PRACTICAL APPLICATION OF WEB SCRAPING TECHNOLOGY IN THE CALCULATION OF THE CONSUMER PRICE INDEX IN UZBEKISTAN

Ismailova Shakhnoza Uktamovna
Independent researcher at the Institute of Personnel Training and Statistical Research of the National Statistics Committee of the Republic of Uzbekistan
E-mail:shaxnoza.ismailova87@bk.ru

Abstract: *This thesis examines the practical application of web scraping technology in the compilation of the Consumer Price Index (CPI) in Uzbekistan. The expansion of the digital economy and the rapid growth of online retail have increased the need for official statistics to use alternative data sources. The paper outlines the stages of collecting online price data with Python-based tools, matching product identifiers to COICOP 2018 codes, building a dedicated classifier, integrating web-scraped data with other price observations, and calculating final indices in Stata. Drawing on both international experience and current practice in Uzbekistan, the study argues that web scraping is an effective instrument for complementing conventional price observation, improving timeliness, and expanding data coverage.*

Keywords: *consumer price index, web scraping, official statistics, online prices, Python, Stata, COICOP 2018, classifier, inflation, price observation.*

INTRODUCTION

The Consumer Price Index is one of the most important indicators in macroeconomic analysis. It is widely used to measure inflation, inform monetary policy, and assess changes in household real income and living standards. For this reason, the quality, reliability, and representativeness of the underlying price data are of central importance.

In conventional practice, CPI price data are collected through direct observation in retail outlets and service establishments. However, the digitalization of the economy, the expansion of online retail platforms, and the growing volume of price information published on the internet have created a need for new approaches to price observation. In particular, the active formation of prices for certain goods and services in the online segment makes it necessary to enrich traditional statistical observation with digital data sources.

Against this background, web scraping has emerged as a promising instrument for official statistics. It enables the automated collection, structuring, and preparation of online price information for statistical use. Eurostat's practical guidelines for the HICP also present web scraping as a systematic approach that combines legal, technological, classificatory, and validation stages [1].

The essence and statistical significance of web scraping technology

Web scraping is a method of automatically extracting data from websites, online platforms, electronic catalogues, and other digital sources using specialized software tools. In price statistics, this technology makes it possible to collect product names, prices, identifiers, units of measurement, brands, packaging formats, and other descriptive attributes on a large scale.

The statistical importance of this approach lies primarily in its ability to increase both the volume of data and the frequency of data collection. Whereas traditional observation records prices at selected outlets and at predetermined intervals, web scraping can retrieve online prices from a much broader set of sources and on a more frequent basis. This makes it possible to monitor price dynamics in greater detail, identify short-term fluctuations, and carry out more timely analysis for specific product groups.

At the same time, data obtained through web scraping are not immediately ready for direct use in official statistics. They must be cleaned, deduplicated, standardized, classified, and statistically validated. In this sense, web scraping should not be viewed as a stand-alone substitute for traditional observation, but rather as a complementary approach that strengthens and enriches the existing system.

International experience

International practice shows that web scraping is becoming an increasingly important data source in the compilation of consumer price indices. Eurostat's practical guidelines for the HICP explain the main stages of introducing web scraping into price statistics, including legal foundations, technological tools, product coverage, classification, validation, and the integration of scraped data into index compilation [1].

The UK Office for National Statistics began methodological work on web scraping in the 2010s and developed experimental indices for categories such as clothing. ONS materials emphasize that web-scraped data can improve both the quality and the efficiency of consumer price statistics, particularly in segments characterized by strong seasonality and rapid product turnover [2].

Statistics Netherlands has gone further by discontinuing manual in-store retail price observation and using scanner data together with web scrapers for CPI compilation. According to CBS, this transition improved index quality while also reducing costs [3].

The Statistics Bureau of Japan also openly reports the use of web scraping in certain service segments. Its official CPI Q&A states that the index for overseas package tours has once again been based on prices collected through web scraping since the January 2024 results [4].

In addition, ESCAP has developed Python-based learning resources on web scraping for CPI statistics, covering both methodological principles and practical aspects of data collection and processing [8]. Taken together, these examples show that web scraping is no longer merely experimental in official price statistics; it is increasingly becoming an institutionalized source of statistical information.

The mechanism for applying web scraping in Uzbekistan's practice

In Uzbekistan, the use of web scraping in CPI compilation broadens the price information base and adapts the observation system to the digital environment. This approach is particularly relevant for monitoring prices published on online retail platforms, the websites of large retail chains, and electronic catalogues.

In practice, online price data are collected using web scraping tools developed in Python. In particular, price information is automatically extracted from the Media Park website on a weekly basis. At the same time, work is underway to expand this approach through agreements with more than ten additional internet resources, and the technical infrastructure and software solutions required for such expansion have already been prepared.

To ensure the statistical usability of the collected data, the identifier assigned to each product item on the website is extracted and matched to COICOP 2018 codes. COICOP 2018 is recommended by the United Nations Statistics Division as an international reference classification for grouping household consumption expenditures by purpose [5]. This matching process makes it possible to build a dedicated classifier that organizes online price data by consumption group.

The web-scraped price data are then integrated into a unified database together with other collected price observations. At the next stage, Stata is used for statistical cleaning, grouping, and averaging. On the basis of these processed aggregates, the final indices are also calculated in Stata. This ensures that price information derived from different sources is consolidated within a single methodological framework and that the CPI compilation process remains internally consistent.

Thus, in Uzbekistan’s practice, web scraping is not merely a technical data-collection element. Rather, it is taking shape as a comprehensive statistical mechanism that encompasses Python-based automated extraction, classification in accordance with COICOP 2018, the construction of a dedicated classifier, integration into a unified database, and the calculation of final indices in Stata.

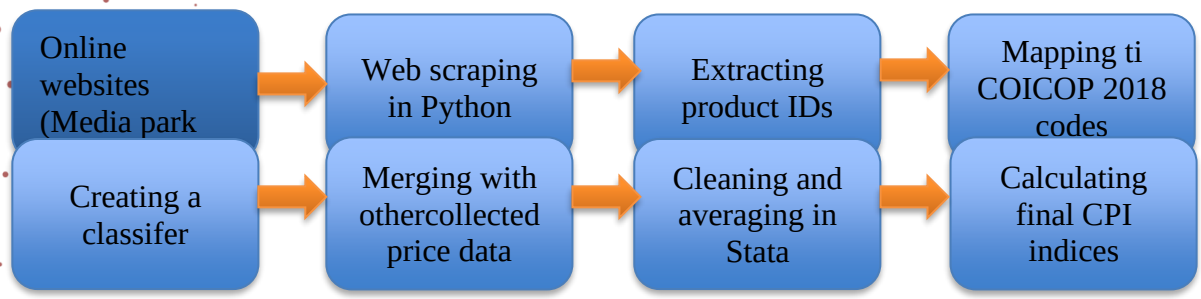

Figure 1. *Stages of integrating web-scraped data into CPI compilation in Uzbekistan*

Advantages and methodological limitations

The practical application of web scraping offers several important advantages. First, it increases the speed of data collection and allows prices to be updated more regularly. Second, it expands coverage by bringing prices formed on online retail platforms and in internet catalogues into the statistical observation system. Third, large-scale digital data make it possible to analyze price dynamics in greater depth across regions, assortments, and time periods. Fourth, integrating web-scraped data with other sources and processing them in Stata creates a unified technological chain for statistical aggregation.

At the same time, this approach also involves methodological limitations. Online prices do not always fully reflect the final transaction price. In some cases, the listed price may be promotional, product availability may be limited, or the description may be incomplete. In addition, the same product may be presented differently across websites in terms of naming, packaging, or specification. For this reason, precise methodological rules are needed for matching, classification, and validation. A further challenge is the changing technical structure of websites, since modifications in webpage design or layout require scraping algorithms to be updated on a regular basis.

CONCLUSION

In conclusion, the practical application of web scraping technology in the compilation of Uzbekistan's CPI represents an important direction in the modernization of official statistics. It improves the timeliness of data collection, broadens the scope of observation, incorporates the online market segment into statistical measurement, and strengthens the analytical capacity of the CPI information base.

Importantly, this process in Uzbekistan is no longer merely a theoretical proposal. It is already taking shape as a real technological chain: price data are collected through Python-based web scraping, product identifiers are matched to COICOP 2018 codes, a dedicated classifier is created, the resulting data are integrated with other price observations, averaged in Stata, and used to calculate

final indices. The existence of weekly practical experience based on the Media Park website, together with ongoing preparations to extend the system to more than ten additional internet resources, demonstrates the institutional development potential of this approach.

Thus, the integration of web scraping into Uzbekistan's CPI practice expands the coverage, timeliness, and analytical possibilities of price observation. In particular, the fact that Python-based data collection, COICOP 2018 classification, and the calculation of final indices in Stata have been linked within a single methodological chain demonstrates both the practical viability and the future development potential of this approach.

At the same time, effective application requires a strong methodological framework for cleaning, standardizing, classifying, matching, and validating data. Therefore, web scraping should be understood not as a replacement for conventional observation, but as a complementary instrument that enriches it in qualitative terms. International experience supports this conclusion [1; 2; 3; 4; 8].

REFERENCES:

[1] Eurostat. Practical Guidelines on Web Scraping for the HICP. Luxembourg: Publications Office of the European Union, 2020.

[2] Office for National Statistics. Research Indices Using Web Scraped Price Data. London, 2017.

[3] Statistics Netherlands (CBS). Manual Retail Price Observations Discontinued. The Hague, 2020.

[4] Statistics Bureau of Japan. Q&A about the Consumer Price Index (Answers). Tokyo, 2024.

[5] United Nations Statistics Division. Classification of Individual Consumption According to Purpose (COICOP 2018). New York, 2018.

[6] Statistics Agency under the President of the Republic of Uzbekistan. On Key Changes in the Methodology for Compiling Price Indices in the Republic of Uzbekistan. Tashkent, 28 January 2026.

[7] Statistics Agency under the President of the Republic of Uzbekistan. Consumer Price Index for January 2026. Press release. Tashkent, 5 February 2026.

[8] United Nations Economic and Social Commission for Asia and the Pacific (ESCAP). Web Scraping for CPI: Learning Resources. Statistics Division's Community of Practices, 2025.