Observations in web scraped data

The scraping process

Web scraping within this project has been performed through a SAS and JSON combined application developed by Statistics Sweden for this specific target web site. The application has been executed manually once a week at this well-known price comparison site. The web site, which has been subject to other studies preceding this (c.f. Swedish Competition Authority, 2019), offers a well-structured data source and comprises practically all relevant providers to the consumer market.

Related data collection and computation methods in the Swedish CPI

The four product categories in the web scraping study, Cell phones, Computers, Televisions and Washing machines are subject to slightly different but yet similar data collection methods – physical on site or in-house inline. Also, their quality adjustments methods differ in the Swedish CPI. For cell phones and computers, the current replacement bridging approach is by *implicit* quality adjustments through monthly chaining with resampling (c.f. the MCR method in the HICP Methodological Manual, 2018). On the other hand, for televisions and washing machines a fixed basket with replacements through *explicit* quality adjustments is undertaken to bridge the in- and outgoing items.

Data contents & page structure

Items on the web page in scraping are identified through a key ID number, a unique *model ID*. This is specific to such extent that each model, from some brand, is identified by certain specifications or characteristics. Storage capacity, for instance, is one such divisive tag that renders a unique *model ID* number between two otherwise identical items, whereas color does not. Also, the *popularity* measure as displayed on the web page relates to the item on that specificity and not the product as such in these cases.

One variable of importance which is non-accessible through web scraping is actual *sales volumes*. It is unfortunate that any study on mere offer prices, such as this based, will be highly restricted regarding inference unless actual purchases are available. Some studies disregard this rather significant information (e.g. Chessa & Griffioen, 2019) whereas in some studies (e.g. Jolivet & Turon, 2014) information on purchases is explicitly accounted for when modeling consumer behavior.

Description of data

Table 1 Aggregate descriptive information on the obtained data

Type	*Cell phones*	*Computers*	*Televisions*	*Washing machines*
Outlets	*155*	*83*	*124*	*53*
Brands	*108*	*21*	*59*	*47*
Models	*1 701*	*2 985*	*2 079*	*1 255*
Price range, SEK	*29 630*	*163 182*	*1 194 837*	*62 202*

As can be seen in Table 1, the number of outlets vary between product categories in the web scraping. This may indicate that consumer choice possibilities differ among different categories of products, which in turn may cause other influential factors regarding choice – e.g. transportation or installation and support, i.e. dependent on type of product.

Stock status:
The data provides information on current *stock status*, i.e. availability. This is on *product offer* level since it relates to the specific store's stock accounts. It may also serve as an indicator on *product level* on where consumers actually direct themselves, i.e. *price* versus *availability and price*. From a consumption perspective, temporarily non-available products cannot constitute consumption, whereas from a price index perspective this feature is more an open-ended question regarding the validity of prices of currently non-available items – this is the feature that distinguishes nominal/shelf-price data from actual transaction data and thus basket treatment thereof.

For analysis, the stock status can be interpreted either a dichotomous variable indicating availability (yes/no) or it can provide information on the number of available items in stock according to display on site. The distribution between available and non-available observations by product category is given in Table 2.

Table 2 Available and non-available items

Type	*Cell phones*	*Computers*	*Televisions*	*Washing machines*
Available	*73.9*	*75.1*	*73.0*	*67.5*
Non-available	*26.1*	*24.9*	*27.0*	*32.5*

Seen in Table 2, a rather substantial share of items, product offers, appear non-available at some time point. This invites to a more stringent analysis of the actual duration of non-availability should this be merely a temporary feature for some items and outlets or being perhaps a strategic or deliberate approach towards marketing. This is however not elaborated on further in this study.

Analysis

Duration: observable time on market

The overall time items are observable in the scraping is referred to as *duration*. This is a restricted measure due to the short time span of scraping and the obvious shortcoming that market entries are unknown (potentially detectable merely for a minor share of products), however the measure may still render some information.

Items (product offers or even products) can enter and leave the market at different times in different outlets and exist for some time. In some cases, items temporarily disappear and then return, i.e. they are discontinued on a priori unknown term and are then re-appearing. Such interrupted durations are referred to as *split* durations and this is a similar account as undertaken in the price collection for the CPI: product offers with non-available prices are indicated, signaled, and replaced should the discontinuity be of permanent nature.

In the following, the average duration in months, from first appearance to last, is computed for the web scraped data, with and without the account for split durations. The caveat is that the history of items prior to the collection start date is unknown, hence duration holds somewhat limited information and could not be remedied by merely counting new introductions due to the short time span in the study (less than 12 months).

As data collection is bounded to a specific time span, interpretations should be careful and results taken as indicative as it is a limited approach downwards, especially for long-lasting products due to the time window for web scraping –later appearing items' life spans are "shortened" by the truncation.

Two cases are given for each product category: the duration of product offers and the duration of products, displayed in Tables 3a, 3b and 4.

Table 3a Average duration in months on product offer level

Type	*Cell phones*	*Computers*	*Televisions*	*Washing machines*
Product offer observations	*13 004*	*11 623*	*11 791*	*4 328*
Average time, all	*4.1*	*3.1*	*3.7*	*6.3*
- Avg. time, non-split durations	*4.6*	*3.3*	*4.0*	*6.7*
- Avg. time, split durations	*2.7*	*2.2*	*2.54*	*3.6*
Share between split/non-split	*22.6 / 77.4*	*18.2 / 81.8*	*20.5 / 79.5*	*14 / 86*

N.b. Displayed current stock status of product offers has not been taken into account.

Table 3b Average duration in months on product offer level, merely available items w.r.t. stock status

Type	*Cell phones*	*Computers*	*Televisions*	*Washing machines*
Product offer observations	*11 553*	*10 725*	*9 807*	*3 848*
Average time, all	*3.7*	*2.8*	*3.4*	*5.2*
- Avg. time, non-split durations	*4.4*	*2.9*	*3.7*	*6.3*
- Avg. time, split durations	*2.7*	*2.2*	*2.6*	*3.4*
Share between split/non-split	*32.5 / 67.5*	*22.9/77.1*	*26.7/73.3*	*39.1/60.9*

When Tables 3a and 3b are compared, the decline in durations when availability is accounted for in Table 3b seems not very sharp. However, the share of division between split and non-split durations is altered to a larger extent indicating a tendency that some outlets have longer relative durations (non-split, continuous). Those appear more robust when offering items and hence preferable for sampling should the *product offer* levels be adhered to. Overall, it appears that products in the observed categories, when controlled for split durations and availability, have a rather short life average.

Table 4 Average duration (in months) on product level

Type	*Cell phones*	*Computers*	*Televisions*	*Washing machines*
Product observations	*2 131*	*3 478*	*2 489*	*1 321*
Average time	*5.3*	*3.6*	*4.7*	*6.2*
- Avg. time, non-split durations	*6.7*	*4*	*5.6*	*6.6*
- Avg. time, split durations	*2.9*	*2.5*	*2.8*	*2.9*
Share split/non-split	*36.6 / 63.4*	*25.8 / 74.2*	*30.4 / 69.6*	*9.7 / 90.3*

N.b. Displayed current stock status of underlying product offers has not been taken into account.

As expected, durations on *product* level (Table 4) are slightly longer than on *product offer* level. The corresponding distribution of duration for *products* over total number of months is given in Table 5.

Table 5 Duration distribution over months in the sample, *product* level

Type:	*Cell phones*		*Computers*		*Televisions*		*Washing machines*	
Months	***Count***	***Percent***	***Count***	***Percent***	***Count***	***Percent***	***Count***	***Percent***
1	*301*	*14.1*	*662*	*19.0*	*399*	*16.0*	*289*	*21.9*
2	*253*	*11.9*	*702*	*20.2*	*367*	*14.7*	*85*	*6.4*
3	*198*	*9.3*	*457*	*13.1*	*351*	*14.1*	*72*	*5.5*
4	*171*	*8.0*	*417*	*12.0*	*247*	*9.9*	*113*	*8.6*
5	*180*	*8.4*	*297*	*8.5*	*135*	*5.4*	*72*	*5.5*
6	*123*	*5.8*	*231*	*6.6*	*110*	*4.4*	*56*	*4.2*
7	*132*	*6.2*	*260*	*7.5*	*131*	*5.3*	*55*	*4.2*
8	*106*	*5.0*	*116*	*3.3*	*108*	*4.3*	*35*	*2.6*
9	*91*	*4.3*	*127*	*3.7*	*121*	*4.9*	*42*	*3.2*
10	*153*	*7.2*	*86*	*2.5*	*94*	*3.8*	*42*	*3.2*
11	*423*	*19.8*	*123*	*3.5*	*426*	*17.1*	*460*	*34.8*

N.b1. Split durations are included, hence higher concentration at fewer months.
N.b2. Displayed current stock status of underlying product offers has not been taken into account.

One conclusion that can be drawn from Table 5 is that duration is either short, especially due to splitting, or rather long (11 months). This may indicate that many items are “fresh” somewhere in-between endpoints, which however must be supported with purchase information to be made valid inference.

Market entries and exits during the year: “churn”

The presence of items, be it either on the level of product or product offer, between months and in relation to some specific base month (instead of non-bounded duration) can be considered through the “churn” measure. The “churn” is terminology well used for scanner data (c.f. Eurostat Practical Guide for Processing Supermarket Scanner data, September 2017) and addresses the number of items entering and leaving the market as observed at the index compilation instances (i.e. monthly). The measure describes basket similarity between adjacent months as well as basket attrition in relation to the base point in which the basket “starts”.

In Tables 6a and 6b, *product* churn is displayed with August 2018 as baseline period for the four product categories. Three measures are given:

1) products in the baseline basket that are still observable, expressed as a share of the current basket from scraping,

2) the inflow of new items as share of the current basket from scraping, and

3) the basket attrition, i.e. the currently observable baseline basket as share of the baseline basket (by definition 0 % attrition in base period).

Table 6a Churn/basket attrition, August 2019 as baseline. Cell phones and Computers

Type	***Cell phones***				***Computers***		
Month	**Share of base products**	**Inflow**	**Attrition**		**Share of base products**	**Inflow**	**Attrition**
0	*1.00*	*1.00*	*0.00*		*1.00*	*1.00*	*0.00*
1	*0.91*	*0.09*	*0.12*		*0.81*	*0.19*	*0.19*
2	*0.81*	*0.12*	*0.18*		*0.65*	*0.19*	*0.35*
3	*0.79*	*0.04*	*0.24*		*0.62*	*0.07*	*0.44*
4	*0.73*	*0.11*	*0.23*		*0.45*	*0.30*	*0.54*
5	*0.71*	*0.05*	*0.26*		*0.39*	*0.14*	*0.61*
6	*0.67*	*0.09*	*0.30*		*0.31*	*0.17*	*0.68*
7	*0.64*	*0.05*	*0.31*		*0.27*	*0.14*	*0.74*
8	*0.61*	*0.04*	*0.36*		*0.24*	*0.13*	*0.78*
9	*0.59*	*0.04*	*0.35*		*0.20*	*0.14*	*0.81*
10	*0.54*	*0.07*	*0.44*		*0.18*	*0.16*	*0.81*

Table 6b Churn/basket attrition, August 2019 as baseline. Televisions and Washing machines

Type	***Televisions***				***Washing machines***		
Month	**Share of base products**	**Inflow**	**Attrition**		**Share of base products**	**Inflow**	**Attrition**
0	*1.00*	*1.00*	*0.00*		*1.00*	*1.00*	*0.00*
1	*0.90*	*0.10*	*0.12*		*0.94*	*0.06*	*0.11*
2	*0.81*	*0.12*	*0.19*		*0.77*	*0.19*	*0.14*
3	*0.80*	*0.04*	*0.26*		*0.89*	*0.03*	*0.17*
4	*0.71*	*0.16*	*0.30*		*0.83*	*0.06*	*0.25*
5	*0.69*	*0.06*	*0.31*		*0.82*	*0.04*	*0.26*
6	*0.68*	*0.04*	*0.33*		*0.75*	*0.08*	*0.30*
7	*0.62*	*0.10*	*0.37*		*0.69*	*0.07*	*0.33*
8	*0.56*	*0.15*	*0.44*		*0.68*	*0.03*	*0.35*
9	*0.52*	*0.09*	*0.49*		*0.66*	*0.04*	*0.36*
10	*0.48*	*0.09*	*0.51*		*0.64*	*0.04*	*0.36*

When interpreting Tables 6a and 6b, it should be noted that since products may re-appear the attrition may decrease temporarily. It should also be noted, however, that results do not necessarily reflect a normal churn - the markets are affected by the ongoing economic crisis and this likely transfers to these numbers. One observation made from this period is that inflow appears to vary between product groups and accumulates as implied by the *duration*, with somewhat stable attrition rates with respect to the base basket.

Variation analysis

Multiple characteristics in data may contribute to explaining price/price variations. Starting with the product itself, the brand is likely to be highly relevant. Similarly, the specific type or model is likely to influence price. The outlet may also be influential on price. Two dependent variables could be analyzed; price *levels* or price *changes,* the latter being perhaps more relevant from a CPI point of view, c.f. Lach (2002) for an outline of this kind a model.

In order to obtain more robust interpretations, outlets that occurred less than 10 months (of 11 in total) were omitted from further analysis. No similar restriction was placed on product offer occurrences, hence they are fully included. Table 7 gives the share of omitted data prior to analysis.

Table 7 Filtering out of non-continuous outlets (=presence in sample is less than 10 months)

Type: observations	***Cell phones***	***Computers***	***Televisions***	***Washing machines***
Remaining obs.	*134 768*	*89 853*	*107 674*	*88 284*
Prior to filtering	*164 116*	*102 218*	*126 909*	*89 114*
Share omitted	*18.12%*	*12.1%*	*15.15%*	*0.93%*

To proceed, effects from product model (the specific variety), brand and outlet can be examined through a variation analysis. It may be motivated to group outlets after some characteristics. As one practical approach applied here, outlets that operate with both online and multiple physical outlet presence (and not merely single-store advertisers) were grouped in order to assert a possible effect from such a distinct feature: *well-known/well-established* & *multi-channel* marketers.

Regarding the *price* variable, some 100 observations were discarded from further analysis of *cell phones* data due to errors. No other data cleanup was necessary and it may be possible that such imperfections are already taken care of prior to marketing on the site.

In Tables 8 and 9, price changes and the price level are subject to the variation analysis as dependent variable (in log scale). Five models (sets of effects) were individually used as regressors to explain the variation in the dependent variable.

Table 8 Explained degree of variation (adjusted R2) in *price changes* due to five models

Type: effect	***Cell phones***	***Computers***	***Televisions***	***Washing machines***
Outlets	*0.074*	*0.074*	*0.135*	*0.071*
Brand	*0.096*	*0.035*	*0.118*	*0.065*
Brand &Dual outlets	*0.104*	*0.038*	*0.132*	*0.065*
Model	*0.608*	*0.525*	*0.684*	*0.565*
Model & Dual outlets	*0.616*	*0.527*	*0.69*	*0.565*

N.b. Dependent variable in logarithmic scale (natural). Time period (YYMM) is controlled for, as well as availability in stock (dichotomous 1/0) in all cases.
**Dual outlets* refers to the effect from grouping outlets into those with both physical and online stores.

Table 9 Explained degree of variation (adjusted R2) in *price* levels due to five models

Type: effect	***Cell phones***	***Computers***	***Televisions***	***Washing machines***
Outlets	*0.206*	*0.069*	*0.194*	*0.086*
Brand	*0.667*	*0.287*	*0.268*	*0.350*
Brand &Dual outlets	*0.667*	*0.290*	*0.27*	*0.352*
Model	*0.973*	*0.952*	*0.961*	*0.917*
*Model & Dual outlets**	*0.974*	*0.952*	*0.961*	*0.917*

N.b. Dependent variable in logarithmic scale (natural). Time period (YYYYMM) is controlled for, as well as availability in stock (dichotomous 1/0) in all cases.
**Dual outlets* refers to the effect from grouping outlets into those with both physical and online stores.

Interpreting Tables 8 and 9, *brand* is a stronger explanatory variable than is *outlet.* The distinction into *outlet type* does not provide much more explanatory power for explaining price variation. The product *model* has higher a degree of explanation than both brand and outlet, as expected, and once again the *outlet type* has not much more contribution.

Regarding the outlet dimension, a counter-question is on what terms outlets do differ – if not through price? Such non-observable or even "soft" facts may be present without being assessable in the numeric analysis undertaken here, i.e. there may be some subjectivity after all in the purchase decision by consumers. This topic has been subject to discussions, as found in e.g. Jolivet & Turon (2014) as a seller differentiation mechanism, but appears hard to explore further without actual purchase quantities and if the channel, a price comparison site, serves to assess this kind a question should there be purchase quantities.

Price spread and assortments

At each time point, the price of a product most often differs between outlets' offers. This has been a parameter of interest in several studies (c.f. Swedish Competition Authority, 2019). Such price spread is reported in Tables 10a and 10b by all product offers and separately as well for the dual outlet product offers. The average price spread is computed for product offers found in at least 5 outlets at each collection time point in the sample. Price spread is formulated as a coefficient of variation, i.e. the standard deviation as percent of the average price for each product.

Table 10a Price spread as mean or median *coefficient of variation (std. dev./avg. price)*, abbrev. ***c.v.***

Type: overall spread	*Cell phones*	*Computers*	*Televisions*	*Washing machines*
Mean c.v.	11.39	7.08	12	9.18
Median c.v.	9.90	6.55	11.45	7.48
Mean c.v., dual outlets	3.88	2.6	7.05	8.06
Median c.v., dual outlets	2.98	1.52	5.97	5.4

N.b1. Minimum requirement 5 observations (=outlets) per time point except for *Washing machines* in which the dual outlets were fewer on the price comparison site, thus accounted for unrestricted.
N.b2. Displayed current stock status of underlying product offers has not been taken into account.

Table 10b Price spread as mean or median *coefficient of variation (std. dev./avg. price)*, abbrev. ***c.v.*, **merely product offers that are available.

Type: overall spread	*Cell phones*	*Computers*	*Televisions*	*Washing machines*
Mean c.v.	10.15	6.54	11.51	9.23
Median c.v.	8.84	5.89	10.94	7.39
Mean c.v., dual outlets	4.15	2.54	5.84	7.35
Median c.v., dual outlets	3.34	1.4	2.38	4.16

N.b. Minimum requirement 5 observations (=outlets) per time point except for *Washing machines* in which the dual outlets were fewer on the price comparison site, thus accounted for unrestricted.

Displayed in Tables 10a and 10b, average price spread is in general smaller for the dual outlets, which indicates that perhaps assortments differ and are well-maintained, which adds to the inference from Tables 8 and 9 that variability is slightly less in general for these. This interprets perhaps such that well known outlets may have more similar discount patterns and thus similar and smaller spread as a common feature than the other smaller and more ad-hoc offering outlets.

One surprising finding is that for *cell phones*, the condition of availability renders an increases in the price spread, while for the three other product categories the price spread decreases in that circumstance.

In Table 11, the average number of brands and models are reported for the two outlet types (dual, non-dual).

Table 11 Assortments offered: average number of varieties per type of outlet (dual/non-dual)

Type: assortment by outlet type	**Cell phones**	**Computers**	**Televisions**	**Washing machines**
Brands, non-dual	7	4	5	8
Brands, dual	10	8	7	10
Models, non-dual	61	111	71	68
Models, dual	125	230	202	88

N.b. Merely available products regarding stock status are included.

It is seen from Table 11 that assortments in general are larger for the dual outlets, both regarding brands and models.

Relative price movements

One topic that may be interesting from a CPI point of view is the observed *relative* price changes. This can be assessed by examining the changes of prices for product offers with respect to their relative position to each other. The topic has been surveyed by Lach (2002) through an application to the residuals from an effects model analogous to the specification variance analysis in the preceding section.

The approach is through asserting the price an initial relative position and a one time point head final relative position. This can be seen as a transition matrix with one-step price movements (or non-movements, i.e. inertia) in relative terms in which relativity is effectuated by dividing the prices into quartiles at both initial and final time points. If this is taken as an average of all such one step movements during the sample (1->2, 2->3,..,t-1->t) over all product offers, the following transition matrices can be obtained for the products.

Table 12 Transition matrix, one-step movements (t->t+1), by product category

Type: Cell phones	***p25 (t+1)***	***p50 (t+1)***	***p75 (t+1)***	∞
p25 (t)	0.88	0.09	0.02	0.02
p50 (t)	0.10	0.82	0.06	0.02
p75 (t)	0.04	0.07	0.83	0.06
∞	0.03	0.04	0.07	0.86
Type: Computers	***p25 (t+1)***	***p50 (t+1)***	***p75 (t+1)***	∞
p25 (t)	0.87	0.09	0.03	0.01
p50 (t)	0.14	0.75	0.08	0.02
p75 (t)	0.07	0.08	0.79	0.06
∞	0.04	0.04	0.09	0.83
Type: Televisions	**p25 (t+1)**	**p50 (t+1)**	**p75 (t+1)**	∞
p25 (t)	0.92	0.05	0.02	0.01
p50 (t)	0.10	0.83	0.05	0.02
p75 (t)	0.06	0.07	0.81	0.06
∞	0.05	0.04	0.08	0.83
Type: Washing mach.	**p25 (t+1)**	**p50 (t+1)**	**p75 (t+1)**	∞
p25 (t)	0.84	0.09	0.04	0.02
p50 (t)	0.13	0.79	0.06	0.02
p75 (t)	0.07	0.07	0.80	0.05
∞	0.05	0.04	0.07	0.84

N.b1. The boundary *p25* indicates all prices up until the first quartile, *p50* indicates those above the first quartile and up until the median, *p75* indicates prices above the median up until the third quartile, and ∞ indicates the prices above the third quartile.
N.b2. Merely available product offer included. Analysis restricted to those products with price variation between simultaneous offers and in both time points (t, t+1).

It is seen in Table 12 that diagonal entries are high – indicating that relative positions do not change for some 80 percent of product offers. Also notable, diagonal endpoints, the first and the last entries, are always the highest/never exceeded by the two diagonal entries in between. This interprets such that low-priced offers remain low, whereas high-priced offers remain high, in relative terms. From a sampling point of view, a precautious interpretation could be that the outlets differs regarding their positioning on the market, but on the other hand are rather steady regarding their prices in 8 of 10 time points or 8 of 10 products (both may be interchangeably).

Conclusions

Duration of online product offers online can be rather short. This could be explored more whether this refers to some specific kind of variety within the product groups and if this transfers to all outlets and if it could be a feature to take in consideration when designing outlet sampling.

Some differences appear to be between outlets with mere online sales and the dual-channel outlets, the so called well-known retailers. A main conclusion to be drawn is that there appears no notable disadvantages in choosing such well known outlets to sample products from. On the contrary, those outlets offer less variable / more stable prices for their given assortments. Also, they can be more suitable for exhaustive sampling of *different* products, i.e. for sampling broader selections as they offer more products.

The limitation of not having purchase quantities does not allow for more elaborate conclusions regarding actual purchase decisions, i.e. the consumer choice. Additionally, although price differentiation may exist that leads to volatility, it is unclear if such actions render any broader outcomes (significant turnover increase) or if they are applied for stock clearance. So summarize, the underlying mechanisms cannot be assessed further here. As expected, regarding the variation analysis, the fixed effects accounted for do exhibit differences in explanatory power, with the product *model* being the most influential effect on price.

Regarding the *relative* price changes, the most notable observation from the transition matrices is that relative positions of prices most often do not change: lower prices tend to continue be low and vice versa. From a sampling point of view, this may indicate that some outlets could be exchangeable with each other.

References

Swedish Competition Authority (Konkurrensverket) (2019). *Prisspridning på e-handelsmarknader med låga sökkostnader.* Report in the series “Uppdragsforskning 2019:1”, published 2019-02-21. Retrievable from
http://www.konkurrensverket.se/publikationer/prisspridning-pa-e-handelsmarknader-med-laga-sokkostnader/

Chessa, A.G & Griffioen, R. (2019). Comparing Price Indices and Footwear for Scanner Data and Web Scraped Data, in *Economie et Statistique/Economics and Statistics* no. 509, 2019, pp. 49-68. Retrievable from
https://www.persee.fr/doc/estat_0336-1454_2019_num_509_1_10891

HICP Methodological Manual, Eurostat 2018. Retrievable from
https://ec.europa.eu/eurostat/documents/3859598/9479325/KS-GQ-17-015-EN-N.pdf/d5e63427-c588-479f-9b19-f4b4d698f2a2

Jolivet, G. & Turon, H. (2014). Consumer Search Costs and Preferences on the Internet. IZA DP No. 8643, November 2014. In *Discussion Paper Series*, Institute for the Study of Labor, Bonn.

Lach, S. (2002). Existence and persistence of price dispersion: an empirical analysis. NBER Working paper 8737.