Bank of Japan Working Paper Series

Compilation of Experimental Price Indices Using Big Data and Machine Learning: A Comparative Analysis and Validity Verification of Quality Adjustments

Nobuhiro Abe*
nobuhiro.abe@boj.or.jp

Kimiaki Shinozaki*
kimiaki.shinozaki@boj.or.jp

No.18-E-13
August 2018

Bank of Japan
2-1-1 Nihonbashi-Hongokucho, Chuo-ku, Tokyo 103-0021, Japan

* Research and Statistics Department

Compilation of Experimental Price Indices Using Big Data and Machine Learning: A Comparative Analysis and Validity Verification of Quality Adjustments[*]

Nobuhiro Abe[†] and *Kimiaki Shinozaki*[‡]

August 2018

Abstract

This paper compiles experimental price indices for 20 home electrical appliances and digital consumer electronic products using big data obtained from *Kakaku.com*, the largest price comparison website in Japan, and a machine-learning algorithm which pairs legacy and successor products with high precision. In so doing, authors examine the validity of quality adjustment methods by performing comparative analyses on the difference these methods have on price indices. Findings from the analyses are as follows: Indices applied with the *Webscraped Prices Comparison Method*—the quality adjustment method newly developed and introduced by the Bank of Japan—are more cost-effective than those applied with the *Hedonic Regression Method* which is known to possess high accuracy in index creation. Indices applied with the *Matched-Model Method*, which is frequently applied to price indices using big data is unable to precisely reflect price increases intended to ensure the profitability often seen in home electronics at time of product turnover. This indicates the significant downward bias in price indices. These findings once again highlight the importance of selecting the appropriate quality adjustment method when compiling price indices.

Keywords: price index, quality adjustment method, hedonic approach, support vector machine

JEL Classification: C43, C45, E31

[*] This paper was presented at the Meeting of the Group of Experts on Consumer Price Indices held in Geneva, Switzerland on 7-9 May 2018. The authors would like to thank staff members of the Bank of Japan for their useful comments; however, the opinions expressed here, as well as any remaining errors, are those of the authors and should not be ascribed to the Bank.

[†] Research and Statistics Department, Bank of Japan (Email: nobuhiro.abe@boj.or.jp)

[‡] Research and Statistics Department, Bank of Japan (Email: kimiaki.shinozaki@boj.or.jp)

I. Introduction

A price index is constructed with the primary aim of understanding the fluctuations in general price levels by indexing the constant-quality price of goods and services by setting the price at the base point in time as 100. The index is created by selecting representative products in the market and by continuously surveying their transaction prices during each period.

However, there are several cases where products lose their representativeness. When products become discontinued as a result of technological innovations or products are no longer *strong-sellers* due to a decrease in transaction volumes resulting from the appearance of successor products. In order to maintain accuracy of price indices, it is necessary to ensure that the surveyed products are representative in the market. Representativeness is secured by performing a *change of sample prices* at an appropriate frequency, and adopting strong-seller goods to be surveyed. When linking old and new products that do not have identical quality specifications, the way of evaluating the differences between them needs to be addressed.

Price statisticians have traditionally adopted a method of processing quality differences between old and new products by splitting the difference in prices at the same point in time into "price change resulting from quality changes" and "pure price change". By eliminating the former, the price index only reflects the latter. Such methods are called *quality adjustments* (Chart 1). Price statisticians use various quality adjustment methods at the time of product turnover, regarding the specification of the products or the feasibility of conducting price surveys. Referring to best practices stated in price index manuals established by international organizations such as the IMF or OECD, statisticians strive to compile consistent and highly-precise price indices[1].

In recent years, owing to the advance of big data analysis, price statisticians and economists

---

[1] Consumer Price Index (CPI) published by the Ministry of Internal Affairs and Communications, measuring price changes of goods and services purchased by consumers (household) nationwide, Corporate Goods Price Index (CGPI) and Services Producer Price Index (SPPI) published by the Bank of Japan, measuring price changes of goods and services traded between firms, are representative price indices in Japan. CGPI consists of Producer Price Index (PPI), Export Price Index (EPI), and Import Price Index (IPI), and reference indices.

have become capable of creating price indices based on scanner data (POS data), or webscraped data posted on online stores' websites[2]. These methods, however, are still in the experimental stage. Meanwhile, the traditional methods are still regarded as the standard compilation methods for price statistics. However, with the use of the vast amounts of data, which were previously considered ineligible for use in the traditional methods, some improvements have been made in terms of frequency of release of the index, easing the burden on both price statisticians and reporting firms.

In this paper, authors point out the problems inherent in the *traditional* approach, in which a price index is created by carrying out changes of sample prices reflecting the life-cycles of products and quality adjustments between old and new products; and in the *non-traditional* approach, in which a price index is compiled by making use of the vast amounts of data and by improving computing capabilities via advanced knowledge in the data sciences. Then we compile experimental price indices using big data obtained from Japan's leading price comparison website *Kakaku.com* and the machine-learning method to imitate the expertise of price statisticians. We also compare the difference of the selection of quality adjustment methods on price indices and verify the validity of those methods. Finally, based on the results of the comparative analyses, we emphasize the importance of conducting appropriate quality adjustments when compiling price indices.

II. Comparison of Approaches for Compiling Price Indices

(1) *Traditional* Approach of Price Statistics Agencies

In order to create consistent price indices, the representativeness of the products needs to be maintained by punctually performing a change of sample prices and choosing new products to be surveyed when the currently surveyed products approaches the end of their product life-cycles.

Price statisticians well-versed in statistical practices and industry customs select representative products to be surveyed, keeping in mind changes in product specifications

[2] Examples of efforts by price statisticians overseas: the U.K. (Office for National Statistics (2017)), the Netherlands (Chessa, Verburg, and Willenborg (2017)), New Zealand (Bentley and Krsinich (2017)), etc. Examples of efforts by economists: Cavallo and Rigobon (2016), Ueda, Watanabe, and Watanabe (2016), Abe, Enda, Inakura, and Tonogi (2015), etc.

and data availability. They then choose an optimal quality adjustment method to remove price change arising from changes in quality[3,4]. In this paper, we call this a *traditional* approach of price index compilation. Employing this approach, qualitative changes are eliminated from nominal price changes in order to facilitate price comparisons across the actual life-cycles of the products. Furthermore, due to resource constraints at price statistics agencies and reporting firms, the number of sample prices that can be examined is unavoidably limited (Chart 2(1)).

(2) *Non-Traditional* Approach Using Big Data

Methods that compile price indices using big data such as scanner data or webscraping data, which we call the *non-traditional* approach enables to enhance the efficient compilation of price indices, without relying on the knowledge or expertise of statisticians.

The *Matched-Model Method* (hereinafter, MMM) calculates and reflects the percentage of price change to the index for products existing in the market in both survey period t and t+1. This enables the continuous survey of products with constant quality (Chart 2(2)). This method, which mitigates the burden on both price statistics agencies and reporting firms by making use of big data, is expected to improve statistical practices.

However, in cases where price increases (price pushbacks) are common practice when launching new products, the index cannot properly reflect the impact of such price pushbacks, and may possibly show a downward bias. If the *non-traditional* approach is adopted for consumer durables such as home electronic products, there is a possibility of non-negligible downward bias[5]. In fact, prices of home electronics are tend to take the

---

[3] For details of quality adjustment methods in addition to section IV (1) of this paper, refer to ILO et al. (2004a), ILO et al. (2004b), Triplett (2006), etc.

[4] For example, Price Statistics Division of the Bank of Japan attentively checks the appropriateness of sample prices every month. Checking focuses especially on whether questionnaires are returned properly from reporting firms and price data are collected; and whether products to be surveyed continues to have large transaction volumes (whether those products are strong-sellers). If thorough examination is necessary, expert meetings would be held monthly in order to promptly decide the implementation needed to ensure accuracy of price indices.

[5] For example, Gowrisankaran and Rysman (2012) points out for camcorders, and Melser and Syed (2014) points out for non-durable consumer goods sold in supermarkets, the possibility of quality improvement being assessed excessively if the impact of price pushback is disregarded.

highest price immediately after the release, and steadily decline thereafter. As results, firms implement price pushbacks at the time of model upgrades to maximize their profit[6].

(3) The Approach We Take in This Paper

Both of these two existing approaches have their own challenges. Issues related to the approach using traditional data are as follows. First, the quality difference reflected in price tends to be an all-or-nothing, i.e. 0% or 100%, whereas in fact values anywhere within the range could be included. Second, the number of sample prices is insufficient due to the limited number of representative products. Third, the selection of successor products and quality adjustment methods are subjective. On the other hand, the approach using non-traditional data also has its issues. For example, if MMM is applied, price pushbacks cannot be adequately reflected for, thus causing a downward bias to the index.

In this paper, we have combined the *traditional* and *non-traditional* approaches to create an index which properly reflects the impact of price pushbacks. Each time old products are replaced by new products, the old are paired with the new, using big data. Once the products are successfully paired, the most appropriate quality adjustment method can be carried out.

We developed a supervised machine-learning algorithm, which pairs old and new products with high precision. Applying this algorithm to *Kakaku.com*'s big data on the prices and specifications of durable products enables the incorporation of a vast amount of data, otherwise neglected under the *traditional* approach. With the enhanced use of machine, we aim to simultaneously improve both accuracy and efficiency by replacing expertise of price statisticians (Chart 2(3)).

Verification of the Webscraped Prices Comparison Method

The previous study, Abe, Ito, Munakata, Ohyama, and Shinozaki (2016), shows the quality improvement ratios (ratio of price difference arising from quality differences of old and new products), measured immediately after the release of new products, of 0.5 to 0.6 for home

[6] For price setting actions by means of reducing product content of food or beverage at product turnover while maintaining price (real price increase), Abe et al. (2015) has managed to reflect quality (content) changes in price index to some extent. Loon and Roels (2018) has advocated a method called *non-matched model approach* in order to eliminate a downward bias inherent in MMM.

electrical appliances, and of 0.6 to 0.7 for digital consumer electronic products. Based on the results of Abe et al. (2016), the Bank introduced the *Webscraped Prices Comparison Method* (hereinafter, WSM) as one of its quality adjustment methods. Under this method, 50% of the retail price difference between old and new products sold at online stores is regarded as quality growth. The Bank has adopted WSM at the time of rebasing of the Corporate Goods Price Index (CGPI), carried out in February 2017. WSM is used only for home electronic products where frequent model upgrades occur (Bank of Japan (2017))[7].

"Price change due to quality changes account for approximately half of the price difference between old and new products", was the conclusion obtained from cross-sectional analysis in the previous study. However, resources available to conduct a time-series analysis in order to verify the accuracy of the indices created using WSM were insufficient and the analysis has been tabled for a future time. In this paper we attempt to verify the appropriateness of WSM by comparing the trends of indices created using the highly accurate *Hedonic Regression Method* (hereinafter, HRM).

Understanding of Price Pushback Effect at Product Turnover

By observing trends of indices created by combining features of the *traditional* and *non-traditional* approaches, we can quantitatively analyze the impact of price pushbacks.

In cases of food products or daily necessities, where the pace of obsolescence is relatively slow, there is little incentive for manufacturers to bring out new products to push back the price. For these products, sample prices change less frequently and the impact of price pushback on indices is negligible. Therefore, no matter what method is adopted, the effect on the indices is relatively small[8]. On the other hand, for the home electronic products, which are targeted for analyses in this paper, it is necessary to examine whether the *non-traditional* approach is suitable as it cannot take price pushbacks into consideration.

---

[7] Price statistic agencies overseas have applied quality adjustment method, as a kind of experts judgment, to regard 50% of the price difference between old and new product as contribution from quality improvement. See Dalen and Tarassiouk (2013), Hoven (1999), Hoffmann (1999). In Japan, Ohta (1977) has proposed to use the 50% rule for quality adjustment based on the principle of risk minimization under uncertainty where understanding of product quality is insufficient.

[8] Office for National Statistics (2017) is working on the Grocery Prices Scraping Project, which compiles indices using online store price of three supermarkets. If there is no quality changes, indices applied with HRM and MMM are to coincide (Aizcorbe, Corrado, and Doms (2003)).

III. Making Old and New Product Pairs Using Machine Learning Method

In this section, after outlining the dataset obtained from *Kakaku.com*, we explain the machine-learning method developed to effectively pair legacy products that have come to the end of its life-cycle with successor products at the beginning of its life-cycle. We then compile experimental indices by applying five different quality adjustment methods, *Direct Comparison Method* (DCM), *Overlap Method* (OLM), HRM, WSM and MMM at the time of product turnover, and conduct comparative analyses of those indices[9]. Finally, after organizing the facts obtained from the analyses, we refer to our aspiration for the future research.

(1) Outline of Dataset

The dataset used in this paper needs to include both frequently revised price information and a wide variety of specification information for the purpose of implementing proper quality adjustments for each product.

In order to satisfy these requirements, we used the same dataset from the previous study of Abe et al. (2016). The dataset contains the following information: One is specification data for 20 major home electrical appliances and digital consumer electronic products registered in *Kakaku.com* during the three years from December 2012 to December 2015[10]. The other is weekly average price (tax exclusive) data of individual products in the two years from December 2013 to December 2015, provided by paid marketing service *Kakaku.com Trend Search Enterprise version*, offered by *Kakaku.com, Inc.*, the operating company of the website[11].

---

[9] Aizcorbe and Pho (2005) performed a comparative analysis of indices applied with HRM and MMM for home electrical appliances and digital consumer electronic products. However, due to dataset constraints, the analysis of impact of the product life-cycle on the index is insufficiently.

[10] The dataset consists of eight home electrical appliances (air conditioners, refrigerators and freezers, washers and dryers, rice cookers, vacuum cleaners, microwaves, hair dryers and curling irons, air purifiers) and twelve digital consumer electronic products (GPS navigations, external hard drives, LCD TVs, LCD monitors, printers, Blu-ray and DVD recorders, headphones, camcorders, laptops, desktops, point-and-shoot cameras, DSLR and mirrorless cameras).

[11] In order to eliminate direct impacts of increases in consumption tax rate in April 2014 from the analyses results, tax-exclusive dataset was prepared in this paper.

The number of products included in the dataset is approximately 4,500; and the number of samples after multiplying the number of products with the corresponding weekly price data is approximately 150,000. Moreover, the total amount of data, obtained by multiplying the number of samples with the corresponding specification series is approximately 5.6 million[12].

(2) Creation of Product Pairs

When changing products to be surveyed due to decline in representativeness of the product, it is common to choose products with similar specifications which determine product value. For example, when representativeness of the old product is lost at the time of product turnover, it is considered more desirable to select the successor product from the same manufacturer and the same lineup, rather than selecting from other manufacturers or lineups in order to ensure continuity and stability of the index.

However, in the data from *Kakaku.com*, explicit information to specify which products are the legacy or successor products of individual products is not always available. In the *traditional* approach, price statisticians need to manually match the legacy and successor products. This is not a feasible measure applied to such a huge dataset used in this paper[13].

In order to overcome the difficulties, we attempt to choose successor products efficiently using the method of *supervised machine learning* algorithm in this paper. First, we make a lot of paired products by generating combinations of two products from the whole dataset exhaustively. Next, for the entire dataset, we narrow down the product pairs by imposing the following three necessary conditions and obtain approximately 92,000 pairs of products.

---

[12] If average price data of continuously sold product is temporarily missing, in principle the price is imputed by the price reported immediately before the missing occurred.

[13] For your reference, looking at the Bank's CGPI, in year 2015 (the 2010 base), there were over 1,800 cases of sample price changes. The frequency of change is equivalent to 0.21 times per year for every sample price. On groups basis, for material related groups such as "Petroleum & coal products", "Iron & steel", and "Agriculture, forestry & fishery products" the frequency was less than 0.1 times per year. On the contrary, for machinery related groups where product life-cycle is relatively short and model upgrade occurs frequently, such as "Information & communications equipment", "Transportation equipment", "Electrical machinery & equipment" and "Business oriented machinery", the frequency of change tends to be high at around 0.3-0.7 times per year.

Necessary conditions to compose old and new product pairs	
Condition 1	The release date (registration date) of the new product is later than that of the old product.
Condition 2	The old and new products are made by the same manufacturer.
Condition 3	Release date of the new products is prior to or within in 1 week of the end of sales date of the old product (sales interval between products is not so long).

In order to create supervised data used in machine-learning, we randomly selected 512 product pairs for each individual item, and categorized all the extracted data one-by-one into "old and new product pairs that seem to belong to the same manufacturer and same lineup" and "pairs that cannot be regarded as old and new product pairs", utilizing detailed information described in manufacturers' catalogs, images of product appearance, etc.[14] However, for the four items (air purifiers, Blu-ray and DVD recorders, camcorders, DSLR and mirrorless cameras) for which the total number of product pairs is relatively small, we extract old and new product pairs from all product pairs manually without applying machine-learning methods. As a result, we identified 551 pairs as old and new product pairs, out of 8,192 randomly selected pairs which were created as supervised data (Chart 3(1)).

(3) Outline of Characteristics and Classifiers

When distinguishing old and new product pairs using machine-learning methods, it is necessary to specify labels which serve as indicators, so-called *characteristics* in the field of machine-learning. In this paper, we carried out verification for a large volume of product pair included in the dataset, and we extracted the following three labels as characteristics, while taking a trade-off between usefulness and computational burden into account.

Characteristics used for detecting old and new product pairs		
Characteristic 1	Jaro-Winkler distance of Product Names	Whether the names (i.e. product codes) of the paired products are relatively similar.
Characteristic 2	Zone of product price	Whether it is possible to say that the paired products belong to nearly the same price zone.
Characteristic 3	Product launch interval	Whether there is a reasonable interval between the release dates of the paired products

[14] We confirmed that even if we increased the amount of supervised data to 1,024, the improvement of classification performance is limited. Thus considering cost-effectiveness, we set the amount of product pair data used as supervised data to 512 (Chart 3(2)).

The *Jaro-Winkler distance* is an indicator to quantitatively evaluate similarity levels of two letter strings. It integrates the number of common letters from the first four letters of both strings (Winkler (1990)). The pairing accuracy was higher for the *Jaro-Winkler distance*, compared to the *Levenshtein distance* which was used in Abe et al. (2016)[15]. Thus we decided to use the *Jaro-Winkler distance* as one of the characteristics, as well as zone of product price and product launch interval.

There are numerous methods to solve binary classification problems relying on machine-learning. In this paper, we adopted the *Support Vector Machine* (SVM) as the classifier which has a balance between robustness against noise included in supervised data and calculation speed, as pointed out by Wu et al. (2008)[16]. SVM is an algorithm for obtaining the classification boundary as separating hyperplanes in the characteristic space, specifying the data which is the closest to identification boundary in supervised data and maximizing the Euclidean distance between the sample and identification boundary. (For further details, see the mathematical appendix at the end of this paper.)

Suppose $\boldsymbol{w}$ is coefficient vector of hyperplanes and $\boldsymbol{x}_i$ is characteristic vector, with in mind that $1/\|\boldsymbol{w}\|$ shows the margin between the closest data and identification boundary, it is formulated as the following minimization problem with inequality constraints:

$$\min L(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{s.t.} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 \tag{1}$$

By solving the dual problem of the equation (1), the optimal separating hyperplane is expressed as the following:

$$y^* = \operatorname{sign}\left(\sum_{i \in S} \lambda_i^* y_i \boldsymbol{x}_i^T \boldsymbol{x} + b^*\right) \tag{2}$$

---

[15] The *Levenshtein distance* is also known as minimum edit distance which quantifies the extent of variation between two strings by measuring the minimum number of editing operations (insertion, deletion, substitution) needed to transform one string into the other.

[16] In general, compared to other machine-learning methods, SVM has superior performance in classification. Classification performance of *random forest* declines when explanatory variables are relatively small whereas SVM is capable of maintaining a certain standard. This paper tackled binary classification issue using not only SVM but also *decision tree* and *random forest*. When evaluating each method based on the F-measure, which is an indicator to represent the classification performance, the best result was obtained by SVM.

Here λ_i represent Lagrangean multipliers, $y_i \in \{+1, -1\}$ imply class labels, and $\mathrm{sign}(u)$ is a sign function that takes 1 when $u > 0$ and -1 otherwise.

In reality, it is extremely rare case that all the samples are linearly separable while it may be natural that the separating hyperplane indicates non-linearity in the characteristic space. Therefore, in this paper, a *soft margin non-linear SVM* with the *kernel trick* is used as classifier. Linear separation is conducted after mapping the characteristic space to higher dimension space while relaxing the constraints allowing some extent of identification errors. Inverse mapping is again conducted to bring back to the original space (Chart 4(1)). In conducting *kernel trick*, we adopted the general-purpose RBF (Radial Basis Function) kernel when calculating inner product of characteristic vector.

At that time, the optimal separating hyperplane is obtained by using the kernel function $k(\cdot)$ as follows:

$$y^* = \mathrm{sign}\left(\sum_{i \in S} \lambda_i^* y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b^*\right) \quad \text{where} \quad k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(\frac{-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2}{2\sigma^2}\right)$$

Upon implementation of the algorithm, we used *Python*, which has strengths in big data analysis and scientific technology calculations.

(4) Creation of Classifiers

In order to improve the classification performance of the non-linear SVM using RBF kernel, it is important to properly configure the extent to which the complexity of the data boundary surface will be reflected in classifiers (set by kernel parameters σ), and the extent to which faulty identification is allowed (set by penalty parameters C). With classifiers excessively fitting the given supervised data, there is a possibility of harming classification performance for unknown data. Such phenomenon is called *overfitting*. In order to improve the accuracy of pairing old products with new one, it is necessary to restrain this overfitting.

In this paper, for the 16 items applying machine-learning method, hyper parameters (σ, C) are computed to maximize the *F-measure* which represents performance of classifiers. In order to compute the F-measure, we use the *10-fold cross-validation* and the *grid search method* targeting the lattice of $\log(\sigma)$ and $\log(C)$ with 0.50 and 0.25 increments (Hsu, Chang and Lin (2016), Powers (2011))[17].

The F-measure is defined as a harmonic mean of *precision ratios* (the ratio those judged to be "true" by classifier and is actually "true") and *recall ratios* (the ratio of those actually "true" and is classified as "true") (Chart 4(2)). Using the hyper parameters (σ, C) which maximize the F-measure (Chart 5 and 6), we obtain separating hyperplanes on the characteristic space, and used it as a classifier (Chart 7).

IV. Compilation of Quality-Adjusted Experimental Price Indices

In this section, we compile experimental price indices for 20 home electrical appliances and digital consumer electronic products using old and new product pairs prepared in the previous section, and conduct a comparative analyses on the impact different quality adjustment methods have on indices. For each old and new product pair, we implement both sample price changes and quality adjustments under the interpretation that it is the time when old products come to the end of the life-cycle and switches over to new products, and create uninterrupted price indices. The analysis is conducted paying attention to following two points: First, comparing how price indices change depending on the applied quality adjustment method, and second, observing the features of the indices applied with WSM newly introduced by the Bank.

(1) Outline of Major Quality Adjustment Methods

Here, we compare indices compiled using five quality adjustment methods; DCM, OLM, HRM, WSM, and MMM, for which changes of sample prices and quality adjustments are unnecessary.

As previously stated, MMM is a method which calculates the price changes for products existing in the market at both survey period and the following period in percentage, linking the ratio to the index. Therefore, neither changes of sample prices nor quality adjustments

[17] *K-fold cross-validation* is a method to evaluate classification performance, by dividing subject data into K units, and using 1 unit as test data and the remaining $K-1$ units as supervised data, repeat learning and verification while changing test data for a total of K times. When determining K, trade-off of bias which impacts generalization performance (the difference between the average value estimated by model and true average value) and variance (randomness originating from differences in supervised data) needs to be taken into consideration. $K = 10$ is commonly used.

are necessary to compile index based on MMM. For that reason, we use individual product data taken from *Kakaku.com* directly rather than old and new product pairs made in the previous section. Conversely, since DCM, OLM, HRM and WSM are all quality adjustment methods applied at sample price replacement in order to eliminate the impact of price changes due to quality changes, it is necessary to utilize old and new product pair data[18].

Outline of quality adjustment method subject to comparative analyses	
Direct comparison method (DCM)	Method which assumes that quality difference between old and new products is ignorable and thus processes "price change due to quality changes" as zero. Therefore all price difference between old and new products is regarded as "pure price change".
Overlap method (OLM)	Method which assumes that all price difference between old and new products as "price change due to quality changes" and there is no "pure price change".
Hedonic regression method (HRM)	Method which assumes that price difference between old and new products is partially due to quality difference arising from product specification. By econometric analysis, using large scale dataset, the method estimates "price change due to quality changes" and processes the remainder as "pure price change." Accuracy of the method is relatively high but the estimation burden is heavy.
Webscraped prices comparison method (WSM)	Applicable to products for which product turnover accompanying quality improvements are conducted frequently. Based on results of empirical analysis stating "price change due to quality changes account for approximately 50% of the price differences between old and new products", the method assumes the portion equivalent to 50% of the webscraped retail price difference as "price change due to quality changes" and the remainder as "pure price change".
Matched-model method (MMM)[19]	(Irrespective of whether old and new product pairs exist or not) the method calculates the percentage change of price for products which exist in the market in both survey period t and following period $t+1$ to compile an index.

[18] *Production cost method*, which assumes a portion equivalent to cost required for quality improvement obtained through interviews to firms, as "price change due to quality change," and the rest as "pure price change", is also frequently used when compiling Producer Price Index. We exclude this method because it requires interviewing and collecting cost information from firms.

[19] In the case of using a dataset furnished with information on sales volume such as scanner data, it is possible to create a weighted average price index such as Törnqvist index. However, the dataset used in this paper does not include sales volume information, thus we compile Laspeyres index upon the assumption that weights of individual products would not change over time.

As Bank of Japan (2017) noted that WSM is positioned as second-best method since the possibility of inferior accuracy compared with the other quality adjustment methods can not necessarily be denied. As a result of the analyses in this paper, if the validity of WSM can be verified, it is expected that to be widely used as a cost-effective and highly useful method.

(2) Estimation of Hedonic Functions

Under the same setup as Abe et al. (2016), in preparation for compiling indices applied with HRM, we estimate a semi-logarithmic linear (log-lin) hedonic function using the product prices as the dependent variable, the product specifications as the explanatory variable, the elapsed weeks dummy variable to control for product obsolescence, and the time dummy variable to control for macroeconomic environment.

$$\ln(p_{i,t}) = \alpha + \sum_{k} \beta_k X_{i,k} + \sum_{\tau} \gamma_\tau D_t(\tau_i + \tau) + \sum_{\tau} \delta_\tau D_t(\tau) + \varepsilon_{i,t}$$

$D_t(T)$ is a discrete delta function to satisfy the following conditions.

$$D_t(T) = \begin{cases} 1 & (\text{if} \quad t = T) \\ 0 & (\text{if} \quad t \neq T) \end{cases}$$

The hedonic function consists of $p_{i,t}$, price of product i as of the point in time t; $X_{i,k}$, kth specification of product i; dummy variables $D_t(\tau_i + \tau)$ which controls the number of elapsed weeks from the launch of each product at τ_i and $D_t(\tau)$ which controls macroeconomic shocks such as price level fluctuations in each quarter during the data period, respectively[20]. $\varepsilon_{i,t}$ is an unobserved random disturbance term. We used robust estimation for autocorrelation of error terms and heteroscedasticity. Estimation was conducted by excluding specifications that induced strong multicollinearity with other explanatory variables and specification with coefficient that does not satisfy 5% significance level, or sign condition[21].

---

[20] For the elapsed weeks dummy, total elapsed days from the launch of each product is divided by 7. For the time dummy, orthogonality with the elapsed weeks dummy was secured by identifying the quarter which includes the point of time in accordance with the calendar date.

[21] The products which lacked the detailed data on specifications were excluded from the estimation. However, if there were too many products which lacked a certain specification, the corresponding specification was excluded, in order to secure sufficient amount of data for the estimation.

We found out that adjusted R-squared for individual items secured a high level of 0.8 to 0.9, and major specifications of each item also showed high degrees of significance in general (Chart 8). Judging from the above, it is reasonable to consider that indices compiled by applied with HRM have relatively high accuracy.

(3) Compilation of Experimental Price Indices

Based on the above estimation results, we compiled experimental price indices for 20 items, and weighted average synthesis indices for both home electrical appliances and digital consumer electronic products using the *Census of Manufacture* by the Ministry of Economy, Trade and Industry, *Trade Statistics of Japan* by the Ministry of Finance, etc. (Chart 9). Observing the indices, we can point out the following three points.

First, indices applied with DCM and OLM bring a significant deviation in their trends. For durable consumer goods which tend to price the new products higher than the old, it is not surprising that the indices applied with DCM show higher price level than that with OLM. Our concern lies in the expanding pace of deviation between these two indices.

Observing the indices levels for two years, at the end of data period (December 2015), up to 50 points of deviations arise compared to the beginning (December 2013) when prices are set as 100. For example, in the case of refrigerators and freezers, indices using DCM increases to 105.9, meanwhile indices using OLM declines to 60.2. The deviation is also apparent for home electrical appliances, e.g., washers and dryers, microwaves[22]. If such a large deviation arises in just two years, we need to take into consideration that the indices level may be significantly biased and index accuracy cannot be assured if price index is compiled without examining the quality adjustment method mindfully.

Second, the trend of indices applied with WSM in general matched indices applied with HRM. When we conduct periodic averaging using RMSE (Root Mean Squared Error) and

[22] Presumably, reason why the deviation in index levels is prominent in home electrical appliances compared to digital consumer electronic products is due to difference in evaluating viewpoints of consumers. For home electrical appliances, there are factors other than quality that can be quantified (for example, product design, product images evoked via advertising media).These factors tend to be assessed by consumers, allowing product differentiation. Price competitions of these products tend to be more moderate compared to digital consumer electronic products. As a result, there is a possibility of relatively high level of price pushback (Abe et al. (2016)).

MAE (Mean Absolute Error) on the deviation between HRM applied indices and indices which applied with other quality adjustment methods, WSM trends closest to HRM except for two exceptional items, camcorders and desktops (Chart 10)[23].

Since HRM is a method to quantitatively estimate the impact of quality change on price, index accuracy tends to be higher compared to other quality adjustment methods. However, index compilation cost is high with large burden on price statisticians, as large datasets with price and specification information are needed every time the estimation is conducted. Meanwhile, WSM is a convenient method which assumes half of the price difference between old and new products to be price change due to quality change whereas the remainder to be pure price change, and does not need to conduct functional estimation on periodic basis, thus it makes the compilation cost low compared to HRM. Taking this into consideration, if difference in trend of price indices between these two methods is not significant, this can validate WSM as an appropriate quality adjustment method. Especially, considering its cost-effectiveness, it may be possible to say that making use of WSM is a wise strategy for price statistics agencies under resource constraints.

Third, there is a high tendency for indices applied with MMM to be biased downward. MMM cannot reflect price pushbacks, resulting from the revision of profitability conducted at product turnover. Therefore, for home electronic products, where products tends to obsolete easily and model upgrades occur frequently, there is a possibility of harming index accuracy due to downward bias arising from incapability to reflect price pushback. Needless to say, there is a possibility that factors unique to the estimation period are affecting the results, thus some room for allowances are necessary. However, possibility of bias still needs to be taken into consideration[24].

---

[23] For camcorders and desktops, the trends of indices applied with DCM are closer to that with HRM. We assume that the pace of technology advancement is slowing down for the relevant items, while replacement of functionalities by innovative products such as smartphones or tablet PCs is progressing. Consequently, difference in quality between old and new products have a shrinking trend, and the relevance of DCM, which assumes the price difference as pure price change, becomes larger in this case.

[24] When focusing on the fact that MMM does not reflect price pushback in indices, MMM has a similar feature to OLM, which assumes all price differences are due to price changes. In fact, when comparing indices compiled with MMM and with OLM, both indices have a tendency to decline. However, in the framework of this paper, there are differences between the two in reflecting price

V. Final Remarks

The methods used to compile price indices are broadly divided into the *traditional* and *non-traditional* approaches. The *traditional* approach uses the expertise of price statisticians when changing the surveyed target products, and the *non-traditional* approach is compiled using big data such as scanner or webscraping data.

In this research, we have combined features of both approaches by applying machine-learning methods in order to pair old and new products to big data from *Kakaku.com* for 20 major home electrical appliances and digital consumer electronic products. Using these compiled indices, we state the appropriateness of WSM, the Bank's newly developed and introduced quality adjustment method, by observing the impact individual quality adjustment methods have on indices. We also verified that applying MMM to products frequently pushing back price at product turnover may cause a downward bias.

The intended contribution of this paper is to present a new price index compilation method using big data, and once again indicating the possibility of potential bias.

In recent years, in addition to economists and data scientists, price statisticians, too, have begun to show interest in augmenting methods of compiling a price index using big data. The combined methodology addressed in this paper shows promise for use, as it has high compatibility with methods traditionally adopted by price statisticians and ensures the validity of the index. The objective of this method is not limited to the compilation of price indices. Instead of demonstrating that indices may differ significantly depending on the selection of quality adjustment method, we aim to raise an alarm over the habitual use of big data in statistical compilation, and state the importance of selecting the method appropriately. Regardless of approaches, unceasing effort to capture and suitably process qualitative change in surveyed target products is critically important in compiling price indices.

It should be noted that the machine-learning methods used in this paper are still under development. When adjusting the quality differences, the same quality adjustment method

---

trends for irregular products and supplementation of lacking prices. Therefore, MMM which only reflects the impact of product obsolescence tends to decline even further than OLM.

is applied uniformly to all pairs. Although we are, at this point in time, technologically limited to employ useful characteristics to decide upon the best quality adjustment method over the others, our future aspirations include decision-capable machine algorithms that can identify method-appropriateness depending on the product set examined.

References

Abe, N., T. Enda, N. Inakura, and A. Tonogi (2015), "Effects of New Goods and Product Turnover on Price Indexes," RCESR Discussion Paper Series No.DP15-2.

Abe, N., Y. Ito, K. Munakata, S. Ohyama, and K. Shinozaki (2016), "Pricing Patterns over Product Life-Cycle and Quality Growth at Product Turnover: Empirical Evidence from Japan" Bank of Japan Working Paper Series No.16-E-5.

Aizcorbe, A. M., C. A. Corrado, and M. E. Doms (2003), "When Do Matched-Model and Hedonic Techniques Yield Similar Measures?" FRB of San Francisco Working Paper No. 2003-14.

Aizcorbe, A. M. and Y. Pho (2005), "Differences in Hedonic and Matched-Model Price Indexes: Do the Weights Matter?" U.S. Bureau of Economic Analysis WP2005-06.

Bank of Japan (2017), "Rebasing the Corporate Goods Price Index to the Base Year 2015 —Main Features of the Rebasing and Price Developments in the 2015 Base Index—," *BOJ Reports & Research Papers*, Research and Statistics Department, Bank of Japan.

Bentley, A. and F. Krsinich (2017), "Towards a big data CPI for New Zealand," presented at the 15th Meeting of the Ottawa Group on Price Indices, Eltville am Rhein.

Cavallo, A. and R. Rigobon (2016), "The Billion Prices Project: Using Online Prices for Measurement and Research," *Journal of Economic Perspectives*, 30.2, pp.151-178.

Chessa, A. G., J. Verburg, and L. Willenborg (2017), "A Comparison of Price Index Methods for Scanner Data," presented at the 15th Meeting of the Ottawa Group on Price Indices, Eltville am Rhein.

Cortes, C. and V. Vapnik (1995), "Support-Vector Networks," *Machine Learning*, 20, pp.273-297.

Dalen, J. and O. Tarassiouk (2013), "Replacements, Quality Adjustments and Sales Prices," presented at the 13th meeting of the International Working Group on Price Indexes, Copenhagen.

Gowrisankaran, G. and M. Rysman (2012), "Dynamics of Consumer Demand for New Durable Goods," *Journal of Political Economy*, 120.6, pp.1173-1219.

Hoffmann, J. (1999), "The Treatment of Quality Changes in the German Consumer Price Index," presented at the 5th meeting of the International Working Group on Price Indexes, Reykjavik.

Hoven, L. (1999), "Some Observations on Quality Adjustment in the Netherlands," presented at the 5th meeting of the International Working Group on Price Indexes, Reykjavik.

Hsu, C.-W., C.-C. Chang, and C.-J. Lin (2016), "A Practical Guide to Support Vector Classification," Department of Computer Science, National Taiwan University. Available online at https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

ILO, IMF, OECD, UNECE, Eurostat, World Bank (eds.) (2004a), "Consumer Price Index Manual: Theory and Practice," International Labour Office, Geneva.

ILO, IMF, OECD, UNECE, World Bank (eds.) (2004b), "Producer Price Index Manual: Theory and Practice," International Monetary Fund, Washington, D.C.

Loon, K. V. and D. Roels (2018), "Integrating Big Data in the Belgian CPI," presented at the Meeting of the Group of Experts on Consumer Price Indices, Geneva.

Melser, D. and I. A. Syed (2014), "Life Cycle Price Trends and Product Replacement: Implications for the Measurement of Inflation," UNSW Business School Research Paper No.2014-ECON-40.

Office for National Statistics (2017), "Research Indices using Web Scraped Price Data: August 2017 Update," Office for National Statistics, U.K.

Ohta, Makoto (1977), "A Proposal for Using 50 Percent Rule in Making Quality-Adjusted Price Indexes", *Economic Studies Quarterly*, 28.3, pp.266-269.

Powers, D. M. W. (2011), "Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation," *Journal of Machine Learning Technologies*, 2.1, pp.37-63.

Triplett, J. E. (2006), "Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products," OECD Publishing.

Ueda, K., K. Watanabe, and T. Watanabe (2016), "Product Turnover and Deflation: Evidence from Japan," CARF Working Paper F-400.

Winkler, W. E. (1990), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods*, 76.4, pp.501-505.

Wu, X., V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg (2008), "Top 10 Algorithms in Data Mining," *Knowledge and Information Systems*, 14:1-37.

(Chart 1)

Concept of Quality Growth and Pure Price Increase

(1) Concept of Quality Adjustment

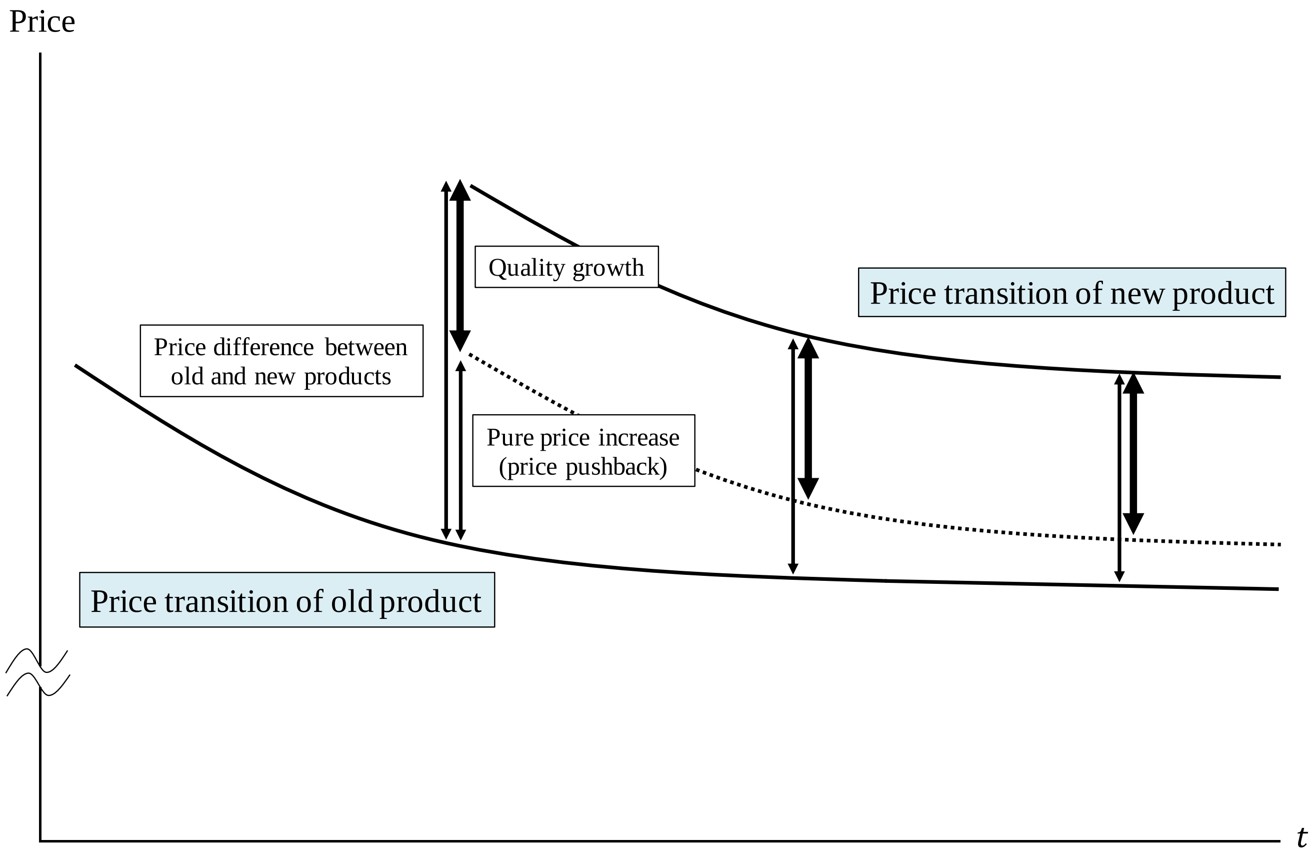
Price difference between old and new product = Price change resulting from quality changes + Pure price change

Price change resulting from quality changes: Used for quality adjusting

Pure price change: Used to reflect to price index

At the time of sample price replacement, it is endeavored to reflect only "Pure price change" to price index, after removing "Price change resulting from quality changes" by using an appropriate quality adjustment method.

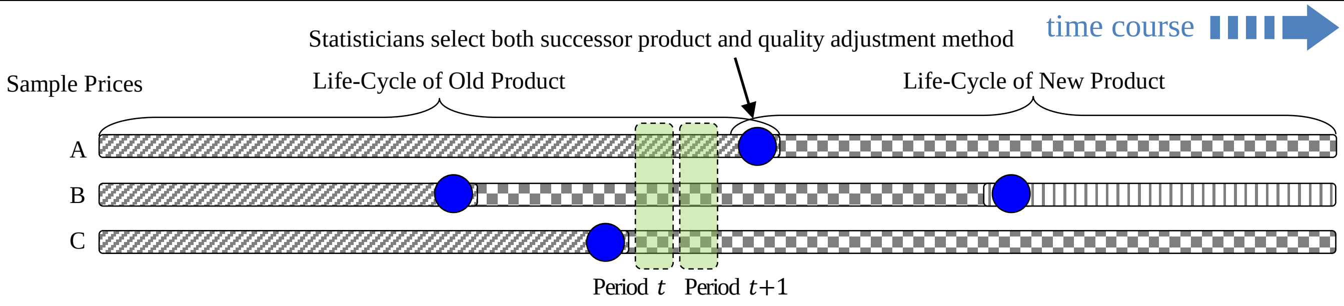
(2) Conceptual Diagram of Quality Growth and Pure Price Increase



(Chart 2)

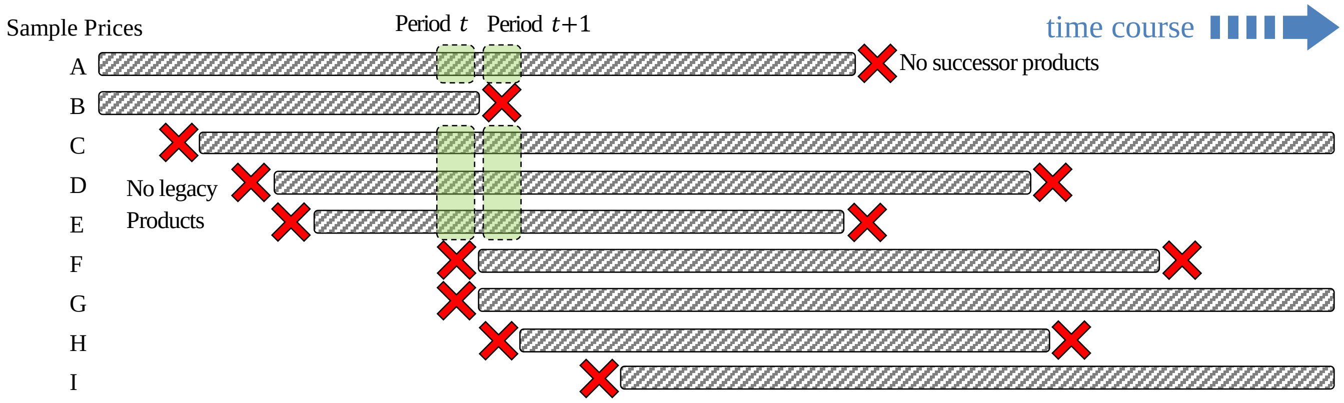
Comparison of Approaches for Compiling Price Indices

(1) *Traditional* Approach of Price Statistics Agencies



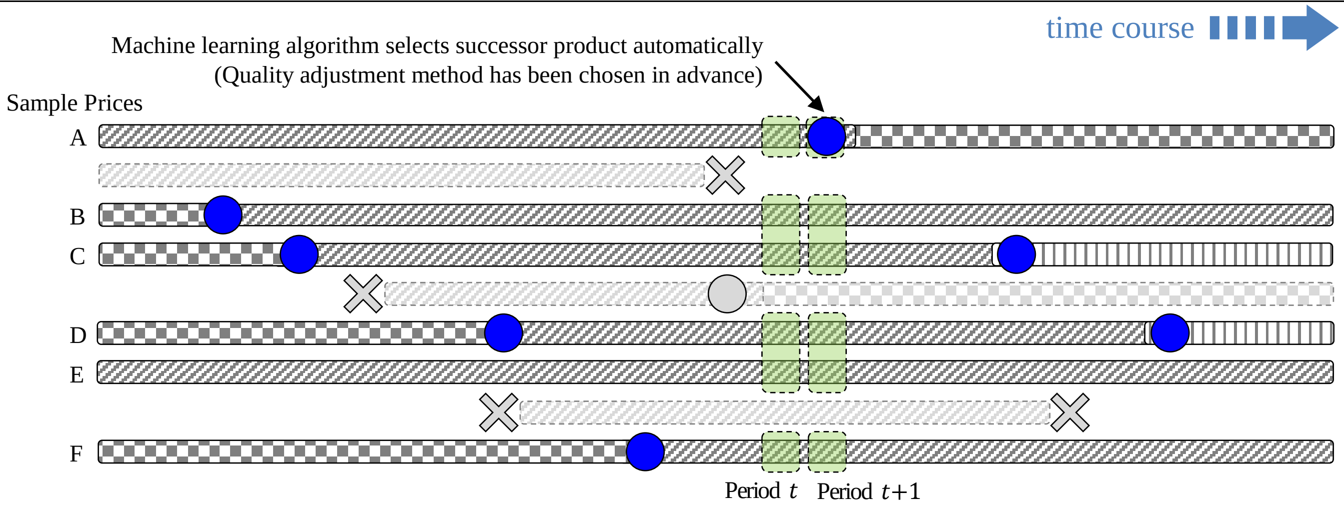
*For relatively small number of pairs of old and new products, statisticians select quality adjustment method.

(2) *Non-Traditional* Approach Using Big Data



*Calculating the percentage change of price for products which exist in the market in both survey period (Period *t*) and following period (Period *t+1*), and reflect the ratio to the index.

(3) Approach We Take in This Paper



*For vast amounts of data, machine learning algorithm makes pairs of old and new products automatically and applies pre-determined quality adjustment method for those pairs.

(Chart 3)

Total Number of Product Pairs and the Amount of Supervised Data

(1) Total Number of Product Pairs

	Total Number of Product Pairs	Number of Pairs in Supervised Data	Number of Old and New Product Pairs in Supervised Data
Home Electrical Appliances	40,205	3,584	254
Air conditioners	23,061	512	49
Refrigerators and freezers	6,287	512	31
Washers and dryers	2,465	512	49
Rice cookers	2,860	512	23
Vacuum cleaners	1,334	512	49
Microwaves	1,306	512	40
Hair dryers and curling irons	1,959	512	13
Air purifiers	933	N/A	N/A
Total	91,758	8,192	551

	Total Number of Product Pairs	Number of Pairs in Supervised Data	Number of Old and New Product Pairs in Supervised Data
Digital Consumer Electronics	51,553	4,608	297
GPS navigations	1,499	512	64
External hard drives	6,739	512	19
LCD TVs	2,346	512	19
LCD monitors	1,999	512	18
Printers	5,286	512	17
Blu-ray and DVD recorders	971	N/A	N/A
Headphones	6,396	512	22
Camcorders	286	N/A	N/A
Laptops	19,791	512	29
Desktops	3,716	512	54
Point-and-shoot cameras	1,496	512	55
DSLR and mirrorless cameras	1,028	N/A	N/A

Note: For the four items (Air purifiers, Blu-ray and DVD recorders, Camcorders, DSLR and mirrorless cameras) for which the total number of product pairs is relatively small, we decided to select the pairs manually without applying machine learning methods.

(2) Determination of the Amount of Supervised Data

When we evaluated the performance of classifiers by F-measure (stated in detail later), we confirmed that even if we increased the amount of supervised data to 1,024, the improvement of performance is limited, so we set the amount of product pair data used as supervised data to 512.

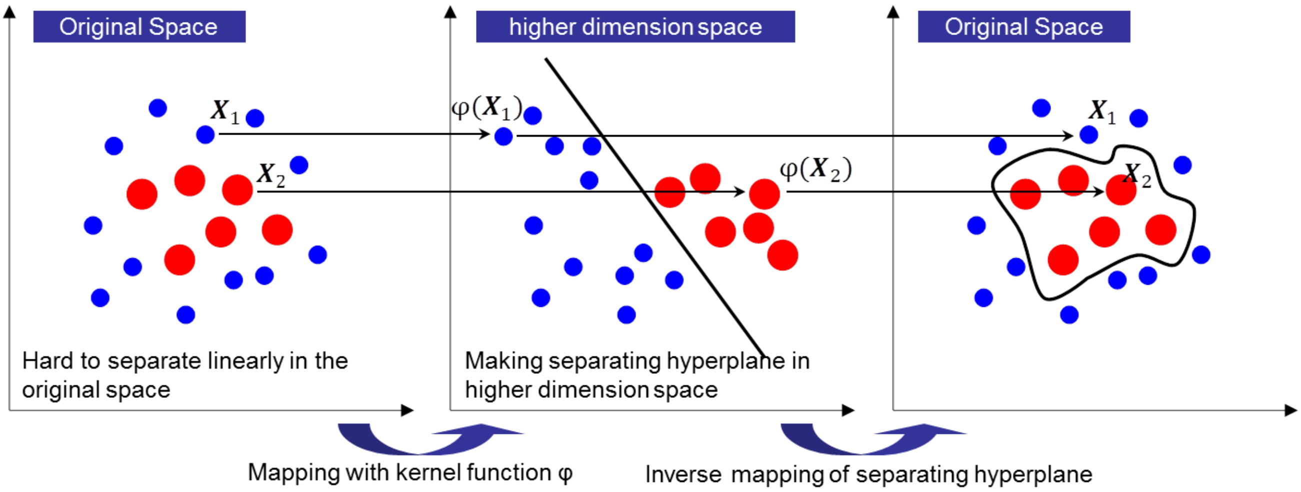
	Total Number of Product Pairs	F-measures Correspond to the Amount of Supervised Data					
		64 (2^6)	128 (2^7)	256 (2^8)	512 (2^9)	1,024 (2^{10})	difference
Air conditioners	23,061	0.33	0.56	0.85	0.85	0.83	▲ 0.02
Refrigerators and freezers	6,287	0.20	0.47	0.67	0.70	0.73	＋ 0.03
External hard drives	6,739	0.10	0.13	0.43	0.71	0.66	▲ 0.05
Laptops	19,791	0.10	0.35	0.70	0.88	0.90	＋ 0.02

(Chart 4)

Understanding Kernel Trick and Performance of Classifiers

(1) Understanding Kernel Trick

Suppose there is a binary classification problem in 2-dimensional space. Although it is difficult to ensure linear separability among samples in the original space, we could derive a separating hyperplane by using a kernel function φ to map the original space to higher dimension space and vice versa. Such technique is called *kernel trick*, and the method to make a non-linear separating hyperplane with the aid of *kernel trick* is called non-linear Support Vector Machines (SVMs). For further details, see Cortes and Vapnik (1995) etc.



(2) Definition of F-measure

		Actual class	
		P	N
Predicted class	P	True Positive (TP)	False Positive (FP)
	N	False Negative (FN)	True Negative (TN)

$$\text{precision} \equiv \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{recall} \equiv \frac{\text{TP}}{\text{TP} + \text{FN}}$$

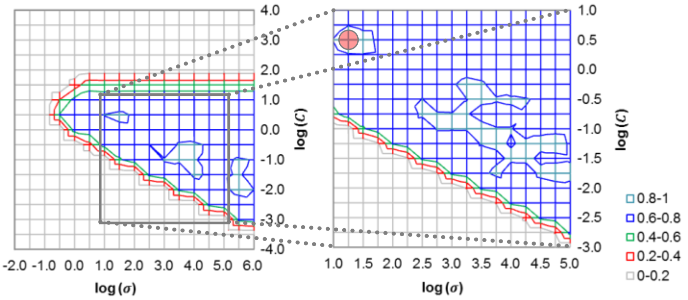
The F-measure indicating the performance of classifiers is defined as follows:

$$\text{F-measure} \equiv \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

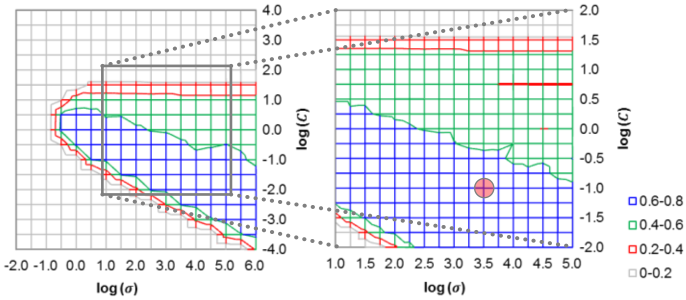
For further details, see Powers (2011) etc.

Hyperparameters Optimization Using 10-fold Cross-Validation and Grid Search [1]

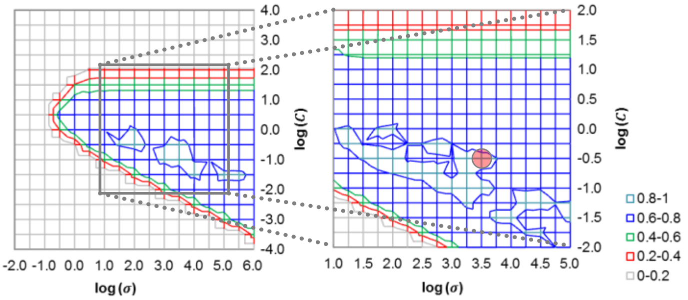
(1) Air conditioners



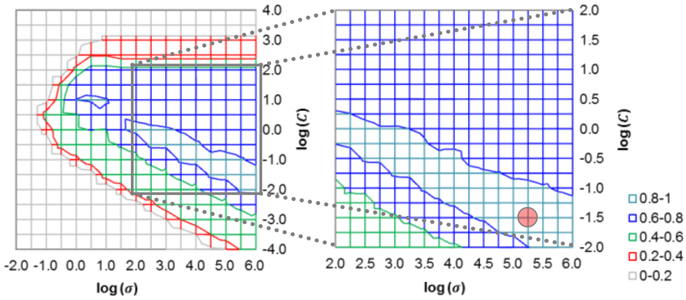
(2) Refrigerators and freezers



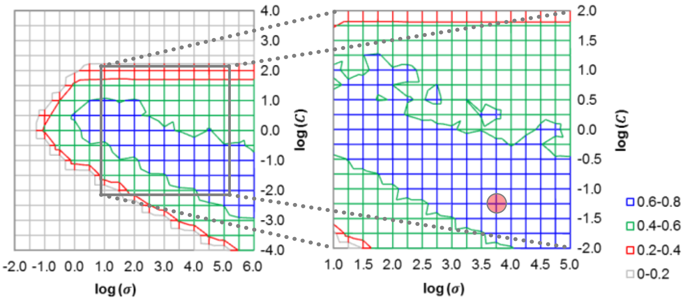
(3) Washers and dryers



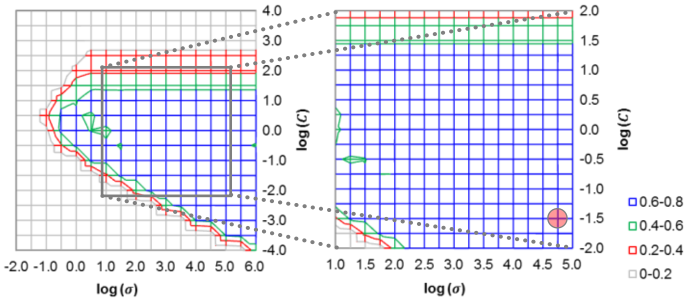
(4) Rice cookers



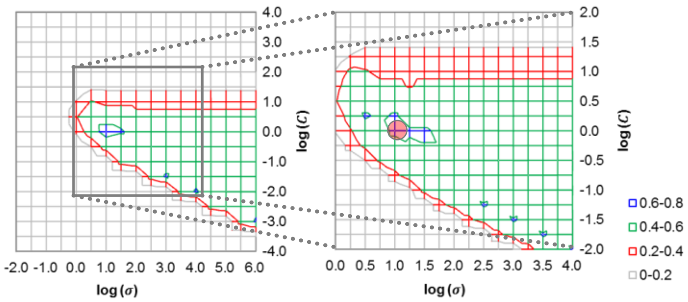
(5) Vacuum cleaners



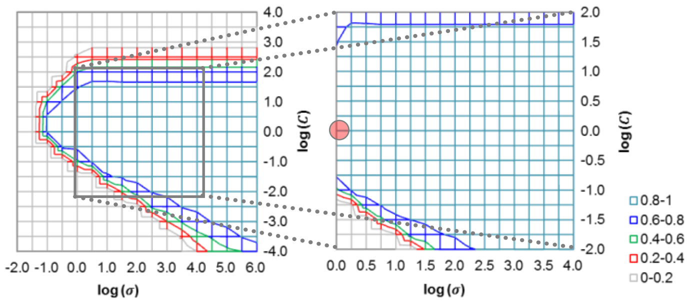
(6) Microwaves



(7) Hair dryers and curling irons



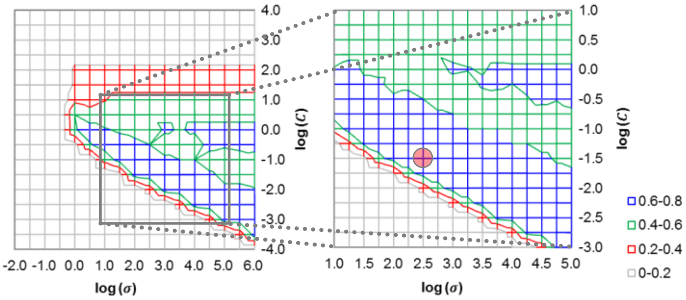
(8) GPS navigations



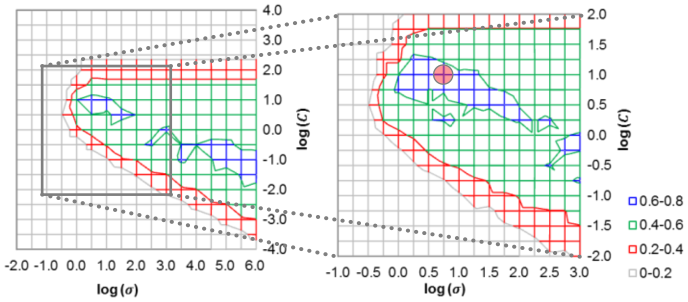
Note: The lattice highlighted in red indicates hyperparameters (σ, C) to maximize F-measure at the time of conducting 10-fold cross-validation obtained by grid search.

Hyperparameters Optimization Using 10-fold Cross-Validation and Grid Search [2]

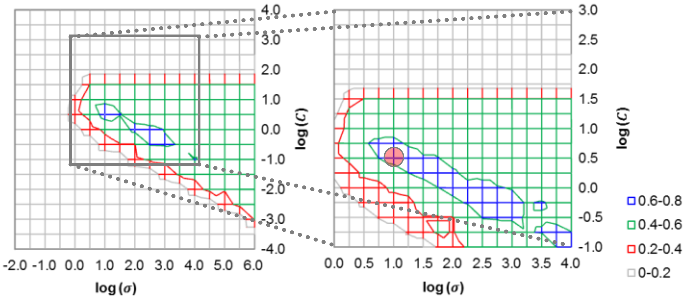
(1) External hard drives



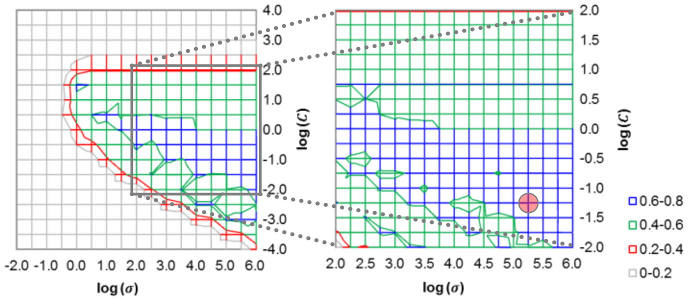
(2) LCD TVs



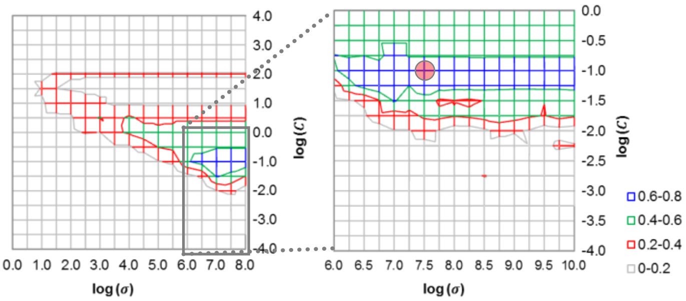
(3) LCD monitors



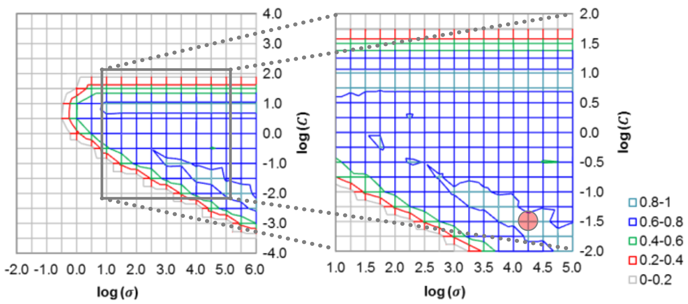
(4) Printers



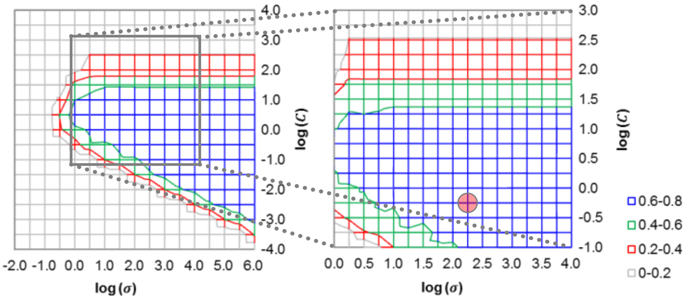
(5) Headphones



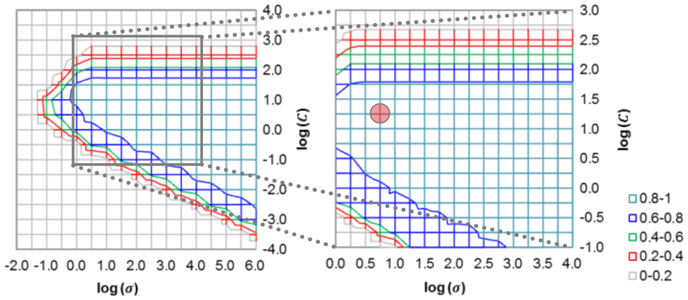
(6) Laptops



(7) Desktops



(8) Point-and-shoot cameras



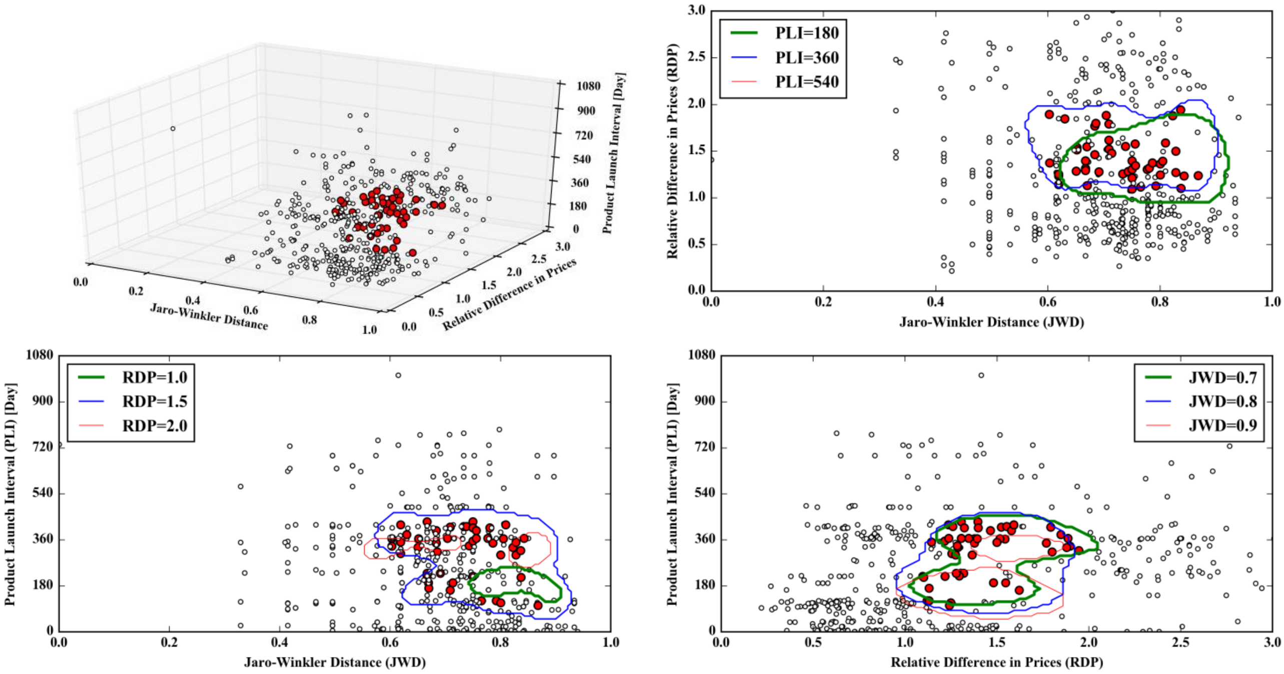
Optimization Results of Hyperparameters (σ, C)

	Hyperparameters obtained by using Grid Search Method				F-measure
	σ	log(σ)	C	log(C)	
Home Electrical Appliances					
Air conditioners	17.78	1.25	3.16	0.50	0.8500
Refrigerators and freezers	3162.28	3.50	0.10	-1.00	0.7026
Washers and dryers	3162.28	3.50	0.32	-0.50	0.8305
Rice cookers	1.78E+05	5.25	0.03	-1.50	0.9023
Vacuum cleaners	5623.41	3.75	0.06	-1.25	0.7219
Microwaves	56234.13	4.75	0.03	-1.50	0.7711
Hair dryers and curling irons	10.00	1.00	1.00	0.00	0.6467
Air purifiers	N/A	N/A	N/A	N/A	N/A
Digital Consumer Electronics					
GPS navigations	1.00	0.00	1.00	0.00	0.8925
External hard drives	316.23	2.50	0.03	-1.50	0.7133
LCD TVs	5.62	0.75	10.00	1.00	0.7005
LCD monitors	10.00	1.00	3.16	0.50	0.6857
Printers	177827.94	5.25	0.06	-1.25	0.7067
Blu-ray and DVD recorders	N/A	N/A	N/A	N/A	N/A
Headphones	3.16E+07	7.50	0.10	-1.00	0.7616
Camcorders	N/A	N/A	N/A	N/A	N/A
Laptops	17782.79	4.25	0.03	-1.50	0.8778
Desktops	177.83	2.25	0.56	-0.25	0.7778
Point-and-shoot cameras	5.62	0.75	17.78	1.25	0.9512
DSLR and mirrorless cameras	N/A	N/A	N/A	N/A	N/A

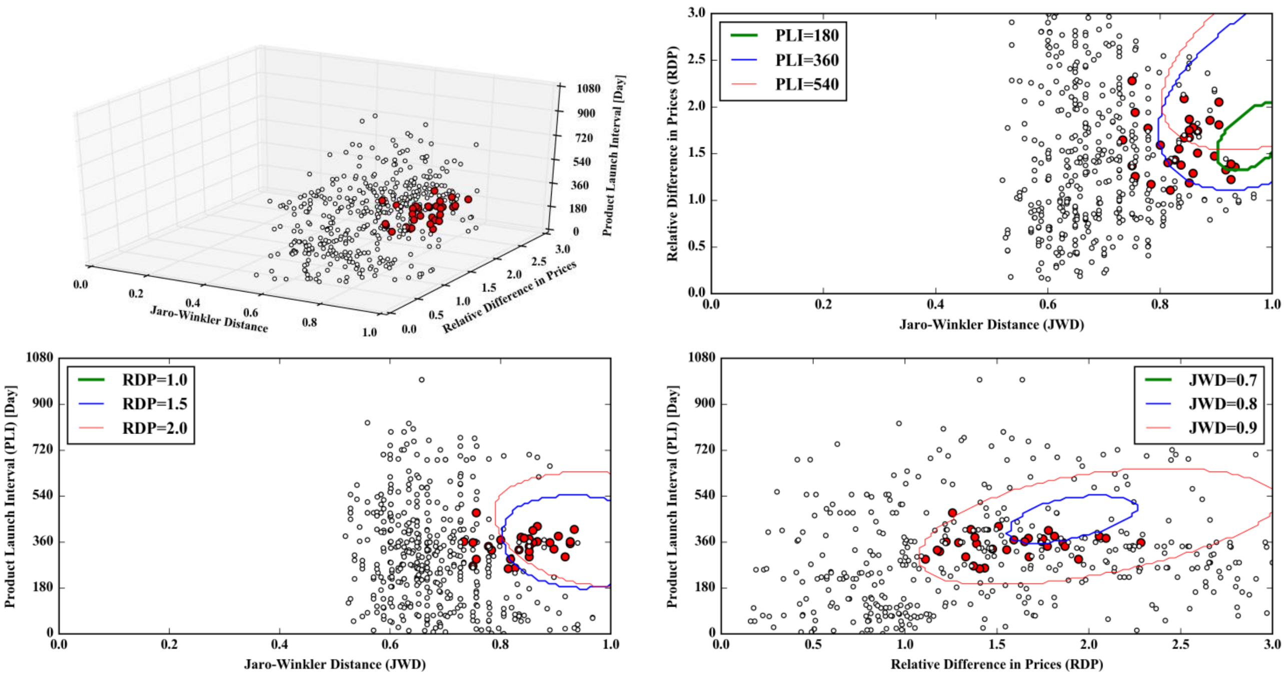
Note: This table organizes the optimal hyperparameters (σ, C) for each item to maximize F-measure at the time of conducting 10-fold cross-validation obtained by grid search. The kernel parameter σ controls the extent to which the complexity of the data boundary surface will be reflected in classifiers, and the penalty parameter C controls the extent to which faulty identification is allowed.

Optimal Hyperplanes Using Non-Linear SVM Classifiers [1]

(1) Air conditioners



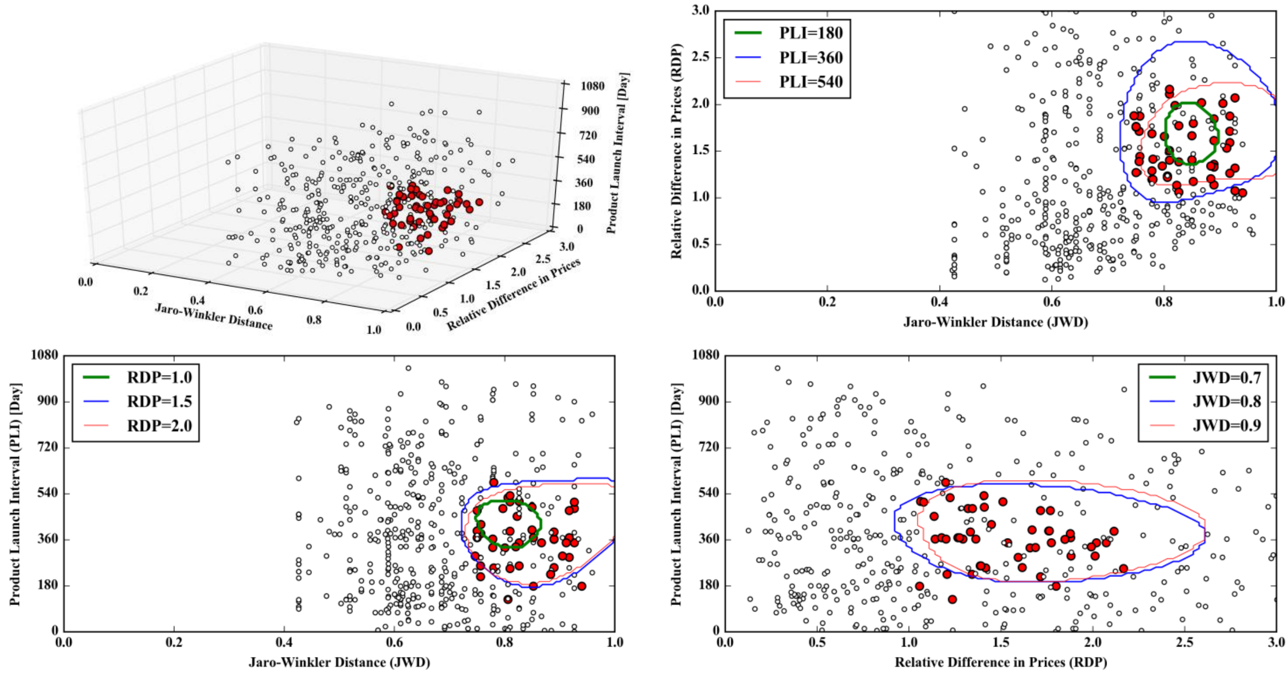
(2) Refrigerators and freezers



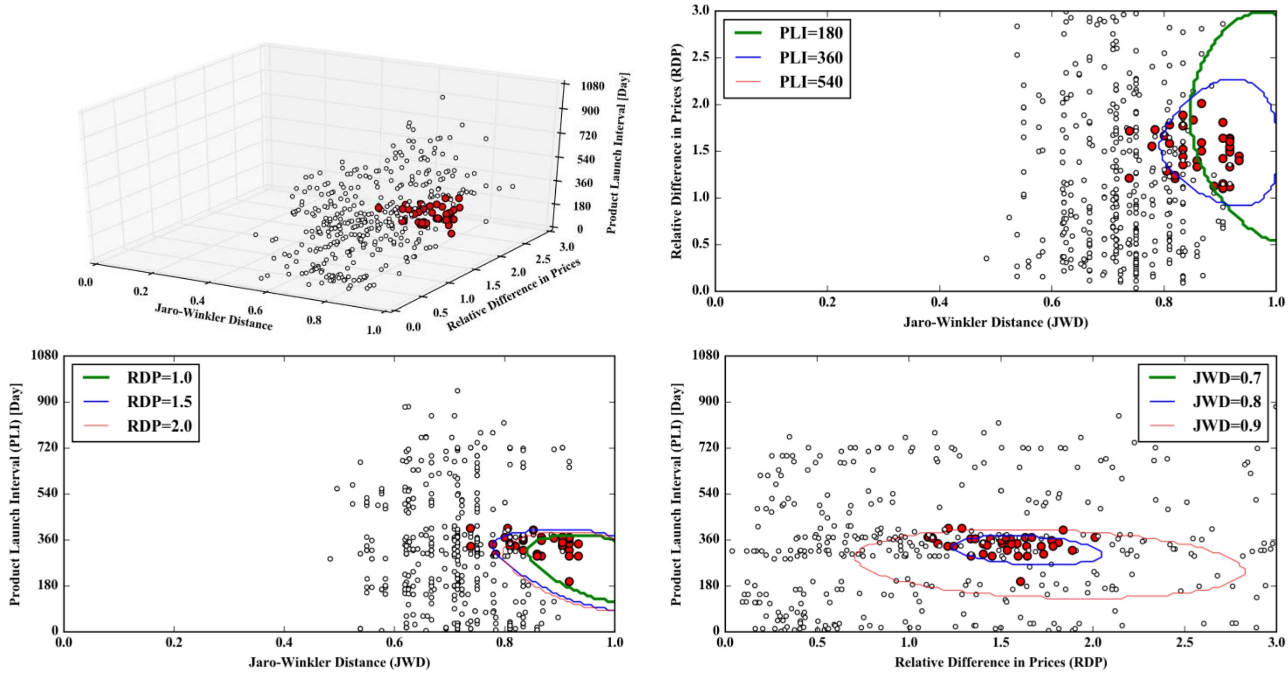
Note: Red dots indicate pairs of old and new products created as supervised data for machine learning, and white dots represent pairs of the irrelevant products. Among the four scatter diagrams illustrated for individual items, the diagram to the upper left indicates 3-dimensional stereogram with three characteristics vectors, and the other three diagrams indicate 2-dimensional sectional views, respectively.

Optimal Hyperplanes Using Non-Linear SVM Classifiers [2]

(3) Washers and dryers



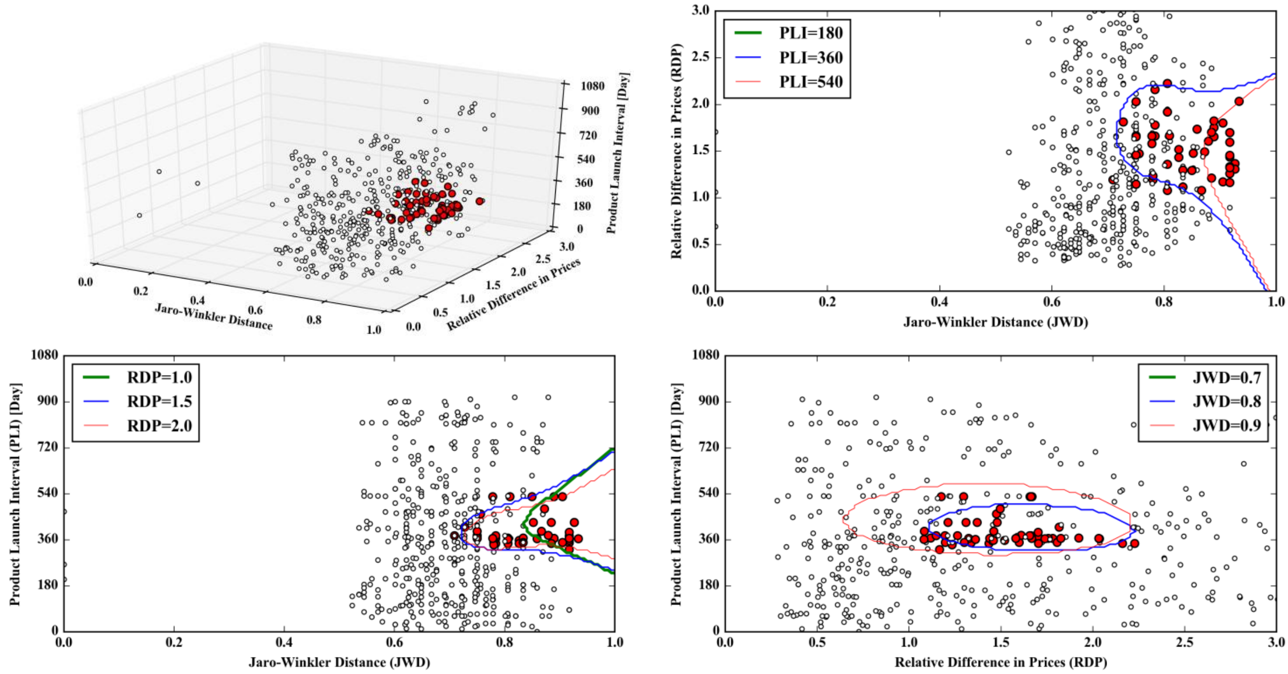
(4) Rice cookers



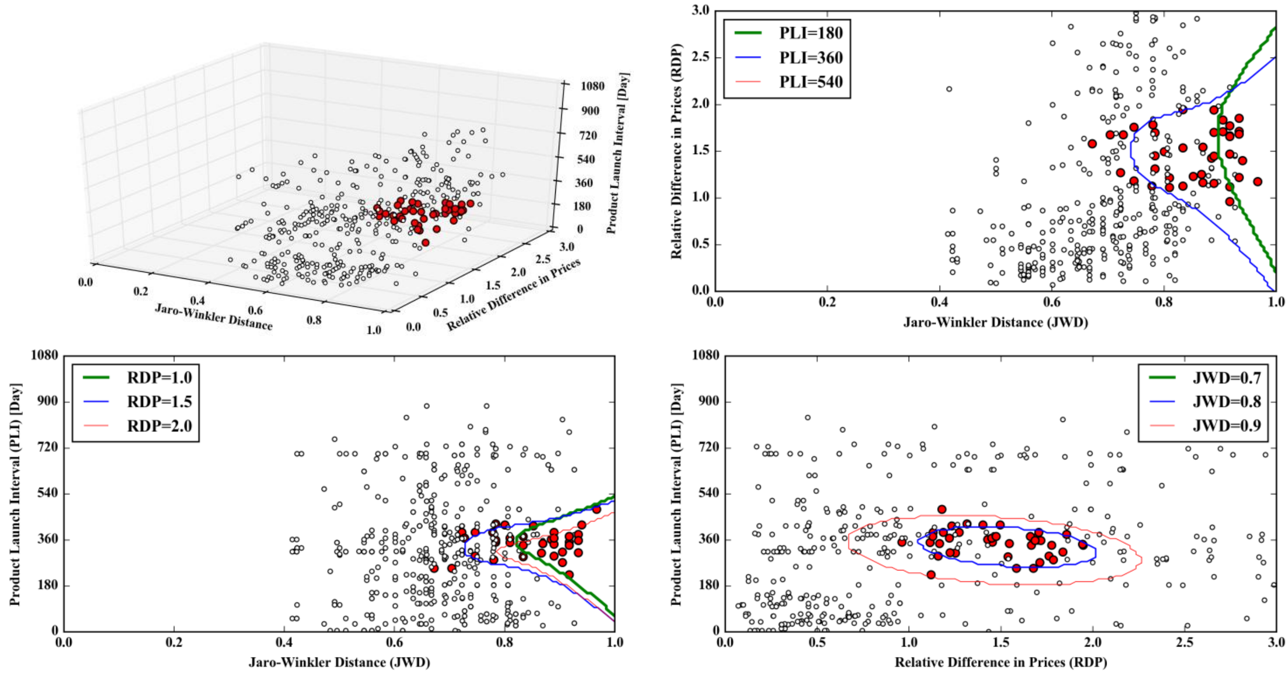
(Chart 7–3)

Optimal Hyperplanes Using Non-Linear SVM Classifiers [3]

(5) Vacuum cleaners

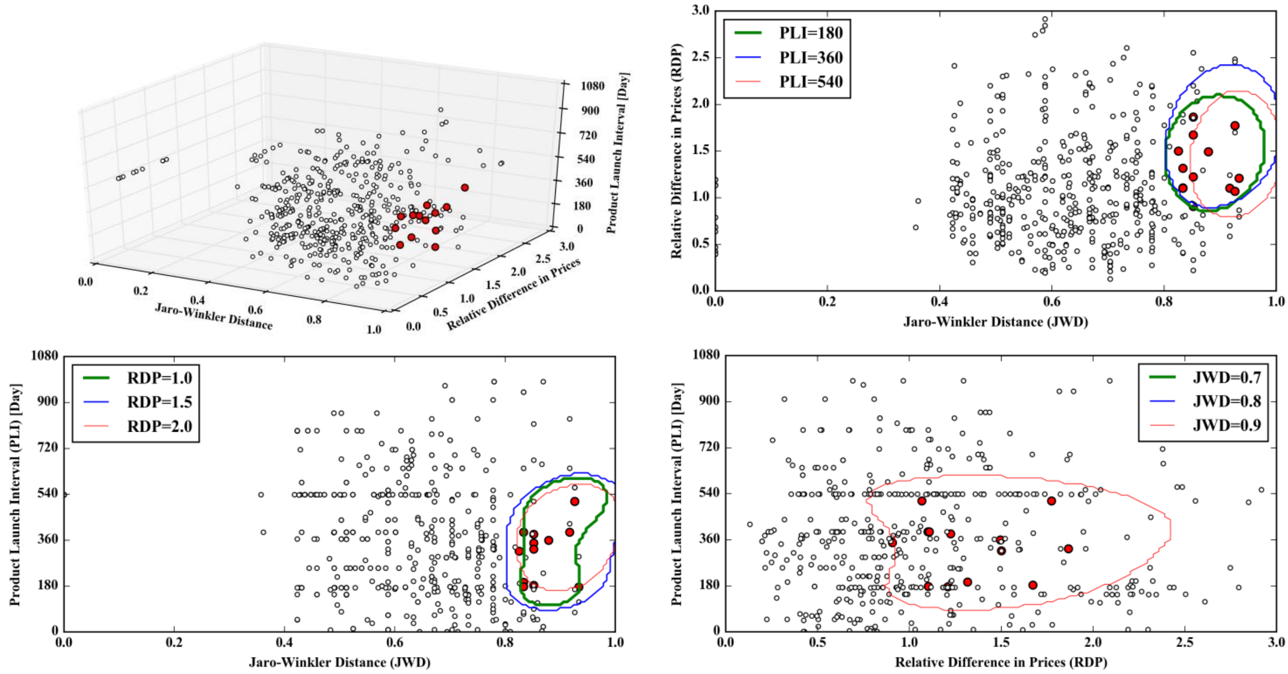


(6) Microwaves

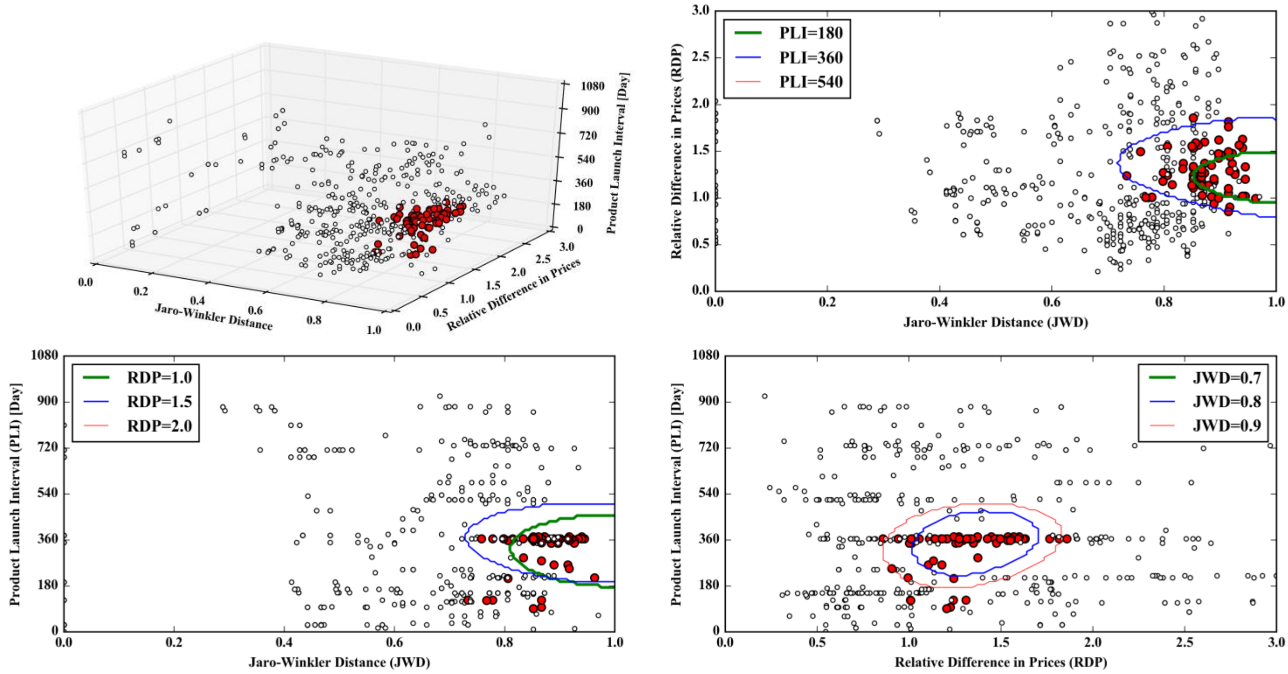


Optimal Hyperplanes Using Non-Linear SVM Classifiers [4]

(7) Hair dryers and curling irons

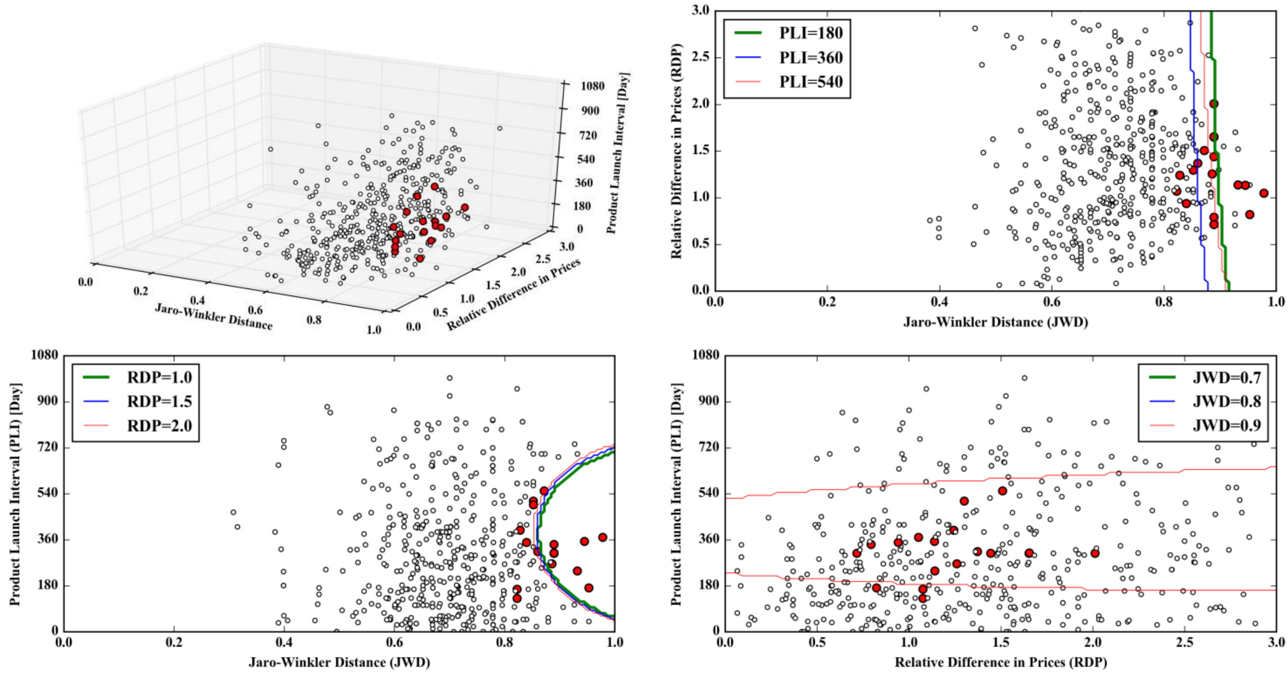


(8) GPS navigations

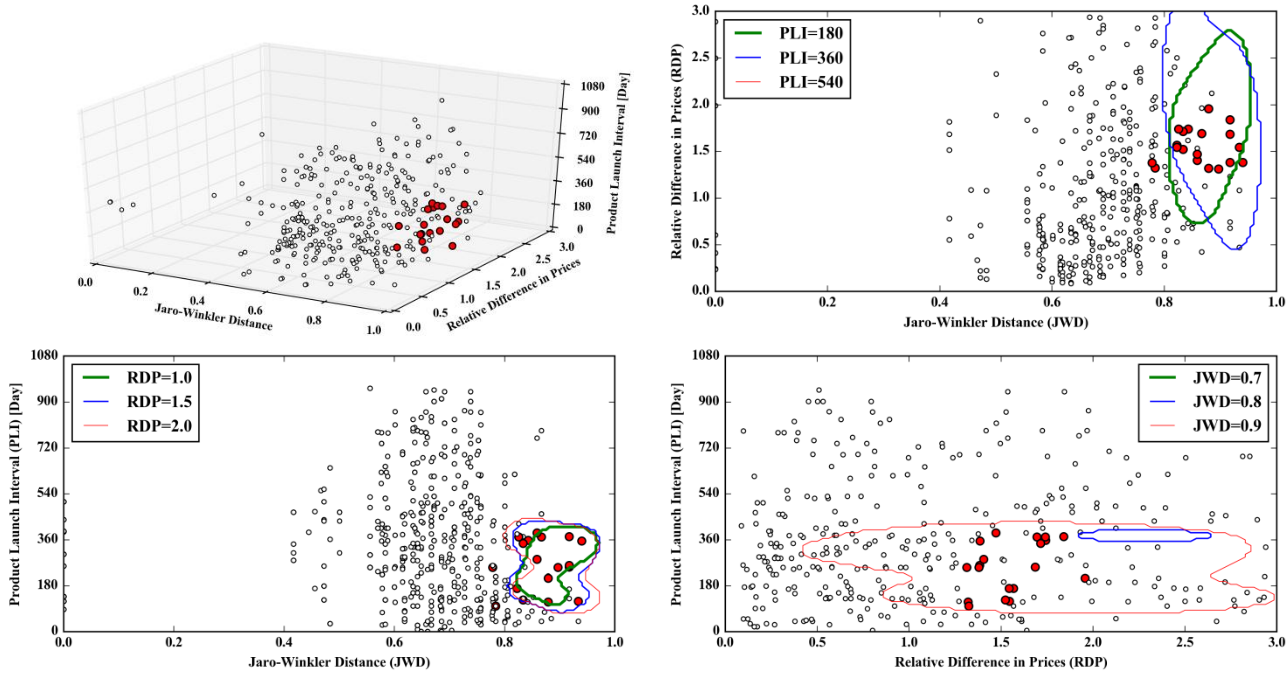


Optimal Hyperplanes Using Non-Linear SVM Classifiers [5]

(9) External hard drives

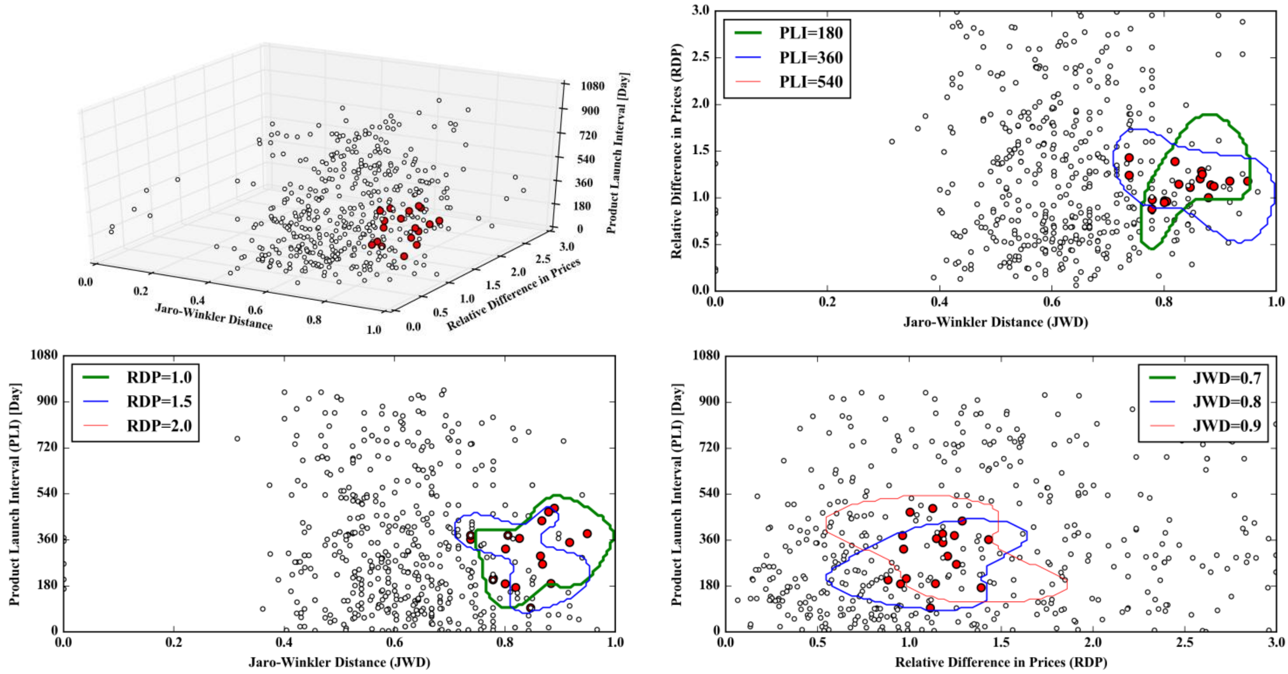


(10) LCD TVs

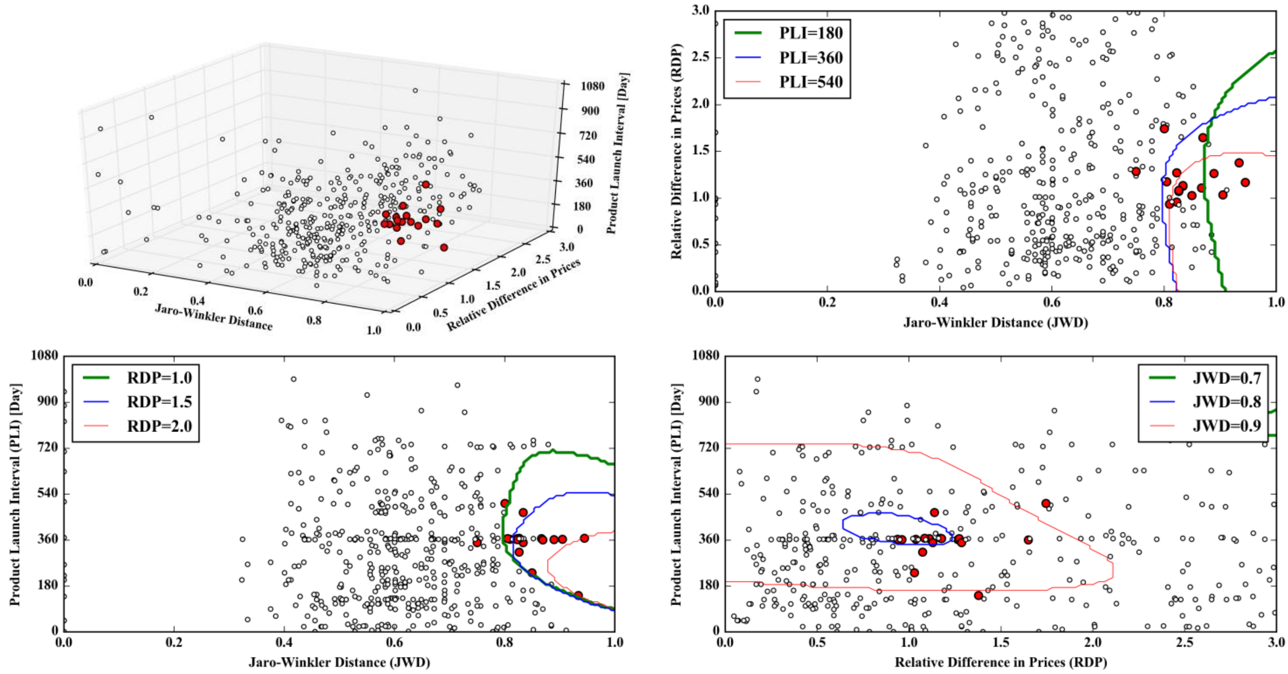


Optimal Hyperplanes Using Non-Linear SVM Classifiers [6]

(11) LCD monitors



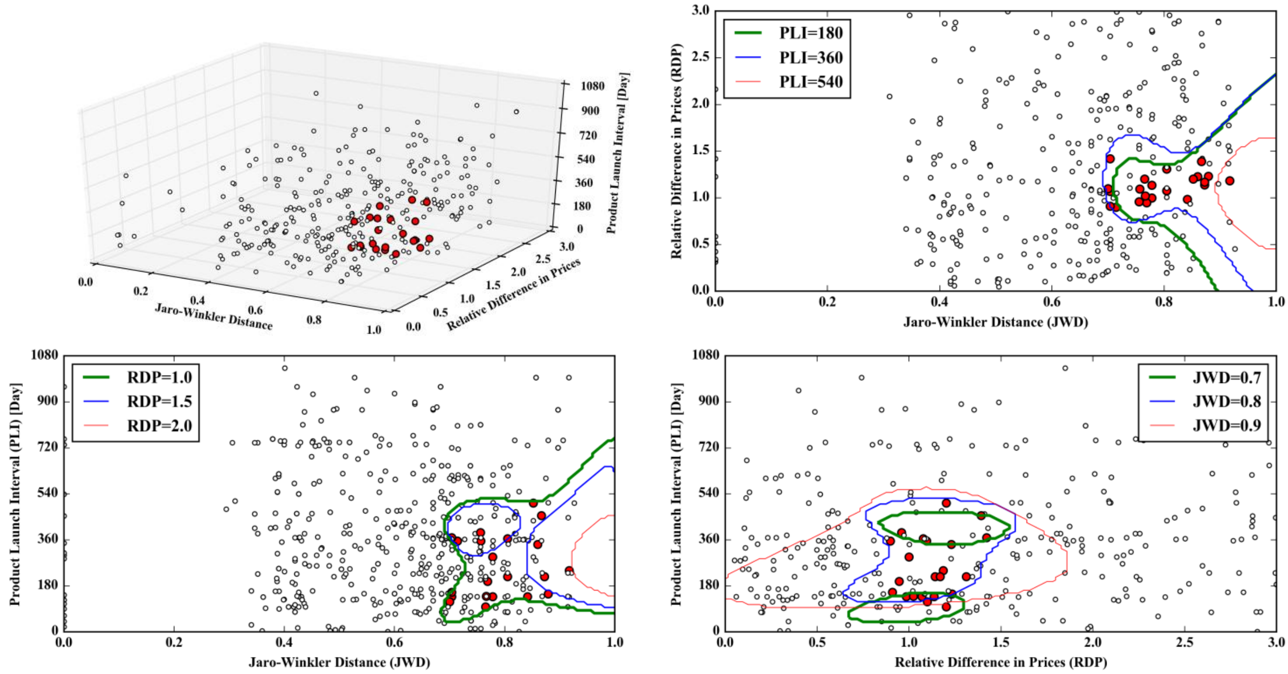
(12) Printers



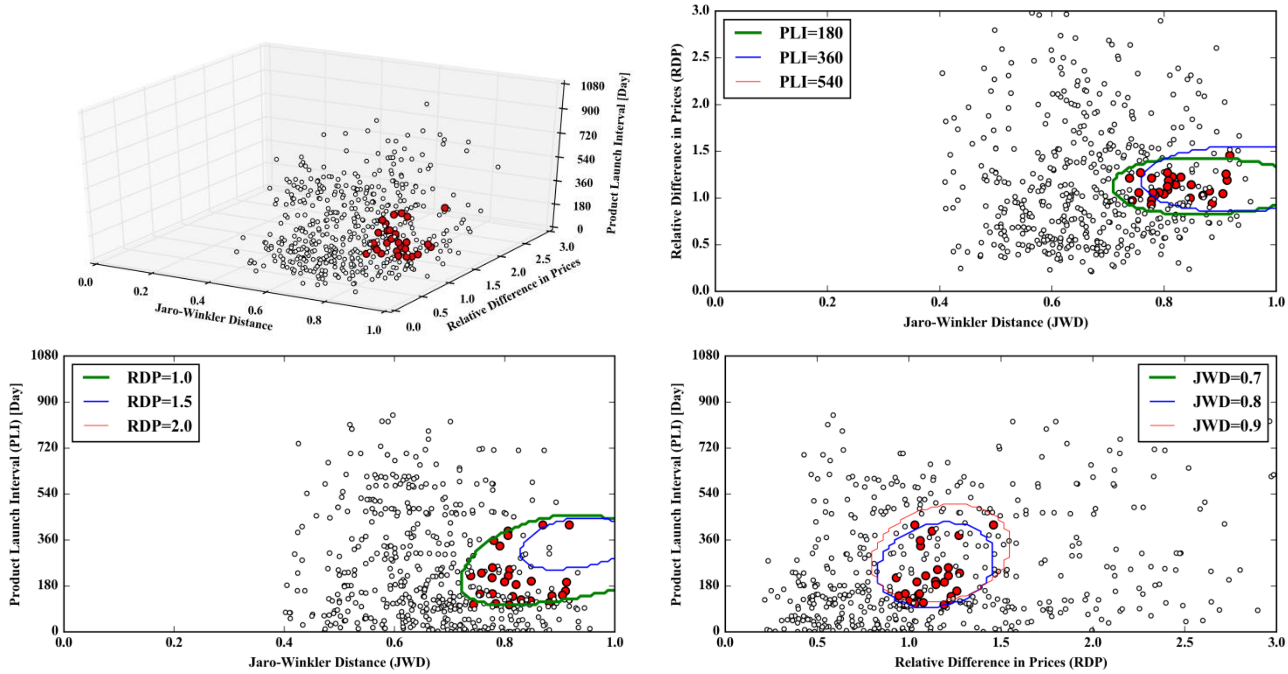
(Chart 7–7)

Optimal Hyperplanes Using Non-Linear SVM Classifiers [7]

(13) Headphones

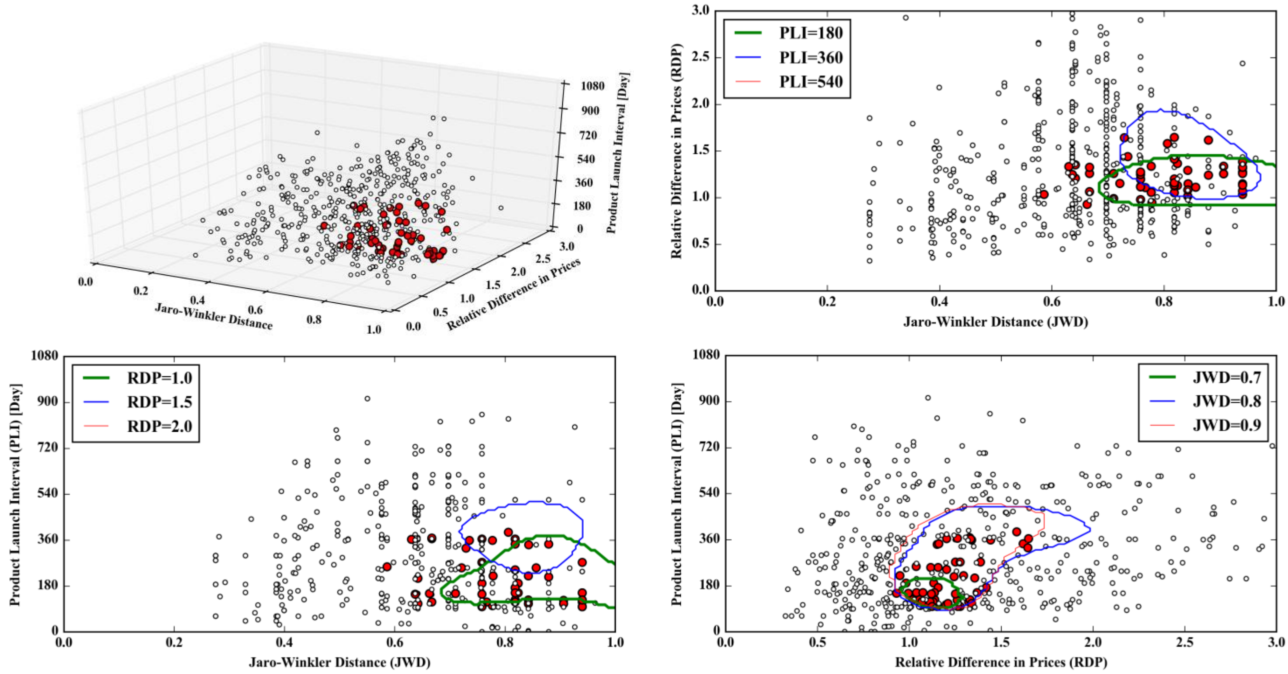


(14) Laptops

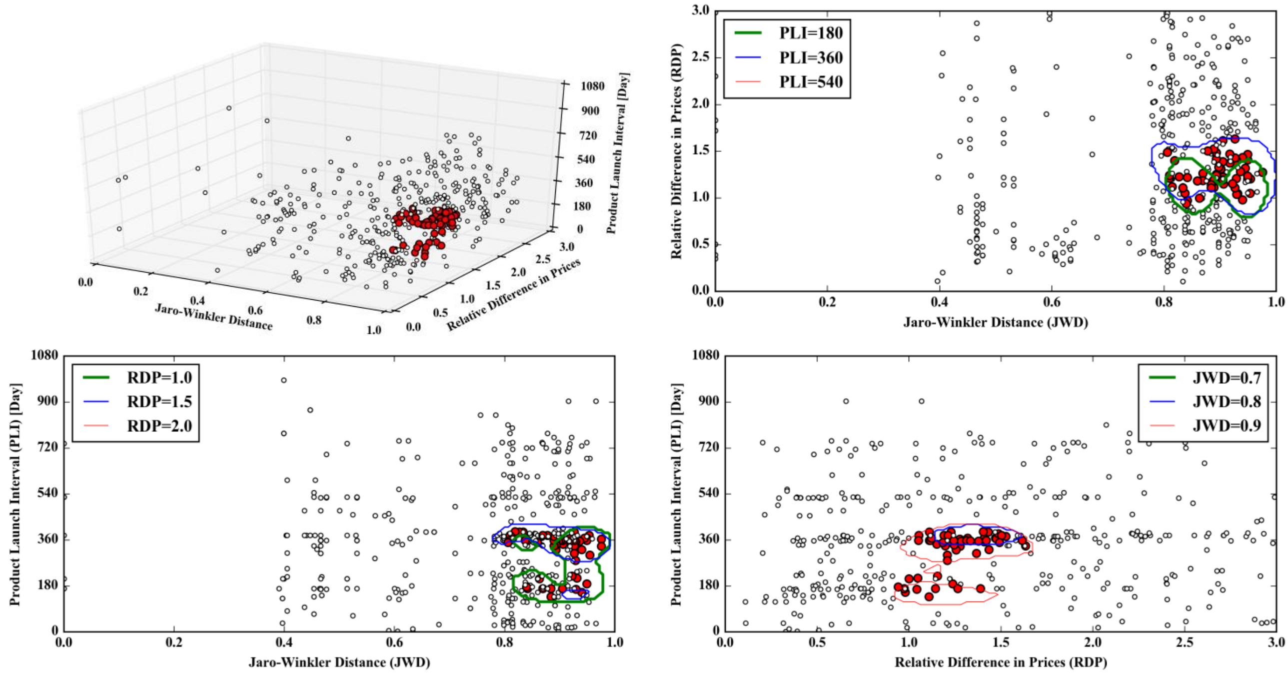


Optimal Hyperplanes Using Non-Linear SVM Classifiers [8]

(15) Desktops



(16) Point-and-shoot cameras



Estimation Results of Hedonic Regression: Home Electrical Appliances [1]

(1) Air conditioners

Dependent Variable: log(average price)		
Intercept	10.239 (0.153)	***
Heating Capacity (mat)	0.041 (0.004)	***
Low-temperature Heating Capacity (kW)	0.025 (0.009)	**
Annual Performance Factor	0.072 (0.021)	***
Dummy Variables		
Human Body Sensitive Sensor		
Body	0.076 (0.021)	***
Remote Control	0.262 (0.066)	***
Air Sterilization System	0.107 (0.030)	***
Clothes Dryer System	0.168 (0.024)	***
Automatic Washing System of Filter	0.162 (0.025)	***
Airflow Control System	0.206 (0.052)	***
The refrigerant circuit R32	0.096 (0.028)	***
Reheating Dehumidifier System	0.078 (0.025)	**
Voice Guide System	0.119 (0.026)	***
Manufacturers		
Manufacturer A	0.148 (0.034)	***
Manufacturer B	0.284 (0.040)	***
Manufacturer C	0.278 (0.034)	***
Manufacturer D	0.146 (0.039)	***
Manufacturer E	0.121 (0.039)	**
Elapsed Weeks		
2nd week	0.006 (0.014)	
3rd week	0.000 (0.019)	
4th week	-0.032 (0.019)	
5th week	-0.042 (0.020)	*
6th week	-0.059 (0.020)	**
7th week	-0.067 (0.020)	***
8th week	-0.079 (0.020)	***
9th week	-0.100 (0.020)	***
10th week	-0.120 (0.020)	***
11th week	-0.130 (0.020)	***
12th week	-0.137 (0.020)	***
13th week	-0.154 (0.020)	***
Adjusted R-squared		0.870
Standard Error of Regression		0.159
Mean of Dependent Variable		11.836
Standard Deviation of Dependent Variable		0.441
Number of products		536
Size of Panel Data		20,135
Number of Specifications Data		30
Volume of Total Data		664,455

Notes: Values in ( ) indicate standard errors.

***, **, * denote significance at the 0.1%, 1%, 5% level.

(2) Refrigerators and freezers

Dependent Variable: log(average price)		
Intercept	9.992 (0.063)	***
Internal Volume (L)	0.003 (0.000)	***
Switching Chamber (L)	0.003 (0.001)	*
Achievement Ratio of the Energy Saving Target	0.001 (0.000)	***
Dummy Variables		
Deodorizing System	0.136 (0.052)	**
Automatic Icemaker System	0.150 (0.024)	***
Manufacturers		
Manufacturer A	0.231 (0.076)	**
Manufacturer B	0.352 (0.082)	***
Manufacturer C	2.191 (0.120)	***
Manufacturer D	0.288 (0.077)	***
Manufacturer E	0.354 (0.080)	***
Manufacturer F	0.366 (0.085)	***
Manufacturer G	0.431 (0.083)	***
Elapsed Weeks		
2nd week	-0.035 (0.016)	*
3rd week	-0.044 (0.017)	*
4th week	-0.068 (0.018)	***
5th week	-0.055 (0.028)	
6th week	-0.111 (0.019)	***
7th week	-0.143 (0.021)	***
8th week	-0.153 (0.025)	***
9th week	-0.188 (0.020)	***
10th week	-0.199 (0.020)	***
11th week	-0.215 (0.020)	***
12th week	-0.229 (0.020)	***
13th week	-0.239 (0.020)	***
Adjusted R-squared		0.940
Standard Error of Regression		0.163
Mean of Dependent Variable		11.745
Standard Deviation of Dependent Variable		0.662
Number of products		321
Size of Panel Data		10,910
Number of Specifications Data		20
Volume of Total Data		250,930

Notes: Values in ( ) indicate standard errors.

***, **, * denote significance at the 0.1%, 1%, 5% level.

Estimation Results of Hedonic Regression: Home Electrical Appliances [2]

(3) Washers and dryers

Dependent Variable: log(average price)		
Intercept	10.304 (0.227)	***
Washing Capacity (kg)	0.123 (0.013)	***
Noise Level (dB)	-0.014 (0.004)	***
Dummy Variables		
Style		
Washer Dryer	0.432 (0.041)	***
Opening and Closing type		
Left-opening	0.368 (0.050)	***
Right-opening	0.503 (0.064)	***
Automatic Cleaning System	0.139 (0.032)	***
Bath Water Drawing Pump System	0.088 (0.043)	*
Manufacturers		
Manufacturer A	0.178 (0.033)	***
Manufacturer B	0.367 (0.058)	***
Manufacturer C	0.239 (0.045)	***
Manufacturer D	0.113 (0.043)	**
Elapsed Weeks		
2nd week	0.009 (0.011)	
3rd week	0.034 (0.016)	*
4th week	0.029 (0.020)	
5th week	0.027 (0.016)	
6th week	0.000 (0.021)	
7th week	0.006 (0.023)	
8th week	0.014 (0.024)	
9th week	0.021 (0.026)	
10th week	0.013 (0.024)	
11th week	0.022 (0.025)	
12th week	0.018 (0.025)	
13th week	0.008 (0.024)	
Adjusted R-squared	0.894	
Standard Error of Regression	0.196	
Mean of Dependent Variable	11.307	
Standard Deviation of Dependent Variable	0.604	
Number of products	154	
Size of Panel Data	3,880	
Number of Specifications Data	21	
Volume of Total Data	93,120	

Notes: Values in ( ) indicate standard errors.
***, **, * denote significance at the 0.1%, 1%, 5% level.

(4) Rice cookers

Dependent Variable: log(average price)		
Intercept	8.217 (0.083)	***
Power Consumption (Wh)	-0.004 (0.001)	***
Thickness of Inner Pot (mm)	0.127 (0.010)	***
Weight (kg)	0.305 (0.018)	***
Dummy Variables		
Type		
IH Rice Cooker	0.713 (0.058)	***
Pressure IH Rice Cooker	0.711 (0.079)	***
Steam Function	0.362 (0.068)	***
Steam Saving System	0.161 (0.052)	**
Manufacturers		
Manufacturer A	0.298 (0.073)	***
Manufacturer B	0.338 (0.060)	***
Manufacturer C	0.231 (0.089)	**
Manufacturer D	0.406 (0.055)	***
Manufacturer E	0.184 (0.060)	**
Manufacturer F	0.257 (0.057)	***
Elapsed Weeks		
2nd week	-0.044 (0.023)	
3rd week	-0.090 (0.024)	***
4th week	-0.122 (0.026)	***
5th week	-0.145 (0.026)	***
6th week	-0.167 (0.026)	***
7th week	-0.179 (0.027)	***
8th week	-0.201 (0.027)	***
9th week	-0.218 (0.027)	***
10th week	-0.233 (0.027)	***
11th week	-0.243 (0.026)	***
12th week	-0.255 (0.027)	***
13th week	-0.268 (0.026)	***
Adjusted R-squared	0.906	
Standard Error of Regression	0.227	
Mean of Dependent Variable	10.348	
Standard Deviation of Dependent Variable	0.741	
Number of products	191	
Size of Panel Data	7,349	
Number of Specifications Data	19	
Volume of Total Data	161,678	

Notes: Values in ( ) indicate standard errors.
***, **, * denote significance at the 0.1%, 1%, 5% level.

Estimation Results of Hedonic Regression: Home Electrical Appliances [3]

(5) Vacuum cleaners

Dependent Variable: log(average price)		
Intercept	11.684 (0.577)	***
Suction Power (W)	-0.001 (0.000)	***
Noise Level (dB)	-0.042 (0.007)	***
Weight (kg)	0.130 (0.058)	*
Dummy Variables		
Cordless Device	0.663 (0.156)	***
Manufacturers		
Manufacturer A	0.826 (0.249)	***
Manufacturer B	2.212 (0.175)	***
Manufacturer C	1.025 (0.127)	***
Manufacturer D	1.398 (0.168)	***
Manufacturer E	0.570 (0.157)	***
Manufacturer F	1.134 (0.210)	***
Manufacturer G	1.115 (0.207)	***
Manufacturer H	1.334 (0.162)	***
Manufacturer I	1.399 (0.153)	***
Manufacturer J	0.791 (0.127)	***
Manufacturer K	1.417 (0.153)	***
Manufacturer L	1.525 (0.138)	***
Manufacturer M	0.762 (0.159)	***
Elapsed Weeks		
2nd week	-0.112 (0.038)	**
3rd week	-0.137 (0.039)	***
4th week	-0.171 (0.040)	***
5th week	-0.210 (0.041)	***
6th week	-0.233 (0.041)	***
7th week	-0.263 (0.041)	***
8th week	-0.274 (0.041)	***
9th week	-0.287 (0.040)	***
10th week	-0.301 (0.041)	***
11th week	-0.318 (0.041)	***
12th week	-0.324 (0.041)	***
13th week	-0.329 (0.042)	***
Adjusted R-squared		0.741
Standard Error of Regression		0.324
Mean of Dependent Variable		9.989
Standard Deviation of Dependent Variable		0.638
Number of products		150
Size of Panel Data		5,302
Number of Specifications Data		20
Volume of Total Data		121,946

Notes: Values in ( ) indicate standard errors.

***, **, * denote significance at the 0.1%, 1%, 5% level.

(6) Microwaves

Dependent Variable: log(average price)		
Intercept	4.643 (0.260)	***
Maximum Output (W)	0.001 (0.000)	***
Height (mm)	0.013 (0.001)	***
Dummy Variables		
Type		
Microwave Oven	0.246 (0.087)	**
Weight Sensor System	0.306 (0.076)	***
Flat Table	0.170 (0.076)	*
Manufacturers		
Manufacturer A	0.718 (0.100)	***
Manufacturer B	0.330 (0.066)	***
Manufacturer C	0.255 (0.060)	***
Manufacturer D	0.264 (0.063)	***
Manufacturer E	0.341 (0.157)	*
Elapsed Weeks		
2nd week	-0.028 (0.017)	
3rd week	-0.069 (0.018)	***
4th week	-0.118 (0.021)	***
5th week	-0.149 (0.023)	***
6th week	-0.167 (0.021)	***
7th week	-0.195 (0.023)	***
8th week	-0.215 (0.023)	***
9th week	-0.230 (0.023)	***
10th week	-0.203 (0.035)	***
11th week	-0.172 (0.043)	***
12th week	-0.180 (0.041)	***
13th week	-0.187 (0.041)	***
Adjusted R-squared		0.914
Standard Error of Regression		0.259
Mean of Dependent Variable		10.323
Standard Deviation of Dependent Variable		0.886
Number of products		140
Size of Panel Data		4,847
Number of Specifications Data		23
Volume of Total Data		126,022

Notes: Values in ( ) indicate standard errors.

***, **, * denote significance at the 0.1%, 1%, 5% level.

Estimation Results of Hedonic Regression: Home Electrical Appliances [4]

(7) Hair dryers and curling irons

Dependent Variable: log(average price)		
Intercept	6.345 (0.227)	***
Hot Air Temperature (degree)	0.005 (0.001)	***
Weight (g)	0.003 (0.000)	***
Dummy Variables		
Manufacturers		
Manufacturer A	0.700 (0.173)	***
Manufacturer B	1.326 (0.151)	***
Manufacturer C	0.990 (0.136)	***
Manufacturer D	1.032 (0.107)	***
Manufacturer E	0.685 (0.099)	***
Manufacturer F	0.530 (0.156)	***
Manufacturer G	0.316 (0.091)	***
Manufacturer H	0.525 (0.103)	***
Manufacturer I	0.277 (0.061)	***
Manufacturer J	0.618 (0.112)	***
Manufacturer K	1.324 (0.135)	***
Manufacturer L	0.168 (0.073)	*
Manufacturer M	0.768 (0.076)	***
Manufacturer N	0.305 (0.085)	***
Manufacturer O	0.594 (0.078)	***
Manufacturer P	1.212 (0.070)	***
Elapsed Weeks		
2nd week	0.056 (0.025)	*
3rd week	0.063 (0.030)	*
4th week	0.031 (0.031)	
5th week	0.006 (0.031)	
6th week	-0.003 (0.032)	
7th week	-0.019 (0.032)	
8th week	-0.047 (0.033)	
9th week	-0.054 (0.034)	
10th week	-0.060 (0.035)	
11th week	-0.061 (0.037)	
12th week	-0.071 (0.038)	
13th week	-0.071 (0.038)	
Adjusted R-squared		0.675
Standard Error of Regression		0.350
Mean of Dependent Variable		8.437
Standard Deviation of Dependent Variable		0.614
Number of products		203
Size of Panel Data		7,314
Number of Specifications Data		8
Volume of Total Data		80,454

Notes: Values in ( ) indicate standard errors.
***, **, * denote significance at the 0.1%, 1%, 5% level.

(8) Air purifiers

Dependent Variable: log(average price)		
Intercept	8.596 (0.249)	***
Effective Floor Area (mat)	0.018 (0.003)	***
Height (mm)	0.001 (0.000)	**
Dummy Variables		
Humidification Function	0.225 (0.042)	***
Dehumidifying Function	0.683 (0.048)	***
Deodorizing Function	0.213 (0.055)	***
Wall Mount Function	0.973 (0.183)	***
Automatic Power Saving System	0.405 (0.053)	***
Concentrated Ion Generating Function	0.346 (0.040)	***
Automatic Cleaning System	0.364 (0.079)	***
Manufacturers		
Manufacturer A	0.389 (0.081)	***
Manufacturer B	1.262 (0.064)	***
Manufacturer C	0.964 (0.202)	***
Manufacturer D	0.473 (0.059)	***
Manufacturer E	0.300 (0.073)	***
Manufacturer F	0.586 (0.075)	***
Manufacturer G	0.695 (0.129)	***
Manufacturer H	0.208 (0.053)	***
Manufacturer I	0.393 (0.063)	***
Manufacturer J	0.631 (0.067)	***
Elapsed Weeks		
2nd week	0.012 (0.019)	
3rd week	0.002 (0.022)	
4th week	-0.015 (0.023)	
5th week	-0.030 (0.026)	
6th week	-0.038 (0.032)	
7th week	-0.051 (0.033)	
8th week	-0.060 (0.033)	
9th week	-0.080 (0.042)	
10th week	-0.088 (0.042)	*
11th week	-0.085 (0.045)	
12th week	-0.091 (0.046)	*
13th week	-0.110 (0.045)	*
Adjusted R-squared		0.914
Standard Error of Regression		0.149
Mean of Dependent Variable		10.573
Standard Deviation of Dependent Variable		0.507
Number of products		103
Size of Panel Data		3,291
Number of Specifications Data		32
Volume of Total Data		115,185

Notes: Values in ( ) indicate standard errors.
***, **, * denote significance at the 0.1%, 1%, 5% level.

Estimation Results of Hedonic Regression: Digital Consumer Electronics [1]

(1) GPS navigations

Dependent Variable: log(average price)		
Intercept	8.058 (0.174)	***
Screen Size (inch)	0.331 (0.019)	***
Dummy Variables		
Recording Medium Type		
HDD	0.413 (0.080)	***
SSD	0.181 (0.063)	**
Rear Monitor Device	0.405 (0.028)	***
Terrestrial Digital Tuner	0.624 (0.084)	***
Vehicle Information and Communication System	0.232 (0.045)	***
Blu-ray Disk Device	0.491 (0.073)	***
Voice Recognition System	0.160 (0.036)	***
High-resolution Audio Device	0.428 (0.051)	***
Manufacturers		
Manufacturer A	0.589 (0.059)	***
Manufacturer B	0.344 (0.122)	**
Manufacturer C	0.202 (0.071)	**
Manufacturer D	0.797 (0.123)	***
Manufacturer E	0.585 (0.127)	***
Manufacturer F	0.825 (0.123)	***
Manufacturer G	0.730 (0.139)	***
Manufacturer H	0.273 (0.067)	***
Manufacturer I	0.384 (0.073)	***
Manufacturer J	0.744 (0.120)	***
Elapsed Weeks		
2nd week	-0.045 (0.011)	***
3rd week	-0.074 (0.015)	***
4th week	-0.096 (0.016)	***
5th week	-0.112 (0.016)	***
6th week	-0.128 (0.016)	***
7th week	-0.147 (0.016)	***
8th week	-0.160 (0.017)	***
9th week	-0.162 (0.017)	***
10th week	-0.171 (0.017)	***
11th week	-0.174 (0.017)	***
12th week	-0.181 (0.017)	***
13th week	-0.176 (0.017)	***
Adjusted R-squared		0.896
Standard Error of Regression		0.184
Mean of Dependent Variable		11.418
Standard Deviation of Dependent Variable		0.571
Number of products		152
Size of Panel Data		4,891
Number of Specifications Data		30
Volume of Total Data		161,403

Notes: Values in ( ) indicate standard errors.
***, **, * denote significance at the 0.1%, 1%, 5% level.

(2) External hard drives

Dependent Variable: log(average price)		
Intercept	8.961 (0.078)	***
Memory Capacity (TB)	0.174 (0.000)	***
Dummy Variables		
Cooling Fan Device	0.263 (0.056)	***
IEEE1394b	0.674 (0.069)	***
LAN	0.553 (0.155)	***
Thunderbolt	0.821 (0.123)	***
Manufacturers		
Manufacturer A	0.187 (0.045)	***
Manufacturer B	0.164 (0.041)	***
Manufacturer C	0.252 (0.102)	*
Manufacturer D	0.135 (0.061)	*
Manufacturer E	0.191 (0.070)	**
Elapsed Weeks		
2nd week	-0.006 (0.014)	
3rd week	-0.004 (0.015)	
4th week	-0.010 (0.015)	
5th week	-0.009 (0.015)	
6th week	-0.013 (0.015)	
7th week	-0.018 (0.016)	
8th week	-0.019 (0.017)	
9th week	-0.022 (0.017)	
10th week	-0.029 (0.017)	
11th week	-0.039 (0.018)	*
12th week	-0.047 (0.018)	**
13th week	-0.054 (0.018)	**
Adjusted R-squared		0.850
Standard Error of Regression		0.260
Mean of Dependent Variable		9.727
Standard Deviation of Dependent Variable		0.671
Number of products		303
Size of Panel Data		10,908
Number of Specifications Data		13
Volume of Total Data		174,528

Notes: Values in ( ) indicate standard errors.
***, **, * denote significance at the 0.1%, 1%, 5% level.

Estimation Results of Hedonic Regression: Digital Consumer Electronics [2]

(3) LCD TVs

Dependent Variable: log(average price)		
Intercept	9.327 (0.048)	***
Screen Size (inch)	0.034 (0.001)	***
Pixel Number (million pixels)	0.059 (0.000)	***
Dummy Variables		
IPS system	0.123 (0.035)	***
3D Television	0.124 (0.028)	***
Screen Split Display System	0.105 (0.035)	**
Speed Converting Circuit		
4 times	0.141 (0.033)	***
16 times	0.271 (0.079)	***
20 times	0.562 (0.068)	***
Digital Tuner 9 Channels	0.195 (0.041)	***
Internal Blu-ray Function	0.550 (0.054)	***
HDMI 4 terminals	0.148 (0.031)	***
ARC Function	0.084 (0.032)	**
Manufacturers		
Manufacturer A	0.268 (0.030)	***
Manufacturer B	0.155 (0.023)	***
Manufacturer C	0.181 (0.045)	***
Manufacturer D	0.161 (0.035)	***
Manufacturer E	0.688 (0.068)	***
Manufacturer F	0.217 (0.055)	***
Manufacturer G	0.486 (0.065)	***
Manufacturer H	0.297 (0.056)	***
Manufacturer I	0.406 (0.047)	***
Manufacturer J	0.300 (0.042)	***
Manufacturer K	0.282 (0.055)	***
Manufacturer L	0.323 (0.041)	***
Elapsed Weeks		
2nd week	-0.049 (0.009)	***
3rd week	-0.080 (0.010)	***
4th week	-0.109 (0.011)	***
5th week	-0.135 (0.011)	***
6th week	-0.162 (0.011)	***
7th week	-0.178 (0.011)	***
8th week	-0.199 (0.012)	***
9th week	-0.216 (0.012)	***
10th week	-0.231 (0.012)	***
11th week	-0.242 (0.012)	***
12th week	-0.249 (0.013)	***
13th week	-0.259 (0.013)	***
Adjusted R-squared		0.981
Standard Error of Regression		0.120
Mean of Dependent Variable		11.605
Standard Deviation of Dependent Variable		0.872
Number of products		188
Size of Panel Data		6,666
Number of Specifications Data		39
Volume of Total Data		279,972

Notes: Values in ( ) indicate standard errors.
***, **, * denote significance at the 0.1%, 1%, 5% level.

(4) LCD monitors

Dependent Variable: log(average price)		
Intercept	6.690 (0.209)	***
Screen Size (inch)	0.061 (0.007)	***
Resolution (dpi)	0.000 (0.000)	***
Response Speed (ms)	0.038 (0.007)	***
Luminance (cd/m2)	0.004 (0.001)	***
Dummy Variables		
Monitor Type		
Square	0.379 (0.098)	***
3D Function	0.433 (0.111)	***
Micro USB	0.196 (0.072)	**
Panel Type		
AH-IPS	0.415 (0.065)	***
IPS	0.287 (0.073)	***
Touch Panel Function	0.805 (0.095)	***
USB Hub	0.260 (0.040)	***
Manufacturers		
Manufacturer A	0.233 (0.041)	***
Manufacturer B	0.234 (0.055)	***
Manufacturer C	0.245 (0.079)	**
Manufacturer D	0.576 (0.072)	***
Manufacturer E	0.182 (0.054)	***
Manufacturer F	0.444 (0.078)	***
Elapsed Weeks		
2nd week	-0.009 (0.010)	
3rd week	-0.015 (0.011)	
4th week	-0.023 (0.011)	*
5th week	-0.029 (0.015)	
6th week	-0.035 (0.015)	*
7th week	-0.039 (0.015)	**
8th week	-0.043 (0.015)	**
9th week	-0.042 (0.015)	**
10th week	-0.044 (0.016)	**
11th week	-0.049 (0.016)	**
12th week	-0.056 (0.017)	***
13th week	-0.053 (0.017)	**
Adjusted R-squared		0.907
Standard Error of Regression		0.234
Mean of Dependent Variable		10.604
Standard Deviation of Dependent Variable		0.769
Number of products		193
Size of Panel Data		6,566
Number of Specifications Data		46
Volume of Total Data		321,734

Notes: Values in ( ) indicate standard errors.
***, **, * denote significance at the 0.1%, 1%, 5% level.

Estimation Results of Hedonic Regression: Digital Consumer Electronics [3]

(5) Printers

Dependent Variable: log(average price)		
Intercept	6.794 (0.204)	***
Maximum Number of Layered Sheets (sheet)	0.001 (0.000)	***
Width (mm)	0.004 (0.001)	***
Depth (mm)	0.002 (0.000)	***
Dummy Variables		
Printer Type		
Color Laser	0.835 (0.105)	***
Monochrome Laser	0.862 (0.107)	***
Mobile Function	1.376 (0.086)	***
FAX Function	0.283 (0.054)	***
Direct Printing System	0.307 (0.075)	***
Label Printing System	0.257 (0.082)	**
Manufacturers		
Manufacturer A	0.626 (0.184)	***
Manufacturer B	0.386 (0.072)	***
Manufacturer C	0.745 (0.088)	***
Manufacturer D	0.545 (0.131)	***
Manufacturer E	0.572 (0.134)	***
Manufacturer F	0.428 (0.131)	***
Manufacturer G	0.571 (0.178)	**
Elapsed Weeks		
2nd week	-0.001 (0.013)	
3rd week	-0.010 (0.014)	
4th week	-0.008 (0.016)	
5th week	-0.012 (0.018)	
6th week	-0.017 (0.021)	
7th week	-0.023 (0.021)	
8th week	-0.037 (0.022)	
9th week	-0.047 (0.026)	
10th week	-0.044 (0.028)	
11th week	-0.048 (0.028)	
12th week	-0.050 (0.028)	
13th week	-0.060 (0.030)	*
Adjusted R-squared		0.826
Standard Error of Regression		0.403
Mean of Dependent Variable		10.605
Standard Deviation of Dependent Variable		0.965
Number of products		264
Size of Panel Data		9,983
Number of Specifications Data		32
Volume of Total Data		349,405

Notes: Values in ( ) indicate standard errors.
***, **, * denote significance at the 0.1%, 1%, 5% level.

(6) Blu-ray and DVD recorders

Dependent Variable: log(average price)		
Intercept	10.662 (0.047)	***
HDD Capacity (TB)	0.222 (0.000)	***
Simultaneously Recordable Number of Programs	0.117 (0.014)	***
Recording Capacity for a long time (times)	0.007 (0.002)	**
Dummy Variables		
Coaxial Digital Audio Output Terminal	1.127 (0.028)	***
Ultra HD Blu-ray Function	0.194 (0.036)	***
Manufacturers		
Manufacturer A	0.088 (0.032)	**
Manufacturer B	0.086 (0.033)	**
Elapsed Weeks		
2nd week	-0.013 (0.014)	
3rd week	-0.086 (0.015)	***
4th week	-0.118 (0.016)	***
5th week	-0.148 (0.017)	***
6th week	-0.185 (0.016)	***
7th week	-0.219 (0.016)	***
8th week	-0.235 (0.021)	***
9th week	-0.243 (0.021)	***
10th week	-0.248 (0.021)	***
11th week	-0.252 (0.021)	***
12th week	-0.262 (0.020)	***
13th week	-0.262 (0.019)	***
Adjusted R-squared		0.874
Standard Error of Regression		0.153
Mean of Dependent Variable		10.999
Standard Deviation of Dependent Variable		0.430
Number of products		90
Size of Panel Data		3,143
Number of Specifications Data		47
Volume of Total Data		157,150

Notes: Values in ( ) indicate standard errors.
***, **, * denote significance at the 0.1%, 1%, 5% level.

Estimation Results of Hedonic Regression: Digital Consumer Electronics [4]

(7) Headphones

Dependent Variable: log(average price)		
Intercept	5.150 (0.962)	***
Minimum Reproduction Frequency (Hz)	-0.040 (0.007)	***
Impedance (ohm)	0.002 (0.000)	***
Sound Pressure Sensitivity (dB)	0.026 (0.009)	**
Weight (g)	0.004 (0.001)	***
Dummy Variables		
Type		
Canal-type	0.504 (0.123)	***
Ear-hooking	0.832 (0.263)	**
Standard Plug Device	0.320 (0.117)	**
Noise Cancel System	0.497 (0.192)	**
High Resolution Function	1.121 (0.100)	***
Remote Control Cable Device	0.645 (0.097)	***
Wireless System	0.736 (0.125)	***
Manufacturers		
Manufacturer A	1.302 (0.308)	***
Manufacturer B	0.921 (0.149)	***
Manufacturer C	0.399 (0.108)	***
Manufacturer D	2.648 (0.110)	***
Manufacturer E	0.624 (0.152)	***
Manufacturer F	2.535 (0.153)	***
Manufacturer G	0.943 (0.127)	***
Manufacturer H	2.073 (0.150)	***
Manufacturer I	1.384 (0.156)	***
Manufacturer J	3.723 (0.236)	***
Elapsed Weeks		
2nd week	-0.015 (0.013)	
3rd week	-0.023 (0.014)	
4th week	-0.026 (0.015)	
5th week	-0.045 (0.017)	**
6th week	-0.054 (0.018)	**
7th week	-0.052 (0.019)	**
8th week	-0.044 (0.020)	*
9th week	-0.038 (0.023)	
10th week	-0.046 (0.023)	*
11th week	-0.061 (0.024)	**
12th week	-0.072 (0.025)	**
13th week	-0.073 (0.025)	**
Adjusted R-squared		0.818
Standard Error of Regression		0.516
Mean of Dependent Variable		8.880
Standard Deviation of Dependent Variable		1.210
Number of products		429
Size of Panel Data		15,186
Number of Specifications Data		23
Volume of Total Data		394,836

Notes: Values in ( ) indicate standard errors.
***, **, * denote significance at the 0.1%, 1%, 5% level.
Other 16 manufacturers are significance at the 0.1% level.

(8) Camcorders

Dependent Variable: log(average price)		
Intercept	8.801 (0.211)	***
Pixel Number (million pixels)	0.034 (0.004)	***
Photographable Time (minute)	0.004 (0.001)	***
Weight (g)	0.000 (0.000)	**
Dummy Variables		
Finder Device	0.507 (0.199)	*
AV Output Function	0.860 (0.098)	***
DC Input Funtion	0.781 (0.131)	***
Micro USB 2.0	0.198 (0.082)	*
Manufacturers		
Manufacturer A	0.634 (0.120)	***
Manufacturer B	0.322 (0.115)	**
Manufacturer C	0.654 (0.096)	***
Elapsed Weeks		
2nd week	-0.025 (0.023)	
3rd week	-0.043 (0.022)	
4th week	-0.057 (0.025)	*
5th week	-0.078 (0.027)	**
6th week	-0.096 (0.022)	***
7th week	-0.116 (0.021)	***
8th week	-0.140 (0.021)	***
9th week	-0.157 (0.019)	***
10th week	-0.172 (0.019)	***
11th week	-0.180 (0.020)	***
12th week	-0.211 (0.022)	***
13th week	-0.216 (0.023)	***
Adjusted R-squared		0.935
Standard Error of Regression		0.162
Mean of Dependent Variable		10.938
Standard Deviation of Dependent Variable		0.639
Number of products		51
Size of Panel Data		1,523
Number of Specifications Data		45
Volume of Total Data		73,104

Notes: Values in ( ) indicate standard errors.
***, **, * denote significance at the 0.1%, 1%, 5% level.

Estimation Results of Hedonic Regression: Digital Consumer Electronics [5]

(9) Laptops

Dependent Variable: log(average price)		
Intercept	9.215 (0.358)	***
Display Size (inch)	0.052 (0.021)	*
Resolution (dpi)	0.000 (0.000)	***
SSD Capacity (TB)	0.869 (0.000)	*
HDD Capacity (TB)	0.280 (0.000)	***
Revolution Speed (rpm)	0.000 (0.000)	***
Memory Capacity (GB)	0.014 (0.005)	**
Number of Memory Slot	0.150 (0.028)	***
Video Memory (MB)	0.000 (0.000)	***
Battery Drive Time (h)	0.018 (0.003)	***
Depth (mm)	-0.004 (0.001)	**
Dummy Variables		
Touch Panel Corresponding to Windows 8	0.088 (0.018)	***
CPU		
Core i3/2 Cores	0.177 (0.018)	***
Core i5/2 Cores	0.268 (0.024)	***
Core i7/2 Cores	0.413 (0.038)	***
Core i7/4 Cores	0.343 (0.028)	***
CD Drive	0.366 (0.054)	***
LAN System	0.191 (0.089)	*
Wi-Fi Direct System	0.212 (0.023)	***
WiDi System	0.064 (0.031)	*
Bluetooth System	0.071 (0.025)	**
3D Acceleration Sensor Device	0.138 (0.057)	*
Acceleration Sensor Device	0.194 (0.029)	***
OS		
Windows 10	0.312 (0.032)	***
Windows 7	0.085 (0.026)	**
Microsoft Office Integrated Software System	0.259 (0.018)	***
Manufacturers		
Manufacturer A	0.180 (0.026)	***
Manufacturer B	0.082 (0.036)	*
Manufacturer C	0.716 (0.062)	***
Manufacturer D	0.161 (0.031)	***
Elapsed Weeks		
2nd week	-0.021 (0.004)	***
3rd week	-0.031 (0.004)	***
4th week	-0.040 (0.005)	***
5th week	-0.036 (0.005)	***
6th week	-0.035 (0.006)	***
7th week	-0.033 (0.007)	***
8th week	-0.036 (0.007)	***
9th week	-0.045 (0.008)	***
10th week	-0.052 (0.008)	***
11th week	-0.053 (0.008)	***
12th week	-0.065 (0.008)	***
13th week	-0.071 (0.008)	***
Adjusted R-squared		0.882
Standard Error of Regression		0.152
Mean of Dependent Variable		11.422
Standard Deviation of Dependent Variable		0.443
Number of products		527
Size of Panel Data		14,716
Number of Specifications Data		66
Volume of Total Data		1,015,404

Notes: Values in ( ) indicate standard errors.
***, **, * denote significance at the 0.1%, 1%, 5% level.

(10) Desktops

Dependent Variable: log(average price)		
Intercept	9.694 (0.210)	***
CPU Frequency (GHz)	0.144 (0.032)	***
Memory Capacity (GB)	0.018 (0.006)	**
HDD Capacity (TB)	0.045 (0.000)	**
Screen Size (inch)	0.021 (0.008)	**
Resolution (dpi)	0.000 (0.000)	***
Dummy Variables		
Case Structure		
Integrated Liquid Crystal Display	0.149 (0.050)	**
Tower Type	0.102 (0.038)	**
CPU		
Core i3	0.121 (0.044)	**
Core i5	0.175 (0.030)	***
Core i7	0.182 (0.035)	***
DDR4 Memory System	0.160 (0.073)	*
Hybrid HDD System	0.466 (0.068)	***
Integrated Software System		
Office Home and Business 2013	0.232 (0.039)	***
Office Home and Business Premium	0.312 (0.043)	***
Office Personal 2013	0.248 (0.043)	***
Office Personal Premium	0.304 (0.053)	***
Touch Panel Corresponding to Windows 8	0.130 (0.024)	***
3D Function	0.184 (0.026)	***
4K Output Function	0.070 (0.027)	*
Manufacturers		
Manufacturer A	0.169 (0.046)	***
Manufacturer B	0.334 (0.035)	***
Manufacturer C	0.177 (0.055)	**
Manufacturer D	0.346 (0.031)	***
Manufacturer E	0.440 (0.042)	***
Manufacturer F	0.324 (0.030)	***
Elapsed Weeks		
2nd week	-0.008 (0.006)	
3rd week	-0.027 (0.007)	***
4th week	-0.028 (0.008)	***
5th week	-0.033 (0.009)	***
6th week	-0.043 (0.009)	***
7th week	-0.054 (0.009)	***
8th week	-0.064 (0.010)	***
9th week	-0.070 (0.011)	***
10th week	-0.083 (0.011)	***
11th week	-0.083 (0.012)	***
12th week	-0.103 (0.012)	***
13th week	-0.112 (0.013)	***
Adjusted R-squared		0.892
Standard Error of Regression		0.125
Mean of Dependent Variable		11.778
Standard Deviation of Dependent Variable		0.381
Number of products		213
Size of Panel Data		6,323
Number of Specifications Data		45
Volume of Total Data		303,504

Notes: Values in ( ) indicate standard errors.
***, **, * denote significance at the 0.1%, 1%, 5% level.

Estimation Results of Hedonic Regression: Digital Consumer Electronics [6]

(11) Point-and-shoot cameras

Dependent Variable: log(average price)		
Intercept	4.958 (0.303)	***
Waterproof Performance (m)	0.011 (0.002)	***
Internal Memory Capacity (MB)	0.000 (0.000)	***
Liquid Crystal Monitor Size (inch)	0.912 (0.116)	***
Finder (million pixels)	0.002 (0.000)	***
Weight (g)	0.001 (0.000)	***
Dummy Variables		
Mannual Focus Function	0.110 (0.042)	**
Consecutive Imaging Function	1.415 (0.092)	***
AF Automatic Tracking Function	0.340 (0.046)	***
Liquid Crystal Tilt Monitor	0.223 (0.035)	***
Touch Panel Function	0.106 (0.045)	*
Image Element CMOS Device	0.344 (0.052)	***
RAW Function	0.289 (0.044)	***
RAW(DNG) Function	1.106 (0.154)	***
Optical Media Device	0.590 (0.074)	***
Micro SDHC System	0.182 (0.082)	*
Memory Stick Duo Function	0.413 (0.047)	***
Manufacturers		
Manufacturer A	0.167 (0.048)	***
Manufacturer B	0.217 (0.063)	***
Manufacturer C	1.386 (0.119)	***
Manufacturer D	0.195 (0.047)	***
Manufacturer E	0.524 (0.069)	***
Elapsed Weeks		
2nd week	-0.015 (0.003)	***
3rd week	-0.028 (0.004)	***
4th week	-0.044 (0.004)	***
5th week	-0.055 (0.005)	***
6th week	-0.067 (0.005)	***
7th week	-0.076 (0.005)	***
8th week	-0.092 (0.007)	***
9th week	-0.104 (0.008)	***
10th week	-0.110 (0.009)	***
11th week	-0.120 (0.009)	***
12th week	-0.130 (0.010)	***
13th week	-0.135 (0.010)	***
Adjusted R-squared		0.952
Standard Error of Regression		0.156
Mean of Dependent Variable		10.193
Standard Deviation of Dependent Variable		0.715
Number of products		149
Size of Panel Data		5,206
Number of Specifications Data		80
Volume of Total Data		432,098

Notes: Values in ( ) indicate standard errors.

***, **, * denote significance at the 0.1%, 1%, 5% level.

(12) DSLR and mirrorless cameras

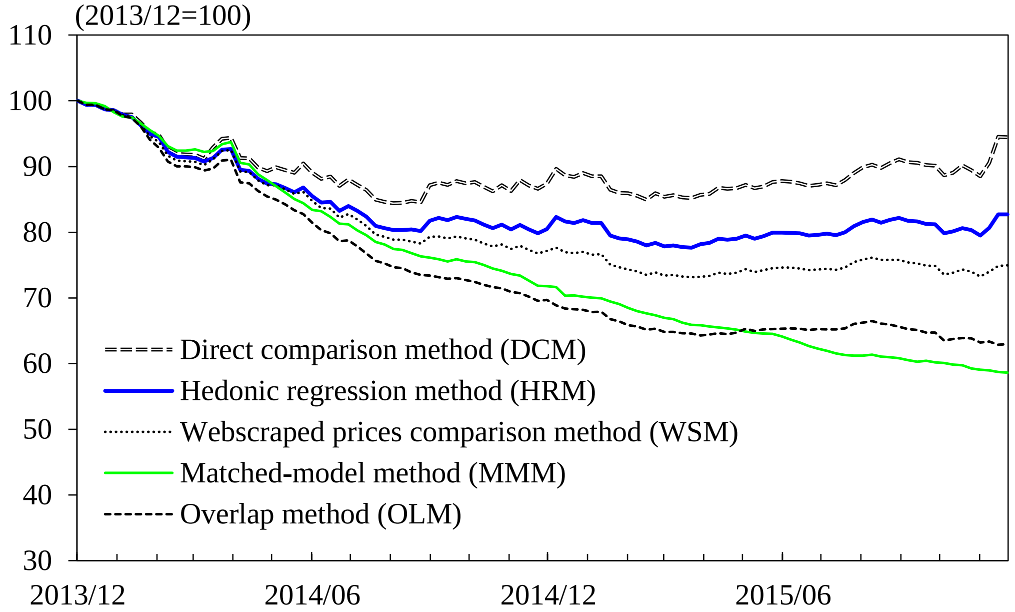
Dependent Variable: log(average price)		
Intercept	4.551 (0.997)	***
Pixel Number (million pixels)	0.019 (0.004)	***
Image Element (mm2)	0.001 (0.000)	***
Photographic Sensitivity (ISO)	0.000 (0.000)	***
Liquid Crystal Monitor Size (inch)	1.401 (0.316)	***
Finder Visual Field Ratio	0.002 (0.001)	***
Height (mm)	0.013 (0.002)	***
Movie Recording Pixel Number (million pixels)	0.054 (0.020)	**
Dummy Variables		
Micro SDHC System	0.290 (0.141)	*
Lap Time Measuring System	0.343 (0.068)	***
Lens Attachment Structure	0.224 (0.038)	***
Manufacturers		
Manufacturer A	0.387 (0.069)	***
Manufacturer B	0.827 (0.092)	***
Manufacturer C	0.458 (0.110)	***
Manufacturer D	0.408 (0.123)	***
Manufacturer E	0.141 (0.066)	*
Manufacturer F	0.487 (0.093)	***
Elapsed Weeks		
2nd week	-0.004 (0.005)	
3rd week	-0.008 (0.006)	
4th week	-0.019 (0.006)	**
5th week	-0.025 (0.007)	***
6th week	-0.031 (0.007)	***
7th week	-0.034 (0.008)	***
8th week	-0.040 (0.009)	***
9th week	-0.046 (0.011)	***
10th week	-0.052 (0.011)	***
11th week	-0.057 (0.011)	***
12th week	-0.057 (0.010)	***
13th week	-0.053 (0.012)	***
Adjusted R-squared		0.874
Standard Error of Regression		0.224
Mean of Dependent Variable		11.494
Standard Deviation of Dependent Variable		0.630
Number of products		138
Size of Panel Data		5,489
Number of Specifications Data		52
Volume of Total Data		301,895

Notes: Values in ( ) indicate standard errors.

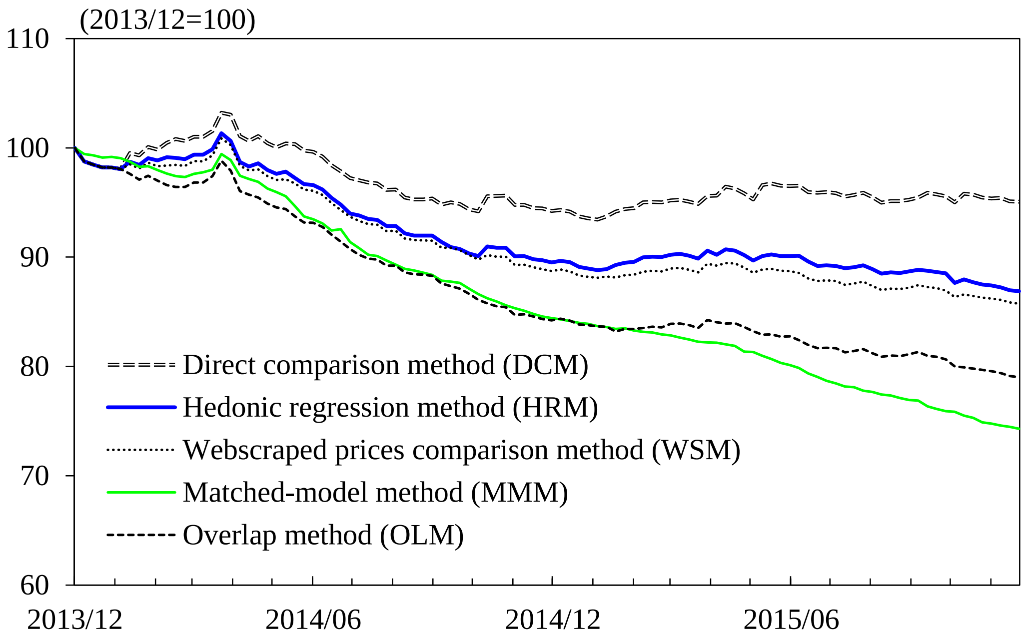
***, **, * denote significance at the 0.1%, 1%, 5% level.

Comparative Analysis of Experimental Price Indices: Overview

(1) Home Electrical Appliances (Total)

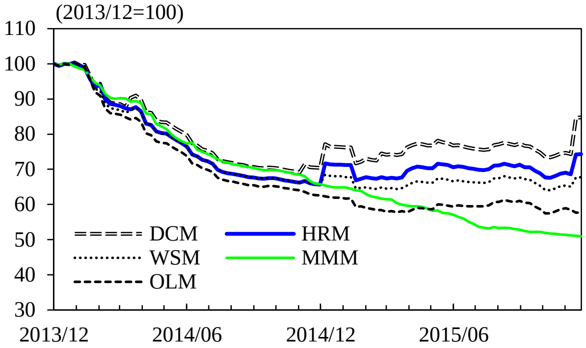


(2) Digital Consumer Electronics (Total)

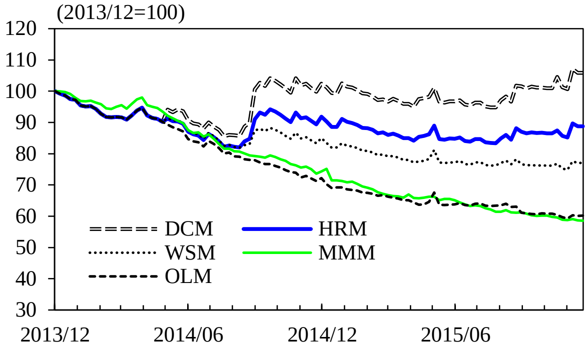


Comparative Analysis of Experimental Price Indices: Home Electrical Appliances

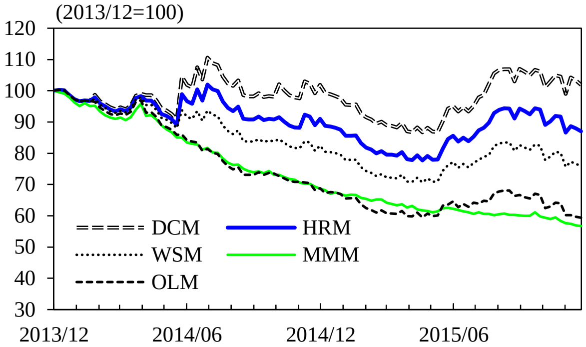
(1) Air conditioners



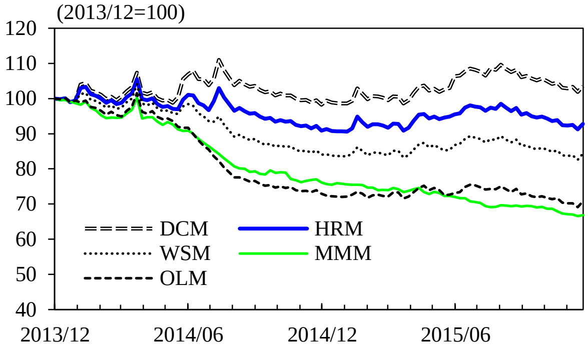
(2) Refrigerators and freezers



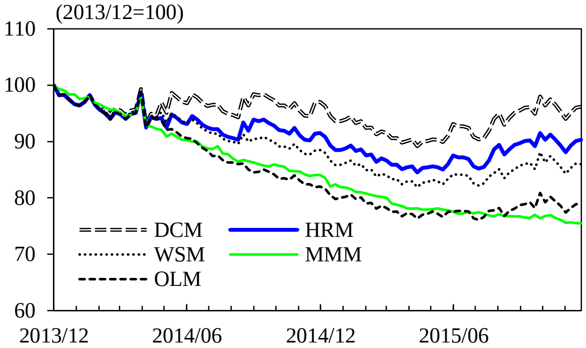
(3) Washers and dryers



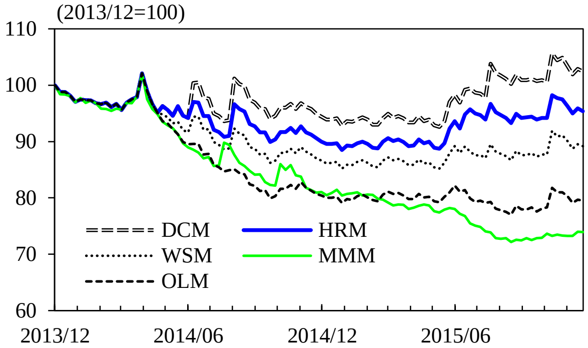
(4) Rice cookers



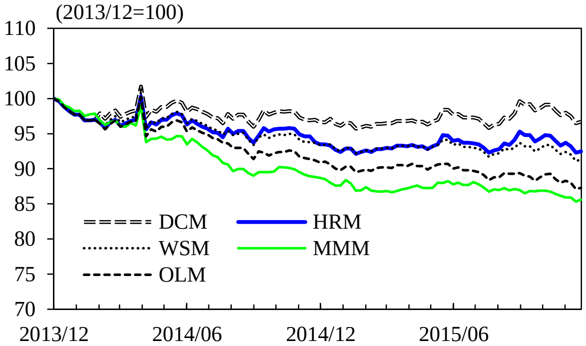
(5) Vacuum cleaners



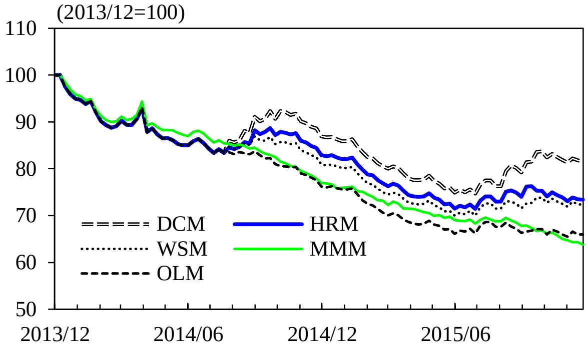
(6) Microwaves



(7) Hair dryers and curling irons

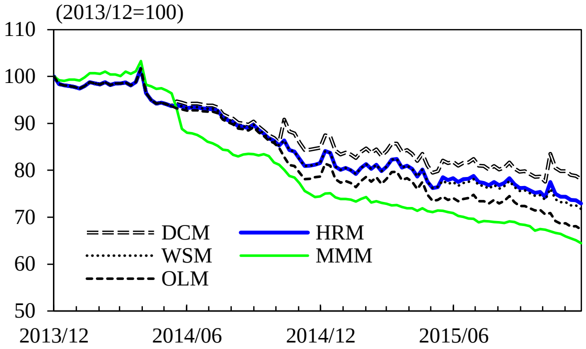


(8) Air purifiers

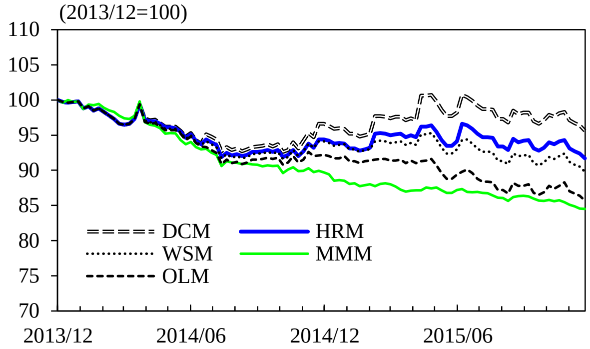


Comparative Analysis of Experimental Price Indices: Digital Consumer Electronics [1]

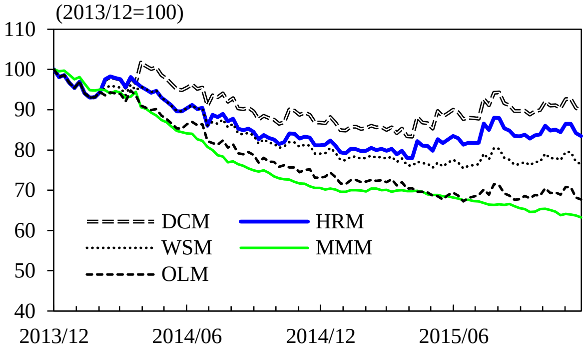
(1) GPS navigations



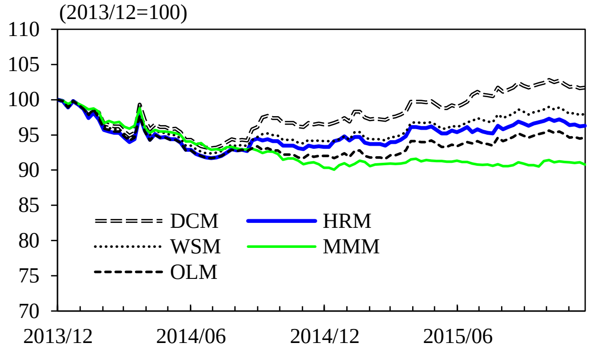
(2) External hard drives



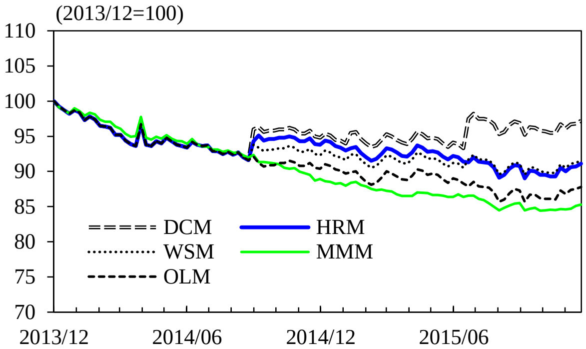
(3) LCD TVs



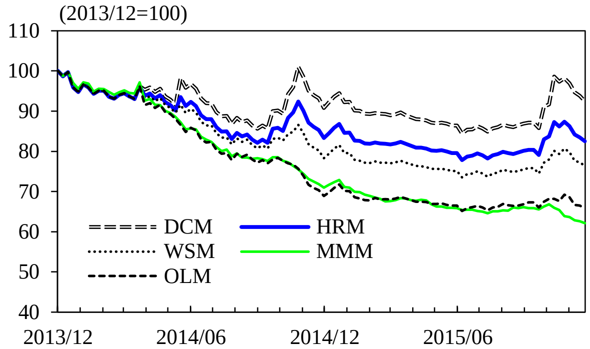
(4) LCD monitors



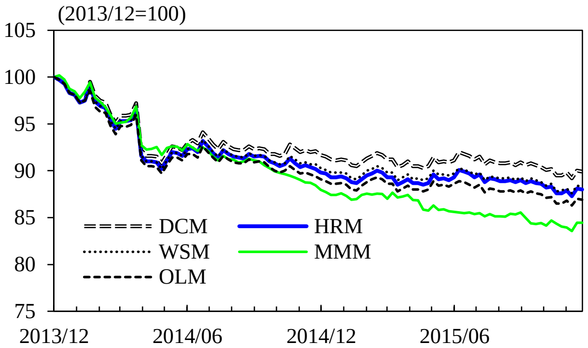
(5) Printers



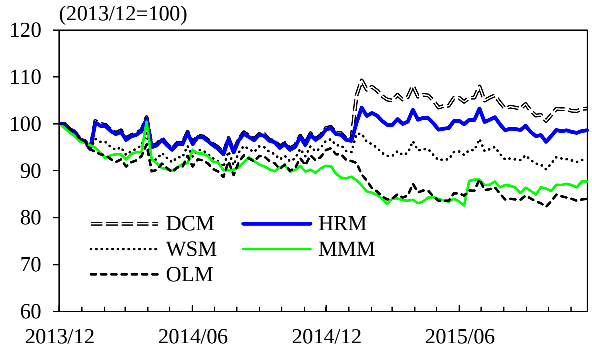
(6) Blu-ray and DVD recorders



(7) Headphones

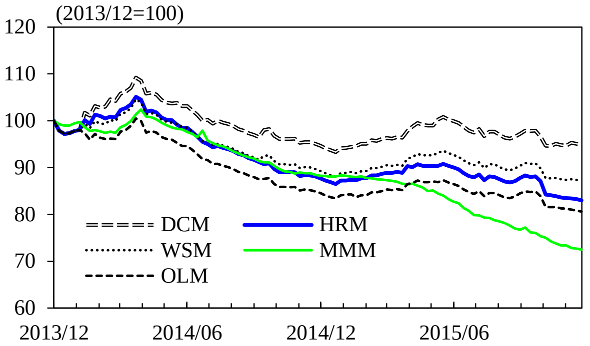


(8) Camcorders

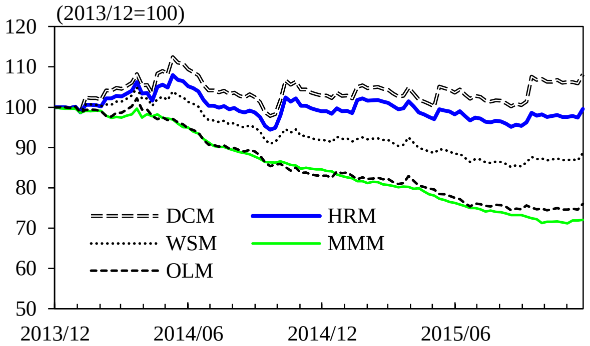


Comparative Analysis of Experimental Price Indices: Digital Consumer Electronics [2]

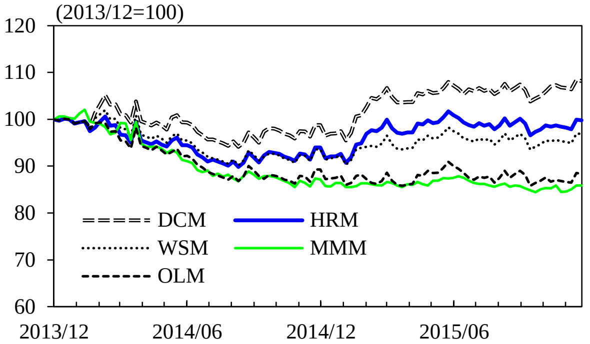
(9) Laptops



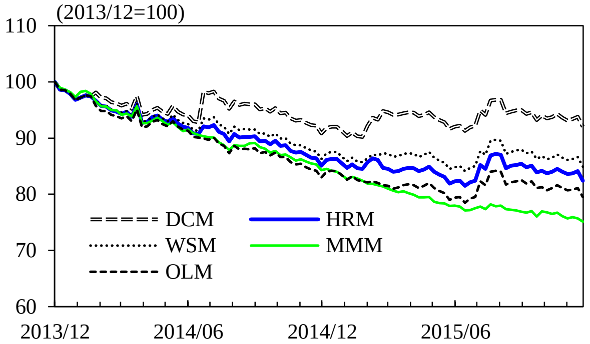
(10) Desktops



(11) Point-and-shoot cameras



(12) DSLR and mirrorless cameras



(Chart 10)

Comparison of Deviations between Indices Applied with HRM and the Others

	DCM		WSM		MMM		OLM	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Home Electrical Appliances	6.31	5.52	4.02 *	3.29 *	11.96	9.30	11.17	9.52
Air conditioners	4.75	4.21	2.63 *	1.91 *	9.77	7.14	7.58	6.21
Refrigerators and freezers	9.70	8.08	6.28 *	4.90 *	16.56	13.46	17.02	13.72
Washers and dryers	8.85	7.72	7.44 *	6.46 *	19.80	17.12	18.54	16.23
Rice cookers	7.65	6.84	6.84 *	6.08 *	17.88	15.87	17.54	15.73
Vacuum cleaners	4.42	3.97	2.54 *	2.11 *	7.64	6.43	7.61	6.48
Microwaves	4.25	3.53	4.09 *	3.41 *	12.47	9.96	10.35	8.73
Hair dryers and curling irons	2.97	2.70	0.63 *	0.46 *	5.13	4.57	3.07	2.55
Air purifiers	3.81	2.89	1.45 *	1.12 *	4.69	3.90	5.24	4.16
Digital Consumer Electronics	4.88	4.37	0.96 *	0.85 *	7.15	5.93	5.48	4.97
GPS navigations	2.83	2.23	0.44 *	0.23 *	6.39	5.86	2.79	2.12
External hard drives	2.52	1.92	1.11 *	0.76 *	5.36	4.27	3.60	2.66
LCD TVs	5.21	4.76	3.83 *	2.82 *	11.94	10.45	9.65	8.26
LCD monitors	3.38	2.93	0.94 *	0.81 *	3.40	2.84	1.40	1.13
Printers	3.08	2.00	0.80 *	0.58 *	4.09	3.36	2.67	2.11
Blu-ray and DVD recorders	6.07	5.20	3.87 *	3.20 *	11.56	9.68	11.14	9.42
Headphones	1.45	1.30	0.29 *	0.22 *	2.33	1.87	0.76	0.71
Camcorders	3.25 *	2.29 *	4.62	4.11	10.39	8.91	10.49	8.96
Laptops	7.27	6.62	1.93 *	1.59 *	5.25	3.65	3.46	3.32
Desktops	4.51 *	4.02 *	7.32	6.30	16.73	14.75	15.73	13.94
Point-and-shoot cameras	5.67	5.29	2.17 *	1.73 *	8.74	7.20	7.58	6.20
DSLR and mirrorless cameras	6.99	6.06	1.75 *	1.45 *	4.22	3.12	2.29	2.04

Note: Since HRM is a method to quantitatively estimate the impact of quality change on price, indices accuracy tends to be higher compared to other quality adjustment methods. We conduct periodic averaging using RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) on the deviation between indices applied with HRM and the other quality adjustment methods. The numbers with asterisk imply the smallest deviation from the results of HRM.

Outline of Support Vector Machine

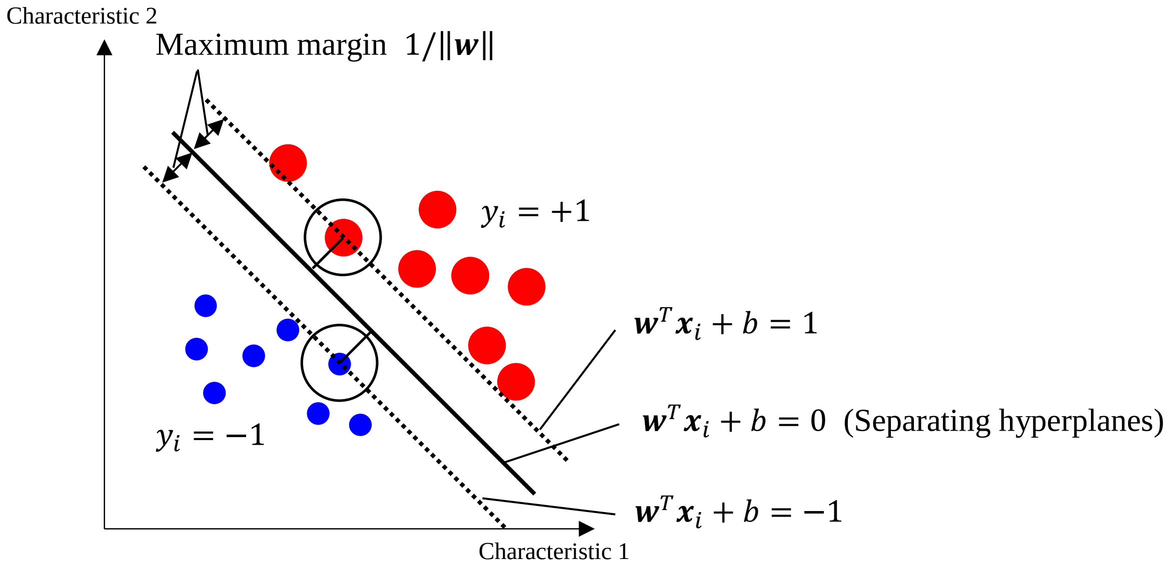
1. Derivation of Hard Margin SVM

Support Vector Machine (SVM) is a supervised machine learning algorithm which shows high performance in separating data into two classes (Cortes and Vapnik (1995)). This appendix explains the points in brief. Supervised data are defined as follows:

$\{(y_i, \boldsymbol{x}_i)\},\ \ i = 1, \ldots, N$	:	Set of supervised data with class label
$y_i \in \{+1, -1\}$	:	Class label[a]
$\boldsymbol{x}_i$	:	Characteristic vector[b]

In order to separate a class involving unknown data, boundary which divides the characteristic space into two, needs to be identified. SVM is a method which creates classifiers to solve a binary classification problem, by identifying the boundary (separating hyperplanes) which maximizes the Euclidean distance from the closest supervised data.

To make explanations simple, we assume 2-dimensional characteristic vectors and supervised data linearly classifiable (hard margin). The graph below shows the distribution of data on characteristic space in relation with the separating hyperplanes.



[a] For graphs in this appendix, label "+1" corresponds to old and new product pairs (old and new product pairs that seem to belong to the same manufacturer and same lineup) and label "-1" corresponds to irrelevant product pairs (pairs that cannot be regarded as old and new product pairs).

[b] In this analysis, supervised data is characterized with 3-dimensional characteristic vector of "Jaro-Winkler distance of product names", "Zone of product price" and "Product launch interval".

Class label y_i can be expressed as linear discriminant function of characteristics vector $\boldsymbol{x}_i$.

$$y_i = \mathrm{sign}(\boldsymbol{w}^T \boldsymbol{x}_i + b)$$

$\boldsymbol{w}$ is coefficient vector of separating hyperplanes. Function $\mathrm{sign}(u)$ is a sign function which takes 1 when $u > 0$ and -1 when $u \leq 0$. If supervised data is linearly classifiable, there is a parameter $\boldsymbol{w}$ and b which satisfy the following:

$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1, \qquad i = 1, \dots, N$$

The margin between separating hyperplane and closest neighboring supervised data is denoted as $1/\|\boldsymbol{w}\|$. Therefore, separating hyperplane derived will be the solution of the following minimization problem with inequality constraint conditions.

$$\min L(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T \boldsymbol{w} \left(= \frac{1}{2}\|\boldsymbol{w}\|^2 \right) \quad \mathrm{s.t.} \quad y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1$$

Replacing into dual problem using Lagrangean method of undermined multiplier $\boldsymbol{\lambda}$ $(\lambda_i \geq 0, i = 1, \dots, N)$

$$L(\boldsymbol{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{w}^T \boldsymbol{w} - \sum_{i=1}^{N} \lambda_i \{ y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1 \}$$

When L takes extreme value, the following first order condition is satisfied.

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{x}_i = 0, \qquad \frac{\partial L}{\partial b} = -\sum_{i=1}^{N} \lambda_i y_i = 0$$

Separating hyperplane is denoted using the optimal solution of the following dual problem including Kuhn-Tucker's complementary conditions.

$$\max L(\lambda_i) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \quad \mathrm{s.t.} \quad \sum_{i=1}^{N} \lambda_i y_i = 0, \quad \lambda_i \geq 0$$

Assuming S as set of supervised data (support vector) closest to separating hyperplane, optimal separating hyperplane obtained from optimal $\boldsymbol{w}^*$ and b^* is as follows:

$$y^* = \mathrm{sign}(\boldsymbol{w}^{*T}\boldsymbol{x} + b^*)$$

$$= \mathrm{sign}\left(\sum_{i\in S} \lambda_i^* y_i \boldsymbol{x}_i^T \boldsymbol{x} + b^*\right) \quad \text{where} \quad \boldsymbol{w}^* = \sum_{i\in S} \lambda_i^* y_i \boldsymbol{x}_i \qquad \blacksquare$$

2. Expansion to Soft Margin SVM

As previously stated, hard margin SVM assumes all supervised data to be linearly classifiable. However, in reality, it is extremely rare case for data to be linearly classifiable. Classifier conditions therefore needs to be loosened to account for practical matters, such as allowing some supervised data to cross the separating hyperplane to the opposite side of the boundary. SVM with these expansions are called *soft margin SVM*.

For data that entered the other side, Euclidean distance from separating hyperplane is expressed with ξ_i, and extent of allowance for misidentification is provided by hyperparameter C (cost parameter). Optimal margin for soft margin SVM can be formulated as follow:

$$\min L(\boldsymbol{w}, \boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} \xi_i \quad \mathrm{s.t.} \quad \xi_i \geq 0, \;\; t_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i$$

Dual expression using Lagrangean method of undermined multiplier $\boldsymbol{\lambda}\ (\lambda_i \geq 0, i = 1, \dots, N)$ and $\boldsymbol{\mu}\ (\mu_i \geq 0, i = 1, \dots, N)$:

$$L(\boldsymbol{w}, b, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \lambda_i\{y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - (1 - \xi_i)\} - \sum_{i=1}^{N} \mu_i \xi_i$$

Applying first order condition gives the following:

$$\max L(\lambda_i) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \quad \mathrm{s.t.} \quad \sum_{i=1}^{N} \lambda_i y_i = 0, \;\; C \geq \lambda_i \geq 0$$

Distinctive point of soft margin SVM is that the range derivable by optimal solution λ_i^* is limited by cost parameter C. $\blacksquare$

3. Expansion to Kernel Trick and Non-linear SVM

Soft margin SVM improves the practicability compared to hard margin SVM but is fundamentally difficult to derive separating hyperplanes from intricate class of supervised data. In such cases where identification boundary is non-linear, *kernel trick* method is applied. This method maps data to a higher dimensional space using kernel function where linear classification is conducted and inverse mapped to original space.

In case of non-linear conversion of characteristic vector $\boldsymbol{x}_i$ by mapping $\varphi(\boldsymbol{x}_i)$, it is easier to calculate the inner product of two characteristic vector mapping $\varphi(\boldsymbol{x}_i)$ and $\varphi(\boldsymbol{x}_j)$ using kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ instead of calculating the characteristic vector mapping individually. If general purpose Gaussian kernel (RBF kernel) is used as kernel function, inner product of characteristic vector mapping is denoted as follows:

$$\varphi(\boldsymbol{x}_i)^T \varphi(\boldsymbol{x}_j) \equiv k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(\frac{-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2}{2\sigma^2}\right)$$

σ is a hyperparameter (kernel parameter) to specify the extent of reflecting the complexity of the identification boundary to the separating hyperplane. Describing the previously stated optimal problem using kernel trick method, inner product of the preciously stated SVM is replaced by kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

$$\max L(\lambda_i) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \quad \text{s.t.} \quad \sum_{i=1}^{N} \lambda_i y_i = 0, \quad C \geq \lambda_i \geq 0$$

$$y^* = \operatorname{sign}\left(\sum_{i \in S} \lambda_i^* y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b^*\right) \qquad \blacksquare$$