Statistisk sentralbyrå
Statistics Norway

The use of online prices in the Norwegian Consumer Price Index

Paper[1] written for the 14th meeting of the Ottawa Group, Tokyo, Japan, 20-22 May 2015.

Ragnhild Nygaard
ragnhild.nygaard@ssb.no
Statistics Norway
Division for Price Statistics

Abstract

Ordering and buying goods or services for private use over the internet is increasing in popularity. The Nordic consumers are among those who shop the most online. During the last ten years e-commerce has changed from a phenomenon for the few to something almost every Norwegian is familiar to. And the e-commerce is growing fast. This paper starts by giving a general overview of the e-commerce in Norway. As the internet is an increasingly important purchaser channel it is important that a representative and relevant CPI/HICP has adequate coverage of online purchases. This paper therefore also looks into how Statistics Norway is working in order to increase the share of online prices in the CPI/HICP as well as the efforts for making efficient online data collection. Up until today prices collected online for the Norwegian CPI/HICP have been collected manually, thus a resource demanding process. Now Statistics Norway is testing and making use of data extraction techniques. In this paper we'll present our experiences with web scraping in addition to some preliminary calculations and conclusions based on collected online data.

[1] This paper is based on the work done by Kjersti Nyborg Hov, Anna Korlyuk, Leiv Tore Salte Rønneberg and Ragnhild Nygaard in the Division for Price Statistics.

1. Introduction

In 2014 the Division for Price Statistics in Statistics Norway initiated a work towards the use of online prices and the possibility of increasing the use of online prices in the Norwegian CPI/HICP. The project is partly financed by Eurostat. To our knowledge, several other European statistical offices have also started similar online data projects. There are several reasons for focusing on online data and the use of online data in the price indices;

More and more consumers prefer to purchase goods and services online and the Norwegians are among those who shop the most. There are several obvious factors that make the Norwegians among the most experienced online shoppers; high-living standards, strong consumer power and high availability of internet[2]. More than 1 out of 3 Norwegian consumers purchase goods or services online at least once a month according to a recent report from Postnord.com. Furthermore, the number of registered Norwegian online stores is increasing.

The internet is already an important purchaser channel and it is important that a representative CPI/HICP not only cover the online purchases and their price movements, but also have the right impact of online prices. When goods and services purchased online and price levels are different from traditional stores this is of course important to reflect in the price indices. In the cases where the prices of products purchased online also have different price development compared to the traditional stores, including these prices correctly in the CPI/HICP become even more important.

Statistics Norway has, like many other statistical offices, a clear strategy that data shall be collected efficiently with reduced costs and with a burden as low as possible for the data providers. As a consequence, there is a strong focus on alternative data sources. Statistics Norway has traditionally collected a high share of the CPI/HICP data through the means of paper questionnaires[3]. Now web questionnaires, scanner data and online data have taken over as important data sources.

According to Eurostat recommendation on internet purchases the e-commerce is not to be ignored in the computation of the HICP. The recommendation says that the "Internet purchases should be taken into account when designing the outlet samples and included in the HICP according to their significance". Statistics Norway's online data project has a two-sided aim; first of all, we want to increase the impact of online prices in the Norwegian CPI/HICP. With the strong increase in online purchases, the prices of tangible goods purchased online are most likely underestimated in the price indices. At the same time we want to make the online data collection as efficient as possible, following general strategies in Statistics Norway. Thus, in order to achieve the aims we are looking into different data extraction techniques. There are many different technical solutions on the internet for data extraction, many free of charge. To our knowledge, also other statistical agencies are testing different software for the same purpose.

The main advantage of online data is the increased coverage through the great amount of data available, but at the same time the high number of price observations is also the major challenge. Unlike scanner data, we only have the offer prices available for online data and no information on quantity sold. In this online data project we want to look into how we best can make use of this data.

This paper documents the work done in this project so far[4]. The paper is structured as follows; chapter 2 gives a general overview of the e-commerce in Norway including the size and general patterns. In chapter 3 we look at the present online data collection in the Norwegian CPI/HICP and its

[2] According to Northern Europe B2C E-commerce Report, Norway is on the top of the list among the European countries when it comes to internet access as share of the population (96 per cent).
[3] Price collectors have not been used for the monthly Norwegian CPI/HICP data collection.
[4] Statistics Norway has a grant agreement with Eurostat for the period January 2014-March 2016.

impact while in chapter 4 our experience with automated extraction of data from the internet is described. In chapter 5 some test calculations are commented, while some concluding remarks are given in chapter 6.

2. E-commerce in Norway

2.1 The Norwegians are on the top

The term "E-commerce" in this paper refers to commerce conducted over the internet via the World Wide Web where consumers can order goods and services from vendors' web sites.

Ordering and buying goods or services for private use over the internet is growing in popularity. For most Nordic consumers e-commerce has become a natural way to buy goods and services. And the e-commerce is growing fast. Rapid development of information technology and secure payment systems has expanded the households' internet purchases.

Different figures[5] illustrate the size of the e-commerce in Norway, and they all show that the numbers of internet purchases are increasing. Statistics Norway's latest survey of ICT[6] usage in households shows that 77 per cent of the private households have used the internet for buying or ordering goods and services during the last 12 months.

According to E-commerce Europe and their Northern Europe B2C E-commerce Report 2014, the Norwegian purchases of goods and services were estimated to EUR 8.9 billion in 2013. With this total of online sales, Norway showed the strongest e-commerce in Northern Europe. The estimated share of online goods in the total retail of goods was estimated to 7.2 per cent for the Northern European countries where the Nordic countries are on the top. Even though the share of online purchases is increasing and has become a natural part of the retail industry, the traditional retail stores still maintain a very strong position.

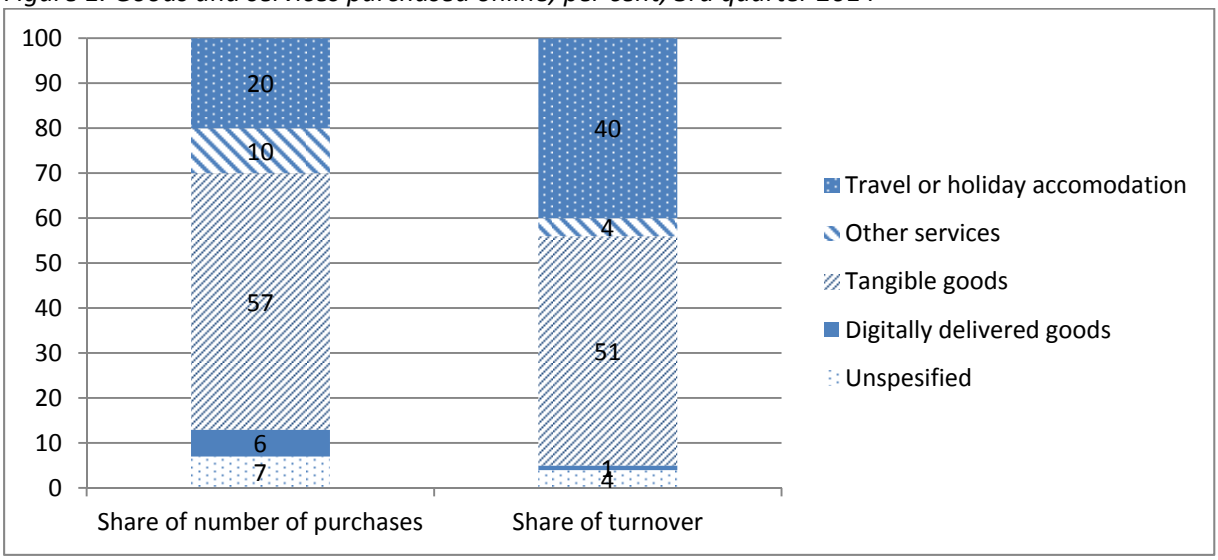
The number of online stores registered in Norway is increasing. According to Statistics Norway's Register of Establishments and Enterprises there are roughly 3 300 online stores actively registered in Norway. Latest figures from Statistics Norway's own retail trade sales statistics also indicate a strong increase in online purchases. Stores selling goods over the internet or by mail orders showed a much stronger growth in turnover compared to the traditional retail industry, with sales of NOK 10.9 billion (roughly EUR 1.3 billion) during the first ten months of 2014 which correspond to an increase of 17.8 per cent compared to the same period last year. Due to differences in scope these figures are not comparable to the figures in the E-commerce Europe report.

2.2 The holidays and leisure travel category is dominating

The category with most purchases in Norway is holidays and leisure travel. According to the latest 2014 figures from the Enterprise Federation of Norway, services make out 44 per cent of the total online turnover, where over 90 per cent of all the online purchases of services are related to the holidays and leisure travel category. Goods are estimated to make out 52 per cent of the total turnover. Digital goods[7] account for only about 1 per cent of the total turnover, as many low value purchases are made within this category, see figure 1 below.

[5] The turnover figures of the overall e-commerce are deviating mainly due to different definitions of e-commerce and what it is supposed to include.
[6] The notion ICT covers technology related to processing, presentation and storing of information, in addition to technology for communication and exchange of information.
[7] Goods that are delivered digitally like e-books, music download, app's and games.

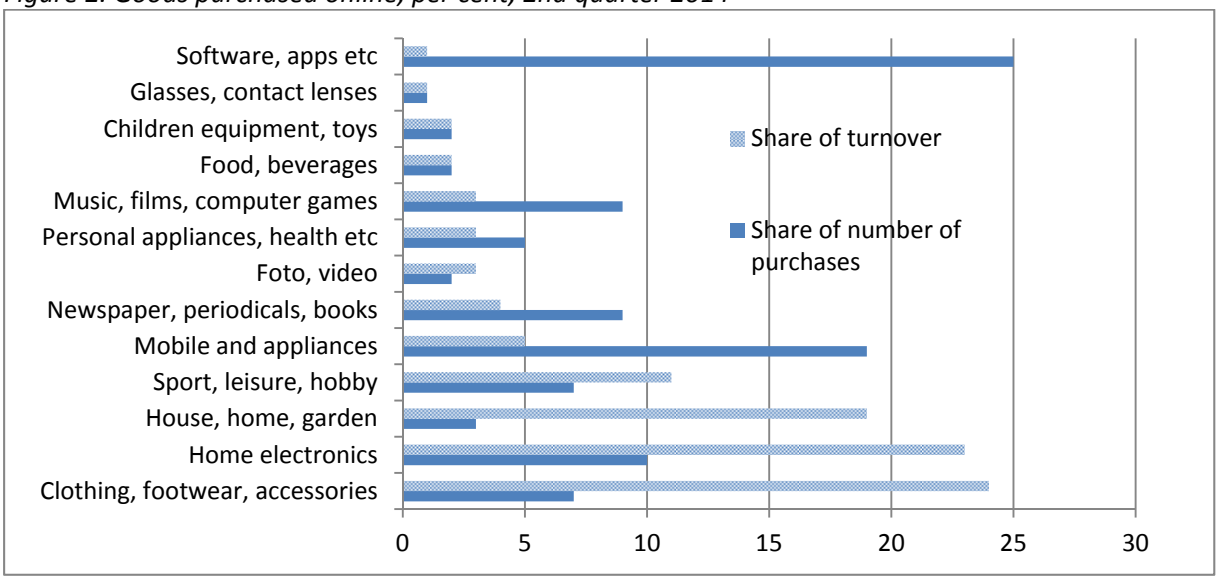
Figure 1. Goods and services purchased online, per cent, 3rd quarter 2014



Source: The Enterprise Federation of Norway

Goods that the Norwegians buy online are for instance home electronics, clothing, books and personal care products. Online grocery shopping currently accounts for only a small part of the market, but is expected to grow in the future. Figure 2 summarizes the purchases of goods online, both as a share of number of purchases made and as a share of turnover.

Figure 2. Goods purchased online, per cent, 2nd quarter 2014



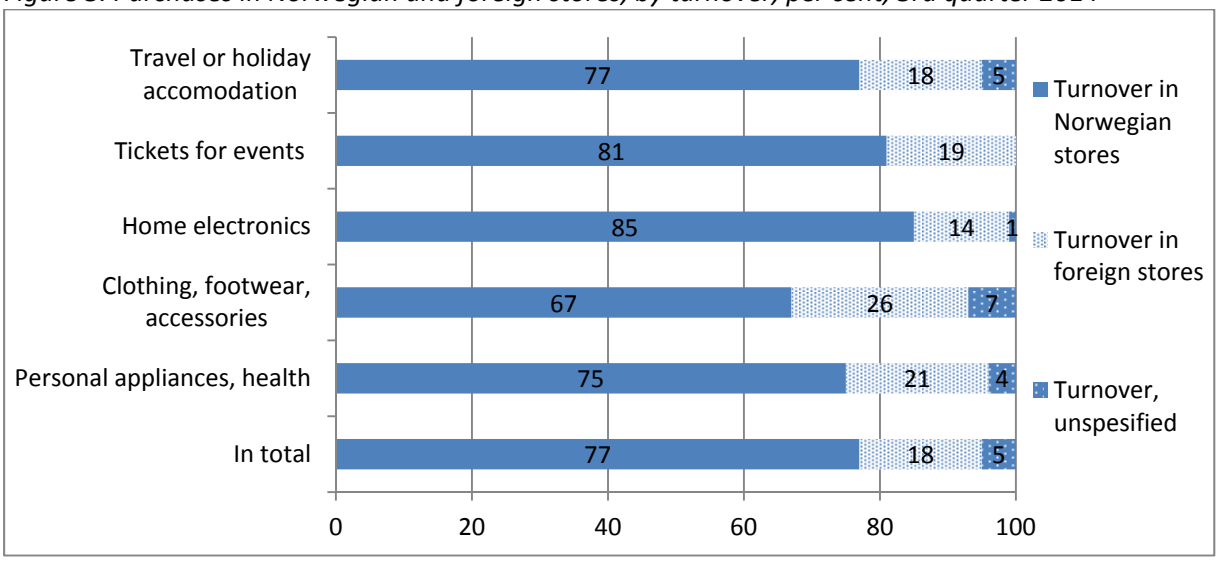
Source: The Enterprise Federation of Norway

2.3 Online shopping from foreign stores

Norway is the Nordic country (and even the European country) with the highest proportion of consumers that shop from foreign stores. The online purchases made by Norwegians from foreign stores are estimated to approximately EUR 1.2 billion in 2014, which is estimated to be around 23 per cent of the total online purchases of tangible goods (The Enterprise Federation of Norway, 2014). According to the leading Norwegian mail distribution company, lower prices and greater freedom of

choice are the main reasons for choosing foreign stores. Figure 3 shows the share of Norwegian and foreign shopping measured by turnover.

Figure 3. Purchases in Norwegian and foreign stores, by turnover, per cent, 3rd quarter 2014



Source: The Enterprise Federation of Norway

In the Norwegian National Budget for 2015 the Government increased the tax collection threshold for low value imports from NOK 200 to NOK 350 (approximately EUR 23 – EUR 40). The threshold for simplified customs clearance for private goods imports was also increased. Even though this appears to be rather small changes, this political move caused a great debate concerning increased online shopping and the possible threat to traditional trade or physical store shopping. The threshold change is, all else equal, expected to increase the incentives for internet shopping from foreign stores[8].

The share of foreign shopping among Norwegians is growing and expected to grow, but still the Norwegian stores are dominating. In other words, even though 1 out of 3 online purchases are made in foreign stores, close to 75 per cent of the turnover still goes to Norwegian online stores (The Enterprise Federation of Norway, 2014). This demonstrates that there are many low value purchases made in foreign stores, more precisely about 50 per cent of all the purchases made are priced less than EUR 23. In this online data project we concentrate on increasing the impact of prices from online stores registered in Norway. Foreign online shopping is probably one of the most complicated areas when it comes to price statistics where traditional sampling and methodology are strongly challenged.

3. Online data in the Norwegian CPI/HICP

3.1 Centralized online data collection – mostly services

The Norwegian CPI/HICP data collection has gradually changed over time and represents today a wide variety in data sources such as online data, scanner data, web questionnaires and different electronical data files. Statistics Norway has collected and used prices from the internet since the mid - 90s, either by simulating online consumer purchases for airline fares or by copying the prices of goods and services from different web sites. Today, prices collected from the internet make out

[8] There are however several factors affecting foreign shopping. During the second half year of 2014 the Norwegian currency depreciated strongly compared to the Norwegian trading partners, mainly as a result of lower crude oil prices, making it less attractive to buy from foreign online stores.

about 18 per cent of the total sample of goods and services measured by the CPI weight shares. The prices are, to a large degree, related to services and are manually collected from the internet, thus a rather labor intensive process.

The prices that are collected online are mainly characterized by a) being administratively determined prices or prices that are the same across the country, for instance prices of health- or telecom services, or b) prices being regarded as representative for all other purchasing channels like for instance airline fares. The leisure travel industry is one of the industries where digitalization and online shopping have led to extensive changes - the major parts of the sales have been twisted towards the internet and in many cases one is dependent on doing the sales online in order to get hold of the services. By far, most of the Norwegian consumers purchase their airline tickets online.

Other consumer areas where prices are collected manually online, with varying frequencies, are prices of package holidays, rents for secondary residences, prices of dental services, electricity tariffs and heat energy, prices of other transport services like bus, railway and boat as well as postal-, sports and recreational- and educational services.

3.2 Store sample based data collection – mostly tangible goods

The traditional way of collecting prices, mostly of tangible goods, is by web questionnaires[9] sent out to a representative sample of stores located in different parts of the country. The share of prices collected by questionnaires has been reduced over the last 10-15 years and today these prices make out roughly less than 30 per cent of the total basket of goods and services, measured by CPI weights. The sample, which is based on Statistics Norway's Register of Establishments and Enterprises and its defined industries, is covering both physical and online stores. Due to our sampling design and the fact that the industry categories are aggregated, we have not always managed to draw and include the most important online stores resulting in an underestimation of the online stores in the sample. This is not a problem if the price developments in online and physical stores are equivalent, but that may not be the case. An important element of the project is therefore to increase the number of important online stores in the CPI/HICP and to achieve a better representation and an improved mix of traditional versus online stores to better reflect the price movements from both purchaser channels. The aim is not to replace the prices from physical stores, but to include a higher share of online prices in the price indices.

Instead of the traditional way of sending web questionnaires to online stores, we are now establishing automated technical solutions for extracting online prices directly from web sites reducing the response burden close to zero for the data providers.

4. Web scraping

4.1 Data extraction software

In order to increase the amount of online prices in the Norwegian price indices and to make online data collection efficient, we have been testing various data extraction software available on the internet. There are several different web scraping applications available and they vary widely both in cost and features. Data scraping is a technique in which a computer program or a so called "web crawler" extracts data automatically from the internet. The web crawler systematically downloads all the web resources that it is able to reach from the starting point(s) according to some pre-defined conditions (e.g., types of files to scrape, types of files to ignore, etc.). We have tested different

[9] Statistics Norway has a history of collecting most of the prices to the CPI/HICP through questionnaires. Web questionnaires have now taken over and as of 2015 the data providers no longer fill out paper questionnaires.

software solutions, but we have mainly focused on the services provided by import.io. To our knowledge, also other statistical agencies have been testing the same software package. Import.io offers a free browser version.

To be able to make use of import.io, one must create an account with username and password[10]. By using simple point and click technology we demonstrate to the "crawler" where the data is. The crawler then runs through the web site and extracts information from pages which are similar in regards to the parameters we have set out. The data is structured into rows and columns and stored on cloud servers to be downloaded and easily loaded, for instance, into csv/excel files. Further, we then load the data into suitable software (SAS) for analysis, calculation and storage. The data is collected automated at the same time every morning. The only requirement is that at least one of us colleagues in the project group must be logged on to our computer.

The extraction is easy to build and doesn't require any coding skills. Making use of computer software outside the control of the statistical office is however risky. The service at import.io is constantly being developed and we should expect that the service can change from time to time. Functionality, for instance, might change without any warranty - it can even be discontinued all together. Statistics Norway also has a rather restrictive IT infrastructure environment that makes it difficult to experiment too much with different software solutions available online. To be able to use the services of import.io we needed an acceptance from the Department of IT and an open port through the computer firewall[11].

Our experience is that import.io is an excellent tool for looking into the possibilities online prices may provide and in case import.io should fail or close down, there are several other similar technical solutions available on the internet[12]. We have chosen to use the specific software, but at the same time it is important to work towards other technical solutions as well. We are therefore in dialogue with other units within Statistics Norway such as the Department of IT and the Department of Data Collection in order to also look for other technical solutions. For the time being, the data is being collected by the project group within the Division for Price Statistics and not, maybe more naturally, within the Department of Data Collection.

4.2 Data extraction in practice

In order to navigate on the web sites, the program or the "crawler" looks for defined URLs[13] in the web pages. Our experience is that it is essential to make an as precise extraction as possible by thoroughly defining the different pages that we want to visit and extract prices from, both in order to make the extraction go as fast as possible and to be as robust as possible. It is important to keep the pages the crawler is visiting to a minimum. By doing so, it reduces the chances of having to reprogram/redefine the extraction.

One obvious advantage of using online data is the enormous amount of data available. Normally, a CPI will have to be based on a basket of representative goods and services and a relatively limited amount of price observations. With online data the amount of price observations can be strongly increased. Unlike scanner data, the main weakness is obviously the lack of quantity information. Online data cannot provide us with quantity information unless the data is accompanied with additional information from the site owner. In addition, we do not know whether the products have

[10] We have created a separate user profile for the CPI Unit.
[11] The firewall helps prevent different types of malware from getting to the computers.
[12] Other statistical offices are testing and using different software solutions.
[13] The global address of documents and other resources on the World Wide Web.

actually been purchased during the period in question, this is however not a new challenge for price statisticians.

We see the advantage of making adequate cost-benefit analyses before starting to explore such scraping techniques. One must be willing to invest quite a lot of resources in order to be able to use it successfully. Even though the software itself may not require any coding skills, there is still quite a lot of testing and work to be done. If there are numerous prices that can be extracted from the same sites there may be an advantage to automatically select the prices with the help of a browser software. E.g. we collect prices for dental services from a web consumer portal where nation-wide prices are listed (all dentists are obliged to report their prices to this portal). Automated extraction of this information is of great advantage. However, if only a few prices are available for download on each site, an automated extractor might not be beneficial at all.

In Norway a few consumer-friendly "prices-comparison sites" have been established in the recent years, and the sites gather a large share of the different online product offers and compare prices[14]. These sites may be a rich source of price information. We have however decided to extract data directly from the different online stores as it is technically easier to get the data from the different site owners. It is also safer to extract data directly from the sellers by reducing the number of sites to depend on. The lack of quantity information is also an issue to consider. We could have extracted the prices for all the different variants of shampoos from all the online stores in the Norwegian market (which is quite a large number). Our only option would then be to make a geometric average of all these prices. Without expenditure information many small online providers would then be equally weighted with the major ones. We have therefore decided to only collect prices from the leading online stores registered in Norway (i.e. the stores with the highest turnover registered).

Today, data is extracted daily in order to see the price movements online. Our aim is not to have daily indices in the CPI, but it can be beneficial to see the real short term movements. This will also give us an idea of how a representative monthly price should be calculated. In order to make a monthly estimate one might have to collect prices for at least the midweeks of the month and then make an average of those prices.

4.3 Legal considerations

Extracting data automated from the internet is a new way of collecting prices for statistical purposes, and in order to take advantage of the data it is necessary to address various issues, as for instance legislation. Making frequent extractions of huge amount of data from web sites - is it really legal? Do we need permission to extract data from someone's web site? To our knowledge that depends on what kind of data we are scraping, the amount of information accessed and copied and to which extent the access adversely affects the page owner's system and the use of the data. One important feature to consider is whether web scraping may be against the terms of use of the web sites. We often agree to use a site according to its terms when we access and stay on a specific web site. What is allowed on one web site may be prohibited on another site. In many cases the web sites do not have any terms of use available on the sites at all.

Most of the web sites Statistics Norway has been looking into underline that all information on their sites is protected by copyright law. Many of the sites inform that download or copying of data should not be done without an explicit consent from the site owners. The Norwegian Statistics Act however, clearly states that Statistics Norway may impose an obligation to provide information necessary to produce official statistics; hence we are legally entitled to collect the data without having to alert the owners. We have however chosen to inform the different sites' owners in order to establish an open

[14] For instance www.prisjakt.no and www.kelkoo.no.

dialogue and cooperation (and to avoid technical obstacles). This may also open up for the possibility of receiving electronical data files (including quantity information) directly from the owners in the future. In any case, it is important to follow some kind of "best data extraction practices" to avoid damaging the sites and furthermore the owners of the web sites and their interests.

4.3.1 The use of cookies

Like most stores on the internet, the stores that we extract data from use "cookies" on their web sites. A cookie is a small piece of data, usually a small text file/text string, sent from a web site and stored in the user's web browser while the user is browsing that web site. The cookie (now stored in the user's web browser) stores information of the user's movements on the website, or information typed in by the user, as username, email address, shipping address etc. Every time the user loads the web site, the browser sends the cookie (i.e. the collected data information from the previous visit) back to the server to notify the web site of the user's previous activity. According to online retailers they use cookies in order to gain knowledge of how the consumers use the web sites. The information can be used to improve functionality of the sites and to simplify the online shopping through easier login, individual shopping baskets and custom-made recommendations etc. This gained knowledge can also be used in the general pricing strategy of the stores. It is not unlikely that online retailers can change prices depending on the individual consumer's browser history or even location. This kind of dynamic pricing, or price discrimination, may be a large challenge for CPIs. It is unknown to what extent this actually occurs among the leading Norwegian online retailers that we are looking into, it is plausible to believe that this is still a relatively small and new phenomenon.

5. Experimental online data price indices

5.1 Data extraction specifications

As a starting point, we have chosen to focus on consumer groups that represent a high share of online purchases such as personal care products and home electronics. We have started to collect data from four of the leading online stores registered in Norway[15] with the highest turnover. The main advantage looking into products related to personal care is that this is an area with rather homogenous and long-lived products with less need for quality adjustments compared to other goods as for instance clothing and home electronics. The online purchases for personal care products are estimated to roughly 7 per cent of the total turnover according to the Norwegian Cosmetics Association. Data is also being collected for home electronic products, but we have not yet managed to look at the data. Home electronic products may give us other challenges compared to personal care products and might require other ways of treatment[16].

In order to reduce the weighting problem we do not extract price information from all the products available. We extract price information[17] for only the most sold products (i.e. price observation level). Most sellers online have their products ranked by most sold. Traditionally we make an average of a limited set of price observations per representative item. Using online data we may strongly increase

[15] Home electronics are mainly purchased from stores registered in Norway, the share of online purchases from foreign stores are higher for personal care products, see figure 3.

[16] We are also looking into the data collection of airline fares which is very time consuming. The focus on airline fares is mainly done for efficiency purposes. For extracting airline fares we are now testing automated data extractions from google servers. The data extraction is done by utilizing Google's QPX Express API client library for Java. A script allows us to query the database and parse the results into a readable format for import into SAS for calculation and storage. Google offers a service where they allow 50 searches per day, free of charge. We may specify all price determining characteristics and Goggle may return as many as 500 replies per search. With an automated data collection process for airline fares it also opens up for increasing the scope of the survey. The automated extraction of airline fares is in a very early stage and is not the focus of this paper.

[17] We extract the actual offer price. Several leading online stores (not stores selling home electronics) do not take additional costs concerning the delivery of the products. If such costs exist, they should be included in the price measured.

the number of price observations, but at the same time, by collecting the most sold products, we keep them "representative".

Along with the prices and the short text descriptions also the actual URLs are extracted. The URLs do not define the actual product (the single price observation) itself, but refer to all the products within the product group or, in other words, the COICOP6 group (i.e. all the different types or variants of for instance shampoo). From the URLs and with the help of SAS coding we create the different COICOP6 groups like for instance "shampoo", "body lotion" etc. Having the URLs in the data file also makes it possible to click into the actual products and to see their characteristics (including pictures). We may not be able to use all of the information directly in the calculations, but it might still be helpful e.g. for handling duplicates etc.

For the personal care products the short text descriptions might be used for matching purposes. The personal care products do not have product codes; hence we need to match the samples by the product's description. In order to make it work it may be important to handle differences in spelling (such as lower- and upper- case letters). We use different coding strategy in SAS in order to reduce mismatching in the description text. By using only the most sold products we make the matching process easier as there seem to be good stability among the most popular products. Our experience so far is that the description text seems to be stable. For other consumer groups matching price observations by description text may not be the solution.

5.2 Calculations of daily indices

In order to make test calculations we have built a production system in SAS that to a large extent copies the scanner data production system that we use for the index of food and non-alcoholic beverages where we have full product coverage. For analytical purposes (and while waiting for longer time series[18]) we have started calculating three different series of daily indices;

1. Daily chained indices based on an unweighted geometric average of the price relatives, i.e. the Jevons index based on matched samples. As we only extract and include the most sold products within each COICOP6 group from each online store, price observations may go in and out of the daily baskets. The maximum amount of price observations on the "bestseller lists" within each COICOP6 group is roughly 50 per store which represent approximately 1/4 or 1/5 of the total number for e.g. "shampoo". We see that roughly 5 per cent of the sample consists of new price observations entering the basket each day. Using matched model methodology, this can also mean that among these new price observations there might be some "old" ones due to products changing description text. However, there seem to be rather few of those cases spotted on a daily basis. Roughly 2-3 per cent of the basket are price observations that re-enter after being out of the defined basket (either "old" products moving up on the "bestseller list" or observations that have been taken off the sites for a certain period). The prices are imputed in the period it is not included based on the price development of products within the same COICOP6 group. With monthly indices the share of price observations entering and leaving the basket is expected to be higher.

2. Daily chained indices based on the ratio of geometric average prices of two following periods (here: days) with stratification. We want to see to what extent we can make use of the extracted information on unit or size of the products in the index calculations. As the price of a 1000 ML shampoo bottle is normally higher than 150 ML bottle of shampoo, we try to stratify the products by unit/size. Roughly 75 per cent of all the price observations have

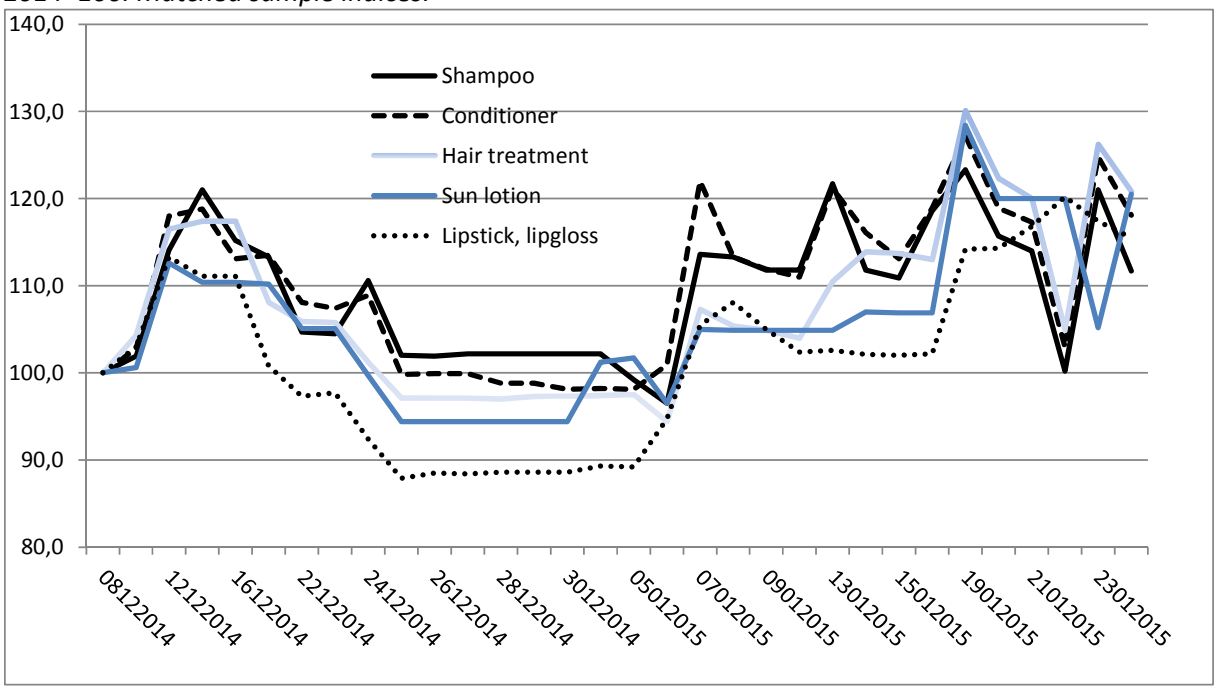
[18] With longer time series we can start calculating monthly indices based on average of daily prices from the midweeks of the months.

information about unit as a part of the text description. A complicating factor is that the unit is not written in a standardized form and in the same way for all the different products. We therefore have to extract the unit value from the text description with the help of SAS coding. Relatives of homogenous unit groups are weighted together[19] into a product group index. New price observations are allowed to enter the basket and others to disappear. There are no imputations of missing prices. These price indices do however not reflect pure price change as quality differences due to e.g. brand is not dealt with.

3. Daily chained indices based on the ratio of geometric average prices of two following periods (here: days) with no stratification within the COICOP6 group/product group. This method does not of course reflect pure price changes. In the average price calculation new price observations enter and old ones disappear without any consideration of quality differences. Outliers defined as 3 standard deviations of the mean are removed. These are pure average price indices.

The last method is unlikely to be used in official statistics. Still it may be interesting to see to what extent the methods differ. Figure 4 shows the daily price movements of some selected COICOP6 groups based on data from two of the largest and leading online stores offering personal care products in Norway. Even though this is a very short data period, we can see that the prices are quite volatile (as opposed to during the Christmas holiday which we easily can spot in the figure).

Figure 4. Price development of selected personal care product groups. Daily chained indices. 8 Dec 2014=100. Matched sample indices.

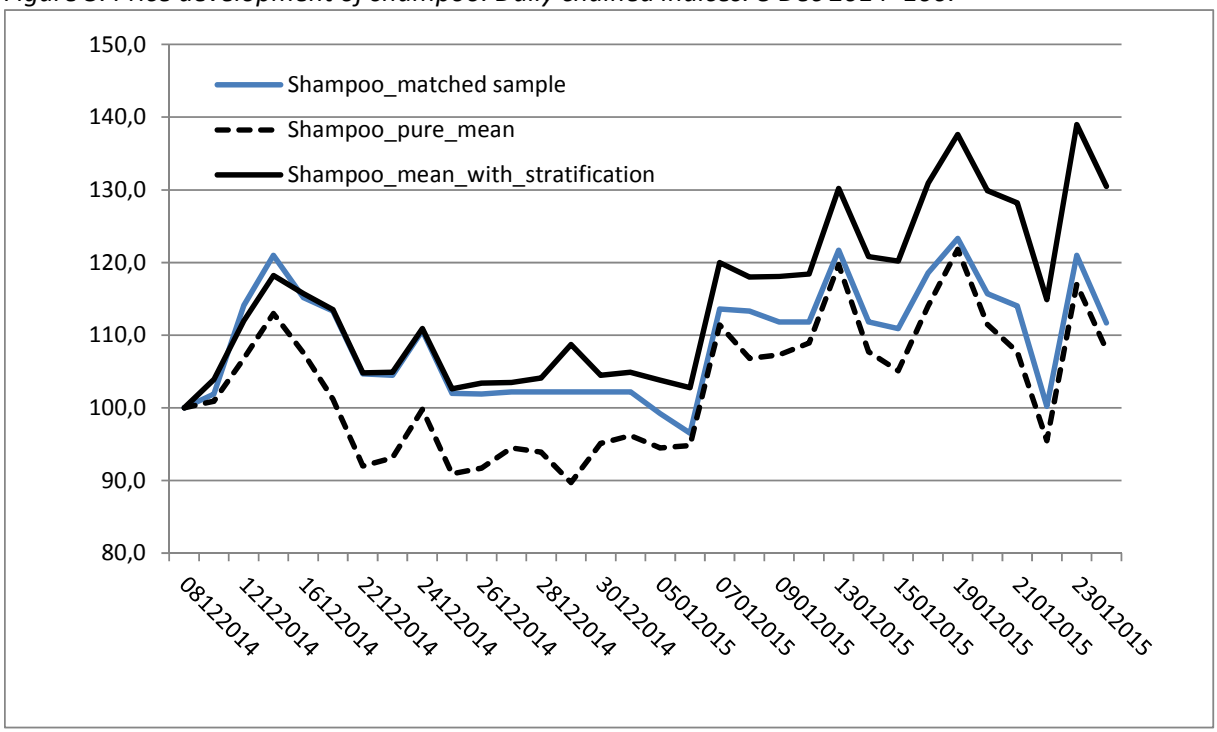


Our test calculations show that the price movements for personal care products seem to be more volatile compared to prices in physical stores which is plausible to believe as it is much more convenient to change prices for online stores compared to traditional stores. It is likely to believe the prices might vary depending on the product's popularity online, cf. chapter. 4.3.1.

[19] The unit relatives are weighted together by the average of the number of price observations on the bestseller list in the two periods. We have also tested to weight the relatives by the "expenditure shares" - the average number of price observations in both the period multiplied with the average price. Different weighting strategies seem to cause only minor changes in the results.

In the figures below we can see how different calculation methods provide different results.
In figure 5 we have calculated daily price indices for shampoo (covering about 100 different products of shampoo) based on data from the two leading online stores for personal care products. The index series are based on the three calculation methods outlined above. We have discovered that some of the deviation in the series is due to error in the data extraction. Twice during the period in question a product group from one of the online stores was missing in the data entirely, see indices from 29 December 2014, where all three indices go in separate ways. This demonstrates the importance of controlling for missing product groups simultaneously as the data is extracted[20]. It is too late to discover this afterwards.

Figure 5. Price development of shampoo. Daily chained indices. 8 Dec 2014=100.



In figure 6 we see rather large differences between the three different methods. The pure mean price index, i.e. the index based on pure average prices without any consideration for quality differences, ends up far below the other two series with a 5 per cent price decrease since the starting point. The other two series show a 3 to 5 per cent price increase during the same period. In figure 7 there seem to be smaller differences between the series.

Both of the average price indices are affected by changes in the composition as well as pure price changes. One could expect, due to differences in composition, greater differences between the matched sample index and the two average price indices. The test calculations however, show that for most of the product groups there are smaller differences between the matched sample index and the pure average price index (pure mean). The index based on average prices combined with unit stratification seems to lie above the other series in most cases. The rather small differences between the series may be due to the daily indices and the high-frequent chaining and relatively small changes in the "best seller lists" on daily basis. The index series show that the matched sample index may work for products related to personal care as the text descriptions (i.e. the linking identifiers) seem to be stable and of good quality. A weakness with this method however is that there is no replacement of permanently disappearing price observations.

[20] We have not yet experienced the need for redefining or reprogramming the extraction code, but this is of course important to discover early in the extraction process.

This is a work in progress. Longer time series will provide a better foundation to conclude on the use of the data. The next steps will be to calculate monthly online price indices and to compare those to the similar product group indices from the physical stores. Weighting online and physical store indices together according to the purchaser channels' significance will then most likely provide improved and more relevant CPI/HICP indices.

Figure 6. Price development of daily cream. Daily chained indices. 8 Dec 2014=100.

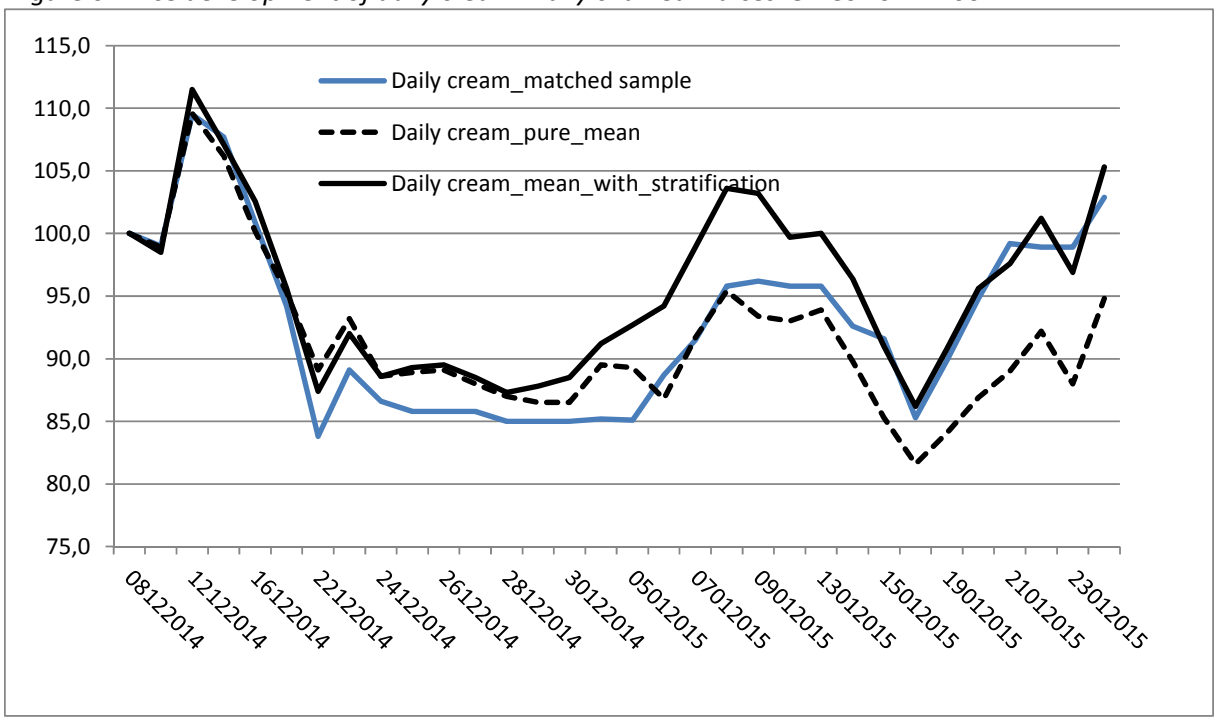
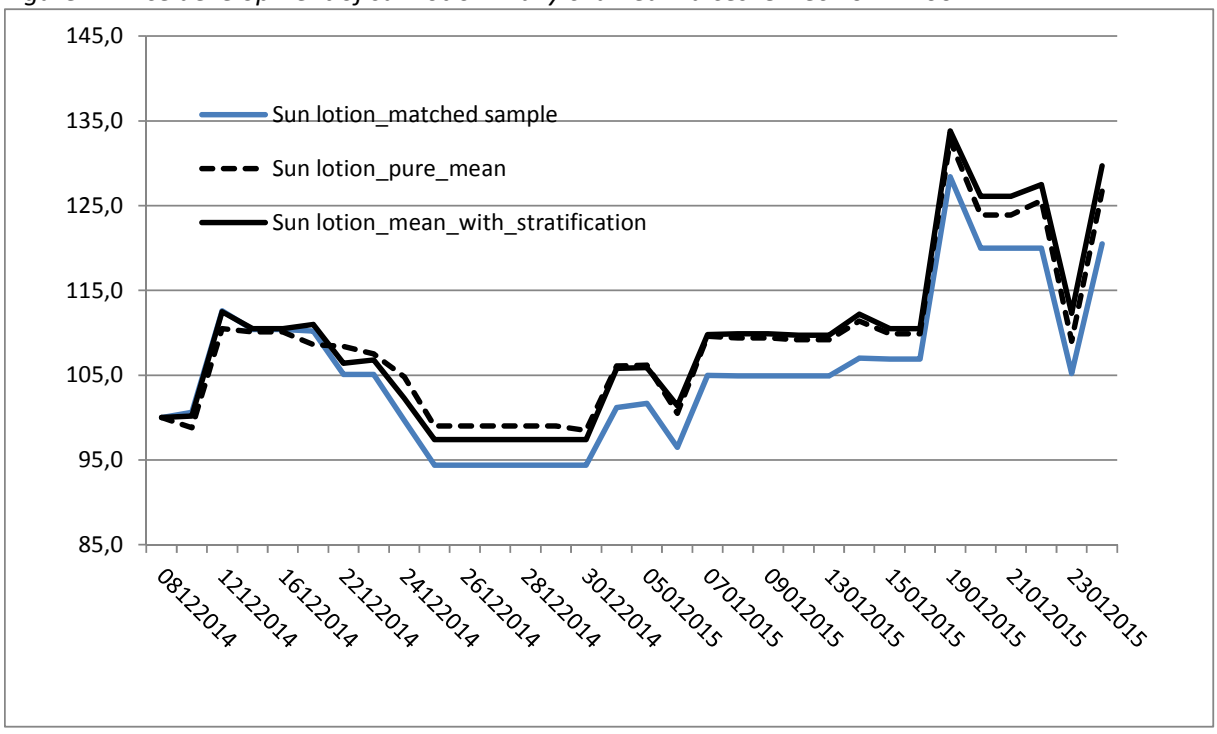


Figure 7. Price development of sun lotion. Daily chained indices. 8 Dec 2014=100.



6. Concluding remarks

The e-commerce has become an important part of daily life for most consumers. Most Norwegian consumers buy both goods and services online. Many national statistical offices are now testing online data for price statistics purposes. This means that national statistical offices must address various issues related to methodology, quality, data access, legislation etc. Traditional methodology and thinking are likely to be challenged.

We see that there is no easy fix making use of different web scraping software for price statistics purposes and many issues, such as both practical and legal, must be sorted out. There are obviously both great opportunities and challenges related to the use of online data in price indices. Our experience is that it is very important to make adequate cost-benefit analyses. It is not obvious that web scraping always will be more efficient compared to the present situation. In any case one must be willing to invest quite a lot of resources in order to be able to use it successfully.

In order to remain relevance in the price indices it is important to include price development from different purchaser channels and to have a proper representation. A preliminary conclusion is that the price movements may be quite different between the different purchaser channels. It is likely that the pricing strategies may be quite different for online stores compared to physical stores. The test calculations made so far are limited and for the moment we have only looked into one consumer group, but the calculations made so far demonstrate that the online prices may be very volatile.

References

Ecommerce Europe.
http://www.ecommerce-europe.eu/facts-figures/infographics/northern-europe-2013

Forbrukerrådet. http://www.hvakostertannlegen.no/

Import.io. https://www.import.io/

Posten Norge AS.
http://www.bring.com/all-of-bring/ecommerce/_attachment/422207?_ts=1416e0aa770

Posten Norge AS. http://www.bring.no/hele-bring/netthandel/ehandelsrapport/_attachment/527116?_ts=14879368fe0

Postnord. http://www.postnord.com/globalassets/global/english/document/publications/2014/e-commerce-in-the-nordics-2014.pdf

Statistics Norway. ICT usage in households, 2014, 2nd quarter.
http://www.ssb.no/en/teknologi-og-innovasjon/statistikker/ikthus

Statistics Norway. The Statistics Act of 1989.
http://www.ssb.no/en/omssb/styringsdokumenter/lover-og-prinsipper/the-statistics-act-of-1989

Statistics Norway. Wholesale and retail trade sales statistics, 5th period 2014.
http://www.ssb.no/en/varehandel-og-tjenesteyting/statistikker/vroms/termin

The Enterprise Federation of Norway.
http://www.virke.no/talloganalyse/Documents/eHandelsbarometeret_Q3_2014.pdf

The Enterprise Federation of Norway.
http://www.virke.no/talloganalyse/Documents/eHandelsbarometeret%20Q2%202014.pdf

The Enterprise Federation of Norway.
http://www.virke.no/talloganalyse/Documents/Ringvirkninger%20av%20netthandel%20fra%20utlandet%20og%20200-kronersgrensen.pdf