Asia–Pacific Economic Statistics Week 2019
Integrating economic statistics in monitoring the Agenda 2030
17-21 June 2019 | Bangkok, Thailand

Seminar Component

Name of author: Reza Hadizadeh, Sedigheh Mirzaei, Hojjat Akbarian, Sara Taati, Leila Teymoorian

Organization: Statistical center of Iran (SCI)

Contact address: Dr. Fatemi Ave, Tehran, 1414663111, Islamic Republic of Iran

Contact phone: +982188959032, +982188959033

Email: reza.h.fuzzy@gmail.com, se_mirzaee@sci.org.ir, akbaryan_2013@yahoo.com, t.s.ec1983@gmail.com, lteymoorian1020@gmail.com

Title of Paper

Calculating PPI for the Arts, Entertainment and Recreation Group of Activities by Using online Data

Abstract

As the technology improves, the new methods and resources for statistical data collection replace the traditional ones. The web data, scanner data, big data and telephone data, all referred to as open data, are among the new sources of price data which have attracted the economists and statistical offices around the world for the advantages they hold. High frequency, easy data collection, low cost and high accuracy and fidelity are the merits of the open data. Given that some major services are marketed online, the available online data could be a valuable source of data collection. The present study is meant to calculate the price index as well as inflation for a number of items from the Arts, Entertainment and Recreation group of services, ISIC Rev. 4, Section R, by the online data. Data capture is conducted by web scrapping.

I. Contents

II. Introduction

Data collection is a major challenge in official statistics production. Nowadays along with progress of science and technology, new sources of data are replacing the old and traditional ones. They hold certain advantages over the traditional ones; they are associated with some pitfalls, however. Lower costs of data collection, higher quality, lesser respondent burden, faster availability and higher accuracy are among the merits of the new sources. The states pioneer in statistics are seeking to use these new sources instead of the old ones. Scanner data, open data and web scraped data are among the new sources of data each of which holds its own merits and demerits. Web scraping is a new method of data collection in which the required data is collected by the web scraping software from the available online data. In price data collection process, the collected data should enjoy some certain specifications. The available online data are qualitative for some goods and services and meet the required specifications. So, they could be used for production of price indices for those goods and services at lower costs and shorter time.

The Arts, entertainment and recreation category is a new category in ISIC classification for which a few countries have so far developed the capacity to produce the related price indices. The data for this category could be easily collected by web scraping method to be used for index production. So, in the present study the web scraping is used to produce the indices for the Arts, entertainment and recreation category as a pilot activity.

The sections of the paper are as follows: in section 2, we provide some empirical investigations on price inflation and web scraping. Section 3 dedicates to theoretical framework (formula of price index and web-scraping methods. Advantages& disadvantages). Section 4 contains description of section R in producer price index; Arts, entertainment and recreation and data. Section 5 we demonstrate the process of calculations and results. Finally, section 6 concludes.

III. Body of Text

A. The literature reviews

Bertolotto and Diego Aparicio (2016) forecast the CPI inflation by using goods' prices collected online through web scraping retailer's websites in multiple countries including the US , the UK, France, Germany, and Netherlands. Its advantage is releasing the non-negligible delay in the official CPI release. This paper recommends that simple autoregressive models augmented especially equal-weighted pooled forecasts as the best performing online models. Alberto Cavallo and Roberto Rigobon (2016) show that online prices can be successfully used as an alternative source of information for constructing consumer price indexes. They describe theirs work with online data at the Billion Prices Project at MIT and discuss key lessons for both inflation measurement and some fundamental research questions in macro and international economics. In particular, they show how online prices can be used to construct daily price indices in multiple countries. Cavallo (2013) uses online prices to study how online indices matchup with official statistics in five Latin American countries. The finds that while in some country such as Brazil, Chile, Columbia, and Venezuela online price indices approximate both the level and the main dynamics of official inflation others like Argentina 's web inflation was nearly three times higher than official statistics. Yukhymenko & others (2018) test the association between online price indices and official statistics. They find that online inflation is generally consistent with official estimates, but the matching capability varies across sub-indexes. Although they find that online prices may indeed represent new information that is not captured by official statistics.

B. Theoretical framework

1. Price index

In this paper we use Laspeyres formula for calculating price index for the Arts, entertainment and recreation category that it is followed below:

$$\frac{\sum_{i=1}^{n} P_i^t Q_i^0}{\sum_{i=1}^{n} P_i^0 Q_i^0} \quad \textbf{(1)}$$

In this formula, quantities are fixed at the base period. For practical use, ***(2***) is transformed as follows;

$$\frac{\sum_i \frac{p_{it}}{p_{i0}} w_{i0}}{\sum_i w_{i0}} \qquad (w_{i0} = p_{i0} q_{i0}) \qquad \textbf{(2)}$$

w_{i0}, represent the weights of services in the base year and $\frac{P_{i,k}^t}{P_{i,k}^0}$ is the relative price being calculated in Jones's methods.

2. Online data

Alternative sources of data have the potential to greatly improve the quality and efficiency of consumer price indices. One such data source is point of *sale scanner data*.

(Breton & others, 2015) Fast growth of mobile applications for various services is providing a big source of price data. Some examples are applications for applying nursing services at home, technical and house maintenance services, transportation and ... In addition, a large and growing share of retail prices are posted *online* all over the world. Retailers show these prices either to sell online or to advertise prices to potential offline customers. This source of data provides an important opportunity for economists who want to study price dynamics, yet it has been largely untapped because the information is widely across thousands of web pages and retailers. Furthermore, there is no historical record of these prices, so they must be continually collected over time. (Cavallo, 2018)

Web scraping is data scraping used for extracting data from websites. (Boeing & Waddell, 2016)

The term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis. Web scraping is used for contact scraping, and as a component of applications used for web indexing, web mining and specially for online price change monitoring and price comparison. Web scraping offers many benefits. It could provide an opportunity for us to *automate* some aspects of price collection. Web scraping also has potential for use in other areas, such as the collection of attribute information for the quality adjustment of technological items, also known as *hedonics*.

Price and attribute collection procedures place a heavy burden on official resources. Web scraping has the potential to offer savings in these areas. However, these savings need to be considered alongside potentially high maintenance costs. Web scraping also provides an opportunity to improve *quality* by increasing the number of price quotes feeding into the index, and to produce indices on a more frequent basis. Perhaps most importantly, it gives us the opportunity to explore big datasets and develop methodologies that are appropriate for the volume of data. These experiences will be invaluable should other sources of big data (e.g. point of sale scanner data) be introduced into consumer price statistics. (Breton& others)

The scraping methodology has three steps. First, at a fixed time each day, a software program downloads a selected list of public web pages where product and price information are shown. These pages are individually retrieved using the same web address (URL) every day. Second, the underlying code is analyzed to locate each piece of relevant information. This is done by using special characters in the code that identify the start and end of each variable, which have been placed by the page programmers to give the website a particular look and feel. For example, prices may be shown with a dollar sign in front of them and enclosed within and tags. Third, the software stores the scraped information in a database that contains one record per product per day. These variables include the product's price, the date, and category information. (Cavallo, 2018)

Scraped data have some important *advantages*. First, these data sets contain posted daily prices that are free from unit values, time averaging, and more imputations. The daily data are also useful to better identify sales and other price changes that might be missed with

monthly data. Second, detailed information can be obtained for all products sold by the sampled retailers instead of a few (as in CPI data) or selected categories (as in scanner data). Third, there are no censored or imputed price spells in scraped data. Prices are recorded from the first day they are offered to consumers until the day they are discontinued from the store. In CPI, by contrast, there are frequent imputations and forced substitutions when the agent surveying prices cannot find the item. Fourth, scraped data can be collected remotely in any country where price information can be found online. Fifth, scraped data sets are comparable across countries, with prices that can be collected for the same to perform simultaneous cross-country analyses. Finally, scraped data are available in real time, without any delays in accessing and processing the information. Eventually this could be used by central banks to obtain real-time estimates of stickiness and related statistics.

There are, have ever, some *disadvantages* with scrape data: First, they typically cover a much smaller set of product categories than CPI prices. While this is enough to demonstrate the effect of measurement errors on pricing statistics, the quantitative findings on stickiness and size of changes shown here should not be viewed as representative of services and other sectors that cannot yet be covered with online data. Second, the data come only from large multichannel retailers that sell both online and offline. Currently the vast majority of retail sales take place in this type of retailer, but in principle, this may represent a form of sampling bias compared to the CPI (though not due to the online nature of the data. Finally, a major disadvantage of scraped data is the lack of information on quantities sold. In measuring stickiness, quantities are useful in obtaining detailed expenditure weights for narrowly defined categories. (Cavallo, 2018)

C. Methodology & data

1. Section R; Arts, entertainment and recreation

The International Standard Industrial Classification of All Economic Activities (ISIC) is the international reference classification of productive activities. Its main purpose is to provide a set of activity categories that can be utilized for the collection and reporting of statistics according to such activities.

This fourth revision of ISIC (ISIC, Rev.4) is the outcome of a review process that spanned several years and involved contributions from many classifications' experts and users around the world. This process resulted in an ISIC structure that is more detailed than the previous version, responding to the need to identify many new industries separately. This is especially applicable in the case of services. Moreover, the relevance of the Classification has been enhanced with the introduction of new high-level categories to better reflect current economic phenomena. A new section entitled "Arts, entertainment and recreation (Section R)" is one such innovation.

This section includes a wide range of activities to meet varied cultural, entertainment and recreational interests of the general public, including live performances, operation of museum sites, gambling, sports and recreation activities.

Table 1. Classification of the section R

Division	Group	Class	Description
Division 90			Creative, arts and entertainment activities
	900	9000	Creative, arts and entertainment activities
Division 91			Libraries, archives, museums and other cultural activities
		9101	Library and archives activities
		9102	Museums activities and operation of historical sites and buildings
		9103	Botanical and zoological gardens and nature reserves activities
Division 92			Gambling and betting activities
	920	9200	Gambling and betting activities
Division 93			Sports activities and amusement and recreation activities
	931		Sports activities
		9311	Operation of sports facilities
		9312	Activities of sports clubs
		9319	Other sports activities
	932		Other amusement and recreation activities
		9321	Activities of amusement parks and theme parks
		9329	Other amusement and recreation activities n.e.c.

2. Data collection

In this section, we will discuss data collection process for section R in Iran. Because gambling is prohibited in Iran, no item or service is selected from the 9200 class (Gambling and betting activities). Therefore, the 9200 class does not participate in the calculation of producer price index of section R.

According to Table 1, the selected items for the 9000 class (Creative, arts and entertainment activities) are theater and concert. The price data for these items are collected by web scraping due to given the advantages and disadvantage of online data mentioned in previous section. The site of Tiwall[1] provide information on theatrical performances and

[1] https://www.tiwall.com/ Tiwall is the most widely-used website in online reservation of theaters and concerts in Iran

concerts in Iran and also sell theater and concert tickets online. These items are chosen which have these characteristics:

- It is possible to access them as online applications nationwide.
- The specifications impacting on their price could be extracted and classified.
- Their online prices are the same as the real market.

The selected item for the 9101 class (Library and archives activities) is library membership charge; and the selected items for the 9102 class (Museums activities and operation of historical sites and buildings) are museum and historical sites. The price data of them are collected from registered data. The selected item for the 9103 class (Botanical and zoological gardens and nature reserves activities) is tickets for recreational sights which is collected through surveys. The selected items for the 9311 class (Operation of sports facilities) are "Bodybuilding club membership", "Pool ticket" and "Rental of mini-soccer fields". The price data for the first two items are collected monthly in Consumer Price Index survey and the data for "Renting mini-soccer fields" is collected through surveys. It is worth mentioning no items are selected for 9312, 9319, 9321 and 9329 classes, as their weights are very low in section R in Iran.

The **Table 1**Figure 1 shows how to select items in the section R.

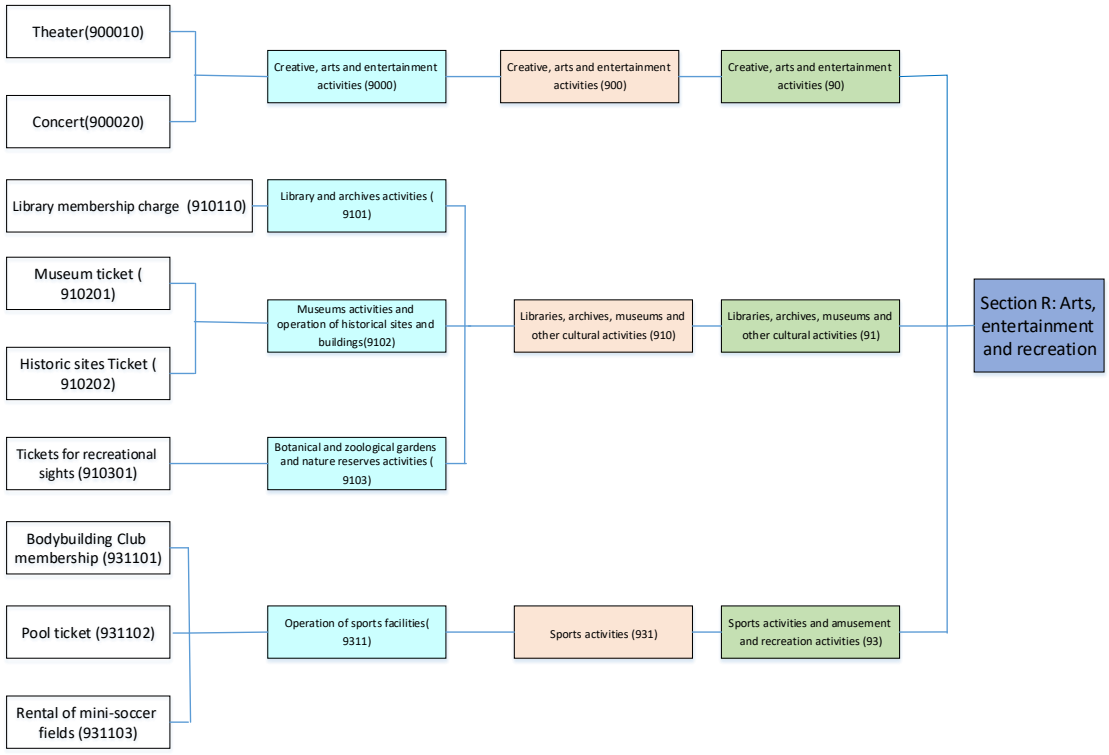

Figure 1: The selected items for section R

Using data of website named Tiwall, the price of show was extracted from website by web scraping method and Scrappy and BeautifulSoup packages in Python. It has planned to run on low website traffic hours. The prices, besides, was derived based on their specified characteristics, for example, hall name and different prices for place in hall.

Since there is a variety of goods and services in section R it is impossible to collect from one data source. Meanwhile it is hard to have geographic coverage for all items; each website has its own way to build its HTML webpage. That is, it needs to design Software separately which is costly and further of this paper.
Figure 2 presents a simplified flow diagram of the current system. It shows that each data flow passes particular cleansing stage and then joins to other data flows.

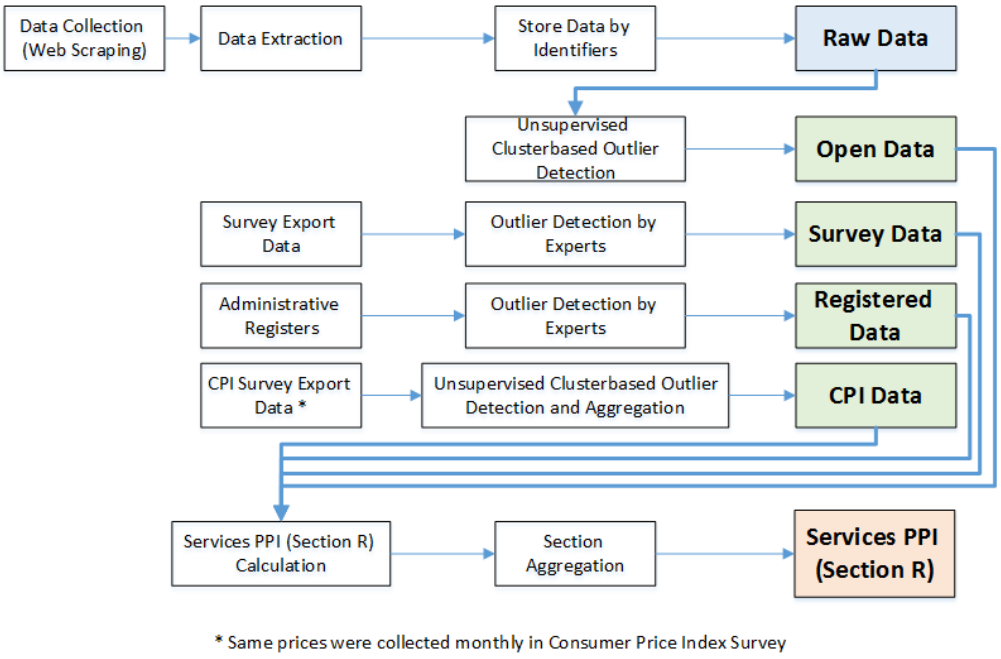

Figure 2: process of collecting price information and generating price index for section R

D. Empirical results

1. Calculation of Price Index

This section presents the results of estimating the producer price index of section "Arts, entertainment and recreation (section R)" at different levels such as items, class, division and the total index of section R. For summarizing in charts and tables, the following (Table 2) abbreviations are used. In this article, the year and chapter are based on the solar calendar.

Table 2. Abbreviation of the items in charts and tables

Item name	ISIC	Item code in software
Theater	900010	1208
Concert	900020	1209
Library membership charge	910110	1207
Museum ticket	910201	1201
Historical sites ticket	910202	1204
Tickets for recreational sights	910301	1205
Bodybuilding club membership	931101	1202
Pool ticket	931102	1203
Rental of mini-soccer fields	931103	1206

In the Table3 summarizes the calculated price index for items of section R.

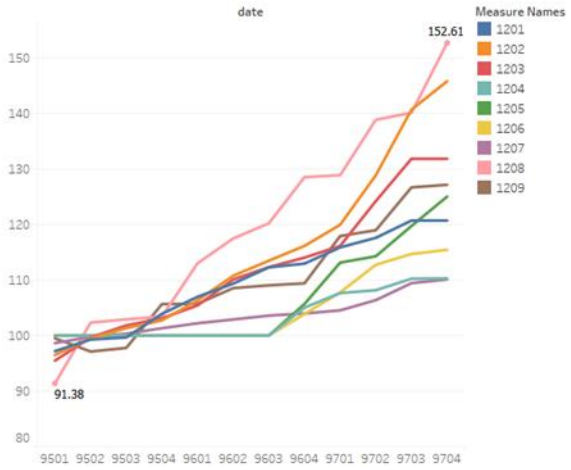

Figure **3** shows the trend price index for these items. The index of these items has ascending trend. In addition, the theater has more acceleration over than the other items in the period 1395 to 1397 and the least growth belongs to the "library membership charge".

Table 3. Seasonally rice index of items (1395-1397)

Items / Date	1208	1209	1207	1201	1204	1205	1202	1203	1206
9501	91.4	99.5	98.6	97.2	100.0	100.0	96.5	95.5	100.0
9502	102.3	97.1	99.7	99.3	100.0	100.0	99.5	99.6	100.0
9503	102.9	97.8	100.3	99.6	100.0	100.0	101.3	101.8	100.0
9504	103.4	105.6	101.3	103.9	100.0	100.0	102.7	103.1	100.0
9601	113.0	105.7	102.2	107.0	100.0	100.0	106.3	105.4	100.0
9602	117.4	108.5	102.9	109.3	100.0	100.0	110.7	110.1	100.0
9603	120.2	109.0	103.6	112.3	100.0	100.0	113.5	112.3	100.0
9604	128.5	109.4	104.0	112.9	105.0	105.7	116.1	114.0	103.9

9701	128.8	117.9	104.5	115.8	107.6	113.1	119.9	116.1	107.8
9702	138.8	118.9	106.4	117.6	108.1	114.3	128.8	124.2	112.7
9703	140.1	126.6	109.4	120.7	110.2	119.7	140.6	131.8	114.7
9704	152.6	127.1	110.1	120.7	110.3	125.0	145.7	131.8	115.4

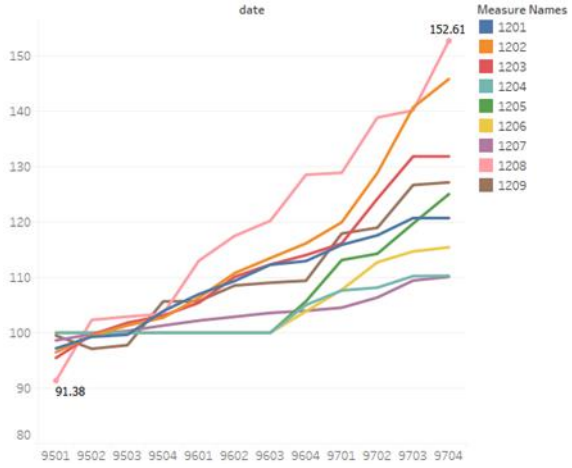

Figure 3: Trend of seasonally price index of Items (1395-1397)

After calculating price index of items of section R, we can calculate price index Class, Division and total index of section R. In the Table 4 summarizes the calculated price index of Class of section R. Figure 4 shows trend of seasonally price index of class (1395-1397). As expected, the Class 9000 has item "theater" so has the most growth compared to other Classes, and Class 9011 has item "Library membership charge" so has the lowest growth rate.

Table 4. Seasonally rice index of Class (1395-1397)

Classes / Date	1208	1209	1207	1201	1206
9501	92.3	98.6	99.0	100.0	96.3
9502	101.7	99.7	99.7	100.0	99.6
9503	102.3	100.3	99.9	100.0	101.4
9504	103.6	101.3	101.4	100.0	102.7
9601	112.2	102.2	102.6	100.0	105.7
9602	116.5	102.9	103.5	100.0	110.0
9603	119.0	103.6	104.6	100.0	112.4
9604	126.4	104.0	108.0	105.7	114.8
9701	127.7	104.5	110.7	113.1	118.0
9702	136.7	106.4	111.6	114.3	126.4
9703	138.6	109.4	114.1	119.7	136.3
9704	149.9	110.1	114.1	125.0	139.4

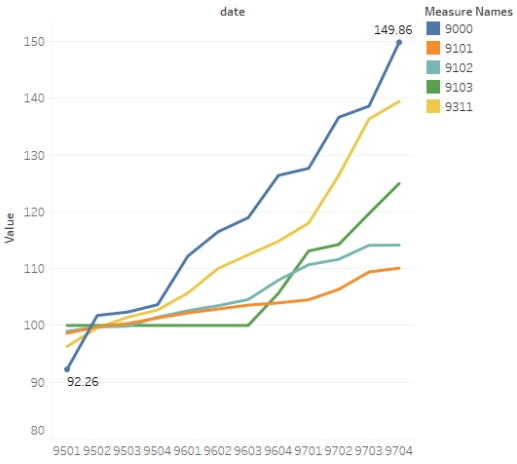

Figure 4: Trend of seasonally price index of Classes (1395-1397)

After calculating price index of Classes of section R, we can generate price index of Division of section R. In the Table 5 summarizes the calculated price index of Divisions. Figure 5 shows trend of seasonally price index of Division (1395-1397). As is remarkable, the Divisions have ascending trend and Division 90 has the most growth.

Table 5. Seasonally price index of Division (1395-1397)

Division / Date	90	91	93
9501	92.3	99.3	96.3
9502	101.7	99.8	99.6
9503	102.3	100.0	101.4
9504	103.6	100.9	102.7
9601	112.2	101.5	105.7
9602	116.5	102.0	110.0
9603	119.0	102.6	112.4
9604	126.4	106.4	114.8
9701	127.7	110.7	118.0
9702	136.7	111.8	126.4
9703	138.6	115.6	136.3
9704	149.9	117.8	139.4

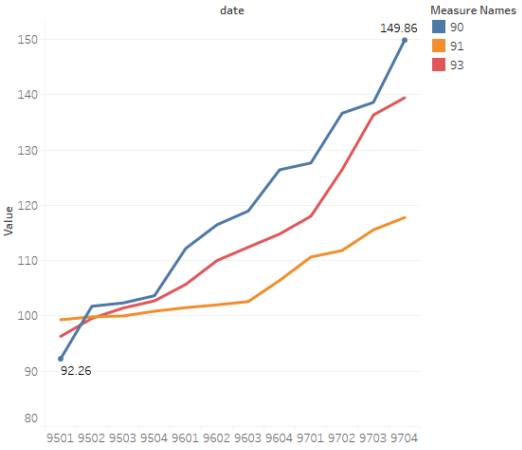

Figure 5: Trend of seasonally price index of Divisions (1395-1397)

The last step is to calculate the producer price index section R. The calculated price index of section R are presented in Table 6 and the Figure 6 shows that price index of section R has risen from 49 to 141 and has ascending trend.

Table 6. Seasonally price index of section R (1395-1397)

Date	Section R
9501	95
9502	100
9503	102
9504	103
9601	107
9602	111
9603	114

9604	118
9701	120
9702	128
9703	135
9704	141

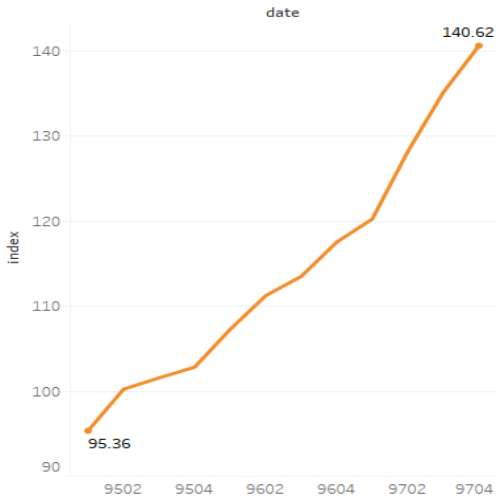

Figure 6: Trend of seasonally price index of section R (1395-1397)

2. Calculation of the inflation rate

After calculating price index, now, we can calculate inflation rate for Items, Classes, Divisions and section R. In the Table 7, Table 8, Table 9 and Table 10 summarized inflation rate of Items, Classes, Divisions and section R, respectively. And also Figure 7, Figure 8, Figure 9 and Figure 10 show that trend of inflation rate of Items, Classes, Divisions and section R, respectively.

The lowest inflation rate is related to the concert item that is located in summer 1395 and the highest inflation rate is related to the theater item in the summer 1395 and equals 12%.

Table 7. Seasonally rice index of items (1395-1397)

Items / Date	1208	1209	1207	1201	1204	1205	1202	1203	1206
9502	12.0	-2.4	1.1	2.2	0.0	0.0	3.1	4.4	0.0
9503	0.6	0.7	0.6	0.4	0.0	0.0	1.8	2.2	0.0
9504	0.5	8.1	1.0	4.3	0.0	0.0	1.4	1.3	0.0
9601	9.3	0.1	0.9	2.9	0.0	0.0	3.5	2.2	0.0
9602	4.0	2.7	0.7	2.2	0.0	0.0	4.2	4.4	0.0
9603	2.3	0.5	0.7	2.7	0.0	0.0	2.5	2.0	0.0
9604	6.9	0.3	0.4	0.6	5.0	5.7	2.3	1.5	3.9
9701	0.3	7.8	0.5	2.6	2.5	7.0	3.3	1.9	3.7
9702	7.7	0.9	1.8	1.5	0.5	1.0	7.4	6.9	4.6
9703	0.9	6.5	2.9	2.6	1.9	4.8	9.2	6.1	1.7
9704	9.0	0.4	0.6	0.0	0.1	4.4	3.6	0.0	0.6

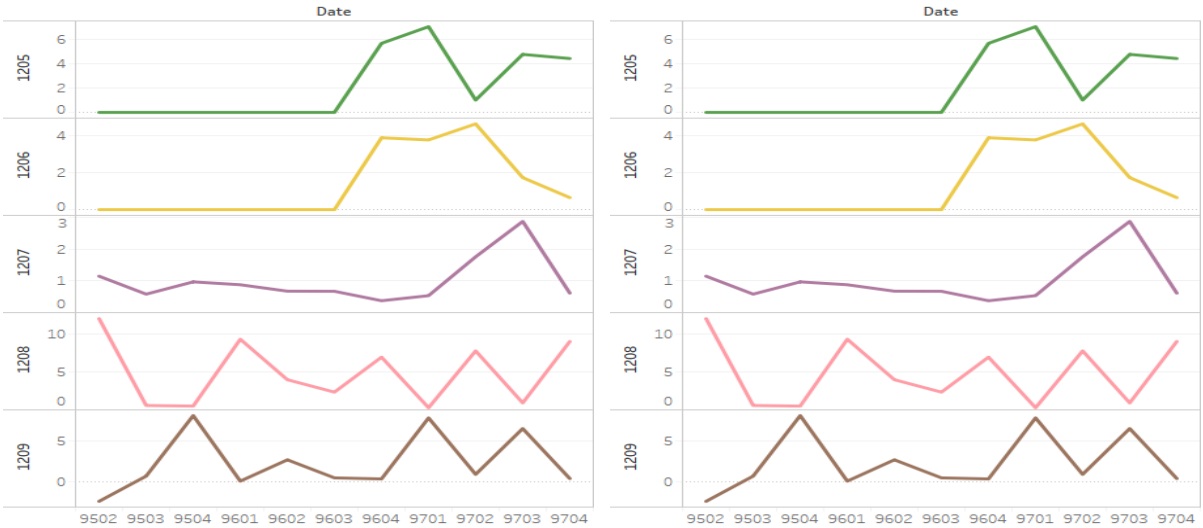

Figure 7: Seasonally rice index of items (1395-1397)

Table 8. Seasonally rice index of Class (1395-1397)

Classes / Date	1208	1209	1207	1201	1206
9502	10.29	1.15	0.79	0.00	3.39
9503	0.58	0.58	0.14	0.00	1.88
9504	1.28	0.97	1.58	0.00	1.28
9601	8.25	0.88	1.12	0.00	2.90
9602	3.82	0.68	0.86	0.00	4.08
9603	2.13	0.67	1.06	0.00	2.19
9604	6.26	0.37	3.25	5.67	2.12
9701	0.98	0.53	2.53	7.04	2.79
9702	7.05	1.76	0.86	1.01	7.14
9703	1.43	2.87	2.20	4.76	7.82
9704	8.12	0.61	0.03	4.42	2.28

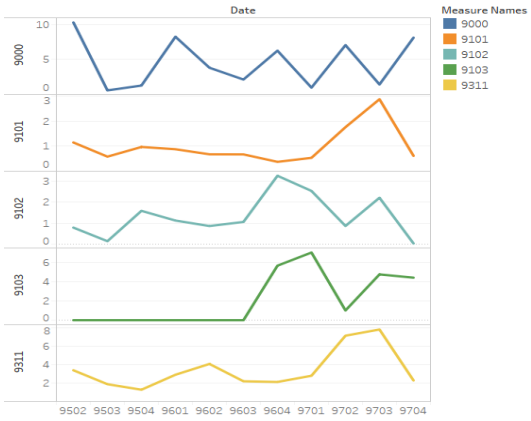

Figure 8: Trend of seasonally price index of Classes (1395-1397)

Table 9. Seasonally price index of Division (1395-1397)

Division / Date	90	91	93
9502	10.3	0.5	3.4
9503	0.6	0.2	1.9

9504	1.3	0.9	1.3
9601	8.2	0.6	2.9
9602	3.8	0.5	4.1
9603	2.1	0.6	2.2
9604	6.3	3.7	2.1
9701	1.0	4.0	2.8
9702	7.1	1.1	7.1
9703	1.4	3.3	7.8
9704	8.1	1.9	2.3

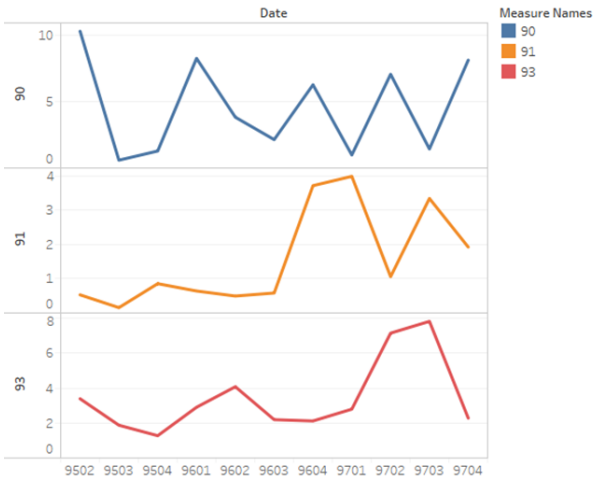

Figure 9: Trend of seasonally price index of Divisions (1395-1397)

Table 10. Seasonally price index of Division (1395-1397)

Date	Section R
9502	5.1
9503	1.3
9504	1.2
9601	4.3
9602	3.7
9603	2.0
9604	3.6
9701	2.3
9702	6.6
9703	5.4
9704	4.1

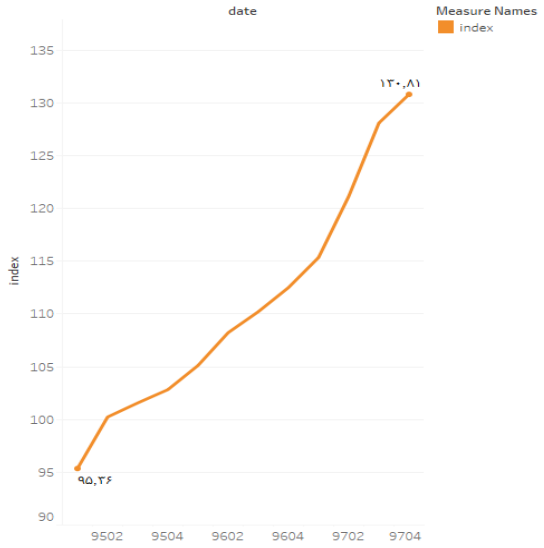

Figure 10: Trend of seasonally price index of section R (1395-1397)

IV. Conclusion

This paper proposes a method for generating the price index in a new section of ISIC rev. 4 using modern methods, so that it does not cost much to the statistical centers. One of these sections is "arts, entertainment and recreation(R)" for which few countries have produced the producer price index. Price information can be collected for items of this section using various sources. Section R includes entertainment and recreation services for which the pricing information could be obtained from the web. Nowadays, most statistical agencies in the world have taken steps to collect prices by this method.

The present paper seeks to find out that how the society could use the web scraping to collect price data. Although this method is applicable for limited goods and services, Advantages of using new methods can be a reason to use them for improving current processes. The new web scraping method used in this paper for collecting prices contains some merits and some demerits:

Benefits:

1. By using different sources to collect the price.
2. Increase the number of price patches per item which will increase the accuracy of results.
3. Reduce the cost of data collection.
4. Optimal use of existing community capacities.

Costs:

1. Providing new and higher-tech software for collecting, cleansing and classifying information.
2. Software maintenance and redevelopment due to changes in source sites.

Opportunities:

1. Ability to use other features along with prices such as descriptions and ...
2. Possibility to generate price index in the new section.

Risks:

1. Copyright must be coordinated with the site or the custodians of the site have the right to provide users with the information to use.
2. The probability of increasing the traffic burden of the source site when collecting price data.

V. References

[1] Bertolotto, M., D. Aparicio (2016). Forecasting Inflation with Online Prices. MIT.

[2] Boeing, G., P. Waddell (2016). New insights into rental housing markets across the United States: Web scraping and analyzing craigslist rental listings. California: Journal of planning education and research.

[3] Breton, R., G. Clews, L. Metcalfe, N. Milliken, Ch. Payne, J. Winton and A. Woods (2015). Research indices using web scraped data. London: Office for National Statistics.

[4] Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. Journal of Monetary Economics, 152–165.

[5] Cavallo, A., R. Rigobon (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. Journal of Economic Perspectives, 30 (2): 151-78.

[6] Cavallo, A. (2018). Scraped data and sticky prices. Cambridge: The Review of Economics and Statistics.

[7] Faryna, O., O. Talavera, T. Yukhymenko (2018). What drives the difference between online and official price indexes? Visnyk of the National Bank of Ukraine.

[8] United Nations (2008). International Standard Industrial Classification of All Economic Activities (ISIC Revision 4). Available from https://unstats.un.org/unsd/publication/seriesM/seriesm_4rev4e.pdf