Web Scraping of Prices of Commodities Included in the Generation of Consumer Price Index (CPI) for the National Capital Region (NCR), Philippines

Divina Gracia L. Del Prado, Ph.D., Elena G. Varona, Desiree R. Robles,
Glen G. Polo, Rosario S. Lodovice, Jo Loiuse L. Buhay

Abstract

Online stores are becoming popular as a new platform for business transactions, not only in the country, but also globally. To take advantage of this new approach, the Philippine Statistics Authority (PSA) started in 2019 the exploration on the use of web scraping as an alternative collection method for prices of commodities included in the computation of Consumer Price Index (CPI) for the National Capital Region (NCR), Philippines. Currently, the PSA uses the traditional face-to-face price collection of commodities from sample outlets or stores. In this paper, prices collected from traditional method or face-to-face method are called offline prices, while web scraped prices are termed online prices. Prices of 514 commodities are web scraped, which comprise about 71 percent of the total commodities in the CPI market basket of NCR.

This study aims to determine if offline prices can be replaced by online prices or by a combination of online and offline prices (hybrid) in computing the CPI for NCR. Results show that the behavior of online and offline prices are comparable for selected commodities that are not highly volatile such as clothing items. However online prices of agricultural commodities, which are highly volatile, do not present the same trend of volatility as that of offline prices. Moreover, for CPI computation, offline prices are more appropriate to use for certain commodity groups, while for others, hybrid prices.

Keywords: Web scraping, CPI

1. Introduction

1.1. Motivation

The Philippine Statistics Authority (PSA) obtains prices of commonly consumed commodities for the monthly generation of the Consumer Price Index (CPI) through the Retail Prices Survey (RPS) of Selected Commodities and Services for the Generation of Consumer Price Index also known as the RPS for the CPI. This survey uses the traditional method of collecting data where the prices are obtained through personal visit to selected sample outlets or stores.

Traditional methods in data collection such as that of the RPS adopt well-established techniques based on statistical theories. Data obtained through this method adhere to standards and are structured for easy to mathematical manipulation. However, the demands for granular and high frequency data are increasing due to the need for timely and evidence-based policies and programs to address various issues and concerns of the country. This sets the traditional data collection on a disadvantage where timeliness and budget are a recurring issue.

In order to address the limitations of traditional methods, and to take advantage of the new technologies available, the PSA conducted a study on web scraping. This aimed at determining the possibility of replacing selected prices collected from sample outlets with the prices collected online and to possibly reduce the cost on the conduct of the survey without sacrificing the timeliness and quality of the data.

1.2. Objectives

The main objective of this study is to know whether the prices collected from website via web scraping can be used as substitute for the data collected via traditional survey in computing the CPI for NCR. To accomplish this objective, two sets of CPIs are computed using online prices only (online) and the combination of online and offline prices (hybrid) and compared with the official CPI.

1.3. Significance

This study will be used as benchmark for the use of big data such as web scraped prices for official statistics. Results of this study will also be beneficial in the research field as it may serve as reference in the future studies related to this topic.

1.4. Scope and Limitations

This study covered the commodities in the 2012-based market basket for CPI in NCR. Websites covered by scraping are selected based on the availability of the commodities within the website, thus, may not be the actual outlets or stores visited for price collection. Meanwhile, the period of data collection through web scraping is from January 2020 to December 2021. All results and conclusions from this study focused on the given geographic domain and time periods only.

2. Relevant Literature

2.1. History of Web Scraping

Web scraping, also known as web crawling, refers to the automatic collection of price quotes and article information from websites (Boettcher, 2015).

The first use of online data in compiling CPI and inflation rate was motivated by the manipulation of inflation statistics in Argentina from 2007 to 2015. Cavallo (2013) conducted a study which automatically collected data from October 2007 to March 2011 from largest supermarkets of selected Latin American countries. Results showed that for Brazil, Chile, Columbia, and Venezuela, the annual movement of the inflation rates between the online and the traditionally collected data are not different. For Argentina, the best approximation to the official numbers is to use one-third of the actual inflation rate observed online.

The study of Cavallo showed the potential of using online prices for inflation measurement applications. In 2008, MIT implemented the Billion Prices Project (BPP) which collected online data on selected retailers' websites from more than 60 countries. Results of this project showed that online prices distribution is strongly bimodal, with very few price changes close to zero percent. Also, online data have the potential to provide datasets with identical sampling characteristics in a large number of countries (Cavallo & Rigobon, 2016).

2.2. Use of Online Data for Official Statistics

The Billion Prices Project showed promising results on the use of big data for official statistics. Various countries adopted the integration of online data in the computation of CPI for their country.

In United States, about two-thirds of their data are collected through personal visits and the remaining data are collected via telephone or the outlets' website. In Ukraine, two types of price collection is being implemented: local – collection through personal visits to sample outlets; and central – collection through head office or through websites. In neighbor ASEAN countries such as Singapore, web scraped data is one of their official data sources of administrative data.

3. Methodology

3.1. 2012-based Market Basket of Commodities for CPI

The market basket for CPI refers to the sample commodities that represent the commonly purchased goods and services by households. The 2012-based CPI market basket was the result of the Survey of Key Informants (SKI) conducted in 2013 to sample stores nationwide. Respondents were store managers, sellers, or proprietors who were asked of the most purchased good and services. In NCR, the total number of commodities in the market basket is 724 representing 1.78 percent of the

total number of the commodities for all income households CPI in the Philippines.

3.2. Web Scraping Application

The web scraping application was developed in-house using Python and Beautiful Soup. The user-interface of the application utilizes Mozilla Firefox where the launcher and uniform resource locator (URL) of the target commodities are saved.

3.3. Web Scraped Commodities

The total number of web scraped commodities for this study is 514. This represents about 71 percent of the commodities in the market basket for all income households in NCR. All commodity divisions or 2-digit Philippine Classification of Individual Consumption according to Purpose (PCOICOP) have representative web scraped prices. Websites of educational institutions do not display tuition fees unless user IDs and passwords are supplied, essentially making the scraping of prices impossible.

The web scraped commodities were classified as exact or equivalent. Exact commodities are those with the same specification found in the market basket in NCR. Equivalent commodities are those with whose specifications are similar, in part, to the specifications of commodities in the market basket.

Table 1 presents the number of commodities scraped by commodity group based on the similarities in specifications.

Table 1. Distribution of Commodities with Exactly Matched and Equivalent Specifications.

Code	Commodity Group	Total	Exactly Matched	Equivalent
	TOTAL	**514**	**218**	**296**
01	Food and Non-Alcoholic Beverages	183	68	115
02	Alcoholic Beverages and Tobacco	10	4	6
03	Clothing and Footwear	41	24	17
04	Housing, Water, Electricity, Gas, and Other Fuels	14	5	9
05	Furnishings, Household Equipment and Routine Household Maintenance	85	29	56
06	Health	41	30	11
07	Transport	2	0	2
08	Communication	2	0	2
09	Recreation and Culture	64	28	36
10	Education			
11	Restaurants and Miscellaneous Goods and Services	72	30	42

3.4. List of Stores and Number of URLs by Division

The study considered 12 online stores namely Abensons, Ace Hardware, Anson's, Lazada, National Book Store, Pushkart, Shopee, Watsons, Western Appliances, Wilcon, Zalora, and Zagana. These online stores were chosen based on the availability of commodities from these websites.

A total of 1,351 URLs were web scraped and the distribution of the URLs per division is shown on Table 2.

Table 2. List of Stores and Number of URLs by Division Code

Name of Online Stores	No. of URLs	Commodity Division Code										
		01	02	03	04	05	06	07	08	09	10	11
Total	**1,351**	**402**	**15**	**94**	**38**	**231**	**74**	**5**	**8**	**233**		**251**
Abenson	16					13				3		
Ace Hardware	15				4	11						
Ansons	12					11				1		
Lazada	550	155	11	39	14	87	16	2	3	107		116
MerryMart	3	1	1				1					
National Book Store	23									23		
Pushcart	74	65	1			1						7
Shopee	538	151	2	43	14	84	16	3	5	96		124
Watsons	45						41					4
Western Appliance	17					16				1		
Wilcon	16				6	8				2		
Zagana	30	30										
Zalora	12			12								

Legend:

01 – Food and Non-Alcoholic Beverages
02 – Alcoholic Beverages and Tobacco
03 – Clothing and Footwear
04 – Housing, Water, Electricity, Gas and Other Fuels
05 – Furnishing, Household Equipment and Routine Household Maintenance
06 – Health
07 – Transport
08 – Communication
09 – Recreation and Culture
10 – Education
11 – Restaurant, Personal Care and Miscellaneous Goods and Services

3.5. Web Scraping Process

Data Collection

Web scraping was done daily except for Saturdays, Sundays, and holidays. For the duration of this study, the frequencies of scraping started from bi-weekly to daily. Shown in Table 3 is the frequency of collection from January 2020 to December 2021. Adjustments in the frequency of collection through web scraping were made to be able to capture prices based on the period of the price surveys for the CPI in NCR.

Table 3. Frequency of Collection through Web Scraping

Month	Period of Price Collection	Remarks
Jan 2020	End of the month (last six days)	Initiated after the general planning on web scraping
Feb 2020	First and Second Phase (Simultaneous with CPI)	Used two computers
Mar 20	Everyday	Used one computer only; Halted for a few days during the ECQ due to new installation of program in one of the staff's laptop
Apr 20		
May 20		Used one laptop (only every 6:00pm onwards)
Jun 20		Used two laptops
Jul 20		Used five laptops
Aug 20		
Sep 20 to Jul 21		
Aug 21 to Dec 21		Used six laptops

Data Cleaning

After a whole month of web scraping, the csv files are compiled by matching the web scraped commodities with the commodities collected offline. These prices underwent data cleaning and validation patterned to the practice regularly done in the CPI.

3.6. Estimation

With the goal of the study to see whether the behavior of the online prices resembles the behavior of the offline prices, the monthly price relatives at the commodity level as well as the indices at the subclass and class level were computed for each set of prices using R Language.

3.6.1 Average Price

In this study, two ways of computing the monthly average prices were explored.

a. Online prices

The average monthly price of a commodity is computed by obtaining the average of the web scraped data according to the price collection schedule in NCR as shown in Table 4.

Table 4. Period of Price Collection for CPI in NCR

Period of Collection in NCR	
Commodity Group	**Schedule**
1. Agricultural food items	Weekly, every Tuesday
2. Processed food, beverages, and tobacco	Weekly, every Friday
3. Non-food	Bi-weekly First phase: Any day during the first five (5) days of the month Second phase: Any day from 15th to 17th day of the month

b. Hybrid

The average monthly price of a commodity was obtained by the following conditions:

i. Without web scraped data: average of prices was obtained from the offline data

ii. With web scraped data: average of prices was computed according to schedule of price collection schedule for CPI in NCR

3.6.2 Price Relatives at the Commodity Level

The price relatives are computed at the commodity level. This is to determine whether the month-on-month movement of prices of each commodity follows that of the offline prices. Line graphs were used to present the movements of price relatives comparing that of online, hybrid and the offline prices.

3.6.3 Index at the Subclass and Class Level

The index at the subclass (5-digit PCOICOP) level were computed for each combination of price quotations regardless the number of successfully web scraped commodities under each subclass. Subsequently, the index at the class (4-digit) level were also computed.

3.7. Absolute Deviation

In order to determine the differences between the computed indices between online, hybrid and offline prices, the absolute deviations were computed. In particular, the minimum, maximum, range and average absolute deviations per exploration were obtained.

4. Results and Discussion

This discussion on the results of this research focused only on four class (4-digit PCOICOP) levels, namely: Fish and Seafood (PCOICOP Code 01.1.3); Vegetables (PCOICOP Code 01.1.7); Tobacco (PCOICOP Code 02.2.0); and Garments (PCOICOP Code 03.1.2). Although the data collection through web scraping was initiated in January 2020, the first month with complete web scraped data was in February 2020. Thus, the computation for CPI using online and hybrid prices started in March 2020. To compete the data for 2020 for the online and hybrid CPI, the offline CPI was used.

Table 5 shows the number of and percentage of web scraped commodities for selected commodity classes.

Table 5. Number and Percentage of Web Scraped Commodities for Selected Classes.

2009 PCOICOP	Description	Percentage of Web Scraped Data	Number of Commodities			
			Market-Basket of NCR	Web Scraped		
				Total	Exactly Matched	Equi-valent
01.1.3	FISH AND SEAFOOD	25%	36	9	3	6
01.1.7	VEGETABLES	59%	39	23	0	23
02.2.0	TOBACCO	25%	4	1	0	1
03.1.2	GARMENTS	63%	27	17	9	8

Among the four selected classes, the computed online and hybrid prices indices of fish and seafood, tobacco, and garments follow the same trend are. On the other hand, vegetables – a class which contains agricultural items was selected to examine whether the volatility of prices for this commodity group is reflected on the computed CPI for online and hybrid prices.

The class, fish and seafood is composed of five subclasses, four of which were successfully represented in the web scraped data. For the class tobacco, only one

subclass represents the whole class which is also included in the web scraped data. For the class garments, only one subclass was represented from the five subclasses. On the other hand, all the five subclasses of the class vegetables were represented.

The computed CPI for fish and seafood, tobacco, and garments is shown in Figures 9, 10, 11, and 12. To further examine whether the computed CPI for the online and hybrid prices follow that of the offline, the absolute deviations were calculated and summarized in Table 6.

Table 6. Summary Statistics on Absolute Deviation of the Computed CPI from the Official CPI

2009 PCOICOP	Description	Absolute Deviation from the Official CPI			
		Online		Hybrid	
		Lowest	Highest	Lowest	Highest
(1)	(2)	(3)	(4)	(5)	(6)
01.1	**FOOD**				
01.1.3	FISH AND SEAFOOD	0.5	34.6	0.4	2.7
01.1.7	VEGETABLES	0.1	48.9	0.1	37.0
02.2	**TOBACCO**				
02.2.0	TOBACCO (ND)	0.0	21.3	0.1	7.4
03.1	**CLOTHING**				
03.1.2	GARMENTS	0.0	2.9	0.0	3.1

Table 6. Summary Statistics on Absolute Deviation of the Computed CPI from the Official CPI *(continued)*

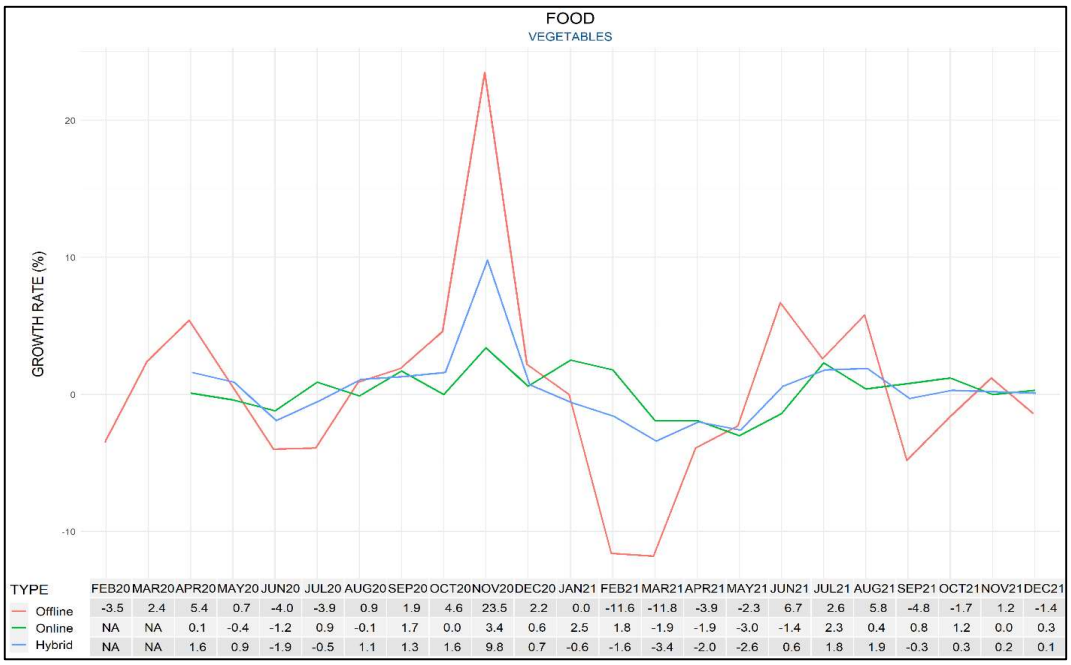
2009 PCOICOP	Description	Absolute Deviation from the Official CPI			
		Online		Hybrid	
		Range	Average	Range	Average
(1)	(2)	(3)	(4)	(5)	(6)
01.1	**FOOD**				
01.1.3	FISH AND SEAFOOD	34.0	12.9	2.2	1.2
01.1.7	VEGETABLES	48.9	11.7	36.9	9.3
02.2	**TOBACCO**				
02.2.0	TOBACCO (ND)	21.3	9.8	7.3	3.9
03.1	**CLOTHING**				
03.1.2	GARMENTS	2.9	1.0	3.1	1.0

Among these classes, the computed CPI for hybrid prices of fish and seafood and tobacco showed relatively low absolute deviation from the official CPI. Meanwhile, online prices are more comparable for items under clothing.

Month-on-month and year-on-year growth rates were obtained and compared. The starting series for growth rates is April 2020. Figures 1, 2, 3, and 4 shows the month-on-month growth rate of CPIs for fish and seafood, vegetables, tobacco, and garments, respectively.

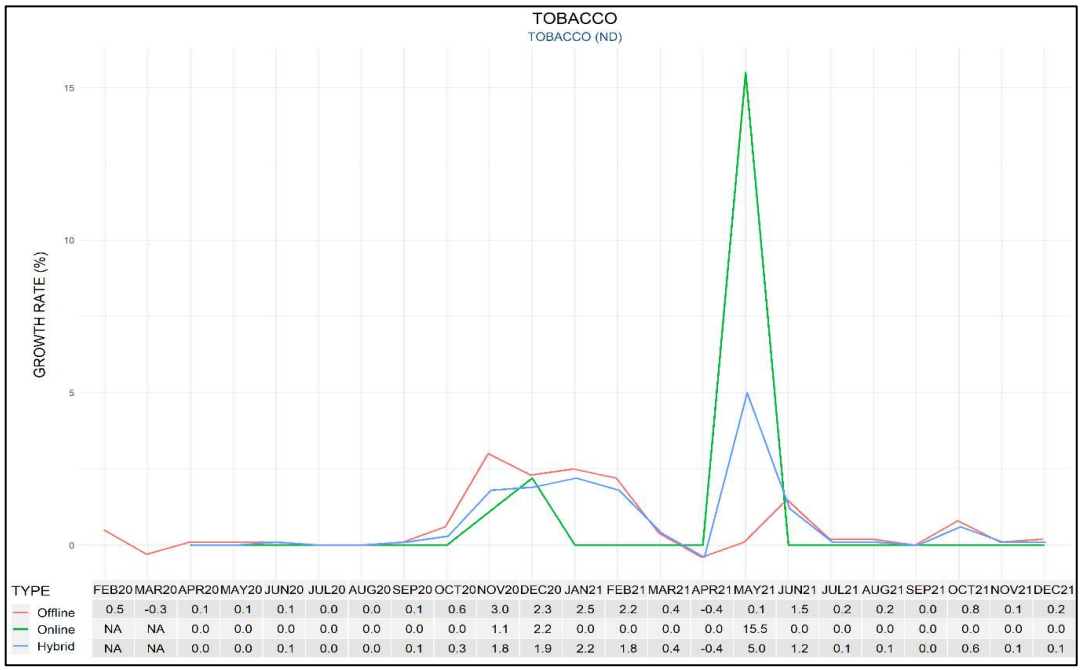
TYPE	FEB20	MAR20	APR20	MAY20	JUN20	JUL20	AUG20	SEP20	OCT20	NOV20	DEC20	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	-0.4	-1.3	-1.9	-4.8	-2.7	-0.9	0.3	0.2	2.4	4.8	3.9	4.1	-0.3	-2.6	-0.7	-1.0	-1.0	0.0	2.5	-0.9	2.0	2.5	2.5
Online	NA	NA	0.2	0.0	1.5	-2.9	0.5	-8.6	0.5	0.6	-0.4	0.0	9.5	0.7	4.0	0.0	-1.0	0.0	0.0	0.0	-0.5	0.0	0.0
Hybrid	NA	NA	-1.8	-4.8	-2.7	-1.0	0.3	-0.1	2.4	4.7	3.8	4.1	0.0	-2.5	-0.5	-1.0	-1.1	0.0	1.7	-1.0	1.9	2.4	2.5

Figure 1. Month-on-Month Growth Rate of CPI for offline, online, and hybrid prices of Fish and Seafood from February 2020 to December 2021



TYPE	FEB20	MAR20	APR20	MAY20	JUN20	JUL20	AUG20	SEP20	OCT20	NOV20	DEC20	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	-3.5	2.4	5.4	0.7	-4.0	-3.9	0.9	1.9	4.6	23.5	2.2	0.0	-11.6	-11.8	-3.9	-2.3	6.7	2.6	5.8	-4.8	-1.7	1.2	-1.4
Online	NA	NA	0.1	-0.4	-1.2	0.9	-0.1	1.7	0.0	3.4	0.6	2.5	1.8	-1.9	-1.9	-3.0	-1.4	2.3	0.4	0.8	1.2	0.0	0.3
Hybrid	NA	NA	1.6	0.9	-1.9	-0.5	1.1	1.3	1.6	9.8	0.7	-0.6	-1.6	-3.4	-2.0	-2.6	0.6	1.8	1.9	-0.3	0.3	0.2	0.1

Figure 2. Month-on-Month Growth Rate of CPI for offline, online, and hybrid prices of Vegetables from February 2020 to December 2021



TYPE	FEB20	MAR20	APR20	MAY20	JUN20	JUL20	AUG20	SEP20	OCT20	NOV20	DEC20	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	0.5	-0.3	0.1	0.1	0.1	0.0	0.0	0.1	0.6	3.0	2.3	2.5	2.2	0.4	-0.4	0.1	1.5	0.2	0.2	0.0	0.8	0.1	0.2
Online	NA	NA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	2.2	0.0	0.0	0.0	0.0	15.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Hybrid	NA	NA	0.0	0.0	0.1	0.0	0.0	0.1	0.3	1.8	1.9	2.2	1.8	0.4	-0.4	5.0	1.2	0.1	0.1	0.0	0.6	0.1	0.1

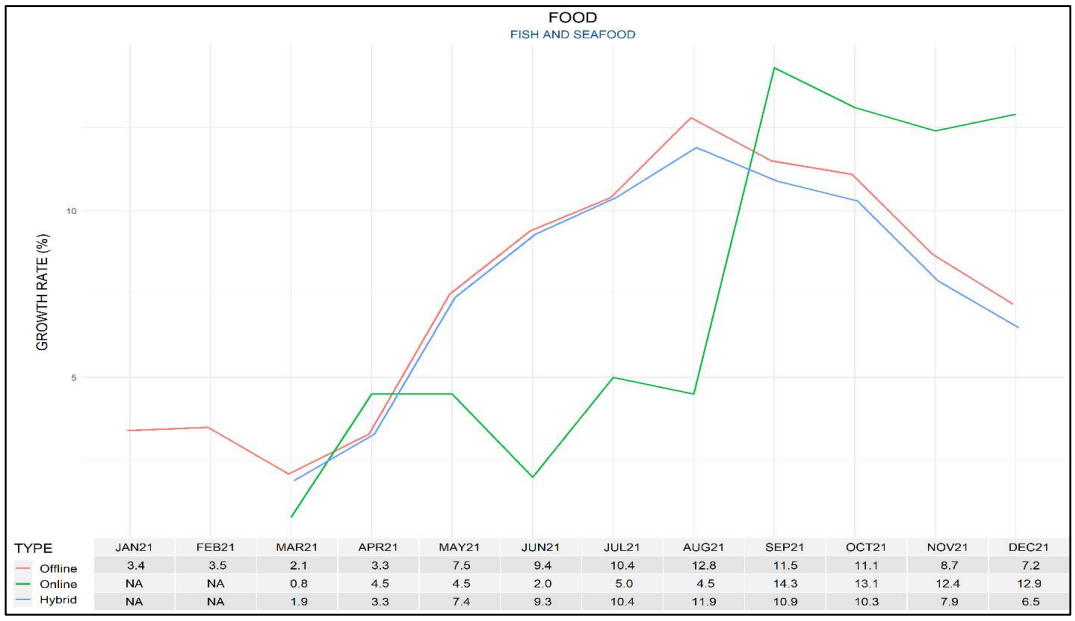
Figure 3. Month-on-Month Growth Rate of CPI for offline, online, and hybrid prices of Tobacco from February 2020 to December 2021

TYPE	FEB20	MAR20	APR20	MAY20	JUN20	JUL20	AUG20	SEP20	OCT20	NOV20	DEC20	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	0.1	0.1	0.1	0.0	0.0	0.2	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.2	0.0	0.0	0.0	0.1	-0.1	0.1	0.1	0.1
Online	NA	NA	-0.5	0.5	0.1	0.3	0.2	0.0	0.3	0.0	0.3	0.1	0.0	0.0	0.1	0.0	1.0	0.9	-0.1	-0.1	-1.0	0.3	0.1
Hybrid	NA	NA	-0.5	0.5	0.1	0.3	0.2	0.0	0.4	0.0	0.3	0.1	0.0	0.0	0.2	0.0	1.0	0.9	-0.1	-0.1	-1.0	0.2	0.1

Figure 4. Month-on-Month Growth Rate of CPI for offline, online, and hybrid prices of Garments from February 2020 to December 2021

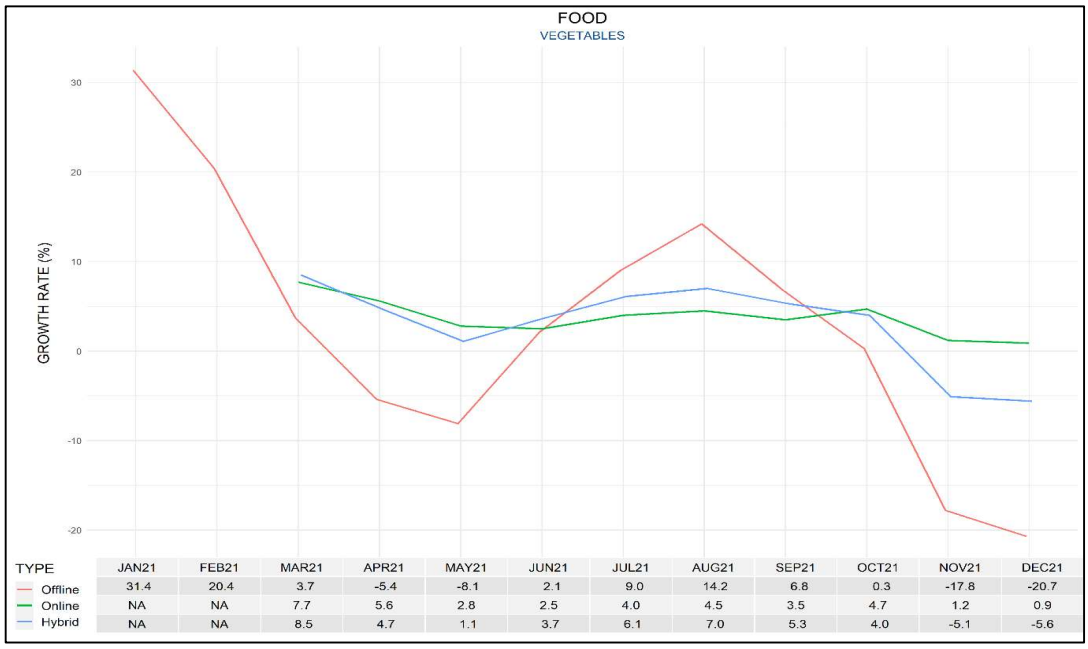
Figure 1 shows that although the growth rate of CPI for online prices fluctuates throughout the series while the growth rate for hybrid prices follows the trend of the offline prices. This, however, is not evident in Figure 2. Meanwhile, the month-on-month growth rates of online and hybrid prices are comparable with that of the offline for tobacco as shown in Figure 3. Moreover, Figure 4 shows that the growth rates for the computed CPI for both online and hybrid prices follow the trend of CPI for offline prices.

To examine the annual rate of change of the computed CPI, their year-on-year growth rates were also obtained. Figures 5, 6, 7, and 8 shows the year-on-year growth rate of fish and seafood, tobacco, and garments.



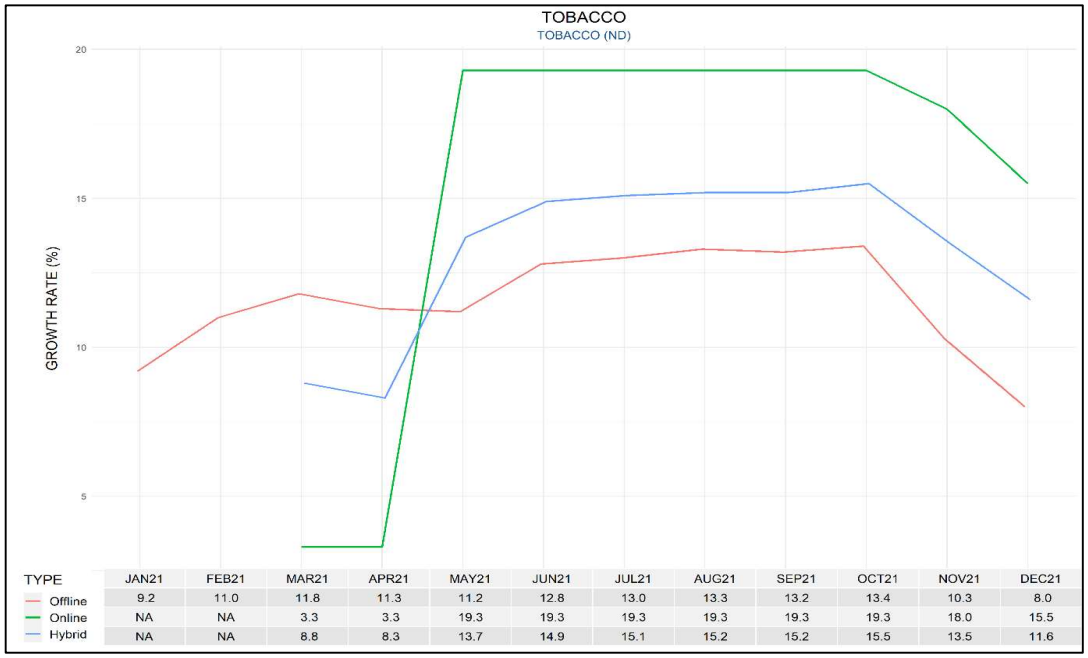
TYPE	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	3.4	3.5	2.1	3.3	7.5	9.4	10.4	12.8	11.5	11.1	8.7	7.2
Online	NA	NA	0.8	4.5	4.5	2.0	5.0	4.5	14.3	13.1	12.4	12.9
Hybrid	NA	NA	1.9	3.3	7.4	9.3	10.4	11.9	10.9	10.3	7.9	6.5

Figure 5. Year-on-Year Growth Rate of CPI for offline, online, and hybrid prices of Fish and Seafood from February 2020 to December 2021



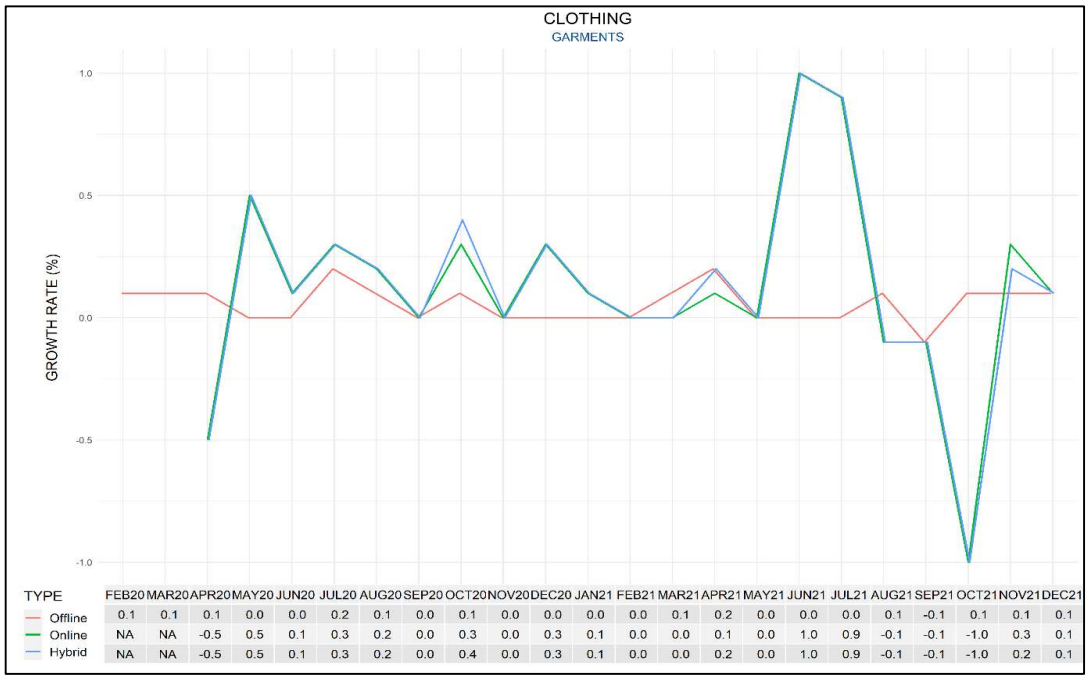
TYPE	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	31.4	20.4	3.7	-5.4	-8.1	2.1	9.0	14.2	6.8	0.3	-17.8	-20.7
Online	NA	NA	7.7	5.6	2.8	2.5	4.0	4.5	3.5	4.7	1.2	0.9
Hybrid	NA	NA	8.5	4.7	1.1	3.7	6.1	7.0	5.3	4.0	-5.1	-5.6

Figure 6. Year-on-Year Growth Rate of CPI for offline, online, and hybrid prices of Vegetables from February 2020 to December 2021



TYPE	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	9.2	11.0	11.8	11.3	11.2	12.8	13.0	13.3	13.2	13.4	10.3	8.0
Online	NA	NA	3.3	3.3	19.3	19.3	19.3	19.3	19.3	19.3	18.0	15.5
Hybrid	NA	NA	8.8	8.3	13.7	14.9	15.1	15.2	15.2	15.5	13.5	11.6

Figure 7. Year-on-Year Growth Rate of CPI for offline, online, and hybrid prices of Tobacco from February 2020 to December 2021



TYPE	FEB20	MAR20	APR20	MAY20	JUN20	JUL20	AUG20	SEP20	OCT20	NOV20	DEC20	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	0.1	0.1	0.1	0.0	0.0	0.2	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.2	0.0	0.0	0.0	0.1	-0.1	0.1	0.1	0.1
Online	NA	NA	-0.5	0.5	0.1	0.3	0.2	0.0	0.3	0.0	0.3	0.1	0.0	0.0	0.1	0.0	1.0	0.9	-0.1	-0.1	-1.0	0.3	0.1
Hybrid	NA	NA	-0.5	0.5	0.1	0.3	0.2	0.0	0.4	0.0	0.3	0.1	0.0	0.0	0.2	0.0	1.0	0.9	-0.1	-0.1	-1.0	0.2	0.1

Figure 8. Year-on-Year Growth Rate of CPI for offline, online, and hybrid prices of Garments from February 2020 to December 2021

Figure 5 shows that as in month-on-month growth rate, the year-on-year growth rate of the CPI for online prices of fish and seafood deviate from that of the offline. In contrast, the year-on-year growth rates for hybrid prices follow the same trend as the offline prices.

The trend of the year-on-year growth rates of the computed CPI for online and hybrid prices for vegetables follow the movement of offline. However, the fluctuations were not as severe as that of the offline CPI as evidenced in the peaks and throughs of the computed CPI for the offline prices as shown in Figure 6.

For tobacco, both online and hybrid CPIs follows the trend of the offline CPI but at varied levels. For garments, while similar trend was observed for online and hybrid CPIs, they both diverge from the offline CPI as shown in Figure 8.

5. Conclusions and Recommendations

Results of this study showed that for web scraped items in the market-basket of CPI for NCR, the least deviation from the offline CPI is obtained by using hybrid prices. It is also shown on the results that commodity groups which showed comparable CPI with that of the offline are those whose web scraped data are with relatively high percentage of exactly matched specifications as compared with the equivalently matched specifications, especially for items under agriculture and other items which have volatile prices.

Although challenging, it is recommended to web scrape data with exactly same specifications in the market-basket of CPI to ensure comparability of the computed indices using online prices and consequently, to capture the same behavior of growth rates of the computed CPI. In case where there are no commodities in the website with exact specification as that in the market basket, the hybrid computation of CPI can be used.

With regard to web scraping as data collection method, it was observed that this method of data collection presents advantages, such as efficiency and the provision of extending the price collection of commodities beyond those listed in CPI market basket. However, there are issues that need to be addressed such as the legality and ethics of web scraping. Although, at present, there are no laws in effect prohibiting the use of web scraper to collect data from the websites, it is still advised to refer to the terms of use of each website to ensure that no law is violated. Another point of consideration is the additional resources needed in terms of manpower who will monitor and process the scraped data, and additional computers dedicated to web scraping.

As this research is an initial attempt of the PSA in exploring Big Data and its potential as input in compiling official statistics, in this case, the CPI, it is recommended that more extensive research be done on this topic. Web scraping should be performed during the start of price collection for the new series of CPI and should cover the websites of the sample outlets or stores for price collection.

This will allow for a direct comparison of the movements of online and offline prices and will present a more reliable analysis of computed CPIs.

6. References

Boettcher, I. (2015). Automatic data collection on the Internet (web scraping). Statistics Austria. https://www.stat.go.jp/english/info/meetings/og2015/pdf/t1s2p6_pap.pdf

Cavallo, A. (2012). Online and official price indexes: Measuring Argentina's inflation. Journal of Monetary Economics (2012), 60(2), 152–165.

Cavallo, A., & Rigobon, R. (2016). The billion prices project: Using online prices for measurement and research. Journal of Economic Perspectives, 30(2), 151–178.

Data Sources : Handbook of Methods: U.S. Bureau of Labor Statistics. (2020, November 24). Retrieved July 01, 2022, from https://www.bls.gov/opub/hom/cpi/data.htm

Data Sources : Handbook of Methods: U.S. Bureau of Labor Statistics. (2020, November 24). Retrieved July 01, 2022, from https://www.bls.gov/opub/hom/cpi/data.htm

FAQ on Consumer Price Index (CPI). (n.d.). Base. Retrieved July 01, 2022, from https://www.singstat.gov.sg/find-data/search-by-theme/economy/prices-and-price-indices/related-info/faq-on-cpi

7. Appendices

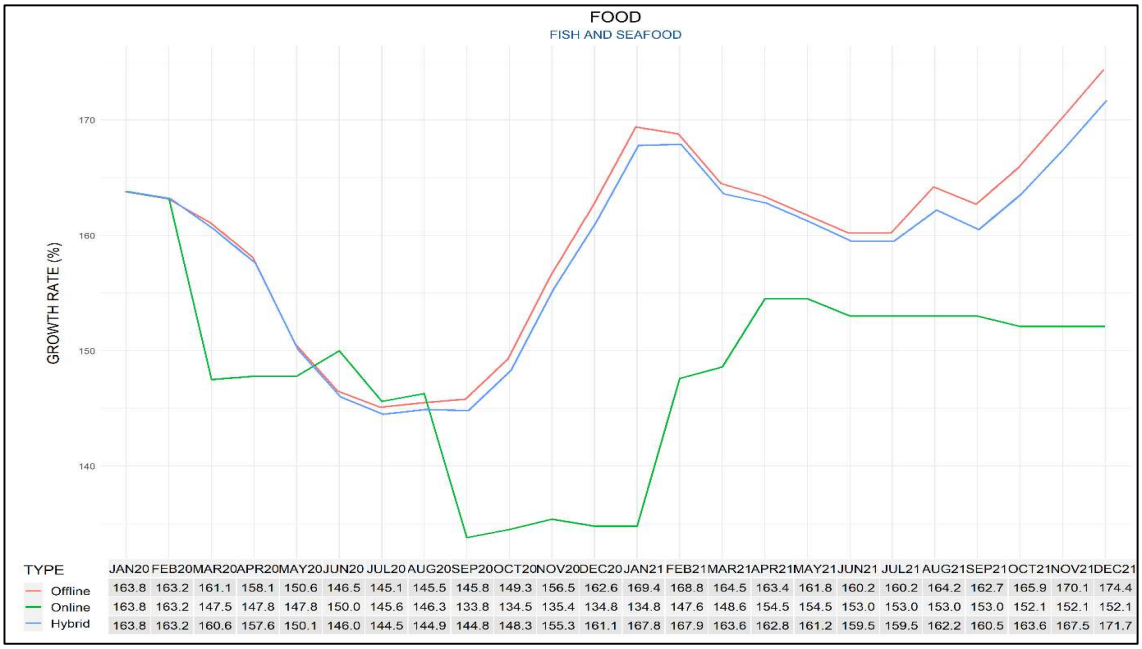


Figure 9. CPI for offline, online, and hybrid prices of Fish and Seafood from January 2020 to December 2021

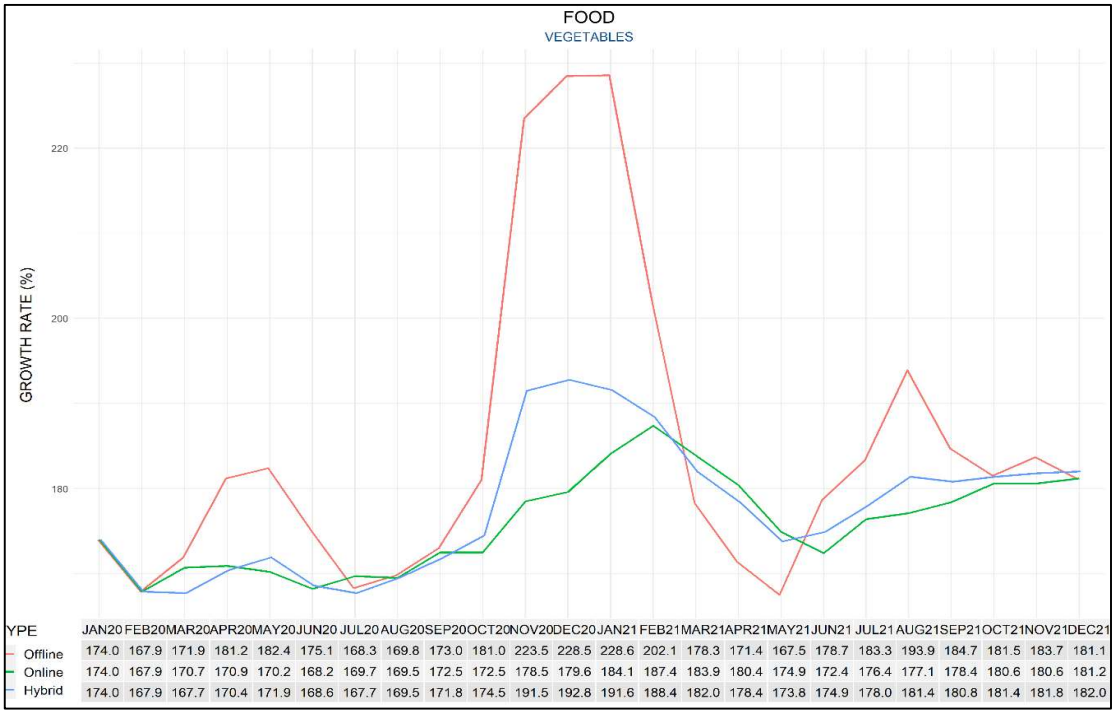
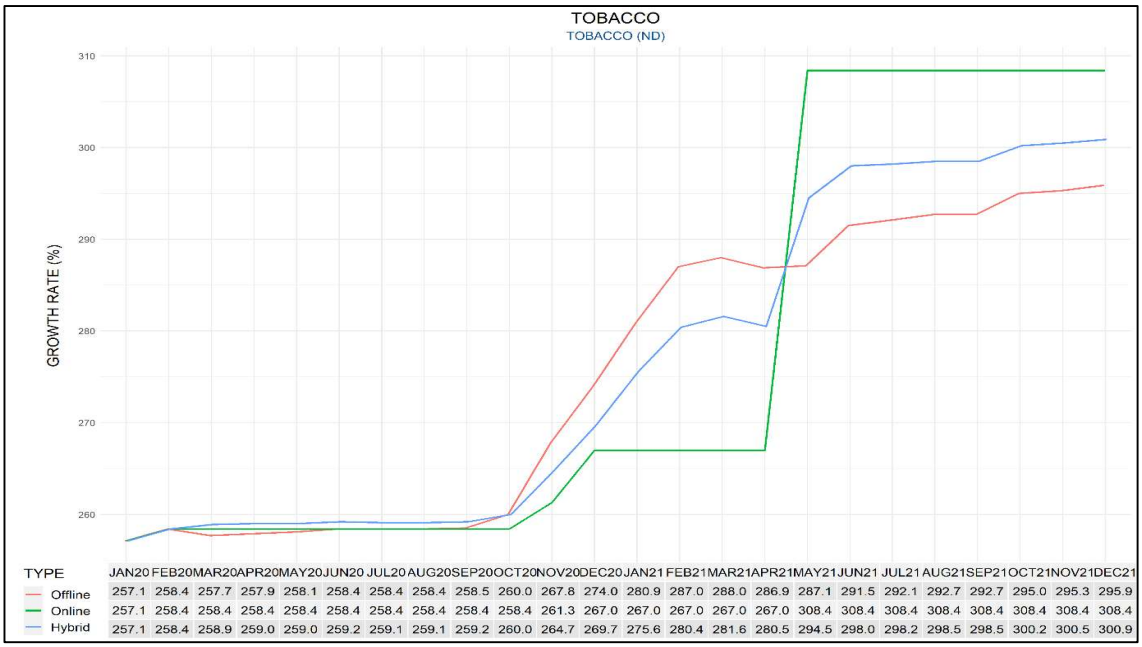
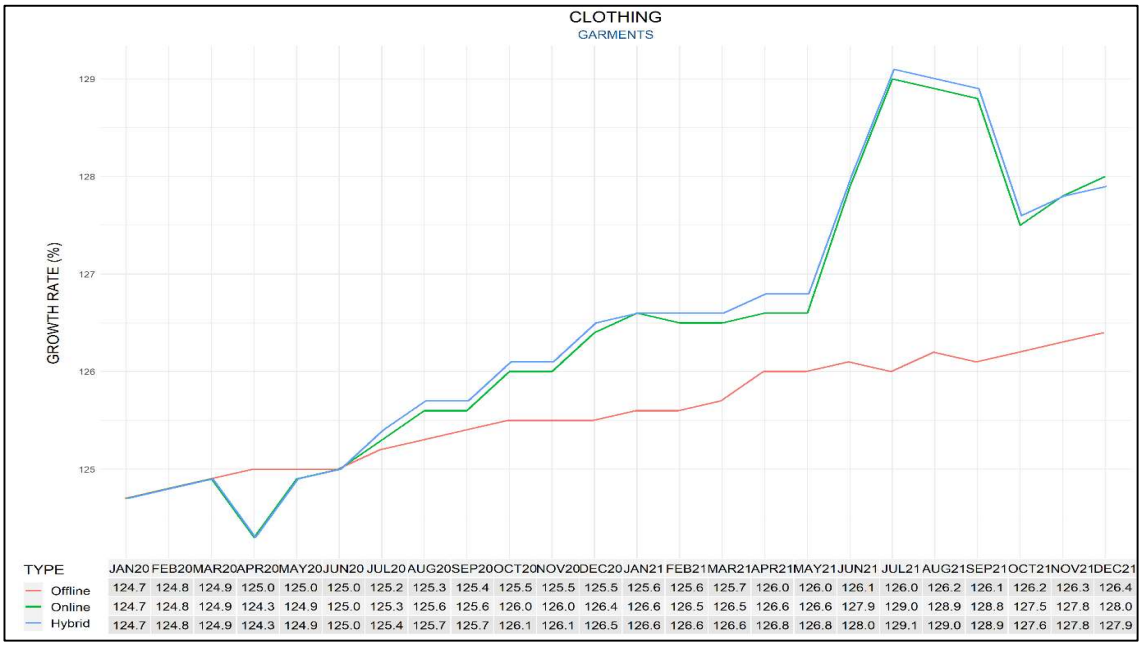


Figure 10. CPI for offline, online, and hybrid prices of Vegetables from January 2020 to December 2021



TYPE	JAN20	FEB20	MAR20	APR20	MAY20	JUN20	JUL20	AUG20	SEP20	OCT20	NOV20	DEC20	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	257.1	258.4	257.7	257.9	258.1	258.4	258.4	258.4	258.5	260.0	267.8	274.0	280.9	287.0	288.0	286.9	287.1	291.5	292.1	292.7	292.7	295.0	295.3	295.9
Online	257.1	258.4	258.4	258.4	258.4	258.4	258.4	258.4	258.4	258.4	261.3	267.0	267.0	267.0	267.0	267.0	308.4	308.4	308.4	308.4	308.4	308.4	308.4	308.4
Hybrid	257.1	258.4	258.9	259.0	259.0	259.2	259.1	259.1	259.2	260.0	264.7	269.7	275.6	280.4	281.6	280.5	294.5	298.0	298.2	298.5	298.5	300.2	300.5	300.9

Figure 11. CPI for offline, online, and hybrid prices of Tobacco from January 2020 to December 2021



TYPE	JAN20	FEB20	MAR20	APR20	MAY20	JUN20	JUL20	AUG20	SEP20	OCT20	NOV20	DEC20	JAN21	FEB21	MAR21	APR21	MAY21	JUN21	JUL21	AUG21	SEP21	OCT21	NOV21	DEC21
Offline	124.7	124.8	124.9	125.0	125.0	125.0	125.2	125.3	125.4	125.5	125.5	125.5	125.6	125.6	125.7	126.0	126.0	126.1	126.0	126.2	126.1	126.2	126.3	126.4
Online	124.7	124.8	124.9	124.3	124.9	125.0	125.3	125.6	125.6	126.0	126.0	126.4	126.6	126.5	126.5	126.6	126.6	127.9	129.0	128.9	128.8	127.5	127.8	128.0
Hybrid	124.7	124.8	124.9	124.3	124.9	125.0	125.4	125.7	125.7	126.1	126.1	126.5	126.6	126.6	126.6	126.8	126.8	128.0	129.1	129.0	128.9	127.6	127.8	127.9

Figure 12. CPI for offline, online, and hybrid prices of Garments from January 2020 to December 2021