Special Proceedings of 19th Annual Conference of SSCA held at SKUAST, Jammu during March 06-08, 2017; pp 31-38

Prices from On-line Shops: Challenge in Measurement of Consumer Price Index

Dilip Kumar Sinha
Price Statistics Division, Central Statistics Office, Government of India

Final Version Received on 19.07.2017

Abstract

Retail inflation computed using CPI must reflect the real situation of price rise for consumers. This requires CPI to be the representative of the prices paid by the consumers at large, incorporating their consumption pattern. With the passage of time, especially in recent years, e-commerce has become parts and parcel of our life. People, especially in urban areas, prefer buying goods and services on-line to visiting physical shops since prices on the web–based markets are more competitive than those of bricks and mortar outlets. Therefore, collection of prices for CPI, only form physical shops, is going to dilute the property of representativeness in the years to come. It is the need of the hour to explore the possibilities of inclusion of prices from on-line stores in compilation of CPI. The paper acknowledges the efforts of CSO, India in bringing out CPI for the entire population. Further, it has attempted to take stock of the development taken place across the globe in the direction of using scanner data and on-line prices in compilation of price indices. Analysis of the opportunities vis-a-vis challenges involved in this activity has also been carried out. Finally, possible solution to meet those challenges in including prices from on-line shops for index compilation has been suggested, with a focus on Indian context.

Key words: Online Prices, Inflation Measurement, Big Data, Scraped Internet Prices

1. Introduction

It was February 2011, when the long awaited Consumer Price Indices (Rural, Urban, Combined), for the general population of India was launched by the Central Statistics Office (CSO) with Base Year 2010. Prior to this, the country had only population segments specific index numbers namely, Consumer Price Index for Industrial Workers, Consumer Price Index for Urban Non-manual Employees, Consumer Price Index for Agricultural Labourers and Consumer Price Index for Rural Labourers. Bringing out an index, covering entire population, was indeed a challenge for the CSO because of varying consumption pattern, preferences, life style etc. of different class of people in different parts of the country. Comprehensive exercise and research were carried out to address those issues and finally the desired index could see the light of the day.

Subsequently, after a few years, the Base Year of the series was revised from 2010 to 2012 by incorporating many methodological improvements which *inter alia* include

Views expressed in the paper are of author only. It does not reflect the opinion of the organization, where the author is working.

Corresponding Author: Dilip Kumar Sinha
E-mail: dilipkumar_sinha@yahoo.co.in

preparation of item basket and weighing diagrams using the Modified Mixed Reference Period (MMRP) data of Household Consumer Expenditure Survey (HCES) in consistent with international practice of shorter reference period for most of the food items and larger reference period for the items of infrequent consumption/purchase, and shifting from the use of Arithmetic Mean (AM) to Geometric Mean (GM) of price relatives for compilation of elementary indices. In addition, items were classified following 'Classification of Individual Consumption according to Purpose (COICOP)', in order to make the index internationally comparable. In this index, item baskets are State-specific and specifications of items have been fixed market-wise based on the popular variety being consumed in that area. Due care was taken while selecting the markets so that the prices are representative of the consumption by the entire population.

'Practical Guide to Producing Consumer Price Index (PGPCPI)', a United Nations document suggests that markets or locations should be selected using Probability Proportional to Size (PPS) sampling with retail turnover or an approximate proxy as a size measure. Since it was difficult to get data on retail turnover, therefore, population was used in allocating markets proportionately to different States in rural as well as urban areas. This provided implicit weight to each market; as a result, simple GM of the Price Relatives of different markets resulted in their weighted average. The issue of change in the quality of product has also been addressed adopting the explicit quality adjustment process, as suggested in the PGPCPI. Thus, CSO has tried its best to make the indicator reliable by addressing the challenges in compilation of Consumer Price Index. Still there are other issues; those are cropping up, to be dealt with in future for making the index more robust.

Before moving further, it is important to understand the issues involved in compilation of this index. These issues are becoming more complex over a period of time because of changing market structure, consumers' buying behaviour, faster substitution of items etc.

2. Issues in Compilation of CPI

A CPI has two components: Weights; and Price Relatives. Weights are generally referred to as expenditure weight, which is computed from the results of HCES. Price Relatives are ratios of prices of current period over base period. For this, prices are collected from a number of markets which are representative of the targeted population.

Using fixed expenditure weights for a certain number of years, till the next base revision is done, has a well known disadvantage of being biased because of not incorporating the price elasticity. There are some methods to address this issue. For instance, Modified Laspeyers' method suggests a way to adjust the expenditure weight with price weight. In this direction, European countries as well as USA have also experimented with the use of high frequency data i.e. bar code scanner data for compiling superlative indices because of availability of real time information on prices and quantity. India also needs to explore the possibility of incorporating suitable mechanism in its methodology to address the issue of fixed expenditure weight.

The second as well as the most important issue is to collect the prices, which are representative of the consumption of the targeted population. In the standard method of CPI, prices are collected from earmarked outlets of a number of markets, selected using appropriate sampling procedure. But over a period of time, structure of markets is changing. Due to faster industrial development, products of different varieties have started coming to the markets resulting in shifting of consumers' preferences. This is ultimately affecting the representativeness of the prices. Substitution of items becomes a challenge because of nonavailability of weights of new products. Use of scanner data has been one of the most important strategies to resolve this issue as well. But this method has also some limitations.

Now, another issue, of inclusion of prices from online shops, is also evolving. Rapid expansions of information technology, awareness of internet among people, and their hectic life style have created a large space for virtual market. People sitting in the remotest place of a country can buy the same thing, which could be purchased by people living in mega cities. If a sizeable number of people shifts from bricks and mortar shops to web based markets, then the prices collected following the classical survey approach are no more representative of consumption. Issues relating to inclusion of prices from on-line shops for compilation of CPI would definitely be examined; before that, it is important to take a stock of the development taken place in the field of using scanner data, in index compilation.

3. Use of scanner data or high frequency bar code data for compilation of CPI

Numerous research works have been undertaken on possibility of using scanner data, analysing its impact, and evaluating its merits and demerits. Some of those papers *(listed in the references)* have been referred while attempting this article. As per available literatures, many countries have experimented with the use of scanner data in compilation of CPI. The staff of the U.S. Bureau of Labour Statistics (BLS) has been in contact with several private vendors of scanner data since 1993, conducting research on the use of this data for compilation of CPI. As on December 1996, supermarkets items (food, housekeeping supplies, toiletries, over the counter health products, tobacco) accounted for 13.0% of the total weight in the CPI-U and these items, covered by scanner data, represented a smaller percentage since some outlets did not have scanners (Ralph Bradley, Bill Cook, Sylvia G. Leaver and Brent R. Moulton). In January 2012, the US BLS purchased scanner data from the Nielsen Company. A paper by Jenny FitzGerald and Owen Shoemaker attempted to evaluate how accurately the CPI reflected the reality. It was found that there was large difference between the traditional CPI and the index compiled using Nielsen scanner data.

Some countries in Europe also are using these high frequency data in compilation of their CPI. Statistics Norway has been applying this data since August 2005. Statistics Netherland introduced supermarket scanner data in June 2002. Statistics Swiden has replaced manually collected prices with scanner data for samples of outlets and products. UK has also attempted to use such data in compilation of CPI.

A number of papers and reports, authored to analyse the cost and benefits of using the high frequency data, has concluded that scanner data provides a number of opportunities, in terms of coverage of all transaction, availability of real time data on daily basis, surveillance of new goods, simultaneous observation of price and quantity information, product attributes and measurement of quality etc. On the other hand, it suffers from many challenges as well. Retailers develop systems suited to them resulting in databases, which are different from company to company. It also depends on cooperation of retailers. Central Bureau of Statistics, Netherlands has mentioned that it can take anything from six months to several years to negotiate the sharing of scanner data. Products continuously disappear and new ones come with new bar codes. This amounts to problems in matching bar codes for the same product over time. Initial Report on Experiences with Scanner Data in ONS by Derek Bird, Robert Breton, Chris Payne and Ainslie Restieaux, Prices Division Office of National Statistics clearly mentions that the availability of information on prices and quantity in the scanner data provides a good option to go for superlative indices e.g. Fishers or Tornqvist index, as these indices are the theoretical method for calculating cost of living price indices because of having provision of incorporating market dynamics, capturing of new goods, enhancing representativeness of the baskets etc. however, the issue is broader than this on the count of practical considerations. Laspeyers-type indices are upper bound indices therefore the superlative indices would lead to lower inflation rates having implication that substitution behaviour would be necessary to keep track with inflation. Another problem with the superlative indices at item level is that it suffers from chain drift. However, there should be a continuous effort to incorporate such high frequency data in compilation of CPI.

4. Advantages and disadvantages in using prices of online shops in compilation of CPI

With the growing use of internet and shifting of people from off-line shops to web based on-line shops, it would be inevitable, in days to come, to include prices from virtual shops also for compilation of CPI in order to keep it relevant. These data have definitely many merits. First, prices of web based stores are available at very low cost, even cheaper than that of scanner data. Second, it is also available at very high frequency like scanner data. Third, it includes detailed information for all products being sold by the sampled retailers, which can be helpful in doing quality adjustments. Fourth, online data can be collected remotely in any country. It also allows us to centralize the data collection and homogenize its characteristics. Fifth, these data are available in real-time, without any delays to access and process the information.

But it suffers from several disadvantages as well. First, it does not cover many service items. Second, they currently cover a much smaller set of retailers and product categories. On-line price data lack information on quantities sold. To obtain expenditure-weighted data, it is necessary to use weights from a government consumer expenditure survey or other sources. Online prices are much volatile than those of off-line stores. Comparison of different characteristics of on-line prices, scanner price data and traditional method price collection for CPI are given in Table1 below:

Table 1. Comparison of different sources of prices for CPI

Attributes	Prices from On-line stores	Scanner Price Data	Tradition survey of prices for CPI
Cost of price collection	Very low	Medium	High
Data frequency	Very High	Very High	Low
All products available with retailers	Yes	No	No
Real time availability	Yes	Closely	No
Comparable across country	Yes	Limited	Limited
Product category covered	Few	Few	Many
Retailers covered	Few	Few	Many
Quantities or expenditure weights	No	Yes	Yes with huge time lag

A paper titled ‘The Billion Prices Project: Using Online Prices for Measurement and Research’ by Alberto Cavallo and Roberto Rigobon, which is a preliminary report on the Billion Price Project created in 2007 in Argentina and further extended to other countries, including US, in 2008, acknowledges aforementioned issues.

As mentioned in another paper titled ‘The use of online prices in the Norwegian Consumer Price Index’ by Ragnhild Nygaard, the Division for Price Statistics in Statistics Norway initiated a work towards the use of online prices in 2014 and the possibility of increasing the use of on-line prices in the Norwegian CPI/HICP (Harmonized Index of Consumer Prices). The project is partly financed by Eurostat. Several other European statistical offices have also started similar online data projects. Situation of Norway is different from that of India. Even in that country, they have focussed on consumer groups that represent a high share of online purchases such as personal care products and home electronics. They have started to collect data from four of the leading online stores registered in Norway15 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. They also have done it on experimental basis. It has been concluded in this paper that there are obviously both great opportunities and challenges related to the use of online data in price indices. Price movements may be quite different between the different purchaser channels. It is likely that the pricing strategies may be quite different for on-line stores compared to those of physical stores. The test calculations made so far are limited and for the moment they have only looked into one consumer group, but the calculations made so far demonstrate that the on-line prices may be very volatile.

5. Challenge in using prices from on-line stores

Apart from aforementioned advantages and disadvantages of on-line prices, the biggest challenge in using this data for compilation of CPI is how to map the prices with locations. In the traditional survey of price collections, prices from a particular market refer to the prices paid by the consumers living in the nearby areas. Generally, the most popular outlet or the outlet having highest retail turnover is selected for price collection within that market so that price data are representative of the most of the consumers living in that area. Let us suppose that 50 markets have been selected in a particular State, following suitable sampling procedure of PPS sampling. Here, these 50 markets are representing the entire population of that State, by giving them implicit weight. As a result, the prices collected from these markets are representative of the prices paid by consumers in that State. Accordingly, even un-weighted GM of the price relatives of these markets, of an item, gives elementary index of that item. Even in using scanner data, mapping of location is possible, but in the case of prices from online stores, it is very difficult to do such mappings. Unless we know the percentage share of population that purchased a particular item on-line during a given reference period, it is difficult to assign weight to the price relative computed from the data of on-line stores. The said percentage share is changing day by day as the number of people, shifting from bricks and mortar shops to web based virtual outlets, are increasing very fast. In order to capture this information, a large scale survey would be required to be conducted throughout the country, which involves huge cost. Therefore, some alternative solution needs to be worked out for including these prices in CPI.

Second issue relates to frequency of price collection. In the standard survey, prices are collected during peak hour on a fixed day of the week. Peak hour is chosen to get prices which are paid by most of the consumers. But, this is not possible in the case of on-line prices as we don’t have information at which price maximum sales took place. Moreover, as mentioned before, prices of web based stores change very frequently, even on hourly basis. In such condition, deciding the time point of data collection becomes very difficult.

Third issue pertain to substitution of items. In standard practice, if an item, of a certain specification, is not available on the selected shop in a market, price collector waits for one or two months. If the situation continues or the shopkeeper gives specific information that the product has disappeared then only the specification of that item is substituted with comparable specification. In the case of on-line shops, the specifications of items change regularly because of some marketing strategy. It does not have to do anything with consumers’ preference. For them, the entire universe is a market place; therefore, they experiment with changing specifications to maximize profits. Thus, collecting prices, of available specifications, would unnecessarily substitute the item and introduce bias in the index.

6. Possible way to include on-line prices in compilation of CPI

It is an established practice that item-wise weight, in the CPI basket, is assigned using the data of HCES where weight of an item is share of expenditure on that item to total expenditure. If one additional column is inserted in the HCES schedule, in which item-wise value of expenditure incurred, on on-line shopping, may be collected, then using this piece of information, we can compute the share of web-based shopping in the weight assigned for an item. Accordingly, item or elementary index would be compiled as weighted GM of the GMs of the price relatives of off-line and on-line shops. This has been mathematically explained below:

Let us suppose that the number of sampled off-line markets = *n* and number of sampled on-line markets = *m*. Prices collected during current period from off-line markets for *i*th item are P_{ij}^{c}; where *j = 1, 2, 3, ...n* and *c* refers to current period and, prices collected during base period from off-line markets for *i*th item are P_{ij}^{0}; where *j = 1, 2, 3, ...n* and *0* refers to base period. Therefore, price relatives of these off-line markets are $\frac{P_{ij}^{c}}{P_{ij}^{0}}$; *j=1,2,3,........n*. And GM of these price relatives would be $G^{off} = (\prod_{j=1}^{n} \frac{P_{ij}^{c}}{P_{ij}^{0}})^{\frac{1}{n}}$.

Similarly, prices collected during current period from on-line markets for *i*th item are p_{ik}^{c}; where *k = 1, 2, 3, ...m* and *c* refers to current period.

And, prices collected during base period from on-line markets for *i*th item are p_{ik}^{0}; where *k = 1, 2, 3, ...m* and *0* refers to base period. Therefore, price relatives of these on-line markets are $\frac{P_{ik}^{c}}{P_{ik}^{0}}$; *j=1,2,3,........n*. And GM of these price relatives would be $G^{on} = (\prod_{k=1}^{m} \frac{P_{ik}^{c}}{P_{ik}^{0}})^{\frac{1}{m}}$.

Further suppose expenditure weight of *i*th item is $w_{i.}$ (percentage share of expenditure of *i*th item to total expenditure). And $w_{i.}$ is distributed between w_i^{on} and w_i^{off}, where these two weights refers to expenditure of *i*th item on on-line and off-line stores respectively.

Therefore,

$$w_{i.} = w_i^{off} + w_i^{on}.$$

Now, the index of item i = $[\ (G^{off})^{w_i^{off}}\ X\ \ (G^{on})^{w_i^{on}}\ ]^{\frac{1}{w_i^{off} + w_i^{on}}}$ .... (1)

his has been further elaborated using hypothetical numerical example as follows:

Table 2. Estimation of share of online expenditure in item weight

Item	**Estimated Expenditure**	**Estimated Expenditure on online shopping**	**CPI Weight** *[w_i = Exp. of an item in Col. (2)/2000]*100*	**Share of online expenditure in item weight** *[w_i^{on} = Col. (3)/2000]*100*	**Share of offline expenditure in item weight** *[w_i^{off} = Col. (4)-Col.(5)]*
(1)	(2)	(3)	(4)	(5)	(6)
Rice	150	50	7.50%	2.50%	5.00%
Wheat/Atta	100	40	5.00%	2.00%	3.00%
......	.....	.....	.....	.....	.....
....	.....	.....	.....	.....	.....
Total	2,000	300	100.00%	15.00%	85.00%

Note: It may happen that for some or many of the items shares of online expenditure may be zero.

Now let us assume that prices are collected from sampled 50 off-line markets (as per the existing practice) and five on-line markets (now proposed to be introduced in CPI). Then the item indices would be compiled as follows:

Table 3. Computation of Item/Elementary Indices

Item	**GM of price relatives of offline markets** (G^{off})	**Item Index** ***[as per existing practice, Col. (2)*100]***	**GM of price relatives of online markets** (G^{on})	**Weighted GM of the GMs of price relatives (*Using the equation 1)***	**Item index**
(1)	***(2)***	***(3)***	***(4)***	***(5)***	***(6)***
Rice	1.45	145	1.25	1.38	138
Wheat/Atta	1.30	130	1.20	1.16	116
......	.....	.....	.....	.....	.....
....	.....	.....	.....	.....	.....
......	.....	.....	.....	.....	.....

As mentioned before, on-line shopping transactions are increasing very fast, the values of w_i^{off} and w_i^{on} needs to be updated at frequent intervals, may be every year. In India, HCES is conducted once in every five years. Therefore, updating these weights annually would require some follow up surveys. In its absence, alternative method may be used, for instance, if the value of total turnover of these web stores is available, then ratio of this turnover to total Private Final Consumption Expenditure (from National Accounts estimates) can be computed every year, including the base year and the w_i^{on} may be moved with the movement of aforementioned ratio. This, of course, needs to be examined empirically.

The second issue of frequency of price collection may be resolved this way. Prices may be collected at the highest possible frequency and then the GM of those prices of a given specification may be considered as representative prices, since GM would exclude extreme values.

The third issue regarding substitution of items is definitely very difficult to address. Any statistical technique does not seem to be appropriate to tackle this problem. One possible way is to have dialogue with the on-line retailers about their strategy to change the items in their catalogue. This strategy may be different from retailer to retailer. Accordingly, retailer specific methods are required to be worked out to incorporate substitution of items while including the on-line prices for compilation of CPI.

7. Conclusion

The e-commerce has become an integral part of our day to day life, especially in urban areas. Though, majority of buyers, are still dependent upon physical shops, the penetration of internet, competitive prices and convenience to buyers is making on-line purchase popular day by day. Moreover there are many items, which are not available in the bricks and mortar shops of nearby market, can easily be purchased on-line with the facility of door step delivery. Because of lower investment in physical infrastructure, cost per unit of item of on-line retailers is very less. As a result, their market shares is perceived to be exponentially increasing (data in this regard is not readily available). Therefore, it is the need of the hour to include prices of on-line stores in compilation of CPI, in order to keep it relevant and representative of the entire population.

Some of the European Countries and United States of America have attempted to move in the direction. Many research papers have done its cost and benefit analysis. There are, of course, a number of advantages of these data as these are available at very low cost, with detailed information and at very high frequency. But, there are many challenges too. It is very difficult to map the price data with locations. Appropriate expenditure weights for items of these web-based stores are not available. Collection of information, on amount spent on on-line shopping, during HCES with some other adjustments may provide some way to move in the direction of including these prices in index compilation.

References:

Cavallo, Alberto and Rigobon, Roberto. *The Billion Prices Project: Using Online Prices for Measurement and Research.*

Nygaard, Ragnhild. *The use of online prices in the Norwegian Consumer Price Index*.

Samar, Muhanad, Norberg, Anders and Tongur, Can (2012). *Issues on the use of scanner data in the CP.* Statistics Sweden.

Initial report on experiences with scanner data in ONS – by Derek Bird, Robert Breton, Chris Payne and Ainslie Restieaux, Prices Division, Office of the National Statistics.

Heymerik van der Grient and Haan, Jan de (2010). The use of supermarket scanner data in the Dutch CPI.

FitzGerald, Jenny and Shoemaker, Owen (2013). *Evaluating the Consumer Price Index Using Neilsen's Scanner Data.*

Bradley, Ralph, Cook, Bill, Leaver, Sylvia G. and Moulton, Brent R. (1997). *An overview of research and potential uses of scanner data in the U.S. CPI.*

Feenstra, Robert C. and Shapiro, Matthew D. (2003). *Scanner Data and Price Indexes.* University of Chicago Press. ISBN: 0-226-23965-9.

Armknecht, Paul A. (2015). *Fixed basket method for compiling consumer price indexes* (2015)., *American International Journal of Contemporary Research,* Vol. **5(5)**, pp 97-106.

Practical Guide to Producing Consumer Price Index (PGPCPI). (2009). United Nations document.