large number of product attributes. For researchers it is difficult to assess which product dimensions are important for consumers to perceive two products as close substitutes. Is it goods of identical size, of similar longevity, or of the same color that consumers find interchangeable? Interestingly, this questions has received limited attention in applied economic research so far. This is surprising, because identifying close substitutes is important for firms to decide on issues related to product innovation (Flach and Irlacher, 2018) or marketing, but also for competition authorities (e.g. market definition, merger analysis).[1] Furthermore, evaluating the degree of product differentiation is necessary to assess the welfare gains of an increase in product variety, for example due to free trade (see Broda and Weinstein, 2006). This article suggests a simple and easy implementable metric to measure horizontal product differentiation of complex products with many product characteristics based on the click behavior of consumers in frequently visited online product comparison platforms (e.g. price comparison sites).

The empirical literature offers only rather costly generalized methods to determine the degree of product differentiation: Estimating cross-price elasticities would be a natural candidate to measure substitutability of products or, similarly, proximity of goods in product space. Ideally, researchers can observe prices and quantities of all products in the market, as well as exogenous supply shifters for a large enough number of products to convincingly identify all cross-price elasticities.[2] Alternatively, one can draw on a panel of

[1] Giraud-Héraud et al. (2003) show theoretically that market power (and thus prices) of a multi-product firm are highest if that firm is able to monopolize a large segment in product space. This aspect is also stressed by Shapiro (1996).

[2] Without exogenous variation one has to impose — often rather restrictive — assumptions on the distribution of consumer preferences in product space to derive cross-price elasticities, as done, e.g., in the seminal contribution by Feenstra and Levinsohn (1995).

individual consumers' purchase decisions for a long time period.[3] These data, however, are unavailable for many markets.

In contrast, information on prices and a (potentially large) number of product attributes are often easily accessible, but evaluating the relative importance of various attributes remains an unresolved issue. Researchers often focus on a single dimension of product differentiation and provide external validation of the importance of this product dimension (Matsa, 2011; Mazzeo, 2003), or show that the particular measure under scrutiny is correlated with other product attributes (Matsa, 2011). Alternatively, statistical methods such as factor analysis (see e.g. Caves and Williamson, 1985) or multidimensional scaling analysis (e.g. Andrews and Manrai, 1999) are used to reduce the number of product dimensions, or tools like cluster analysis are applied to group goods into different categories of similar products.

These approaches, however, are problematic, (i) because structural models in the spirit of Feenstra and Levinsohn (1995) require strong assumptions regarding the distribution of consumer preferences that cannot be tested, (ii) because the number of product attributes is not necessarily exhaustive, and important dimensions may remain unobserved (in particular regarding product quality), and (iii) because the statistical tools discussed previously provide no information regarding the relative importance of each product attribute.

In this article we pursue a different strategy to identify product substitutability by directly observing consumer search behavior. We track individual consumers in a price search engine to isolate individual search spells. Based on consumers' search behavior this study suggests a simple and easily implementable metric to measure horizontal product differentiation. It is very plausible that two differentiated products are considered as close substitutes by consumers, if they survey (or click) both products within one search episode. In contrast to that, unrelated products (this corresponds to large product differentiation) should be rarely clicked conjointly. Hence, we suggest

[3] Meng et al. (2014), for example, draw on survey data on individual purchase decisions to construct a pseudo-panel based on household-level repeated cross-sectional data to estimate price and cross-price elasticities of alcohol beverages. For durable goods, however, the panel has to cover a time span long enough to observe multiple purchases per consumer to be able to control for individual effects.

that the click frequency of customers during search spells represent a suitable measure of product substitutability (= proximity in product space) for each pair of products. Our method has the considerable advantage that we can measure horizontal product differentiation without having to know exactly all the different product characteristics of complex products.

This approach is most closely related to articles using consumers' "second-choice" information, where customers are asked which product they would have purchased (second choice) if their preferred product (first choice) had not been available. Bordley (1989, 1993) and Berry et al. (2004) use this information to estimate substitution patterns for the U. S. automotive industry,[4] and in particular Berry et al. (2004) stress that using second-choice information substantially improves their estimates on substitution patterns.[5]

Using digital cameras as an example of complex products, our measure of distance in product space performs well in empirical applications: (a) We show that differences in product characteristics are significantly and positively correlated with our measure of distance to a surprisingly high degree. The results also suggest that for complex products the number of product dimensions important to (at least a substantial share of) consumers is quite large. (b) Results from hedonic price functions indicate that prices are significantly lower if products have to compete with a larger number of close substitutes. Furthermore, prices seem to be sensitive to a large variety of rival products (rather than depending on very few close substitutes only), indicating that complex goods compete with a large number of other differentiated products.

The remainder of the article is organized as follows: The following Section 2 outlines how we generate our measure of distance in product space based on consumer search behavior. The available data are described in Section 3. Afterwards, our measure of product differentiation is applied to the market for digital cameras. In Section 4 this distance measure is related to differences

[4] The close relationship between cross-price elasticities, diversion ratios and the share of costumers considering two products as their first and second choice is also emphasized by Shapiro (1996).

[5] A more general discussion on the benefits of micro data in general and second-choice information in particular when estimating demand systems is provided in Ackerberg et al. (2007). Rather than relying on survey data, Conlon and Mortimer (2018) construct second-choice data by removing particular products from the consumers' choice sets in an experimental setting.

in product attributes, and in Section 5 we investigate how the number of close substitutes (based on our measure of distance) influences retail prices. The final Section 6 discusses the results and concludes.

2 A metric for product differentiation

The widespread availability of the internet in general, and of price search engines or price comparison sites in particular, enables consumers to efficiently search markets. Consumers now have access to an enormous number of differentiated products, and are able to gain information on a large number of product attributes and to compare products along these dimensions. Offering firms or researchers also observe the products' attributes and thus face the challenge to assess which characteristics are important to consumers to evaluate which products are considered to be close substitutes. In this article we propose an indicator of distance in product space based on observed consumer search behavior.

Challenges in the definition of a measure for distance: For the purpose of illustration we assume a market with only four differentiated products $\{A, B, C, D\}$, characterized by merely two attributes x and y of equal importance, depicted in panel I-a of Figure 1.[6] Consumer preferences are uniformly distributed across the characteristic space, and consumers receive a disutility (often labeled as transportation costs) if a product is not their ideal variety. The disutility increases with distance between a consumer's preferences and a particular product's attributes.

The polygons spanned around the positions of each product indicate the market areas of each good.[7] Two products compete for the same customers if they share a common boundary, i.e. if these products are neighbors in the characteristic space. In this simple example, cross-price elasticities of each pair depend linearly on the length of the border between these two products,

[6]The equal importance is mediated by identical lengths of the x and y axes in the diagram. See Veendorp and Majeed (1995) for details.

[7]Figure 1 is based on equal prices across products, but this assumption is not important for the purpose of illustration.

and are zero if the products' market areas are not adjacent (see Feenstra and Levinsohn, 1995, equation (12) for details).

While it is easy to locate products with observed attributes in the characteristic space, two issues remain unresolved to assess the substitutability (or the proximity) between products: First, the importance of a particular product attribute is difficult to evaluate. Graphically, the importance of one characteristic can be illustrated by altering the length of this dimension (see e.g. Veendorp and Majeed, 1995).[8] In panel I-b (I-c) [I-d] of Figure 1, the importance of characteristic x (on the horizontal axis) relative to attribute y (on the vertical axis) shrinks to 1/2 (to 1/4) [to 0]. Note that the decreasing importance of characteristic x substantially affects the degree of competition between products and thus their proximity in product space. Product pair $\{B, D\}$ clearly shows how the intensity of competition between these products is affected by the relative importance of category x. These products are very similar in dimension x, but different in attribute y. If both dimensions are equally important, product B competes with product D to a substantial degree. Figure 1 supports this claim, because the border between the market areas of these two products is a rather long segment. If x gets less important, the similarity in this attribute becomes less relevant to consumers, and the length of the border decreases (see panel I-b) or disappears entirely (panels I-c and I-d). If, however, only characteristic x is important to consumers, products B and D are perceived to be virtually identical, as illustrated by panel II-a.

Second, the substitutability does not only depend on the length of the border between two market areas, but also on the consumer density in this area. If consumers are distributed with higher density e.g. in the upper half of panel I-a (i.e. more consumers prefer higher values of attribute y), product D will be perceived as a close substitute to A by more consumers than product B. The issue is further complicated if consumers are heterogeneous regarding

[8] To change the importance of a particular product attribute one could alternatively follow Irmen and Thisse (1998) and introduce different parameters characterizing the transportation costs.

the importance of different attributes, i.e. if characteristic y is more important than attribute x for some consumers, but vice versa for others.

We conclude that while it is easy to observe a large number of product attributes and to therefore locate differentiated products in the characteristic space, it is generally difficult to generate an indicator of distance that reflects the consumers' substitutability between pairs of products, because (i) the relative importance of different attributes is typically difficult to quantify and (ii) consumer preferences are usually not uniformly distributed in product space.

A measure based on consumers' search behavior: Instead of drawing on data of product characteristics we thus suggest using information on consumer search behavior on price comparison websites. The website under scrutiny (i.e. www.Geizhals.at) does not sell products itself, but redirects consumers to online retailers.[9] We record whether a consumer requests a referral to the web-shop of a retailer for a particular product, which we define as a "click" (see Section 3 for details). Even though not all clicks lead to actual purchases, clicks on differentiated products indicate the customer's interest and the perceived substitutability of these products.

If the product characteristics available at the price search engine are exhaustive, consumers can precisely locate all products in the characteristic space and can choose the product variant giving the highest utility. There is no need to further search the market in this case. While price search engines provide information in numerous dimensions, consumers might nevertheless gain additional information, for example on the product's design or the product's quality, by visiting the web-shops of online retailers (e.g. by accessing consumer reviews published on the retailers' websites). Therefore, consumers might inspect a number of different product variants more closely. Differentiated products clicked by a consumer during one search episode are usually perceived as closer substitutes than product pairs which are not clicked jointly.

[9] Note that our approach works on any website, on which search behavior among (hierarchically organized) product structure can be observed (e.g. Amazon, Google Shopping). The hierarchy is helpful to identify potentially similar products which are to be analyzed regarding their horizontal product differentiation. In our case we analyze "Digital cameras" as a subcategory of a more general product group of "Video, Cameras & TV".)

Hence, the frequency of clicks on product pairs can be a good predictor for the substitutability of products and therefor for the degree of product differentiation. Referring to the illustration in Figure 1, it seems plausible that customers with preferences located close to the border of the market areas of two goods are likely to click both adjacent products for information purposes.

Equally distributed consumer preferences: For our first measure of distance we implicitly assume consumers to be uniformly distributed in product space. For each product-pair (l, k) we count the absolute frequency, $AF_{l,k}$, of how often both products are clicked conjointly by consumers during single search spells. Based on the absolute frequency, we calculate the (unweighted) distance, $DIST_{l,k}$:

$$DIST_{l,k} = 1 - \frac{AF_{l,k} - \min(AF_{a,b})}{\max(AF_{a,b}) - \min(AF_{a,b})} \tag{1}$$

$\min(AF_{a,b})$ and $\max(AF_{a,b})$ denote the number of clicks for the pairs of products surveyed together least and most frequently, respectively. The metric is normalized such that the measure $DIST_{l,k} \in [0, 1]$, with $DIST_{l,k} = 0$ ($DIST_{l,k} = 1$) indicates that product variants l and k constitute the product pair with the largest (smallest) number of clicks among all possible product pairs.[10] Table 1 reports a numerical example of a market with four products $\{A, B, C, D\}$, the absolute number of clicks, $AF_{l,k}$, for all possible product pairs (column 3) and the corresponding (unweighted) measure of distance, $DIST_{l,k}$ (column 4), based on equation (1).

Unequally distributed consumer preferences: In empirical applications we cannot rule out the possibility of higher consumer masses at some

[10] Two comments on the suggested min-max normalization: (i) Note that with a large number of products usually a substantial share of product pairs are not at all clicked conjointly, thus $\min(AF_{a,b}) = 0$ in many applications. The distance metric is then simplified to $DIST_{l,k} = 1 - AF_{l,k}/\max(AF_{a,b})$. (ii) Equation (1) assumes a minimal horizontal distance of zero for the most frequently clicked product pair. This assumption is not mandatory as we could define our distance also as $DIST_{l,k} = |\alpha + \frac{AF_{l,k} - \min(AF_{a,b})(\omega - \alpha)}{\max(AF_{a,b}) - \min(AF_{a,b})} - \omega| + \alpha$ with α as the minimal horizontal distance between the closest substitutes and ω the assumed maximal distance between unrelated products. Although this alternative measure correlates perfectly with equation (1), it would also allow a graphical true-to-scale representation of horizontal distances in the characteristic space.

locations in the characteristic space and lower masses at others. Without assuming equally distributed consumer preferences, a high click frequency could also be the result of a higher concentration of consumers in the vicinity of these product varieties. Hence, we should control for the unequal distribution of consumer preferences in the distance measure.

As online market places do not only observe the click-frequency of product pairs, but also the total number of clicks for each product i, TC_i, a proxy for the concentration of consumer demand at certain locations in the characteristic space is readily available. We can, therefore, control for the varying consumer masses by weighting the click-frequency for each product pair by (the inverse of) the average number of clicks for each of the product pairs' items. We hence define the weighted absolute click-frequency by $wAF_{l,k} = \frac{AF_{l,k}}{(TC_l+TC_k)/2}$. Correspondingly, the weighted distance between product variants l and k, $wDIST_{l,k}$, can then be calculated as:

$$wDIST_{l,k} = 1 - \frac{wAF_{l,k} - \min(wAF_{a,b})}{\max(wAF_{a,b}) - \min(wAF_{a,b})} \tag{2}$$

The weighted absolute frequency for each product pair, $wAF_{l,k}$, of the numerical example discussed above, as well as the respective weighted distances, $wDIST_{l,k}$, are also reported in Table 1 (see columns 6 and 7). The difference between these two approaches can be illustrated by comparing the product pairs (A, B) and (C, D): Without weights, $DIST_{A,B} < DIST_{C,D}$ (see column 4). Note, however, that product varieties A and B have much higher consumer masses which have their preferences in the vicinity of these products, whereas the consumer masses for product variants C and D are much smaller (the total number of clicks $TC_A > TC_B \gg TC_C > TC_D$). Hence, the higher absolute value of the click-frequency of the product pair (A, B) in contrast to the product pair (C, D) is not driven by a higher substitutability and therefore a smaller distance between products A and B in the product space, but by the higher concentration of consumer preferences in the vicinity of these products. We control for this concentration of consumers by weighting the absolute number of click-frequencies with the halved sum of the absolute clicks on both products. After accounting for the mass concentration of consumers the weighted distance $wDIST_{A,B} > wDIST_{C,D}$ (see column 7).

Properties and limitations: The purpose of this weighted distance measure $wDIST_{l,k}$ is to provide a simple and intuitive measure of distance in product space based on consumers' observed search behavior, the precision of which may be limited by the lack of comprehensiveness and consistency of the underlying data: (i) In large-scale data with many differentiated products and product characteristics we do not always observe complete and transitive click-frequencies, which would lead to consistent distances in a multi-dimensional characteristic space. (ii) For products with a low degree of substitutability we do not at all observe clicks on the product pairs, but this does not mean that all product pairs without any common clicks are "equally different" (suggested by a distance measure $wDIST_{l,k} = 1$ for all pairs with $wAF_{l,k} = 0$). (iii) There may be consumers clicking on products pairs accidentally. (iv) Consumer search (time) costs are expected to be heterogeneous, and some consumers might thus only click on neighboring products if they are in a narrow corridor close to the product's market border, while others inform themselves about many adjacent products.

Despite these limitations, we are confident that the measure of distance proposed in this article is intuitive and informative. Our indicator of product differentiation is simple, because it is one-dimensioned and easily interpretable. It is comprehensive, because it implicitly takes all (observed and unobserved) product characteristics into account: It draws on consumers' search behavior directly and thus considers the consumers' perceptions of the importance of (a possibly very large number of) different product characteristics. Furthermore, the indicator can be easily calculated even for a large number of products, compared to data and identification problems in estimating a potentially huge number of cross-price elasticities.[11] Finally, our measure of product differentiation performs well in empirical application (see Section 4 and Section 5).

[11] In the empirical applications we use data on 1,642 different digital cameras. This leads to 1.35 million cross-price elasticities to be estimated.

The following section specifies the data available to track consumers' search behavior.

3 Data

To calculate our measure for product differentiation we use data from Geizhals.at. This website is a large and dominant price search engine in Austria, which covers the e-commerce market for more than 100,000 products. This platform does not sell products itself, but lists products offered by (a potentially large number of) online retailers. Like many other online-platforms, the universe of products offered at Geizhals.at can be separated in a hierarchical system of product categories. The most detailed subdivided product groups are subsubcategories, which are typically not related to each other (e.g. camcorder and scanner).

We use subsubcategories as the relevant product group, because they include similar, but not identical products that fulfill a similar purpose. The spectrum of items within each subsubcategory can therefore be interpreted as differentiated products,[12] and we investigate consumer search behavior separately within single product groups. Although products are classified in product groups (in our case subsubcategories), a researcher cannot directly observe a measure for the distinctness (or the degree of substitutability or horizontal differentiation) between products within one of these groups.

The concept of the click-frequencies of product pairs: In this article we suggest the customers' click-frequencies of product pairs to derive an indicator for the degree of horizontal product differentiation. While consumers cannot buy products directly at Geizhals, they can request a referral from the Geizhals-homepage to the web-shop of an online retailer, which we define as a "click". We observe the complete search behavior of each customer. Consumers can either search for the name of a product directly and

[12] For instance products within the scanner subsubcategory (with products Reflecta x8-Scan or a Rollei DF-S 100 SE) or within the camcorder subsubcategory (with products like Vivitar DVR508HD or the Easypix DVC2712) can be considered as similar (differentiated) products, while goods in different subsubcategories are clearly different products.

get a list of the offering online retailers, starting with the retailer charging the lowest product price. Alternatively, consumers can use the hierarchical product structure and select particular attributes. In this case, Geizhals.at displays a listing of different product characteristics at the top of the page and a list of products fulfilling these criteria with the corresponding online retailers below.[13] It has to be mentioned that the website hardly influences the consumers' search processes by other means than providing information in an agnostic way. Manufacturers cannot buy preferential treatment of their products in this price search engine. Furthermore, while many platforms inform their visitors which products were also viewed by customers who clicked a particular product, this is not the case at Geizhals.at.[14]

We define the clickstream as all clicks of a consumer (identified by cookies or IP-addresses[15]) in a particular subsubcategory. As we have data on clickstreams for a longer time period, a clickstream might comprise several search spells. To identify interruptions between two search episodes, we apply a Grubbs (1969) test for outlier detection based on the time intervals between the series of clicks. If the time period between two consecutive clicks is longer than one week and the Grubbs test suggests that this interval is an outlier relative to the time intervals between all other consecutive clicks, we assume that

[13] Consumers can choose how to display these products, which can be ordered alphabetically, by the product price, the number of offers, the first date of appearance, or by relevance.

[14] If at all, the option to sort the products by relevance (when using the hierarchical product structure to search the market) might have an impact on a consumer's search behavior as it is obviously based on other customers' past search activities and might affect the number of click frequencies. Although — as a researcher — we would prefer a product listing in random order, we do not believe that our proposal will be seriously challenged by a relevance listing for the following reasons: (i) We do not use the relevance ranking presented by the website for our measurements, but rather the clicks of consumers on online retailers selling a product. Clicks require an active and independent customers' decision, in which the preferences of the current user and not the history of other users manifest. (ii) Even if we cannot completely rule out the influence of a relevance listing, such an algorithm would only amplify the consumers' signals concerning their preference ranking — products or product-pairs that have already been clicked frequently are clicked on even more often and vice versa. If the relevance ranking would have a strong influence, there is the risk that our measure would rate the distance between very close substitutes as too little. However, as we weight the absolute click frequencies $AF_{l,k}$ with the absolute number of clicks, the argument, that a relevance ranking would invalidate our measure, becomes less important.

[15] This approach does not guarantee a perfect identification of individuals. Cookies might be deleted or IP-addresses can be changed. Moreover, several persons might use the same electronic device. However, despite these difficulties many business activities see cookies or IP-addresses as adequate means to identify consumers.

this long break initiates a new search spell. To calculate the click-frequency of product pairs we count how often a given product pair has been clicked conjointly within the customers' search spells. If a consumer requests referrals to multiple online retailers for the same product within one search spell, we count this as one click for this product only, as we are interested in the number and the identities of all clicked product variants. Figure 2 illustrates clickstreams, search spells and the resulting click-frequencies for three consumers.

An application to digital cameras: We use digital cameras as an example of complex products — characterized by many attributes and a large number of differentiated products — to apply our measure of distance in product space based on equation (2) and to evaluate the metric's performance in empirical applications. We investigate consumers' search behavior over a time interval of four months (September 1, 2012 to December 31, 2012).[16] In this period, we observe 93,535 consumers (IP-addresses) searching for digital cameras, comprising 98,456 search spells. We thus typically observe only one search spell per consumer (1.05 spells on average), which is very plausible for investigating a durable consumer good for a four-month period. Within one search spell, consumers survey 1.64 different products on average (the standard deviation is 2.57).

In our observation period 3,066 different digital cameras are listed in the price search engine. The website provides product information on a large number of predefined characteristics, as illustrated in Figure 3. The manufacturer of each product can be identified by the items' brand names (e.g. Nikon, Sony, Olympus, noname products), and manufacturers typically offer several more or less differentiated products. The website also indicates the number of digital cameras sharing a particular attribute: The red circle in Figure 3, for example, implies that there are 29 digital cameras with at least 40 $\times$ optical

[16] We restrict the time period under scrutiny till the end of 2012, because up to 2012 the website listed the complete set of product features without restrictions. In later versions of the website (starting in 2013), the standard setting has been reduced to the most popular product characteristics and the full list is only accessible on request. As we want to suggest a method which has the potential to detect the relative importance of product features, we want to rely on data whose product features have not been pre-selected by the website.

zoom. The website does not give an indication on the importance of certain product characteristics.

Starting with 3,066 products, we delete all cameras (i) with obvious outliers in important variables (e.g. 100-fold price or zero megapixels), (ii) which are not offered in the middle of our observation period (our cross section results refer to October 31, 2012), (iii) without any clicks at all (these products are obviously not relevant for consumer decisions), and (iv) which are never clicked conjointly with any other product.[17] This leaves 1,642 digital cameras in our data-set, which are clicked on average 73 times and surveyed in 48 different search spells.

Out of the resulting in 1,347,261 product pairs, 1,229,410 (or 91%) pairs are never clicked jointly within a search spell, with $wDIST_{l,k} = 1$ in these cases (see equation (2)). 117,851 (or 9%) do have common clicks (median: 1 click, mean: 2.1 clicks, standard deviation: 4.7 clicks). The top 10 product pairs in the subsubcategory "digital cameras", which are most frequently clicked jointly within particular search episodes by consumers, are reported in Table 2 and contrasted to the top 10 clicked product pairs of all categories available at Geizhals.at. The website records 610 common clicks for the most frequently clicked pair of digital cameras within our observation period. Due to the large number of product pairs which are never clicked conjointly the frequency distribution of the distance measure $wDIST_{l,k}$ is left-skewed with a peak at distance $wDIST_{l,k} = 1$. For distances smaller than one, we observe a mean of 0.985 (median: 0.992) and a standard deviation of 0.0265. Hence, very close distances are rare, which is quite plausible, given that most of the 1641 rival products are substantially different as to a large number of attributes.

4 Importance of product attributes

We expect products with similar attributes to be perceived as close substitutes by consumers. They are likely to be clicked within one search spell, and these product pairs are thus characterized by a small distance in product

[17]Products never clicked conjointly with other products seem to be perceived as "island products" by consumers, unrelated to other goods in this product group.

space, measured by $wDIST_{l,k}$. Consequently, we expect the partial correlation between our distance measure and the difference (distance) in each product attribute to be non-negative. As various product characteristics are perceived by consumers as differently important, the strength of these partial correlations will be heterogeneous, and some of them might well be zero (if attributes are irrelevant to consumers). Non-negative partial correlations (to a substantially high degree) would indicate that our distance measure is a consistent and plausible proxy for horizontal product differentiation. In this case we could interpret the strength of the partial correlations as indicators of the relative importance of specific product characteristics in the characteristic space. This provides valuable insight for managerial product development and marketing decisions, and also allows us to evaluate the plausibility of restricting product differentiation in multiple dimensions to a single dimension only, common in empirical research (see, e.g., Matsa, 2011; Mazzeo, 2003).

Estimation model: Hence, we regress our distance measure $wDIST_{l,k}$ for product pair (l, k) on the differences of this pair's product attributes, $\Delta attributes_{u,l,k}$:

$$wDIST_{l,k} = \alpha_0 + \sum_u \alpha_{1,u} \Delta attributes_{u,l,k} + \sum_v \alpha_{2,v} controls_{v,l,k} + FE_l + FE_k + \epsilon_{l,k}. \tag{3}$$

FE_l and FE_k denote product fixed effects, $\epsilon_{l,k}$ the error term, and α_0, $\alpha_{1,u}$ and $\alpha_{2,v}$ the parameters to be estimated.

For each product characteristic u we include the difference in this attribute between products l and k, $\Delta attributes_{u,l,k}$. The Geizhals.at website reports information on a total of 97 product characteristics (see Table 3 for a complete list). We drop product attributes if characteristics do not vary over the products included in our sample or if the values of these characteristics are observed for less than 2000 product pairs. For variables with only fewer missing values, we interpolate missing values with the sample mean and control for the interpolated values with dummy variables. These dummy variables are included in the $controls_{v,l,k}$ in equation (3). We further eliminate some characteristics in case of high multicollinearity, and drop attributes with a variance inflation factor $VIF > 10$, leaving 85 product features. We aggregate some characteristics covering very similar aspects in the characteristic space (e.g. types of connectors or video formats) to count variables,[18] which reduces the number of product characteristics to 43.[19]

Empirical results: Regression results including differences in these 43 product attributes as explanatory variables are reported in Table 4. The columns in this table vary with the inclusion of different control variables (the number of online retailers offering particular products, and differences in prices and brands). Estimations in all three columns include product fixed effects for products l and k, as well as dummy variables indicating whether missing values of particular variables are imputed. As all variables are standardized, the reported parameter estimates are beta coefficients, allowing us to directly compare the size of the parameter estimates of different variables within one regression.

In the first specification, reported in Column (1) of Table 4, 39 of our 43 differences in product features are significantly positively related to our distance measure $wDIST_{l,k}$. The remaining four characteristics are not significantly different from zero. There is not a single product feature, where the difference between products is negatively related with distance. If we control for additional aspects (see below), the statistics on the significance of coefficients in Column (3) confirm the high correlations of our distance measure with the differences in the product attributes: From a total number of 43 different parameter estimates, 32 are significantly positive at the 1 % level, one is positive at the 5 % level, eight coefficients are not significantly different from zero, and only two parameter estimates are significantly (and unexpectedly) negative (namely sensor size and the number of different special features).[20] We interpret the convincingly high number of (statistically) non-negative coef-

[18] We consolidate information regarding similar attributes by counting (i) the *difference in the absolute number* of available attributes, as well as (ii) the number of *different* attributes (i.e. the number of attributes provided by one but not by the other product in the respective product pair). See Table 3 for details.

[19] We use the full set of 85 product characteristics in the sensitivity analysis, but get very similar results. See below for details.

[20] Negative signs can also be the result of certain production-related trade-offs (e.g. weight and optical zoom).

ficients (between 96 % and 100 % of all parameter estimates, depending on the specification) as strong empirical support that our distance measure is indeed a precise proxy for the degree of horizontal product differentiation.

For that reason we list the most important product features in the upper part of Table 4, ordered in a descending manner based on the point estimates of specification (3). We expect large point estimates for characteristics important to customers, and anticipate that attributes diversifying the products considerably in their intended use to be among them. Not unexpectedly, it turns out that certain special features (e.g. Display at the front, Display 3D technology), as well as strongly advertised product characteristics (like sensor resolution in megapixel, the dimension, or the range of the digital zoom) are the most important product features.

Additional control variables, included in columns (2) and (3) only, take the expected signs and are significantly positive: Products of the same brand, provided by a similar number of online retailers listed at Geizhals.at, and offered for similar (best) prices[21] are characterized by smaller expected values for our distance measure $wDIST_{l,k}$. The latter result indicates that products sold for similar prices are perceived as closer substitutes, and that the variance in prices of similar products is small. Including these additional controls shows that the sign, the point estimates and the statistical significance of the parameter estimates on the differences in product characteristics hardly change for most attributes, increasing our confidence in the plausibility of our distance measure.

Sensitivity analysis: Table 5 shows regression results for the sensitivity analysis, but reports only summarizing figures on the statistical significance of the parameter estimates for brevity. All regressions again include the respective control variables and product fixed effects, as in the main specification reported in Table 4. Instead of aggregating some of the product features, the first panel includes the full set of 85 different product characteristics available. Again, the shares of significantly negative coefficients are very small. Only

[21] The best price is defined as the lowest price of the respective product charged by any retailer listed at Geizhals.at at October 31, 2012.

between four and six out of 85 parameter estimates are significantly negative at the 10 % level,[22] while eight to 14 are not significantly different from zero, and between 65 to 73 estimated coefficients are significantly positive. Thus, including product attributes in a more disaggregated way gives very similar results.

In the second robustness test we eliminate all product pairs with $wDIST_{l,k} = 1$. These are all product pairs which are never clicked conjointly and can therefore add only limited information to our distance measure of horizontal product differentiation. While the amount of negative coefficients remains fairly constant, the number of insignificant parameter estimates increases. This is not surprising, because the reduction in product pairs (and thus also in products) reduces the statistical power of OLS estimates, and the decrease of the variance of observable characteristics results in a higher variance of the estimated coefficients. It is, however, important to notice that restricting the sample to product pairs clicked conjointly at least once increases the R^2 from about 0.06 to 0.60, which is substantial and again confirms the validity of our distance measure. Apparently, the clicks on product pairs summarizes the *wisdom of the crowd*[23] about the horizontal differentiation (substitutability of products) surprisingly well.

In the third sensitivity analysis, we address the statistical properties of the distribution function of our distance measure. 91 % of product pairs are never clicked within one search spell and the distance of these pairs $wDIST_{l,k} = 1$. Although no consumer perceives these pairs as close substitutes, it does not necessarily mean that the distance in product space is identical for these product pairs, implying that our distance measure is censored at $wDIST_{l,k} = 1$. We thus estimate equation (3) by a Tobit model (see Tobin, 1958) with an upper limit at one. The last panel of Table 5 shows the results of Tobit estimations, which are computationally rather demanding due to the large

[22] These are the following characteristics: Connector for infrared available, Flash video light, Display CSTN technology, Viewfinder with LCD technology, Video format MOV, Sensor size.

[23] The notion "wisdom of the crowd" refers to Francis Galton's observation that the average values of a crowd at a county fair accurately guessed the true weight of an ox, although none of the individual values came close to the true value (see Galton, 1907). Galton is seen as 'father' of the theories on collective intelligence. His work gave the impetus for predictive analytics.

number of fixed effects. Again, the regression results are very similar and provide additional evidence of the plausibility of our distance measure.

5 Hedonic price functions

Based on hedonic or reduced form price equations, the empirical industrial organization literature provides widespread evidence that products in areas of the characteristic space with more competitors tend to be sold at lower prices (Ackerberg et al., 2007). Our measure of product differentiation, $wDIST_{l,l}$, provides a way to identify the proximity of rival products and thus the number of close substitutes.

Estimation model: To reproduce these results, we estimate the following hedonic price function for all 1,642 digital cameras in our sample:

$$Price_i = \beta_0 + \beta_1 NC_i + \sum_p \beta_{2,p} controls_{p,i} + \epsilon_i, \tag{4}$$

with the variable $Price_i$ as the lowest price for digital camera i of all online retailers offering the product at Geizhals.at on October 31, 2012. NC_i ("number of competitors") counts the number of products that are close substitutes[24] and $controls_{p,i}$ include the full set of product attributes discussed above, brand fixed effects, as well as the number of online retailer selling the respective product via Geizhals.at. ϵ_i denotes the error term, and β_0, β_1 and $\beta_{2,p}$ the parameters to be estimated.

To calculate the number of products perceived as close substitutes by consumers, NC_i, we determine a threshold value for our distance measure, $\overline{wDIST}$, and count the number of products with a distance measure below this threshold level, i.e. $wDIST_{i,k} \leq \overline{wDIST}$.[25] In different model specifications we set the threshold distances such that the number of products within this distance is 1, 5, ..., 100 *on average*, and denote this as "Radius 1", "Radius 5", ..., "Radius 100".

[24] Below we will also differentiate between the number of own-brand and foreign-brand substitutes.

[25] We use the distance measure to the h^{th}-nearest neighbor directly in the subsection "Distance to rivals" rather than "number of close substitutes". See below for details.

Empirical results: The regression results of the hedonic price function are reported in Table 6, where Columns (1) to (6) account for the number of products within the different radii "Radius 1" to "Radius 100". The coefficients show how the price (in Euros) is related to the number of rival products within the respective threshold distance. An additional product within the "Radius 10", for example, is associated with a price discount of 0.389 Euro. With the exception of "Radius 1" reported in Column (1), we see that the parameter estimates are significantly negative and tend to decline in absolute values when we increase the threshold distances to identify close substitutes: While an additional product within "Radius 30" is associated with a 22 Cent lower price, this figure declines to 16 Cent and 9 Cent for "Radius 50" and "Radius 100", respectively. Furthermore, a larger number of offering retailers increases retail market competition and thus leads to lower prices of the cheapest supplier. With R^2 values above 0.8 the models explain a substantial part of the price variation of digital cameras.

The pattern documented in Columns (2) to (6) is confirmed when we include the number of rival products in different distance bands (rather than distance rings) together in one regression, as reported in Column (7): The parameter estimate of rival products within "Radius 10" is about twice as high compared to the coefficients on products *between* radius 10 and 30 or *between* radius 30 and 50, while the estimated parameter on the number of products located further away in product space is negligibly small and not significantly different from zero. These results suggest (i) that retailers who want to sell their merchandise in the warehouse set lower prices if there are many similar products, (ii) that this price effect is the larger the closer the substitutes are, and (iii) that retailers are sensitive to the proximity of a remarkably large number of rival products.

Own versus foreign brand products: The theory of horizontal product differentiation suggests to distinguish between products of the same and different producers (which we can identify by the products' brands). Giraud-Héraud et al. (2003) show that multiproduct firms can charge substantial markups if they monopolize parts of the product space with their products. A new product introduced by a rival firm should increase competition and lower prices in this area of the product space, whereas this should not be the case for newly introduced same-brand products.

In the first block of Table 7 we thus split the number of products within the radii “Radius 30” and “Radius 50” in own brand and foreign brand products.[26] Each block represents a separate set of regressions, for which we show non-standardized and beta coefficients. Contradicting our expectations, we do not find increasing prices for a higher number of own-brand products in the vicinity of our products. On the contrary, the price-decreasing effect for additional own brand products is even stronger than the price reductions found for additional foreign brand products.[27]

This result delivers important and interesting insights into the vertical oligopoly structure of e-commerce markets: In e-commerce upstream manufacturers use downstream retailers to distribute their products to consumers. But apparently manufacturers do not have enough market power in the market for digital cameras to influence the final consumer prices. A substantial part of the retailers buy their merchandise in stock. Own brand products obviously pose an even greater threat to the stored products than additional foreign brand products. From the perspective of both retailers and customers, brand is obviously just an additional product attribute. This presumption is confirmed by a glance at the beta coefficients in Table 4, which reveals that brand is among the most important product characteristics. Hence, in vertical oligopol structures the standard assumptions concerning the effects of own and foreign brand products might be questioned, and depend on the manufacturer’s power to influence (downstream) retail prices.

We illustrate the presumption, that brand is nothing more than another product characteristic — at least from the viewpoint of retailers and consumers — with the similarity to other important product characteristics in

[26]The qualitative results do not change if we use other radii.

[27]A technical note: There is a systematic difference in the underlying means of the explanatory variables. The number of rival products (within both “Radius 30” and “Radius 50”) of foreign brands is about four times as large compared to the number of same brand products. Hence, this has to be taken into account when interpreting the absolute size of the non-standardized coefficients in Columns (1) and (3). To simplify comparisons, we therefore also include the beta coefficients in both Table 7 and Table 8.

the remaining blocks of Table 7. We use the counts of products with identical and with different product characteristics within the respective radii as explanatory variables in the regressions. Again, for all important product features we find evidence that higher numbers of competing products with identical important product features reduce the prices more strongly compared to competing products which differ in this important product characteristic.

Predecessors, successors and product availability: Instead of product characteristics, Table 8 splits the number of products within "Radius 30" and "Radius 50" into predecessor and successor products. We use the initial listing on Geizhals.at as the time of market entry, and define the predecessor (successor) as the product introduced first (last) for each product pair. The results show that the number of close substitutes introduced later (i.e. successors) are negatively and significantly related to the price of the product under scrutiny. Contrariwise, the number of predecessors is not significantly related to prices. This is a plausible result, because retailers want to get rid of their stock and successor products are likely to be perceived as much stronger competitors by retailers than predecessor products.

Furthermore, immediately available products are expected to be perceived as stronger competitors than products which have a longer delivery period. We observe the availability for product listings for each retailer on the Geizhals homepage. We classify a product as "available" if there is at least one retailer who can deliver the product immediately. As expected, prices are more strongly related to the number of available (rather than unavailable) products.[28]

We interpret these meaningful results as additional empirical support for the plausibility of our distance measure for horizontal product differentiation.

"Distance to rivals" rather than "number of close substitutes": Table 9 replicates Table 6 with an alternative concept to measure the closeness

[28]We were not too concerned with endogeneity issues in previous applications due to the timely structure of the decision processes. However, as decisions on prices and availability are made simultaneously by the retailers, we cannot exclude potential endogeneity problems here. We thus interpret these regression results in a descriptive rather than a causal way.

to competitors in the product space. Based on our measure for horizontal product differentiation we use the distance to the h^{th} competitor (5^{th}, 10^{th}, ..., 100^{th}). Whereas Columns (1) to (6) calculate the isolated effects of the h^{th} competitor, Column (7) also accounts for the respective distances between the selected neighboring products. We find significantly positive parameter estimates for all distance measures, suggesting that products are sold at higher prices if rival products are located further away (in product space). This sensitivity analysis thus also confirms our results. Column (7) further shows that even if we control for the distance up to the 20^{th} closest product, the distance between the 20^{th} and the 30^{th} closest rival still influences the retailers' price setting. We therefore again find evidence that digital cameras seem to compete with a quite large number of rival products.

6 Discussion and Conclusion

In this article, we propose a simple and intuitive distance measure of horizontal product differentiation. To construct this measure, we use data on consumer search behavior on an internet platform where costumers inform themselves about the available products and their characteristics. We construct our measure of horizontal product differentiation based on the frequencies of commonly clicked products during consumers' search episodes. Our measure is thus based on the assumption that different product variants clicked jointly by consumers are considered as close substitutes. This metric reduces the degree of product differentiation to one dimension, even if products are characterized by numerous attributes and evaluating the importance of these product features is difficult.

To show that this measure is indeed informative, we provide two applications by using data on consumer search behavior for digital cameras in the Austrian price search engine Geizhals.at. First, we investigate pairs of products and show that our distance measure is positively correlated with the differences in most of the pairs' product attributes. This application shows that more similar products (in a large number of dimensions) are considered as closer substitutes by consumers, and allows us to identify the relative importance of different product characteristics, providing interesting insights for product design and marketing policies. Second, we demonstrate that our proxy for horizontal product differentiation works well to determine the number of close substitutes. The analysis of the competitive pressure exerted by these rival products — identified by our measure of distance in product space — shows that products with a larger number of close substitutes are sold at lower prices, which is in line with empirical evidence (see Ackerberg et al., 2007).

Although the main focus of this article lies on suggesting a measure for horizontal product differentiation, the application of our distance measure reveals two interesting results beyond the main purpose of this article:

(i) For retailers and consumers brand is nothing more than a product characteristic. Manufacturing many very similar products does not necessarily lead to a monopolization of certain areas of the characteristic space and thus to higher product prices (as suggested by Giraud-Héraud et al., 2003), if the manufacturers cannot control retail prices. This is the case here, where manufacturers distribute their products via a competing retailer network in the e-commerce market. In contrast, two differentiated products by the same manufacturer are seen by consumers as closer substitutes compared to a pair of products from different producers, ceteris paribus. As the supply of own-brand substitutes poses a greater threat to the retailers' stock than foreign-brand products, we thus have stronger price-dampening effects on the retailers' prices by own-brand substitutes.

(ii) The literature in industrial organization often argues that few competitors suffice to bring markets close to a competitive equilibrium. Reinhard Selten (1973) summoned this viewpoint with the quote that "four are few and six are many", and spatial competition models with products differentiated in one dimension (see Hotelling, 1929; Salop, 1979) show that firms compete with a maximum of two rivals directly. In a similar vein, empirical models of market entry following the seminal contributions of Bresnahan and Reiss (1991) and Berry (1992) usually find that the competitive effect of new entrants on firm profitability declines quickly if the number of firms active in a market increases, suggesting that each product competes with a few close substitutes only. Economists therefore tend to neglect the additional effects of more than four or five competitors. These results, however, are often based on markets offering products and services that vary in a small number of dimensions only, such as pharmacies, plumbers (Bresnahan and Reiss, 1991) or the airline industry (Berry, 1992). Our results show that this simplification is not valid for complex products in high-dimensioned characteristic spaces. While we find that closer substitutes exert stronger competitive effects, we document statistically significant price effects even for the 30^{th} distant competitor (see Section 5). We attribute this result to the multi-dimensionality of our characteristic space. This interpretation is consistent with the results provided in Section 4, suggesting that dozens of product attributes are important to consumers, and consistent with spatial competition models: From the perspective of a certain product attribute, even the thirtieth neighbor[29] could indeed be the nearest competitor in that dimension.

[29] Note, that this is a conservative assessment. Given the significant coefficient for the variable „Products between 'Radius 30 and 50'" even the 49th neighbor might have impact on the price setting game.

References

Ackerberg, D., Lanier Benkard, C., Berry, S., and Pakes, A. (2007). Econometric tools for analyzing market outcomes. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, volume 6A, chapter 63. Elsevier, 1 edition.

Andrews, R. L. and Manrai, A. K. (1999). Mds maps for product attributes and market response: An application to scanner panel data. *Marketing Science*, 18(4):584 – 604.

Berry, S., Levinsohn, J., and Pakes, A. (2004). Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of Political Economy*, 112(1):68–105.

Berry, S. T. (1992). Estimation of a model of entry in the airline industry. *Econometrica*, 60(4):889–917.

Bordley, R. F. (1989). Generating market elasticity estimates using cross-sectional first-choice and second-choice data. *Journal of Business & Economic Statistics*, 7(1):141–146.

Bordley, R. F. (1993). Estimating automotive elasticities from segment elasticities and first choice/second choice data. *The Review of Economics and Statistics*, 75(3):455–462.

Bresnahan, T. F. and Reiss, P. C. (1991). Entry and competition in concentrated markets. *Journal of Political Economy*, 99(5):977–1009.

Broda, C. and Weinstein, D. E. (2006). Globalization and the Gains From Variety. *The Quarterly Journal of Economics*, 121(2):541–585.

Caves, R. E. and Williamson, P. J. (1985). What is product differentiation, really?. *Journal of Industrial Economics*, 34(2):113 – 132.

Conlon, C. T. and Mortimer, J. H. (2018). Empirical properties of diversion ratios. Working Paper 24816, National Bureau of Economic Research.

Feenstra, R. C. and Levinsohn, J. A. (1995). Estimating markups and market conduct with multidimensional product attributes. *Review of Economic Studies*, 62(1):19 – 52.

Flach, L. and Irlacher, M. (2018). Product versus process: Innovation strategies of multiproduct firms. *American Economic Journal: Microeconomics*, 10(1):236–77.

Galton, F. (1907). Vox populi. *Nature*, 75:450–451.

Giraud-Héraud, E., Hammoudi, H., and Mokrane, M. (2003). Multiproduct firm behaviour in a differentiated market. *Canadian Journal of Economics*, 36(1):41–61.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

Hotelling, H. (1929). Stability in competition. *The Economic Journal*, 39(135):41 – 57.

Irmen, A. and Thisse, J.-F. (1998). Competition in multi-characteristics spaces: Hotelling was almost right. *Journal of Economic Theory*, 78(1):76 – 102.

Matsa, D. A. (2011). Competition and product quality in the supermarket industry. *Quarterly Journal of Economics*, 126(3):1539 – 1591.

Mazzeo, M. J. (2003). Competition and service quality in the U.S. airline industry. *Review of Industrial Organization*, 22(4):275 – 296.

Meng, Y., Brennan, A., Purshouse, R., Hill-McManus, D., Angus, C., Holmes, J., and Meier, P. S. (2014). Estimation of own and cross price elasticities of alcohol demand in the UK - a pseudo-panel approach using the living costs and food survey 2001 - 2009. *Journal of Health Economics*, 34:96 – 103.

Salop, S. C. (1979). Monopolistic competition with outside goods. *Bell Journal of Economics*, 10(1):141 – 156.

Selten, R. (1973). A simple model of imperfect competition, where 4 are few and 6 are many. *International Journal of Game Theory*, 2:141 – 201.

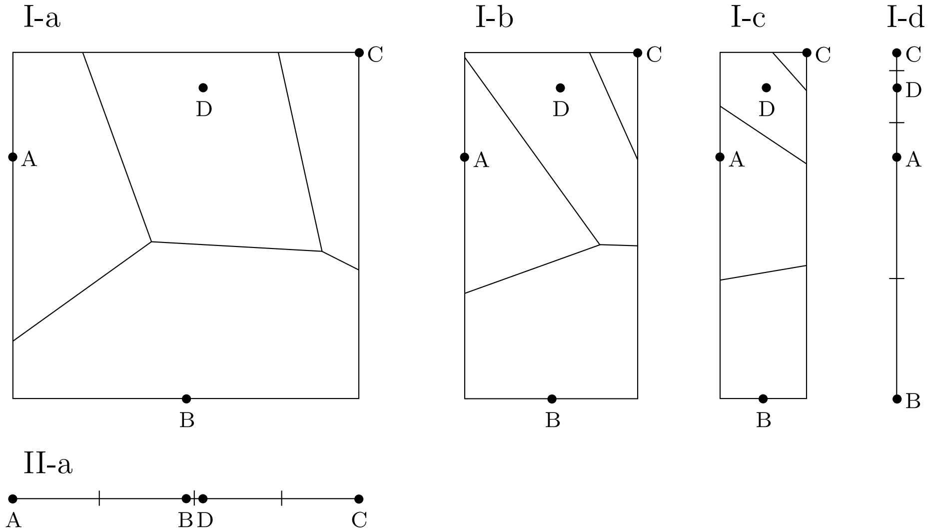
Shapiro, C. (1996). Mergers with differentiated products. *Antitrus*, Spring:23 – 30.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36.

Veendorp, E. C. H. and Majeed, A. (1995). Differentiation in a two-dimensional market. *Regional Science and Urban Economics*, 25(1):75 – 83.

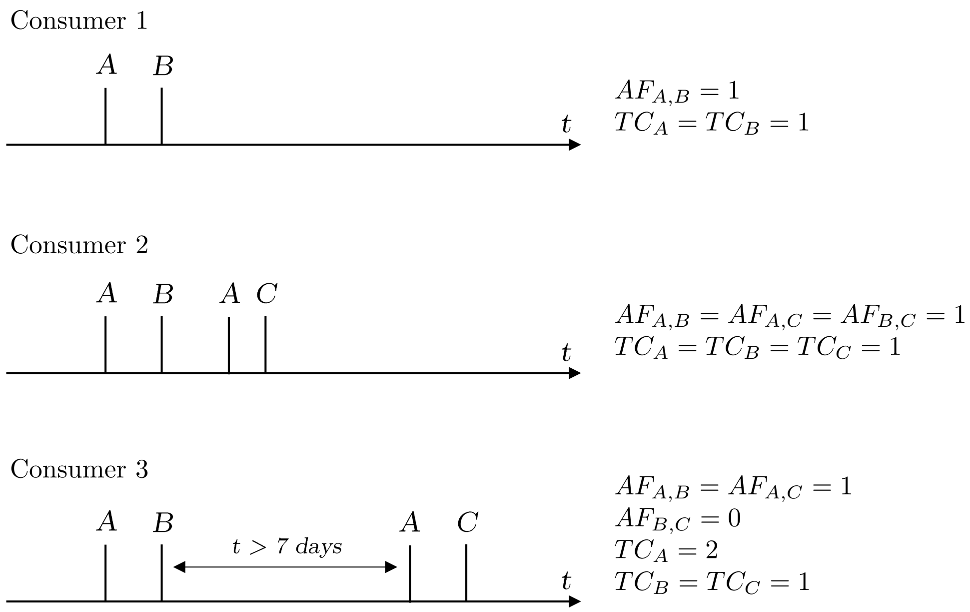
Figures and Tables

Figure 1: Illustration of neighborhood and distance in product space



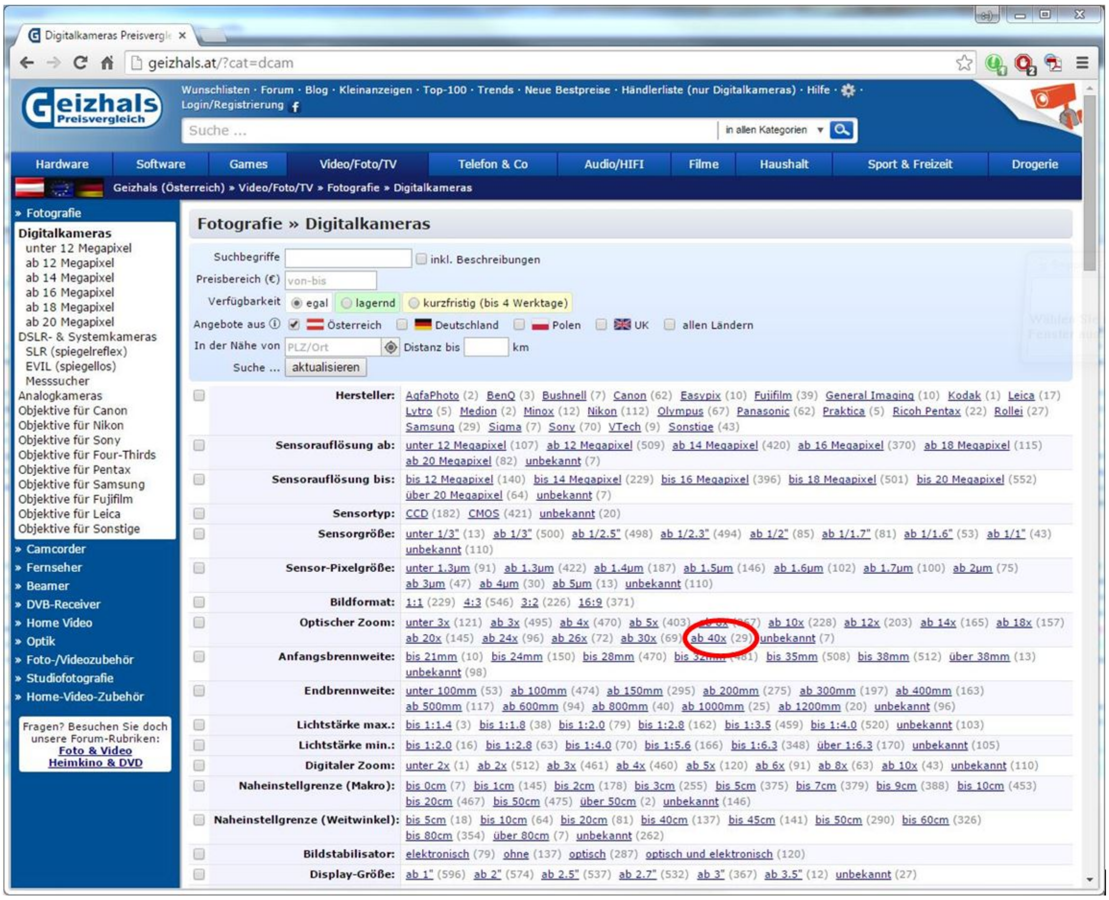
Notes: The figure illustrates markets with four differentiated products $\{A, B, C, D\}$ characterized by only two attributes x and y. The importance of one characteristic can be illustrated by altering the relative length of the dimension: Whereas panel I-a shows product characteristics of identical importance, in panel I-b (I-c) [I-d] the relative importance of characteristic x (on the horizontal axis), relative to attribute y on the vertical axis shrinks to 1/2 (to 1/4) [to 0]. In panel II-a only the attribute x is important.

Figure 2: Illustration of search episodes



Notes: The figure illustrates the clickstream of three consumers. The vertical lines refer to clicks on products at different points in time. The second consumer requests referrals to two online retailers for product A within one search spell, which counts as one click only. The clickstream of the third consumer is divided into two search spells, because the time period between surveying products B and A exceeds 7 days (conditional that this time span is and outlier based on the Grubbs, 1969, test).

Figure 3: Exemplary representation of product attributes on Geizhals.at



Notes: This is a screenshot of the Geizhals website showing the large number of different predefined product characteristics for digital cameras. The website also indicates the number of digital cameras sharing a particular attribute: The red circle, for example, implies that there are 29 digital cameras with at least 40 × optical zoom.

Table 1: Illustration of the measure for product differentiation

Product Pairs l, k		abs. Freq. $AF_{l,k}$	Distance (unweighted) $DIST_{l,k}$	Weight $\frac{TC_l+TC_k}{2}$	weighted abs. Freq. $\frac{AF_{l,k}}{(TC_l+TC_k)/2}$	Distance (weighted) $wDIST_{l,k}$
A	**B**	740	0.00	1,600.0	0.46	0.44
A	**C**	260	1.00	1,295.0	0.20	1.00
A	**D**	520	0.46	1,272.5	0.41	0.56
B	**C**	400	0.71	1,095.0	0.37	0.65
B	**D**	720	0.04	1,072.5	0.67	0.00
C	**D**	500	0.50	767.5	0.65	0.04
Total clicks			TC_A	TC_B	TC_C	TC_D
on product			1,800	1,400	790	745

Notes: Based on observed consumer click behavior, the unweighted and the weighted distances $DIST_{l,k}$ and $wDIST_{l,k}$ have been calculated according to equation (1) and (2), respectively.

Table 2: Top-listed poduct pairs

PANEL A: Top 10 clicked Product Pairs in ALL SUBSUBCATEGORIES

No of clicks	Product pairs	
4,246	Samsung SSD 830 128GB, SATA	Samsung SSD 830 256GB, SATA
4,099	Samsung Galaxy S3 i9300 16GB blue	Samsung Galaxy S3 i9300 16GB white
3,102	Samsung Galaxy S2 i9100 16GB black	Samsung Galaxy S3 i9300 16GB blue
2,909	Samsung Galaxy S3 i9300 16GB blue	Samsung Galaxy Note 2 N7100 16GB grey
2,628	Samsung Galaxy Note 2 N7100 16GB white	Samsung Galaxy Note 2 N7100 16GB grey
2,547	Google Nexus i9250 16GB silver	Samsung Galaxy Nexus i9250 16GB white
2,350	Samsung Galaxy S3 i9300 16GB blue	Apple iPhone 5 16GB black
2,128	Samsung SSD 830 256GB, SATA	Samsung SSD 840 PRO 256GB, SATA
2,115	Samsung Galaxy S3 i9300 16GB blue	Samsung Galaxy S3 i9300 16GB black
1,987	Samsung SSD 830 256GB, SATA	Samsung SSD 840 250GB, SATA

PANEL B: Top 10 clicked product pairs in the SUBSUBCATEGORY DIGITAL CAMERAS

No of clicks	Product pairs	
610	Sony Cyber-shot DSC-RX100 black	Panasonic Lumix DMC-LX7 black
350	Panasonic Lumix DMC-FZ150 black	Panasonic Lumix DMC-FZ200 black
247	Samsung EX2F black Panasonic	Panasonic Lumix DMC-LX7 black
242	Fujifilm FinePix X10 black	Sony Cyber-shot DSC-RX100 black
234	Fujifilm FinePix X10 black	Panasonic Lumix DMC-LX7 black
226	Panasonic Lumix DMC-FZ200 black	Panasonic Lumix DMC-LX7 black
217	Sony Cyber-shot DSC-HX20V black	Sony Cyber-shot DSC-RX100 black
216	Canon PowerShot S100 black	Sony Cyber-shot DSC-RX100 black
208	Sony Cyber-shot DSC-RX100 black	Nikon Coolpix P7700 black
205	Panasonic Lumix DMC-LX7 black	Nikon Coolpix P7700 black

Notes: The table shows the Top 10 product pairs which are most frequently clicked conjointly during a search episode by consumers. In Panel A product pairs in all subsubcategories are counted (not analyzed in this article). Panel B lists product pairs from subsubcategory "digital cameras".

Table 3: Available product characteristics of digital cameras

Variables	**Scale**	**Variables**	**Scale**
3D photo support[f]	dummy	Image stabilizer digital	dummy
3D video support[f]	dummy	Image stabilizer electronic	dummy
Aperture maximal[d]	ordinal	Image stabilizer optical	dummy
Aperture minimal[a,d]	ordinal	Instant camera[f]	dummy
Audioformat AAC supported[e]	dummy	Internal focus[f]	dummy
Audioformat MP3 supported[e]	dummy	Internal projector[f]	dummy
Audioformat PCM supported[e]	dummy	ISO max[d]	ISO value
Audioformat PMP supported[e]	dummy	ISO min[a,d]	ISO value
Audioformat WAV supported[e]	dummy	Optical zoom	x-fold
Audioformat WMA supported[b,e]	dummy	Picture aspect ratio 1:1[e]	dummy
Bridge to professional[f]	dummy	Picture aspect ratio 16:9[e]	dummy
Built-in memory[d]	megabyte	Picture aspect ratio 3:2[e]	dummy
Camera is waterproof[f]	dummy	Picture aspect ratio 4:3[e]	dummy
Camera has GPS[f]	dummy	Removable storage compactflash[b,e]	dummy
Closest focusing distance macro[d]	centimeter	Removable storage memorystick[e]	dummy
Closest focusing distance wide angle[c,d]	centimeter	Removable storage microdrive[b,e]	dummy
Connector for audio available[e]	dummy	Removable storage microSD[a,e]	dummy
Connector for bluetooth available[e]	dummy	Removable storage microSDHC[e]	dummy
Connector for charging available[e]	dummy	Removable storage miniSD[b,e]	dummy
Connector for components available[e]	dummy	Removable storage mircoSDXC[e]	dummy
Connector for docking station available[e]	dummy	Removable storage MMC[e]	dummy
Connector for HDMI available[e]	dummy	Removable storage SDXC[e]	dummy
Connector for headphones available[b,e]	dummy	Removable storage SD[e]	dummy
Connector for infrared available[e]	dummy	Removable storage SDHC[e]	dummy
Connector for loudspeaker available[a,e]	dummy	Remov. storage xD-pict.-card[e]	dummy
Connector for microphon available[e]	dummy	Sensor resolution	megapixel
Connector for USB available[e]	dummy	Sensor size[d]	ordinal
Connector for video available[e]	dummy	Sensor type[c]	ordinal
Connector for WLAN available[e]	dummy	Type of battery	dummy
Digital zoom	x-fold	Video format 3D[b,e]	dummy
Dimensions[a,d]	ccm	Video format AVC[b,e]	dummy
Display 3D technology	dummy	Video format AVCHD[c,e]	dummy
Display CSTN technology[c,h]	dummy	Video format AVI[c,e]	dummy
Display LCD technology[a,h]	dummy	Video format DIVX[b,e]	dummy
Display OLED technology[h]	dummy	Video format H264[c,e]	dummy
Display swiveling	dummy	Video format MJPEG[c,e]	dummy
Display tiltable	dummy	Video format MOV[c,e]	dummy
Display touchscreen	dummy	Video format MPEG[c,e]	dummy
Display at the front	dummy	Video format Quicktime[c,e]	dummy
Display diagonal	inches	Video format VGA[c,e]	dummy
Fineash connector[b,g]	dummy	Video format with sound[e]	dummy
Flash hot shoe[g]	dummy	Video: frames per second[d]	cardinal
Flash infrared[b,g]	dummy	Video: maximal pixel[d]	cardinal
Flash integrated[g]	dummy	Viewfinder available[a]	dummy
Flash video light[g]	dummy	Viewfinder optical	dummy
Focal length maximal[d]	ordinal	Viewfinder with LCD technology	dummy
Focal length minimal[a,d]	ordinal	Warranty	in years
Highspeed continuous shooting[f]	dummy	Weight[d]	in gram
Highspeed video[f]	dummy		

Notes: The table shows available product characteristics for digital cameras. Dummy refers to 1 if the attributes are available and 0 otherwise. [a)] Attribute was dropped as the $VIF > 10$. [b)] Attribute was dropped in some regressions as there was no variation over the included products. [c)] Attribute was not used as individual variable for hedonic pricing because values are observable only for fewer than 1,000 products. [d)] Some values of these attributes have been interpolated. [e)] Attributes have been aggregated to count variables in some regressions (e.g. number of connectors or number of video formats). [f)] Attributes are communicated at the website as special features. These attributes are used in the „Count of special features“ of Table 4 and „Identical special features“ of Table 7. [g)] Attributes are counted in „Flash features“ of Table 4. [h)] Attributes are used to construct the ordinal variable „Display type“ in Table 4 (OLED=2, LCD=1, CSTN=0).

Table 4: Determining the importance of product attributes

Dependent variable	Distance Measure $wDIST_{l,k}$		
	(1)	(2)	(3)
Most Important Product Attributes (Differences in . . .)			
Count of special features	0.0889***	0.0863***	0.0797***
	(0.00245)	(0.00245)	(0.00246)
Display at the front	0.0756***	0.0714***	0.0721***
	(0.00452)	(0.00451)	(0.00451)
Display 3D technology	0.0645***	0.0659***	0.0659***
	(0.0241)	(0.0240)	(0.0240)
Optical zoom	0.0493***	0.0554***	0.0463***
	(0.00294)	(0.00294)	(0.00294)
Display type[a)]	0.0436***	0.0423***	0.0403***
	(0.00485)	(0.00485)	(0.00484)
Dimensions	0.0187***	0.0161***	0.0332***
	(0.00454)	(0.00453)	(0.00454)
Sensor resolution	0.0290***	0.0303***	0.0324***
	(0.00156)	(0.00156)	(0.00155)
Digital zoom	0.0497***	0.0324***	0.0303***
	(0.00132)	(0.00135)	(0.00135)
Flash features[a)]	0.0370***	0.0384***	0.0265***
	(0.00683)	(0.00683)	(0.00683)
Display touchscreen	0.0259***	0.0262***	0.0258***
	(0.00156)	(0.00156)	(0.00156)
. . .	. . .	. . .	. . .
Statistics on Significance of Coefficients			
	pos/neg	pos/neg	pos/neg
Count of coefficients with $p < 0.01$ (***)	37/0	35/1	32/2
Count of coefficients with $p < 0.05$ (**)	0/0	0/0	1/0
Count of coefficients with $p < 0.1$ (*)	2/0	3/0	0/0
Not significant coefficients ($p >= 0.1$)	4	4	8
Total number of coefficients	43	43	43
Additional Controls			
Different brand		0.0583***	0.0591***
		(0.00104)	(0.00104)
Diff in price			0.0689***
			(0.00178)
Diff in # of offering firms			0.0840***
			(0.00163)
Controls for imputed values	X	X	X
Product fixed effects	X	X	X
Number of products	1,642	1,642	1,642
Observations	1,347,261	1,347,261	1,347,261
R^2	0.050	0.052	0.055

Notes: Estimation method: OLS. Observational unit: product pair (l, k). Variables are defined as differences in the product attributes and are standardized with mean zero and standard deviation of one. The table shows beta coefficients for the most important characteristics, i.e. the attributes with the larges parameter estimates (in absolute terms). Top 10 attributes are ordered in a descending manner based on the point estimates of specification (3). Constant suppressed for brevity. Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. [a)] For the definition of these variables see notes [g)] and [h)] in Table 3.

Table 5: Robustness: Determining the importance of product attributes

Dependent variable	Distance Measure $wDIST_{l,k}$		
	(1)	(2)	(3)
Using All Attributes			
	pos/neg	pos/neg	pos/neg
Count of coefficients with $p < 0.01$ (***)	69/2	66/3	62/3
Count of coefficients with $p < 0.05$ (**)	3/1	1/1	2/3
Count of coefficients with $p < 0.1$ (*)	1/1	1/1	1/0
Not significant coefficients ($p >= 0.1$)	8	12	14
Total number of coefficients	85	85	85
Number of Products	1,642	1,642	1,642
Observations	1,347,261	1,347,261	1,347,261
R^2	0.053	0.055	0.058
Using Reduced Sample with $wDIST_{l,k} < 1$			
	pos/neg	pos/neg	pos/neg
Count of coefficients with $p < 0.01$ (***)	27/2	25/2	24/4
Count of coefficients with $p < 0.05$ (**)	0/0	1/1	0/0
Count of coefficients with $p < 0.1$ (*)	1/1	0/2	1/0
Not significant coefficients ($p >= 0.1$)	12	12	14
Total number of coefficients	43	43	43
Number of products	1,637	1,637	1,637
Observations	117,851	117,851	117,851
R^2	0.578	0.578	0.594
Using a Tobit Model			
	pos/neg	pos/neg	pos/neg
Count of coefficients with $p < 0.01$ (***)	33/3	30/3	29/5
Count of coefficients with $p < 0.05$ (**)	0/1	3/1	3/0
Count of coefficients with $p < 0.1$ (*)	1/1	0/1	1/0
Not significant coefficients ($p >= 0.1$)	4	5	5
Total number of coefficients	43	43	43
Number of products	1,642	1,642	1,642
Observations	1,347,261	1,347,261	1,347,261
Controls			
Different brand		X	X
Diff in price			X
Diff in # of offering firms			X
Controls for imputed values	X	X	X
Product fixed effects	X	X	X

Notes: Observational unit: product pair (l, k). The first panel includes all attributes individually rather than in aggregated form. In the second panel all product pairs with zero clicks are eliminated. Whereas the first and second panel use OLS, the third panel uses Tobit regressions with an upper limit at one. Variables are defined as differences in the product attributes and are standardized with mean zero and standard deviation of one.

Table 6: Effect of number of close substitutes on product price

Dependent variable	Price of Product i						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Products in "Radius 1"	1.003*						
	(0.534)						
Products in "Radius 5"		−0.349**					
		(0.158)					
Products in "Radius 10"			−0.389***				
			(0.0880)				
Products in "Radius 30"				−0.224***			
				(0.0335)			
Products in "Radius 50"					−0.163***		
					(0.0215)		
Products in "Radius 100"						−0.0854***	
						(0.0129)	
Products in "Radius 10"							−0.253***
							(0.0973)
Products between "Radius 10 and 30"							−0.121**
							(0.0586)
Products between "Radius 30 and 50"							−0.168***
							(0.0612)
Products between "Radius 50 and 100"							−0.0138
							(0.0263)
No of offering firms	−0.151***	−0.159***	−0.157***	−0.133***	−0.0976***	−0.0565*	−0.0899**
	(0.0312)	(0.0312)	(0.0310)	(0.0309)	(0.0315)	(0.0342)	(0.0352)
Constant	202.9***	198.9***	197.9***	199.0***	199.9***	201.2***	199.5***
	(63.30)	(63.27)	(62.96)	(62.45)	(62.20)	(62.47)	(62.24)
Product attributes	X	X	X	X	X	X	X
Brand dummies	X	X	X	X	X	X	X
Observations	1,642	1,642	1,642	1,642	1,642	1,642	1,642
R^2	0.806	0.806	0.808	0.811	0.813	0.811	0.813

Notes: Observational unit: products i. Estimation method: OLS. The number in "Radius number" refers to the average number of competing products within the given radius. Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

Table 7: Price effects of same brand and foreign brand products

Dependent variable	Price of Product i			
	"Radius 30"		**"Radius 50"**	
	non-stand. coefficients	beta coefficients	non-stand. coefficients	beta coefficients
	(1)	(2)	(3)	(4)
Own brand products	-1.077^{***}	-8.744^{***}	-0.806^{***}	-9.188^{***}
	(0.201)	(1.628)	(0.158)	(1.805)
Foreign brand products	-0.0667	-1.905	-0.0606^{*}	-2.906^{*}
	(0.0494)	(1.412)	(0.0329)	(1.576)
Identical special features$^{a)}$	-0.247^{***}	-5.415^{***}	-0.191^{***}	-5.835^{***}
	(0.0582)	(1.275)	(0.0429)	(1.313)
Different special features$^{a)}$	-0.202^{***}	-4.620^{***}	-0.143^{***}	-6.013^{***}
	(0.0568)	(1.297)	(0.0345)	(1.452)
Identical Display at the front	-0.230^{***}	-7.421^{***}	-0.164^{***}	-8.942^{***}
	(0.0355)	(1.147)	(0.0224)	(1.223)
Different Display at the front	-0.142	-1.230	-0.141	-1.726
	(0.180)	(1.563)	(0.122)	(1.493)
Identical Display 3D technology	-0.221^{***}	-7.448^{***}	-0.161^{***}	-9.009^{***}
	(0.0338)	(1.136)	(0.0217)	(1.215)
Different Display 3D technology	-1.848	-2.167	-1.533	-2.420
	(2.345)	(2.751)	(2.304)	(3.639)
Identical Optical zoom	-0.369^{***}	-4.646^{***}	-0.252^{***}	-4.395^{***}
	(0.112)	(1.413)	(0.0795)	(1.387)
Different Optical zoom	-0.178^{***}	-4.628^{***}	-0.141^{***}	-6.595^{***}
	(0.0491)	(1.276)	(0.0289)	(1.351)
Identical display type$^{a)}$	-0.228^{***}	-7.661^{***}	-0.169^{***}	-9.347^{***}
	(0.0340)	(1.141)	(0.0220)	(1.216)
Different display type$^{a)}$	0.0800	0.296	0.131	0.946
	(0.431)	(1.591)	(0.221)	(1.589)
Constant	*varies with regression*		*varies with regression*	
No of offering firms	X	X	X	X
All attributes	X	X	X	X
Brand dummies	X	X	X	X
Observations	1,642	1,642	1,642	1,642
R^2	0.811-0.813		0.811-0.815	

Notes: Observational unit: products i. Estimation method: OLS. The variables refer to the number of products with the specified properties within the respective radius indicated in the columns. The blocks show the results for different estimations. Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. $^{a)}$Two cameras are identical only if all special features and the display technology match perfectly (see notes $^{f)}$ and $^{h)}$ in Table 3).

Table 8: Influence of predecessors vs. successors and available vs. unavailable products

Dependent variable	Price of Product i			
	"Radius 30"		**"Radius 50"**	
	non-stand. coefficients	beta coefficients	non-stand. coefficients	beta coefficients
	(1)	(2)	(3)	(4)
No of predecessors	0.00506	0.105	0.0134	0.467
	(0.0623)	(1.295)	(0.0388)	(1.349)
No of successors	-0.415^{***}	-8.786^{***}	-0.302^{***}	-10.49^{***}
	(0.0550)	(1.165)	(0.0333)	(1.157)
No of available products	-0.225^{***}	-7.465^{***}	-0.162^{***}	-9.040^{***}
	(0.0342)	(1.137)	(0.0218)	(1.213)
No of unavailable products	-0.205	-0.852	-0.199	-1.035
	(0.269)	(1.120)	(0.212)	(1.105)
Constant	*varies with regression*		*varies with regression*	
No of offering firms	X	X	X	X
All attributes	X	X	X	X
Brand dummies	X	X	X	X
Observations	1,642	1,642	1,642	1,642
R^2	0.811-0.813		0.813-0.816	

Notes: Observational unit: products i. Estimation method: OLS. The variables refer to the number of products with the specified properties within the respective radius indicated in the columns. The blocks show the results for different estimations. Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

Table 9: Effect of distance to close substitutes on product price

Dependent variable	Price of Product i						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Distance to 5^{th} next competitor	2.941**						
	(1.178)						
Distance to 10^{th} next competitor		5.003***					
		(1.141)					
Distance to 20^{th} next competitor			6.516***				
			(1.124)				
Distance to 30^{th} next competitor				7.127***			
				(1.121)			
Distance to 50^{th} next competitor					6.846***		
					(1.140)		
Distance to 100^{th} next competitor						5.718***	
						(1.207)	
Distance to 5^{th} competitor							34.76***
							(9.833)
Distance from 5^{th} to 10^{th} competitor							19.73***
							(5.568)
Distance from 10^{th} to 20^{th} competitor							12.12***
							(3.913)
Distance from 20^{th} to 30^{th} competitor							6.542**
							(2.590)
Distance from 30^{th} to 50^{th} competitor							2.216
							(2.335)
Distance from 50^{th} to 100^{th} competitor							3.123
							(2.498)
No of offering firms	−0.158***	−0.155***	−0.145***	−0.136***	−0.125***	−0.116***	−0.124***
	(0.0312)	(0.0310)	(0.0309)	(0.0309)	(0.0312)	(0.0320)	(0.0320)
Constant	196.3***	193.9***	189.6***	187.4***	188.3***	196.2***	188.7***
	(63.26)	(62.99)	(62.70)	(62.57)	(62.66)	(62.91)	(62.66)
Product attributes	X	X	X	X	X	X	X
Brand dummies	X	X	X	X	X	X	X
Observations	1,642	1,642	1,642	1,642	1,642	1,642	1,642
R^2	0.806	0.808	0.810	0.811	0.810	0.808	0.811

Notes: Observational unit: products i. Estimation method: OLS. The variables refer to the distance of good i to the h^{th} next competitor measured by the distance measure $wDIST_{l,k}$. Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.