ES/PR Rodica Bubuioc
PMU Fredrik Jonsson, Can Tongur

PM till nämnden för KPI
Sammanträde nr 7
2019-10-16

Clothing and footwear - Outlet and brand studies

For discussion

The inception of a new design idea for clothing was presented for the CPI Board at Statistics Sweden in 2018 (c.f. Norberg and Strandberg, 2018), raising insights concerning the contribution of outlet and brand dimensions when explaining the price development. With focus on the outlet and brand-dimensions, and consumer behavior, this memo introduces an extended study based on existing CPI data, as well as a study based on web-scraped data. The studies are part of the Eurostat grants project - "Outlet selection and consumer behavior".

Table of Contents

SCB

Background

Currently, field interviewers and central price collectors using outlet websites perform the price collection for the clothing and footwear surveys in the Swedish CPI. The current practice dates back to 1991, accompanied by a hedonic quality adjustment method for repricing, introduced in 1994. This method relies on information about the fabric, design, brand and other characteristics for clothing. Clothing and footwear, usually characterized by high churn rates, necessitate several substitutions during the year leading to a large variance in price quotes compared to other product groups, as can be concluded from earlier studies (see Norberg and Strandberg, 2018).

As for most consumption areas in the Swedish CPI, there is a two-stage sampling process for clothing and footwear. At the first stage, outlets are sampled and at the second stage, actual items are sampled in each store – with varieties chosen by price collectors. At present (2019), the Swedish CPI for clothing contains 27 product groups, with about 100 unique outlets and 4 300 products. For footwear, there are 11 product groups, with about 50 unique outlets and 1100 products. The weight for clothing and footwear together constitute approximately 5 percent of the total CPI basket.

The collection time devoted to successful registration of all observations, the meticulous work dedicated to finding new replacements (according to earlier established descriptions by CPI experts), and the recording of all new descriptive characteristics (fabric, design brand and others), result in a combination of different challenges that need to be overcome every month. Worthy mentioning as well, the time and effort devoted to the final review of the output, which is in itself another type of challenge.

Nowadays consumers increasingly choose other sales channels than the physical stores for the purchase of clothing and footwear (Carlsson and Strandberg, 2018). The digital economy, known for its fast growth and vast development, has changed the consumer behavior and the companies' price setting strategies. Based on the 2018 "E-barometern" report, clothing and footwear belong to the most popular online shopping items in Sweden. There are many reasons for this growth, where consumer high demand and competition are among the leading factors. For instance, the existence of different comparison websites allow the consumer to choose easily between outlets. This leads to an increase of the e-market positioning quota for all the companies. Nevertheless, the traditional physical store chains are an important complement in the development of e-commerce, as it allows the consumer to easily return or exchange goods, as well as try on various models and sizes.

In attempt to capture the digital economy, Statistics Sweden collects a great part of the prices from different chains' websites, as well as other online fashion platforms. The specific feature with the selected chains is the so-called "central pricing", where the price in a physical store does not differ from the price in an e-store for a specific product. In addition, other methods for price

measurement, such as web-scraped and scanner data are also of focus at Statistics Sweden.

In light with the digital economy, Statistics Sweden sees a need to review and remold the survey design for clothing and footwear. In 2018, a tentative study on clothing, which tested the relevance of brand sampling instead of outlet sampling (see Norberg and Strandberg, 2018), showed the following results: when it comes to clothing, the brand has a high degree of explanation of the price variable. The stores have a relatively low impact on the price, given the brand. The nine classes of brands used by Statistics Sweden today provide much lower explanatory rates than if every commonly used brand may be included in the model. Lastly, there is a correlation between the store group of large multi-store companies and their own brands.

In accordance with the Eurostat funded grants project (210549154: activity 1B) - “Outlet selection and consumer behavior”, Statistics Sweden proposed further studies in order to confirm/disconfirm the earlier achieved results by Norberg and Strandberg (2018), and continue to explore the relevance/necessity of outlet sampling. The grants project is running for a period of two years, 2019 and 2020, and focuses mainly on the study of two different product areas:

1) Clothing and footwear - 2019
2) Other durable goods (e.g. electronics) - 2020

Accordingly, this paper introduces studies done on the first set of products, clothing and footwear. For the second set of products, a report will be presented at the autumn board meeting (2020).

The memo incorporates two studies: in the first study, results from the preliminary findings in Norberg and Strandberg, (2018) are taken in consideration, following more rigorous analyses that are performed based on the existing CPI data; the second analysis introduces a pilot study designed on web-scraped data from a comparison website where analysis of three different products are presented.

Purpose

In this memo, two separate studies on clothing and footwear are presented. Also, previous study conclusions are assessed as well as an attempt to gain new insights about the consumer behavior concerning the outlet and the brand dimensions choice.

Empirical studies

The following empirical studies are divided into two parts: part one assesses manually collected CPI data and part two assesses web scraped data.

Part 1: Manually collected prices - brand vs. retailer

The analysis aims at assessing the influence of brand vis-à-vis outlet (retailer) on the price, for clothing data in the CPI. For this purpose, some delimitations of data are explained below which differ from the preceding study by Norberg & Strandberg (2018). Here, focus is on a certain type of brand - those found in multiple retailers. Retailers are considered as the "parent" company/chain and outlets are the specific store from which prices are collected for the CPI.

Data

The analysis relies on data collected by the CPI department over six years (2013-2018). The data covers merely clothing (and not shoes) for product groups according to the 4-digit Swedish classification which corresponds to COICOP sub-headings, shown in Appendix Table A1.

Approximately 265 000 price quotations were found in the collected data, which roughly implies some 3 600 observations every month. An additional limitation was made by excluding such brands that were rare (less than five unique product offers during the six years) given that the product offer had no other more frequent brands. Such rare-brand product offers constituted approximately 13 000 price quotations and some 250 000 price quotations remained after eliminating these.

Partitioning into brand category

Some categories of brands can be identified in the data, as listed:

1) First, there are well-known brands, often international but not always – some may be well-known national brands.
2) Second, there are retailer-specific brands that are either the only "brand" for the retailer or sometimes one of several brands (i.e. "lines"/ collections.).
3) Third, there are less frequent/less known brands, or so to say brands of somewhat an "ad-hoc" nature regarding presence in the basket. These are usually rare or even non-brands found e.g. in low-cost outlets or sometimes exclusive/specialized brands with however rare occurrence in the CPI.

The first challenge in this analysis has been to identify/allocate the existing brands into one of the three categories. This process was based on judgmental

identification and hence not perfectly exhausting the data regarding this splitting into categories. The most likely error in this manual/judgmental categorization process is that the third category may contain brands that should be assigned to the first category (rare/non-well known but in fact retailer-specific and known for the keen). The third category is for now of mere residual interest for analysis. However, the data amount in the first and the second categories are large.

First category
The first category was the easy/straightforward split of data and constituted half of all the existing price quotations (i.e. half of the 250 000) in the clothing survey data for CPI. This part of the sample was well-known/apparent cases of retailer specific brands as well as shops/retailers devoted to single brands.

Second category
The second category was somewhat not that intrinsic either, and less subject to judgmental calls in the operationalization here: we simply chose the most known brands to our knowledge in the data, adding the criteria that the brands existed in multiple stores/outlets and not merely in their own stores. This latter requirement was to eliminate such brands that, although being international/well-known, would have been inseparable regarding the analysis of brand versus outlet. This second category of easily identified very-well known brands, 25 distinct ones, constituted almost one third of the remaining data - almost 40 000 observations of the remaining 130 000 observations that were à priory identified not belonging to the first category.

Third category
The third category, which roughly accounted for the remaining 90 000 observations, was not elaborated on further but can be done so in the future, should this be of necessity for analytical purposes or to validate estimates and conclusions draw in this study.

Subset of interest in this study: second category

As previously discussed in regarding the brand-outlet (=retailer) symbiosis for some retailer, the first category of brands/outlets is ”endogenous” or so to say self-contained since corresponding brands are only sold by the retailer itself (often in various outlets belonging to the retailer). Hence this first category is not in focus here. The second category is the subject for analysis, and the third category is not in focus at present.

As we do not want to disclose any brand information, we provide a table (Table A1) with the most frequent well-known brands as identified for the second category (well known brands, found in multiple outlets). They are given in ordinal form (Brand 1, Brand 2,.., Brand N) with some process data from the data collection, ordered in descending number of occurrences of distinct retailer names (including Internet retailers).

Results

Variance analysis

A variance analysis is done by modeling known effects (brand and retailer) to explain variation in the regular price, analogous to the hedonic repricing approach employed in the CPI.

Modeling
We are aware of the complexity in the question of brand versus retailer influence on prices, as well as the dependencies between brands and outlets: they appear simultaneously through sampling & price collection, brand appears conditional on (randomly) sampled outlets.

Granularity
The employed approach is similar to the established hedonic method in clothing as used in the Swedish CPI for repricing purposes since mid 1990ies (Norberg, 1993). A technical difference applied here is that all products in clothing are included as effects by assigning dummy variables for each distinct product. This differs from the otherwise applied grouping into clusters of product groups, i.e. combinations of 4 or 5 digit elementary aggregates that correspond to COICOP sub-headings. This more-detailed modeling is a way of approaching/separating effects in the sub-groupings that exist within the product groups, as commented on in the rightmost column of Appendix Table A2.

Further, all unique brands and retailers (in the second category) are assigned dummy variables, i.e. taken as effects when explaining the price (in log scale). Outlets are grouped by their company name (as explained above, the "parent" retailer) whenever possible, which reduces dimensionality in modeling and appears intuitive, especially given the knowledge of "central pricing" for retail chains in clothing.

Delimitations/omitting observations
As a means of mitigating data redundancy, merely the first occurrence of each price quotation is subject for modeling. This is the employed standard approach for the annually estimated hedonic models for clothing in the CPI. This leaves less than 10 000 observations (of almost 40 000 available) in the second category brands/retailers to work with.

In effect, whenever price collectors have sampled an item for the first time or re-sampled the replacement item and it thus enters the basket, it would be included in the analysis data set (such items of the latter kind are flagged as a *change* in our systems). Due to annual basket updates, all items in December that were *not changes* would have been included by construction when assembling data for this study, rendering a repetition of some surviving items between baskets (overlap). Additionally, as price collection starts in October for the new basket in new stores since a few years back, as a means of counteracting entry bias in December from items with discount, "October-items" are sometimes cumbersome to identify à priori and would be included twice (in December as well).

As a remedy, these two issues were mitigated by exclusion of December data (the base period) from the analysis. As a consequence, one third of the remaining data points were omitted, still leaving more than 6 000 data points (price quotations) over the six years in the study. In total, 25 brands and 92 unique retailers are included, covering over 20 clothing product groups in the CPI, c.f. Appendix Table A2.

Data on both physical and internet stores

In Table 1, features as product code, brand and retailer are accounted, in order to explain variation in the price quotations (taken as the quoted *Regular Price* in log scale) for the second category of brand/retailers. The use of product code is because in many cases it entails information about the product type.

Table 1 Degree of explained variation (Adj. R2) in log(*Regular Price*) due to effects. Second category of brands in CPI Clothing, years 2013-2018.

Model	Adj. R2	Effects
1	0.8736	Product code, Brand
2	0.8502	Product code, Retailer
3	0.8729	Product code, Retailer, Brand

N.b. *Product code* refers to the identifier of products according to COICOP categories at lower levels (6-11 digits, given the first 4 from COICOP category). Period and Year are accounted for as effects in all models.

As can be understood from Table 1, the inclusion of brand (Model 1) indicates higher explanatory power than does Retailer (Model 2). Including both retailer and brand appear, in advance, to be the most exhausting way of determining the variations of the dependent variable (Model 3), but given the increase in model parameters, this does not indicate a greater success than using merely brand. The outcomes are indicatively in line with the previous analysis by Norberg & Strandberg (2018) reported in Appendix Table A3.

Sub-selection: internet store prices

In the manually collected data subject to the analysis here, some prices are due to internet pages of retailers (usually the larger ones) as well as some e-trade sites devoted merely to the internet (without physical stores). For the second category here, this data amounts to less than 2 000 price quotations (of the almost 40 000), rendering some 500 unique observations for modeling. This is less than one tenth of the available data for the whole second category when December (as base period) is eliminated. The corresponding variance analysis follows in Table 2a.

Table 2a Internet**:** Degree of explained variation (Adj. R2) in log(*Regular Price*) due to effects. Second category of brands, years 2013-2018.

Model	Adj. R2	Effects
1	0.6644	Product code, Brand
2	0.673	Product code, Retailer
3	0.6527	Product code, Retailer, Brand

N.b. *Product code* refers to the identifier of products according to COICOP categories at lower levels (6-11 digits, given the first 4 from COICOP). Period and Year are accounted for as effects in all models.

Observed in Table 2a, lower explanation of variance is noted. The explanation is the significant decrease in data and the high number of explanatory variables needed for the analysis. However, in this case, the retailer effect is a percentage point higher than for brand, although both explanatory factors are lower than in Table 1.

Sub-selection: non-internet store prices

The majority of price quotations in the second category were physical store prices, after subtracting the internet prices. For this sub-set of data, the corresponding analysis is given in Table 2b.

Table 2b Non-internet**:** Degree of explained variation (Adj. R2) in log(*Regular Price*) due to effects. Second category of brands, years 2013-2018.

Model	Adj. R2	Effects
1	0.8709	Product code, Brand
2	0.8473	Product code, Retailer
3	0,8703	Product code, Retailer, Brand

N.b. *Product code* refers to the identifier of products according to COICOP categories at lower levels (6-11 digits, given the first 4 from COICOP). Period and Year are accounted for as effects in all models.

Findings in Table 2b confirm the results of Table 1.

Part 2: Data from a price comparison web site

In order to have a broader understanding of the consumer's behavior when it comes to the outlet and the brand choice, Statistics Sweden has decided to web-scrape data from a well-known comparison web site in Sweden. The website offers a huge set of products where the consumers can apply comprehensive filters and easily find suitable products. Shop and product reviews are also available, which are of great help in the consumers' decision-making. Prices from all online shops are available and possible to compare, including product price history, popularity and other characteristics.

Data

The realization of the study involved the price scraping for three different products: sneakers, heels, and men's underwear. An elucidation for the product choice is the high fashion character that these products possess as well as the website's limited orientation toward the clothing area.

Using the programming language SAS, the products were scraped from the web site for a period of three months (July-Sept., 2019), where data was respectively scraped once every week and cleaned. July is known for having a high seasonality in the rate of sales, showing an impact on some of the results.

The collected data covered the top 500 products, filtered by popularity (as given by the price comparison web site) with several variables: id number, outlet name, price, product availability, model/brand, quantity price, material type, heel height (for heels), and others. Brand was extracted from a variable string containing information such as model.

Results

2.1 Outlet dimension

Heels
Most products in product area 2 (heels) were covered by 4 outlets, specializing in e-commerce. These 4 outlets covered on average 50-200 products each (among top 500 popularity) per week. The outlets had a rather notable difference in price level, as indicated by mean prices 1120:-, 675:-, 609:- and 579:- SEK, respectively. Table 5 below more broadly confirms this picture, showing large outlet-variability as compared to the within-outlet-variability (columns 3 and 4).

Men's underwear
The average product price per outlet varied between 500:- and 100:- SEK in product area 3 (men's underwear), and most outlets had offers covering that range. 23 out of 74 outlets covered on average less than 2 products (among top 500 popularity) per week, whereas 3 outlets (all specializing in e-commerce) covered on average more than 100 (top 500 popularity) products per week. These 3 outlets were rather similar in price level, as indicated by mean prices 296:-, 290:- and 290:- SEK respectively. More generally, Table 5 indicates less pricing variability between outlets, as compared to product areas 1 and 2.

Sneakers
In product area 1 (sneakers), 33 out of 93 outlets covered less than 2 products (top 500 popularity) on average per week, whereas 8 outlets covered on average more than 100 products (top 500 popularity) per week. These 8 outlets had mean prices varying from 853:- to 600:- SEK. As in product area 2, the between-outlet-variability was larger than the within-outlet-variability (cf. Table 5).

Table 5 Outlet pricing variability in collected data for selected product areas

Product area	No. of outlets	Avg. within outlet coeff. of var.	Coeff. var. betw. outlet means
Sneakers	93	24 %	41 %
Heels	23	27 %	47 %
Men's underwear	74	25 %	31 %

There was in all three product areas a conceivable but weak correlation between outlet mean price and outlet relative pricing (comparing prices to the mean price for a fixed product and point in time).

2.2 Brand dimension

Sneakers
Most offers in product area 1 (sneakers) were related to a top list of 5-10 brands, and in particular two well-known, global brands. These 10 brands were each represented in 100-600 product offers (combination of product and outlet) per week. Average prices among the top 10 brands varied between approximately 500:- to 1000:- SEK. All top brands covered that price range.

Heels
As to product area 2 (heels), a larger amount of available brands could be found in collected data, with mean brand prices ranging from approximately 5000:- to 200:- SEK, with less price variation within brands. Columns 3 and 4 in Table 6 confirm this picture, showing large between-brand-variability of prices as compared to the within-brand-variability. The top 5 brands, in terms of number of product offers (combination of product and outlet) in collected data, were each represented by approximately 25-50 product offers per week.

Men's underwear
Finally, average brand prices for product area 3 (men's underwear) varied between approximately 500:- to 200:- SEK. For most brands, product offers could be found within the entire price range. Table 6 indicates no more price variability between brands as within brands. The top 5 brands, in terms of available product offers in collected data, were each represented by approximately 100 product offers (combination of product and outlet) per week.

Table 6 Brand pricing variability in collected data for selected product areas

Product area	No. of brands	Avg. within brand coeff. of var.	Coeff. var. betw. brand means
Sneakers	64	25 %	59 %
Heels	117	17 %	87 %
Men's underwear	31	28 %	25 %

2.3 Using brand as a primary sampling dimension

The above two sections indicate a somewhat heterogeneous impact of brand and outlet on pricing in selected product areas. Nevertheless, the analysis does not contradict the hypothesis that the outlet dimension can be a weak predictor for pricing, and as such, acting largely through available brands and products.

Brand is a stronger predictor for pricing in some product areas (e.g. heels), while still relatively weak in other product areas (e.g. men's underwear). The potential gain in sampling efficiency of changing primary sampling dimension from outlet to brand could not be immediately assessed by the above analysis.

Discussion

Discussing retailer/outlet dimension vis-à-vis brand has rendered common insights from both data source studies (in-house data and web-scraped data), that to some extent confirm and reconfirm the earlier studies (see Norberg and Strandberg, 2018). When it comes to clothing and footwear, brand has a stronger degree of explanation than the outlet does, but at the same time outlet represents as well a strong predictor on the price. However, outlets are sampled randomly, whereas brands are sampled by interviewers according to what they find, hence this is not independent of the outlets they are assigned. Thus, one may suspect that the assortments are in many cases biased towards large brands and some selected varieties. This assortment selection is unexplored here.

Depending on product areas, the brand and the outlet have different degrees of predictability. There is a distinction between clothing and shoes such that clothing has more exhausting definitions, i.e. separation into product codes. However, to remark on the clothing product codes, they could be more stringently designed in order to serve as stratification for price collection. Perhaps this may reduce the variance contribution from interviewers thanks to narrower definitions. However, this may come at the risk of missing data as interviewers may not be able to encounter items due to too stringent specifications.

An experience concluded from web scraping is the importance of the site to be scraped regarding its content and precision in scraped variables. Especially consistency over time is important and relevance so that identical items can be followed without manual intervention or need for excessive data cleaning.

References

Norberg, A. (1993), “Förslag till reviderad metod för beräkning av prisutvecklingen för kläder i konsumentprisindex”, Report, Statistics Sweden.

Norberg, A. & Strandberg, K. (2018) ”Idé om ny design för KPI kläder”. PM till nämnden för KPI, sammanträde nr 5, 2018-10-17.

Carlsson, E. & Strandberg, K. (2018) ”KPI:s urval och prismätning i en digital ekonomi – nuläge, utmaningar och möjligheter”. PM till nämnden för KPI, sammanträde nr 5, 2018-10-17.

E-barometern. (2018) HUI, Postnord, Svensk Digital Handel

Appendix

Table A1 Top brands with frequencies in the second category in the analysis (Table 1). CPI data 2013-2018.

Brand	#Retailers	#Outlets
Brand 1	36	53
Brand 2	36	53
Brand 3	33	41
Brand 4	29	41
Brand 5	28	43
Brand 6	28	42
Brand 7	28	32
Brand 8	26	38
Brand 9	23	30
Brand 10	22	39
Brand 11	18	24
Brand 12	18	21
Brand 13	17	21
Brand 14	16	19
Brand 15	15	18
Brand 16	14	17
Brand 17	14	25
Brand 18	14	20
Brand 19	14	19
Brand 20	13	15
Brand 21	13	16
Brand 22	13	16
Brand 23	10	10
Brand 24	9	13
Brand 25	8	8

N.b. The brands 1-25 are all non-bounded to single retailers/outlets, i.e. occur in more than one retailer and store.

The corresponding product groups in Table A1 are given below in Table A2. It can be mentioned that practically all brands in Table A1 are “non-domestic” so to say, i.e. known internationally and not merely on the Swedish market.

Table A2 Four digit groupings (headings in line with COICOP) for the data in the second category subject to analysis in Table 1.

No.	Grouping name	Grouping code	Sub-categories*
1	W. underwear	3101	Yes
2	W. trousers	3108	No
3	W. skirt	3110	No
4	W. dress	3111	No
5	W. coat	3114	Yes
6	W. sweater	3118	Yes
7	W. outdoor jacket	3120	Yes
8	W. indoor jacket	3122	No
9	W. blouse	3123	Yes
10	Gloves	3202	Yes
11	M. underwear	3205	Yes
12	M. outdoor jacket	3206	No
13	M. sweater	3207	Yes
14	M. trousers	3210	Yes
15	M. indoor jacket	3212	No
16	M. coat	3214	Yes
17	Jeans	3217	Yes
18	Shirt	3218	Yes
19	M. leather jacket	3219	No
20	Ch. trousers	3304	No
21	Ch. sweater	3306	Yes
22	Ch. body	3307	No
23	Sportswear	3502	Yes

N.b. Abbreviations W. refers Women's, M. to Men's and Ch. to Children's.
N.b2. * Sub-categories may refer to either different products below this level, or even groupings on lower levels l i.e. 5 digits, or both.

Table A3 Results in previous study by Norberg & Strandberg (2018). Denoted "Tabell 4a" in the report (translated).
Regression models with data for Women's clothing 2016-2018. 28 brands that cover 70% of price quotations.

Explanatory variable	Adj.R2	Inflation
All	86.73	2.75
All except stores	86.71	2.79
All except physical attributes	84.46	2.85
9 brand groupings instead of 28 brands	82.45	1.06
All except brands	66.86	-0.85

As a twist to Table A1, the included retailers in the second category can be categorized according to their total number of outlets in the business register (in analogy with Norberg & Strandberg, 2018). The following Table A4 identifies this distribution over the six years of data.

Table A4 Distribution of included retailers by size category of retailer in Table 1

Number of outlets	**1**	**2-9**	**10-49**	**50 and more**
Share	59%	14%	19%	8%

N.b. Based on legal units' and local units' correspondence between the data and the business register.

The distribution presented in Table A4 can be compared with the corresponding Table A5 from Norberg & Strandberg (2018). As noted, since the present study covers the entire CPI Clothing and is not restricted to the largest brands merely within Women's clothing, some non-correspondence between tables is observed. A closer analysis showed the presence of several legal units here that had merely one local unit but still, judging from their retail name, would be considered as belonging to the categories of larger retailers.

Table A5 Distribution of women's clothes in the CPI according to brand frequency and retailer size, year 2016-2018. Caption from the study by Norberg & Strandberg (2018). Denoted "Tabell 3a" in the report, translated and reduced.

Brand frequency	**Single-outlet retailers**	**Retailers with 2-9 outlets**	**Retailers with 10-49 outlets**	**Retailers with > 50-outlets**
28 brands – 70% coverage	2.9	8.9	10.7	47.7
Other less common brands	12.8	2.5	5.5	9.0