Statistical Journal of the IAOS 31 (2015) 165–176
DOI 10.3233/SJI-150901
IOS Press

Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation

Federico Polidoro*, Riccardo Giannini, Rosanna Lo Conte, Stefano Mosca and Francesca Rossetti[1]
Istat, Italian National Statistical Institute, Rome, Italy

Abstract. The paper is focused on the results of testing web scraping techniques in the field of consumer price surveys with specific reference to consumer electronics products (goods) and airfares (services). The paper takes as starting point the work done by Italian National Statistical Institute (Istat), in the context of the European project "Multipurpose Price Statistics" (MPS). Among the different topics covered by MPS are the modernization of data collection and the use of web scraping techniques. Included are the topic of quality (in terms of efficiency and reduction of error) and some preliminary comments about the usability of big data for statistical purposes. The general aims of the paper are described in the introduction (Section 1). In Section 2 the choice of products to test web scraping procedures are explained. In Sections 3 and 4, after a description of the survey for consumer electronics and airfares, the results and/or the issues of testing web scraping techniques are conveyed and discussed. Section 5 stresses some comments about the possible improvements in terms of quality deriving from web scraping for inflation measures. Some conclusive remarks (in Section 6) are drawn with a specific attention to big data issue. In two fact boxes centralised collection of consumer prices in Italy and the IT solutions adopted for web scraping are presented.

Keywords: Multipurpose price statistics, web scraping, big data, Internet as data source

1. Introduction

Modernization of data collection tools for improving quality of Harmonised Index of Consumer Prices (HICP) is one of the pillars of the project "Multipurpose Price Statistics" (MPS), launched by the European Commission. The modernization of data collection mainly consists of three main features: widening the use of electronic devices to collect price information in the field, accessing scanner data as a source for inflation estimates, and enlarging the adoption of web scraping techniques to scrape data from the web for HICP compilation.

Italian National Statistical Institute (Istat) is actively working on all three of these features in its modernization of data collection. Specifically, a pool of statisticians and IT experts has tested the use of web scraping techniques in consumer price survey focusing the attention on two groups of products: "consumer electronics" (goods) and "airfares" (services). Web scraping procedures have been developed and tested for both these groups of products.

The tests were carried out within the present situation of Italian consumer price survey where a part of the data collection is already conducted centrally by Istat (with more than the 21% of the basket of products in terms of weights) and on the web.

Therefore, the aim of these new efforts was twofold: on one hand understanding the on quality in terms of reducing the error of the measures obtained and in

[1]The paper is the result of the work of all five authors, but, in particular, Riccardo Giannini is the author of the Box 2, Rosanna Lo Conte is the author of the Section 2 and the Box 1, Stefano Mosca is the author of the Subsections 4.1 and 4.2, Francesca Rossetti is the author of the Section 3, Federico Polidoro is the author of the abstract, of the Sections 1, 5 and 6, of Subsection 4.3 and the supervisor of the draft of the paper.

*Corresponding author: Francesca Polidoro, Istat, Italian National Statistical Institute, Rome, Italy. Tel.: +39 0646734157; Fax: +39 0646734173; E-mail: polidoro@istat.it.

terms of efficiency and costs of the consumer price survey; on the other hand exploring and analyzing the issues emerging from the use of web scraping opened the way to some thoughts and remarks about the use of web scraping to access "big data" on the web for the measurement of inflation.

Even though the work done so far was conducted in just a limited time span, as you will see in what follows, it did allow us to achieve important, although not definitive, results concerning new efficiencies. Much more is needed. For example, we just started the analysis of the possible improvements in terms of survey quality with some preliminary comments about the use of big data and the consequences for the survey design (Scheveningen Memorandum, 2013).

2. Identification of the products on which testing web scraping techniques

Within the extensive list of the products whose consumer prices are collected centrally by Istat, Istat experts collected price information on the Internet (partially for groups B and D, and wholly for group C; Box 1). For these products, until the start of the MPS project, data collection on the web was mainly carried out manually, through "copy and paste" technique.

Setting up the tasks of developing, testing and implementing the use of web scraping within the European project framework, initially involved carving out the selection of products or groups of products:

i. Representative of both goods and services;
ii. For which the importance of web is paramount, as retail trade channel;
iii. For which the phase of data collection is extremely time consuming;
iv. For which it is important widening the coverage of the sample in both temporal and spatial terms overcoming the constraints due to manual data collection.

For what concerns criterion ii, with reference to the Istat survey on "Aspects of daily life" that provides data about households' behavior and relevant aspects of their daily life, in 2012 the percentage of households owning a personal computer was equal to 59.3% (in 2013 it has grown to 62.8%) and that of households having access to the Internet equal to 55.5% (60.7% in 2013). An indicative estimation of e-commerce was provided by the data of 28.2% of the individuals aged 14 and over who have used the web during the last 12 months and who have bought or ordered goods or services for private use over the Internet in 2012 (it was 26.3% in 2011).

Box 1: Istat centralised data collection of consumer prices

In 2013 and 2014, centralised data collection of consumer prices which produces HICP, has involved more than 21% (in terms of weight) of the basket of products. It is carried out by Istat and it is broken down into four main groups:

A. Acquisition of entire external databases (medicines, school books, household contribution to National Health Service). This first group accounts for around 0.6% of the basket.
B. Central data collection as it is the most efficient way to collect prices necessary for indices compilation. This second group accounts for around 11.6% of the basket and includes:
   i. list prices that could differ from the actual purchase price (i.e. camping, package holidays);
   ii. actual purchase prices for both online purchase and purchases in a real shop (i.e. online payment for tv subscription);
   iii. actual purchase prices for products which consumers are unable to purchase online (i.e. passport fee, highway toll).
C. Acquisition of prices referred to the real purchases on the Internet. The third group accounts for approximately 2.3% of the basket and includes:
   i. actual purchase prices collected by simulation of online purchases (e.g. air tickets, consumer electronics and e-book readers);
   ii. actual purchase prices collected by simulation of online purchases + list prices (i.e. sea transport tariffs).
D. Other prices centrally collected. This last group accounts for around 7.0% of the basket, including:
   i. unique prices in the entire Italian territory (i.e. tobacco and cigarettes);
   ii. data that have been collected using different sources such as magazines, web prices lists, information transmitted by e-mail (i.e. cars, regional railway transport tariffs);
   iii. data coming from other Istat surveys and used as proxies of the actual prices (i.e. hour contractual pay as proxy of the actual wage).

Table 1 highlights (in percentages) the ranking of groups of products which were purchased or ordered online by individuals aged 14 and over, who have used the web during the last 12 months and who have bought or ordered goods or services for private use over the Internet. Holiday accommodation and other travel goods and services were the main areas which households purchased on the Internet in 2012, whereas consumer electronics products were sixth in the ranking.[1]

[1] The relevance of the purchases on web of some products – as Clothing and footwear or Books, newspapers, magazines, including e-books or Tickets for show or Articles for the house, furniture, toys – proposes the topic of the growing importance of the web as retail trade channel for products for which data collection is carried out in the field or making reference not to prices offered on web (i.e. in

Table 1
E-commerce. Individuals aged 14 and over who have used the web during the last 12 months who have bought or ordered goods or services for private use over the Internet, by groups of products purchased or ordered. 2012. *Percentages*

Overnight stays for holidays (hotels, pension etc.).	35.5
Other travel expenditures for holidays (railway and air tickets, rent a car, etc.)	33.5
Clothing and footwear	28.9
Books, newspapers, magazines, including e-books	25.1
Tickets for shows	19.7
Consumer electronics products	18.6
Articles for the house, furniture, toys, etc..	17.9
Others	15.1
Film, music	14.4
Telecommunication services	14.0
Sofware for computer and updates (excluding videogames)	11.5
Hardware for computer	8.4
Videogames and their updates	8.0
Financial and insurance services	6.0
Food products	5.6
Material for e-learning	2.8
Games of chance	1.2
Medicines	0.8

Source: Istat survey on "Aspects of daily life".

Taking into account criteria iii and iv, and taking into account that it was preferable to test web scraping techniques on products for which data collection was already completed centrally and through the web, the two groups of products finally chosen were: consumer electronics (goods) and airfares (services).

3. Testing and implementing web scraping techniques on "consumer electronics" products

3.1. Survey on consumer electronics prices and the adoption of web scraping techniques

The set of consumer electronics products for which data collection is regularly carried out by Istat, consists of: a) Mobile phones, b) Smartphones, c) PC notebook, d) PC desktop, e) PC Tablet, f) Pc peripherals (monitors), g) Pc peripherals (printers), i) Cordless or wired telephones, l) Digital Cameras, m) Video cameras.

The survey design is common to all the products listed before and it is possible to describe it as follows:

Phase 1. Selection of brands and stores (annually specified); around 18 shops (on average) for each product operating at national level.

Phase 2. Market segmentation based on technical specifications and performance (annually fixed).

- Example1 – digital cameras: seg1= 'compact' camera; seg2= 'bridge' camera; seg3= 'Mirrorless' camera; seg4= 'reflex' camera;
- Example2 – PC Monitors: seg1=screen dimensions 19–20 inch; seg2=screen dimensions 21–22 inch;
- Example3 – Mobile phones: seg1=mobile phones with basic functionalities; seg2= mobile phones with sophisticated functionalities;
- Example4 – PC Desktop: seg1=desktop; seg2= all-in-one;

Phase 3. Identification of minimum requirements to be satisfied (annually fixed).

- Example1 – PC Desktop: O.S. at least Windows 7, HD capacity 160 Gb or higher, RAM memory at least 2 Gb, etc..

Phase 4. Monthly data collection of the entire range of models in terms of commercial name and main technical specifications offered on the market by the main brands, within the segments identified at phase 2 and satisfying the minimum requirements identified at phase 3 (monthly observed). In phase 4 the sample is selected for a specific month (continually updated sample with 'automated' replacement of models that are losing importance in the market).

An example of how to carry out phase 4 could be done with regard to tablets. To specify an effective segmentation for index compilation, the main characteristics of tablets offered from leading operators in the Italian market are collected. For each new model the following characteristics are reported: screen charac-

2013 for Clothing and footwear the percentage in table was 31.5%, 2.6 percentage point more than in 2012).

Table 2
Example of the output of the phase 4 of the survey concerning tablet consumer prices

Code	Brand	Type	Memory	Operating system	Cpu	Connectivity	Gps	Screen	Transformer facility
T_Ace029	Acer	ICONIA A211 – HT.HA8ET.001	16	Android	nVidia Tegra T30L Quad-core	3G	1	10.1	0
T_Ace041	Acer	ICONIA A211 – HT.HADET.001	16	Android Ice Crea	nVidia Tegra T30L Quad-core	3G	1	10.1	0
T_Ace035	Acer	ICONIA A511_32s – HT.HA4EE.006	32	Android Ice Crea	nVidia Tegra T30S Quad-core	3G	1	10.1	0
T_Ace037	Acer	ICONIAW511-27602G06iss-NT.L0NET.004	64	Windows 8 Pro	Atom™ Z2760 (1.80 GHz Intel® Burs	3G	0	10.0	1
T_Ace036	Acer	ICONIAW511P-27602G06iss-NT.L0TET.004	64	Windows 8 Pro	Atom™ Z2760 (1.80 GHz Intel® Burs	3G	0	10.0	1
T_Ace045	Acer	Iconia W511-27602G06iss – NT.L0LET.004	64	Windows 8 –	Intel® Atom™ Z2760 (1 MB Cache, 1)	3G	0	10.0	1
T_App032	Apple	Ipad display retina 16gb wi fi + cellular	16	Mac: OS X v10.6.	A6X dual-core	+c	1	9.7	0
T_App033	Apple	Ipad display retina 32gb wi fi + cellular	32	Mac: OS X v10.6.	A6X dual-core	+c	1	9.7	0
T_App034	Apple	Ipad display retina 64gb wi fi + cellular	64	Mac: OS X v10.6.	A6X dual-core	+c	1	9.7	0

Source: Istat.

Table 3
Initial workload to develop web scraping macros (to point web sites and to scrape prices)

Stores website	Number of products	Number of pointing macros	Total time (in minutes) for pointing macros	Number of web scraping macros	Time to develop first web scraping macro	Time to follow web scraping macros (including testing)	Time of macros optimization
www.compushop.it	10	12	$5 \times 12 = 60$	24	60	$5 \times 23 = 115$	–
www.ekey.it	6	11	$5 \times 11 = 55$	22	120	$5 \times 21 = 105$	–
www.keyteckpoint.it	9	16	$5 \times 16 = 90$	16	30	$5 \times 15 = 75$	–
www.misco.it	10	11	$5 \times 11 = 55$	22	45	$5 \times 21 = 105$	–
www.pmistore.it	7	10	$5 \times 10 = 50$	10	30	$5 \times 9 = 45$	–
www.softprice.it	10	22	$5 \times 22 = 110$	46	45	$5 \times 45 = 225$	–
www.syspack.it	8	14	$5 \times 14 = 70$	14	30	$5 \times 13 = 65$	–
Total time		8 hours		6 hours	12 hours	8 hours	

Source: Istat.

teristics, memory, operating system, CPU, connectivity, GPS, transformer facility (Table 2).

Phase 5. This is the phase of the price data collection, for all the models included in the sample, from each web site of the shops considered for the survey (monthly observed). Until the start of the experimentation and implementation of web scraping techniques within Eurostat project, data collection was carried out two ways:

– Manual detection – for some shops (nine) price collectors scanned the corresponding websites manually and registered the price in external files or databases;
– Semi-automatic detection – for other nine shops, list prices were manually downloaded ("copy and paste"), and then formatted and submitted to SAS procedures that linked (automatically) the product codes in the sample (phase 4) to the codes in the list from each store.

Phase 6. Setting up the database for the calculation of consumer prices indices and union of semi-automatic and manually detected data;

Phase 7. Analysis of representativeness of each model and control of outliers (to be considered in the

index calculation each model must have a minimum number of elementary quotes and each segment has to be represented by a minimum number of products);

Phase 8. Average price for each model by geometric mean or median (when a few, even though at least the minimum required, observations are available, median is used);

Phase 9. Each stratum (segment/brand) is represented by the cheapest model, so that the minimum price is used to represent the stratum and to produce the micro-indexes;

Phase 10. Aggregation of micro indexes by geometric means (elementary level) and by weighted arithmetic means (upper levels). Weights (where available) are proportional to market shares of each brand and each segment.

Phase 4 and phase 5 are the most time consuming of the ten phases listed before. In the preliminary steps of the project, it was decided to focus the attention on these two phases, but, in particular, on phase 5 and on the semi-automatic detection of prices for which it appeared to be easier implementing web scraping techniques. As a matter of fact for these prices the aim of web scraping macros was to replace the download of the lists of prices (carried out manually by "copy and paste") with the automatic download (web scraped lists of prices). Therefore the evaluation of the results obtained concerned both the amount of prices downloaded in the lists and the amount of prices that it was possible to link automatically for each store (via SAS procedures) to the product codes in the sample selected in phase 4. Testing and implementing web scraping procedures to replace the manual detection of prices (that implies that, for some shops, price collectors scan the websites manually and register each price in external files or databases) proposed issues that appeared, on the basis of a preliminary analysis, too complex to be solved at this stage of the project.

Thus the online shops chosen for the test were the nine shops for which the semi-automatic price data detection was currently used and the experimentation of detecting prices through web scraping was carried out using the free version of iMacros (the software chosen to develop and test web scraping procedures – see Box 2).

3.2. Improvements obtained in terms of coverage, accuracy and efficiency and the adoption of web scraping within the current survey on consumer electronics prices

Tables 3, 4 and 5 show the achievements and quantify an estimation of the gain obtained in terms of time saving. It has to be noticed that in Tables 3 and 4 the on line shops for which web scraping macros are reduced from nine of the beginning of the experimentation to

Box 2: IT choices adopted to implement web scraping procedures

Web scraping software automatically retrieves and makes recognizable the information off the web page, writing it in local database/data-store/files, eliminating the use of "copy and paste" technique.

Different technical tools and software have been compared before choosing that one to be used for the test. Attention was mainly focused on HTQL, IRobotSoft, iMacros.

HTQL (Hyper-Text Query Language) is mainly aimed at extracting HTML contents from Web pages, building table structures from a Web page or modifying automatically HTML pages.

IRobotSoft is a visual Web automation and Web scraping software using HTQL. It is the best software for market researchers who need to collect frequently market data from public Web domains: i.e. realtors collecting house information in a certain area and researchers continuously tracking certain topics on the Web.

Actually the final choice fell on iMacros and the main reasons of this choice could be resumed as follows.

IMacros is a software solution for Web Automation and Web Testing. IMacros enables users to capture and replay web activity, such as form testing, uploading or downloading text and images, and even importing and exporting data to and from web applications using CSV and XML files, databases, or any other source.

It is a product that allows speed up the acquisition of textual information on the web and above all it can be used with the help of programming languages and scripting (e.g. Java, JavaScript). iMacros tasks can be performed with the most popular browsers. The product is documented with a wiki (i.e. http://wiki.imacros.net/iMacros_for_Firefox) and fora (e.g. http://forum.iopus.com/viewforum.php) which provide code examples and problems that have already been addressed and solved by others, helping in speeding up the development of the macro. Moreover it is possible to take advantage from some projects (e.g. http://sourceforge.net/projects/jacob-project/) for the use of Java, delivering to user a great potential for interface and integration with other solutions software and legacy environments.

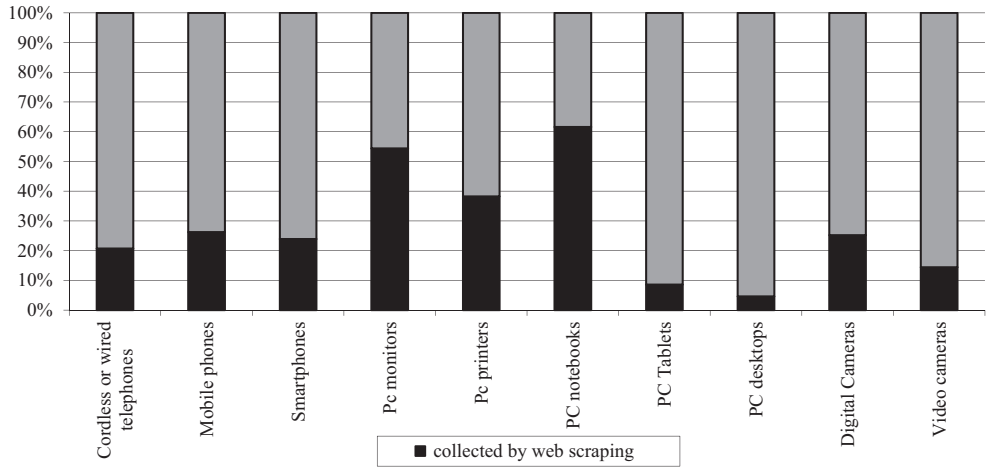
The approach adopted has been implementing two different macros for each survey: pointing and scraping macro. The pointing macro is used with the purpose of reaching the page in which the data to be collected are available. Normally this macro is easy to be built and the collector manages this activity alone. The scraping macro carries out the real work of collecting data and writing them into a flat file.

This approach has showed good results with some important advantages as well as disadvantages. The main advantage consists of easy maintenance. In all cases in which problems reside into pointing macro, there is no need of IT specialist support, because also the collector can regenerate quickly the pointing macro. This is a very relevant advantage because it allows face easily the problem of web sites instability (no that one of the access that is discussed in paragraph 4.3). The main disadvantage is a relatively lower usability, because collectors are forced to use two macros instead of one. In particular the approach of pointing and scraping macros was adopted for the test on consumer electronics products. For the on line shops, for which the experimentation of web scraping techniques was carried out, drawbacks were hugely overcome by benefits and improvements.

Table 4
Current workload for monthly data collection. Comparison between semi-automatic detection and web scraping download

On line shops website	Number of products	Semi-automatic detection: navigation, copy, and paste (minutes)	Semi-automatic detection: standardization of formats (minutes)	iMacros download: macro execution (minutes)	iMacros download: formatting output (minutes)
www.compushop.it	10	50	80	15	50
www.ekey.it	6	30	20	15	70
www.misco.it	10	60	90	10	45
www.pmistore.it	7	40	90	15	20
www.softprice.it	10	90	180	25	40
www.syspack.it	8	45	90	20	45
Total time		5 hours 15 minuts	9 hours 10 minutes	1 hour 40 minutes	4 hours 30 minutes

Source: Istat.



Source: Istat

Fig. 1. Percentages of prices collected by web scraping in the current monthly survey on consumer electronics products, on average, starting from March 2013.

six, for ordinary turnover of data collection units; from January 2014, the online shops, for which web scraping techniques are adopted, are seven (to these shops the data in Table 6 are referred).

Table 3 shows an estimation of the initial workload to develop web scraping macros (in total 34 hours). This workload could be considered the amount of time necessary to implement the macros for the annual changing base when also the sample of data collection units is revised (and then also the sample of shops on line).

Table 4 shows the comparison between the current (monthly) workload of semi-automatic detection ("copy and paste") and web scraping data collection.

Finally, Table 5 compares the workload of semi-automatic data detection techniques and that of web scraping. The advantages coming from the adoption of web scraping techniques for the sample of shops selected (finally six) are clear and they could be summarized as follows: on annual basis the workload necessary to manage the survey is reduced from about 23 working days to 16 working days. This means that the adoption of web scraping techniques for this sub sample of shops saves more than 30% of time and it could increase the amount of elementary quotes usable for the index compilation through the automatic link via SAS procedure.

Table 5
Annual working hours for half of shops sampled for data collection of prices of consumer electronics products. Comparison between semi-automatic detection and web scraping data collection

	Manual detection	Web scraping
Starting workload (annual changing base)	0	34
Current maintenance	0	12
Current data collection	173	74
Total working hours	173	120

Source: Istat.

The previous results impelled Istat from March 2013

Table 6
Sample of models, price quotes scraped and price quotes collected for consumer electronics products survey for Italian HICP compilation. January 2014. *Units and percentages*

Survey	Number of models in the sample	Number of price quotes web scraped	Number of price quotes collected and linked to the sample	Price quotes linked/price quotes scraped (%)
Cordless or wired telephones	190	844	224	26.5
Mobile phones	63	2024	108	5.3
Smartphones	131	2396	187	7.8
Digital Cameras	352	2642	400	15.1
PC desktop	37	1837	81	4.4
PC peripherals: monitors	273	2734	299	10.9
PC notebook	179	3597	288	8.0
PC peripherals: printers	143	5887	370	6.3
PC Tablet	179	1824	42	2.3
Video cameras	152	560	56	10.0
Total	1699	24345	2055	8.4

Source: Istat.

onwards, to adopt a procedure that implements the use of web scraping techniques for all shops, where possible. Navigation of the sites and the collection of information (model number, brand description and price) are automatically recorded through macros produced by iMacros for the online shops for which semi-automatic techniques were previously adopted.

Therefore since March 2013 consumer electronics products survey is carried out in two ways: (i) manual detection and (ii) download via web scraping techniques. Download via web scraping techniques have replaced the semi-automatic ones (Fig. 1).

Table 6 illustrates, for the month of January, the situation of the survey in 2014 for different products of consumer electronics for which web scraping techniques are currently used for price data detection (seven on line shops versus nine shops involving manual detection). The first column reports the number of models selected in the sample in phase 4. The second column reports the amount of elementary price quotes scraped. The third column displays the number of elementary quotes that it was possible to link with the codes of the models selected in phase 4 and in the last column the percentages of the price quotes scraped and linked (and indeed usable for index compilation).

The results clearly indicate that web scraping techniques have many potentials benefits in terms of amount of information captured and in terms of improving efficiency of the data production process with reference to the survey on consumer electronics products for Italian HICP compilation. At the same time crucial issues emerge: is it possible to use these "big data" scraped for improving the quality of the survey on consumer prices? How is it possible to combine this perspective (if feasible) with the other channels of data acquisition for inflation estimation aims (the traditional one, via data collectors and the emerging one, as scanner data)? In the next two Sections this topic will be discussed in light of the results of preliminary testing of web scraping techniques on airfares data collection.

4. Testing web scraping techniques on the survey concerning "airfares"

4.1. Survey on airfares

Istat has been carrying out the specific survey concerning airfares centrally for a very long time. This is due to practical reasons because data collection of airfares in the field is highly inefficient whilst if centrally conducted, it is possible to optimize it, exploiting Internet capabilities. Moreover there are some explicit advantages that are common to all the centralized surveys: direct control on the overall process from the stratification and sampling procedures to the index compilation passing through the data collection; possibility to adopt a very articulated survey design and to quickly adapt methods and procedures; direct control on rules, laws and regulations that can affect prices; good product coverage; engagement of few and specialized human resources.

The reference universe of the survey on airfares consists of passengers transported on scheduled commercial air flights, arriving to or leaving from Italian airports. Charter flights are excluded and only holiday/leisure travels on both traditional (TCs) and low cost carriers (LCCs) are taken into account.

The COICOP class (passenger transport by air, weight on the total HICP basket of products equaled

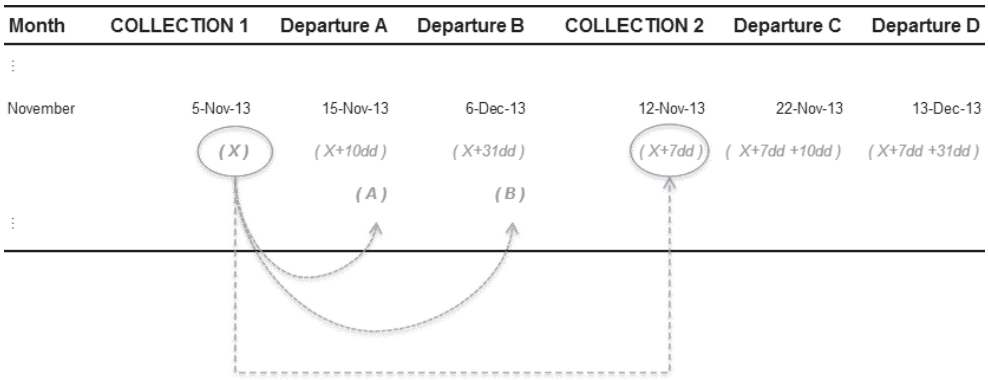


Fig. 2. Example of airfares data collection calendar for Italian CPI/HICP compilation. 2013

0.85% in 2013) is articulated in three main consumption segments, for which a specific index is compiled monthly: Domestic flights, European flights, Intercontinental flights. The three consumption segments are further stratified by type of vector, destination and route (the destination of Intercontinental flights is articulated in continent, sub-continent and extra European area of destination). Monthly prices are collected according to the following product definition: one ticket, economy class, adult, on a fixed route connecting two cities or metropolitan areas, return trip, on fixed departure/return days, final price including airport or agency fees.

In 2013 the size of the final sample consisted of 208 routes (from/to 16 Italian airports): 47 national routes, 97 European routes, 64 intercontinental routes, with 81 routes referred to TCs and 127 routes referred to LCCs.

Data collection for "passenger transport by air" shows specific characteristics and peculiarities: prices are collected by means of purchasing simulations on Internet, according to a pre-fixed yearly calendar. For most routes/type of vector, the frequency of data collection is monthly; data are usually collected on the first Tuesday of the month (day X). The departure day, considered when simulating the purchase of an air ticket, is (A) = X+10 days and (B) = X+1 month, taking into account for the return trip a stay of one week for Domestic/European flights and of two weeks for the Intercontinental ones. For some routes/type of vector, data collection in carried out twice a month (on the first and the second Tuesday, i.e. date X+7 days, of each month,). Figure 2 shows an example of calendar for airfares data collection carried out twice a month.

In 2013, data collection was carried out on 16 LCCs websites and on three web agencies selling air tickets (Opodo, Travelprice and Edreams), where only TCs airfares were collected. More than 960 elementary price quotes were registered monthly, which corresponded to the cheapest economy fare available at the moment of booking for the route and for the dates selected, including taxes and compulsory services charges; as far as LCCs are considered, also the carrier is fixed in the sample.

Two experts, each one working approximately 15 hours per month over a period of three days, carry out the airfares data collection process in Istat.

4.2. *Testing web scraping techniques for airfares survey: Preliminary results*

In the second half of 2013 and the beginning of 2014, attention was focused on the design and development of procedures for applying web scraping to record prices for air transport services.

The aim of testing web scraping techniques on airfares was mainly verifying the possibility of improving the efficiency of the survey (analogously to consumer electronics products).

Characteristics and peculiarities of the survey on airfares of passenger transport (as described in paragraph 4.1) lead to a detailed online purchasing behavior, simulating dates, airports, airline companies and cost definition. Therefore, the activity of testing web scraping techniques on airfares data collection required developing and assembling scraping macros in addition to implementing a multitude of logic controls, relying on the usage of the powerful Scripting Interface that makes possible an interchange of communication between iMacros and every Windows Scripting or programming language used on the involved web sites.

The following low cost airline companies have been scraped first: EasyJet, Ryanair, and Meridiana. Then web scraping techniques have been applied to the

traditional airlines companies using the web agency Opodo.it.

With regard to the LCCs, each airline company site showed its own specific problems. EasyJet did not allow us to scrape directly the prices using the traditional link www.easyjet.com/it/ and required specific airport descriptions (different from the simple airport IATA codes). Ryanair, at the very beginning of the tests, employed CAPTCHA ("**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part"). That system was a type of challenge-response test used to determine whether the web user was human or not; hence, that stopped us from developing a scraping macro at all.

The Meridiana website, in replying to a specific query, showed additional pages offering optional services or asking for travelers' information before displaying the final price, thus obligating us to develop a distinctive and more complex macro to scrape prices.

Finally, attention was focused on EasyJet and the macros developed provided excellent results in correctly replicating manual data collection; however, improvements in terms of time saving have been relatively small. This is due to the time spent in preparing the input files, used by the macros to correctly identify the routes and dates for which scraping the prices and returning a correct output usable for the index compilation; but also given the limited amount of elementary quotes involved (60) only a limited time saving could be derived from the adoption of web scraping techniques, despite the fact that this was a powerful tool to acquire big amount of elementary data in an efficient way.

Then web scraping techniques for airfares offered by traditional airlines companies were applied on the web agency Opodo (www.opodo.it) involving 160 monthly price quotes. The results of the macro were evaluated analyzing both the improvement in terms of efficiency and the coherence with the data manually downloaded. With regards to the improvements in terms of efficiency, also in this case the results obtained are quite small. In the last test on the monthly data collection, Opodo macro took 1 hour and 48 minutes to download the 160 elementary price quotes that were manually downloaded in about 2 hours and half. For Opodo it is necessary to prepare an input file to drive the macro in searching the correct sample of routes and, in addition to Easyjet macro, in managing the distinction between traditional and low cost carriers. Therefore, the length of time necessary for the automatic detection of prices is not very different to that required for the manual detection, considering that additional time is required to update the macro. But it has to be considered that, unlike the case of EasyJet for which the amount of elementary quote was limited (60), if the Opodo macro works correctly and only little check activity is required, since the data collection regards 160 elementary prices, two hours of manual work could be saved and dedicated to other phases of the production process or to improve quality and coverage of the survey.

4.3. Testing web scraping techniques for airfares survey: Issues emerged

It was the use of the Opodo macro that allowed two main issues (and of different nature) to emerge.

The first issue is statistical and concerns the recognition and exclusion of LCC's (as it is carried out with manual data collection). In particular, when, for specific routes and dates, the Opodo macro meets a low cost carrier that shows the cheapest cost within a page where also a traditional carrier offers a flight at the same price, it correctly excludes the LCC but, at the same time, it gets out from the page without detecting the price of the traditional carrier (that should be collected following the rules stated for the survey).

The second issue is of a legal nature. The Opodo web site, as well as that one of edreams (the two sites are connected) setup barriers to Istat automatic procedure as soon it is identified as such. Therefore, Istat drafted a communication, with legal value, addressed to the management of both the web sites, in order to allow Istat procedures of web scraping to detect automatically prices on their sites without posing obstacles. The text of the communication is worded as follows: "the survey is conducted using both automated processes and the intervention of operators that download prices directly from the Internet pages using methods and techniques defined by this Institute. The automated procedures developed by Istat for the acquisition of prices as well as the operators' actions, will be carried out by means of queries on the website www.opodo/edreams.it".

Actually after the sending of the communication both Opodo and edreams re-opened the access to their site but as soon as new robots were launched to detect airfares, automatically Opodo and edreams blocked another access. New sending of the official communication allowed Istat re-access the web sites; actually it emerged that this kind of tool (official communication by NSI with a legal value) is necessary, but further

tools have to be adopted to solve the problem of a stable access to the information on the web sites through automatic procedures implemented by NSIs. This is a crucial critical point because data collection for official statistical purposes can never be prevented. Regarding this issue Istat is exploring the way to establish technical agreements with the main web service providers included in the sample of consumer price survey.

Taking into account the outcomes of the application of web scraping techniques to EasyJet and to Opodo, it is clear that the work to make automatic the data capturing of airfares on the web for the current survey on consumer prices is still under way. The main issues that have to be dealt with in the future could be resumed as follows:

- The use of web scraping techniques exclusively as a tool to make the current survey more efficient produces a few advantages in terms of time saving without expanding the data collection. Nevertheless, improving and maintaining specific macros in order to extend web scraping techniques to the entire data collection of airfares in the web could enhance the quality (more automated and controlled production process) and allow Istat data collectors to move to other activities, even if for a limited amount of hours.
- The challenge for the future concerns the test of web scraping for airfares within a "big data" approach, adopting IT scraping techniques (different from iMacros) that could download a big amount of web pages to retrieve (off line) prices information (Cavallo, 2013). Tests should be addressed to download daily all the prices referred to a wide range of travelling dates for the sampled routes. They could allow explore the possibility of expanding the time sampling (crucial aspect for measuring the temporal evolution of prices that is inflation) and comparing the results obtained with those ones of the current approach. As a consequence statistical issues have to be discussed and solved with respect to the present survey design that could be revised even if with the full compliance with the European Regulation (Eurostat, 2013).
- Sizeable reengineering (i.e. interfacing iMacros or other software with Oracle data base) and reorganization (moving human resources from data collection to data check and analysis) of the production process will be required by a widespread use of web scraping macros in the survey on airfares.
- Finding out a solution to the so relevant problem of stability of access to information on the web.

5. Web scraping techniques and the issue of quality of consumer price survey

Improving quality of consumer price survey to improve quality of HICP is the crucial aim of the MPS European project: the results of the tests on web scraping techniques on consumer electronics products and on airfares raised some very important issues concerning this topic.

The consequences of the use of web scraping techniques on the quality of statistical survey (as it is the consumer price survey) could be analyzed from the main four dimensions of the quality [3] in terms of reduction of the error adopting a logic of total survey error approach [9].

The first dimension regards the coverage error. This error is due to lists, from which the sample is drawn, that do not represent accurately the population on the characteristics that survey is aimed to measure. The use of web scraping techniques should not affect this dimension of the survey error. Nevertheless, the adoption of the web as data source proposes the issue of the list to select the sample in a more complex way with respect to the past. If the list to extract the sample of units to collect information on consumer prices should be a list of all the enterprises/local units that sell the goods and services representative of the household expenditure, this list should include some information concerning the amount of turnover that these enterprises/local units realize by different channel of retailing, distinguishing physical and online shops. The choice of the data collection technique (in the field, on the web, others) should be determined on the basis of the population characteristics, but the adoption of web scraping will not help to solve the coverage error. That, in principle, depends on how representative is the list of the reference population.

The second dimension is the sampling error. This error is intrinsic to the nature of sample survey that has the objective to measure the characteristics of the population on the basis of a subset. Taking into account the objective of consumer price survey (measurement of inflation), time is a dimension of the sample that has to be taken into account beyond the sample unit intended as the combination of enterprise and product. Web scraping techniques have the potentiality to allow a relevant improvement of the time sample

selection that is constrained by the present traditional techniques. With respect to this issue tests carried out within MPS project and described in the paper are inadequate. In the coming steps of the project the relevance of this aspect of the sampling for inflation measurement should be analyzed in depth comparing the results obtained through more frequent than monthly or bimonthly data collection with the present estimation of the temporal evolution of the consumer prices.

The third dimension has to do with nonresponse error that provokes biases in the estimates when the sample units that do not respond are so different with respect to the responding ones in the sample with reference to the characteristics surveyed. This is crucial point for the adoption of web scraping in the field of consumer price survey. If some groups of enterprises/products are exclusively surveyed by web scraping, this choice implies that the prices that are scraped are transaction prices and they are representative of consumer prices of both physical and virtual shops. If a NSI selects a firm that retails consumer electronics products using both the channels (physical shops and the web), it should collect transaction prices referred to both the channels. The exclusive adoption of web scraping tools could potentially introduce non response error if the price evolution in the physical shops is so different with respect to that one recorded on the web channel. It is worth stressing that a) it is very expensive for NSIs to conduct data collection in the field and on line in parallel and b) the nonresponse error already derives from the choice of web as exclusive data source already carried out, even if "manually", for some products for which traditional data collection is really costly. The passage from "copy and paste" to web scraping does not appear to worsen the nonresponse problem, but it could allow investigate more quickly if the temporal evolution of prices online are representative of the evolution of prices in the physical shops. To complete the analysis of nonresponse issue we have to recall the declining responses to national household and business surveys that characterizes the present era [8] and how web scraping techniques, as a tool to access the data, could help face this decline.

The last dimension concerns the error measurement that is caused by inaccurate replies to survey questions. Ceteris paribus (taking for granted the use of the Internet as data source), web scraping techniques should considerably reduce one of the main source of measurement errors that is the human error of data collectors. Also, with respect to this aspect the tests should be extended in order to scrape also information concerning the characteristics (brands, varieties) of the elementary items for which prices are collected and that are crucial to manage one of the main statistical issues of consumer price survey and HICP that is the replacement of the product offers sampled at the beginning of a survey cycle.

Therefore, a) in terms of coverage error, web scraping is essentially neutral; b) in terms of nonresponse error, it is risky, but so is the use of the web as the exclusive data source in the traditional "copy and paste" approach; c) whereas in terms of sampling and measurement errors, web scraping techniques should guarantee improvements.

In Conclusion, if we consider these results and the gains in terms of efficiency documented by the Istat tests, our overall evaluation in terms of net quality is that for web scraping the improvements prevail on the drawbacks, even though both improvements and drawbacks should be better investigated.

6. Possible future developments and conclusive remarks

Developing and testing web scraping procedures to collect data for the Italian consumer price survey have confirmed the greater potential of the use of automatic detection of prices (and of related information) on the web in terms of improving efficiency and quality. These results pose clear challenges for existing statistical practice

With reference to the dimension of efficiency of the production process, the tests developed, for the consumer electronics and airfares, have provided clear evidences of the important improvements that it is possible to achieve. For the time being, these improvements are obvious when data collection is carried out on a few websites with a big amount of information; this could be somewhat different, if it is necessary to collect more prices from several distinct websites. This issue stresses the potential use of web scraping techniques to collect information for Purchasing Power Parity (PPP) or Detailed Average Price (DAP) exercise at international level of comparison but it seems to limit their use for sub national spatial comparison among consumer prices, for which the data collection on a certain amount of websites should be necessary.

For what concern quality, lights prevailing on shadows have been analyzed in paragraph 5.

For what concerns the challenges for the statisticians (and in particular for the official statistics), they have

emerged in terms of use of "big data" for statistical purposes. These challenges were already proposed by the study carried out by economic researchers at the Massachusetts Institute of Technology (MIT), within the project called "The Billion Prices Project @ MIT" that was aimed at monitoring daily price fluctuations of online retailers across the world.

First of all the use of web scraping techniques as a tool to achieve big data for inflation measurement has directly to do with one of the three V's of big data [8] as such ("velocity"), that is a very important dimension for a phenomenon that is characterized by the temporal evolution.

Secondly web scraping techniques for the consumer price statistics are widely available and therefore National Statistical Institutes might no longer be competitive and they are likely to lose their monopoly, currently derived by their official status, over data and information.

Last but not least, web scraping techniques applied to inflation survey could offer access to a bigger amount of data compared to that accessed by the current data collection; hence, with the potentiality of improving the inflation estimation. This issue was briefly discussed in the sections dedicated to both the products analyzed, but actually dealing with this perspective implies a discussion about the survey design that often does not allow or only partially allows using big data methods within the present schemes of sampling. Therefore it emerges clearly that fully exploiting the potentiality of web scraping (and of the "big data" virtually available) raises an important line of research that is worth a deeper look in the next future also with reference to the use of web scraping techniques different from in Macros. Whether a concern or not, we will need to face the new challenge of big data and how to integrate those opportunities with our existing statistical surveys, as they have been conceived traditionally and organized until now.

References

[1] A. Cavallo, *Scraped Data and Sticky Prices*. MIT Sloan Working Paper No. 4976-12. 2013

[2] DGINS, *Scheveningen Memorandum: Big Data and Official* Statistics. 2013

[3] D.A. Dillman, J.D. Smyth and L.M. Christian, *Internet, Phone, Mail and Mixed-Mode surveys, the Tailored Design Method*. Wiley, 4^{th} edition. 2014

[4] Eurostat – *Compendium of HICP reference documents*, Methodologies and Working papers. 2013

[5] Eurostat – *Draft of HICP manual*. 2013

[6] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A.H. Byer, *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. 2011

[7] United Nations, *Big data and modernization of statistical systems. Report of the Secretary-General*. 2014

[8] United Nations *Big Data for Development: Challenges & Opportunities*. 2012

[9] H.F. Weisberg, *The Total Survey Error Approach – A Guide to the New Science of Survey Research*. London, University of Chicago Press. 2005