Q2014
European Conference on quality in official statistics
Vienna, 2-5 June 2014

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

Riccardo Giannini (rigianni@istat.it), Rosanna Lo Conte (rolocont@istat.it), Stefano Mosca (stmosca@istat.it), Federico Polidoro (polidoro@istat.it), Francesca Rossetti (frrosset@istat.it)[1]

Table of contents

Foreword

Modernization of data collection tools for improving quality of Harmonised Index of Consumer Prices (HICP) is one of the pillars of the European project "Multipurpose Price Statistics" (MPS).

The two main aspects of this project, concerning data acquisition, are represented by "scanner data" and by the development of "web scraping" techniques as tools to capture big amount of data useful for the compilation of inflation indices, improving the quality of the statistical information produced and disseminated.

[1] The paper is the result of the work of all five authors, but, in particular, Riccardo Giannini is the author of paragraph 4, Rosanna Lo Conte is the author of paragraph 1, Stefano Mosca is the author of paragraph 3, Federico Polidoro is the author of the Foreword and of paragraph 5 and the supervisor of the draft of the paper, Francesca Rossetti is the author of paragraph 2.

For what concerns the latter issue (web scraping) Istat is actively participating in the project and implementing the development of procedures to "scrape" from web sites a large amount of data for HICP compilation, using the Internet as data source.

Attention was focused on two groups of products: "consumer electronics" (goods) and "airfares" (services); for both of them web scraping procedures have been developed and tested.

The paper deals with the results obtained in statistical and IT terms from web scraping activities on these two groups of products, focusing the attention on the issues emerging about the general topic of the usability of big data for statistical purposes. It is articulated in five paragraphs. In paragraph 1 a brief description of the centralised data collection, within the frame of the Italian consumer price survey, is sketched and the reasons of the choice of specific products to test web scraping to detect elementary price quotes information are given. In paragraphs 2 and 3, after a description of the survey for each selected group of products (consumer electronics and airfares), the results obtained in testing web scraping techniques are illustrated and discussed. In paragraph 4, the main IT solutions adopted are presented. In paragraph 5 some conclusive remarks and perspectives are discussed.

1. Centralised data collection for Italian HICP estimation and identification of the products on which testing web scraping techniques

In 2013 and 2014, centralised data collection within the survey on consumer prices, that produces Italian Consumer Price Index (CPI) and HICP, concerns more than 21% (in terms of weights) of the basket of products.

Central data collection carried out by the National Institute of Statistics of Italy (Istat) is broken down in four main groups:

A. Acquisition of entire external data bases (medicines, school books, household contribution to National Health Service). This first block accounts for about 0.6% of the basket.
B. Central data collection because it is the most efficient way to collect prices necessary for indices compilation. This second block accounts for about 11.6% of the basket and it includes:
   - ✓ list prices that could differ from the actual purchase price (i.e. camping, package holidays);
   - ✓ actual purchase prices for both on line purchase and purchase in a real shop (i.e. pay tv subscription);
   - ✓ actual purchase prices but for products not purchasable on line (i.e. passport fee, highway toll).

C. Acquisition of prices referred to the real purchases on the Internet. This third block accounts for about 2.3% of the basket and includes:
   - ✓ actual purchase prices collected by simulation of on line purchase (e.g. air tickets, consumer electronics and e-book readers);
   - ✓ actual purchase prices collected by simulation of on line purchase + list prices (i.e. sea transport tariffs).

D. Other prices centrally collected. This last block accounts for about 7.0% of the basket, including:
   - ✓ unique prices in the entire Italian territory (i.e. tobacco and cigarettes);
   - ✓ data collected using different sources as magazines, web prices lists, information transmitted by e-mail (i.e. cars, regional railway transport tariffs);
   - ✓ 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).

With the exception of the group A, for the remaining groups, price information are collected by Istat on the Internet partially for the groups B and D, totally for the group C. Until the start of the MPS project, data collection on Internet was mainly carried out manually.

Setting up the work of developing, testing and implementing web scraping techniques within the frame of European project, first of all it was carried the selection of products or groups of products:

i. representative of both goods and services;
ii. for which the importance of web as retail trade channel is relevant;
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 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 become 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).

Table 1 shows the ranking of groups of products purchased or ordered, in terms of percentages of 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 reasons of households Internet purchases in 2012, whereas consumer electronics products were sixth in the ranking[2].

Taking into account criteria iii and iv and that it was preferable testing web scraping techniques on products for which data collection was already done centrally and through web, the choice has finally fallen on two groups of products: consumer electronics (goods) and airfares (services).

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"

Developing and testing web scraping techniques on these two groups of products was aimed first of all at making the on line data collection more efficient. Then, it was aimed at exploring the potentialities of web scraping techniques to allow a better coverage of the reference population using such an innovative tool. The latter objective is strictly linked to a more general question about the use of big data for statistical purposes and the consequences of this use on the traditional sampling methodologies.

[2] 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 2013 for Clothing and footwear the percentage in table was 31.5%, 2.6 percentage point more than in 2012).

2. Testing and implementing web scraping techniques on the survey concerning prices of "consumer electronics" products

2.1 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:

- Mobile phones
- Smartphones
- PC notebook
- PC desktop
- PC Tablet
- Pc peripherals: monitors
- Pc peripherals: printers
- Cordless or wired telephones
- Digital Cameras
- Video cameras

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

Phase 1. Selection of brands and stores (annually specified); about 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 XP, HD capacity 160 Gb or higher, RAM memory at least 2 Gb, etc..

Phase 4. Monthly data collection of all the 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 as regards tablets. To specify an effective segmentation for index compilation, the main characteristics of tablets offered from leading operators in Italian market are collected. For each new model the following characteristics are reported: screen characteristics, memory, operating system, CPU, connectivity, GPS, transformer facility (table 2).

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 (1.80GHz Intel® Burs	3G	0	10	1
T_Ace036	Acer	ICONIAW511P-27602G06iss-NT.L0TET.004	64	Windows 8 Pro	Atom™ Z2760 (1.80GHz Intel® Burs	3G	0	10	1
T_Ace045	Acer	Iconia W511-27602G06iss - NT.L0LET.004	64	Windows 8 -	Intel® Atom™ Z2760 (1MB Cache, 1	3G	0	10	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

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 by two ways:

- Manual detection - for some shops (9) price collectors scanned the corresponding websites manually and registered the price in external files or databases;
- Semi - automatic detection - for other 9 shops price lists 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.

As example in table 3 the amount of elementary price quotes collected (and linked to the products codes coming from phase 4) both by semi-automatic and manual way for the different consumer electronics products are reported with reference to January 2013.

Table 3. Survey concerning tablet consumer prices: elementary price quotes collected and linked to the products codes coming from phase 4. January 2013

Survey	Semi – automatic	Manual	Semi-auto/man
Cordless or wired telephones	156	722	21,6
Mobile phones	103	314	32,8
Smartphones	173	448	38,6
Digital Cameras	212	457	46,4
PC desktop	478	990	48,3
PC peripherals: monitors	418	372	112,4
PC notebook	102	279	36,6
PC peripherals: printers	125	359	34,8
PC Tablet	428	1135	37,7
Video cameras	92	301	30,6
TOTAL	2287	5423	42,2

Source: Istat

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

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

Phase 10. Aggregation of micro indexes by weighted arithmetic means (upper levels) and by geometric means (elementary level). 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 manual download of the lists of prices ("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 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 on line 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.

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

Table 4 shows, for some of the electronic consumer products, a comparison, in two following months, between the amounts of elementary price quotes linked to the sample of product codes selected in phase 4: in one case the amounts are derived from the manually detected lists of prices ("copy and paste"), in the other case they are derived from the web scraped lists of prices.

Taking into account that the greater is the number of elementary price quotes usable for the index compilation more robust is the result obtained, it is to be stressed the increase of elementary price quotes for PC monitors, notebook PCs and printers, products for which the codes of the models are complex to be linked with the codes archived, so that, in these cases, it is clearly better the performance of web scraping application.

Table 4. Number of elementary price quotes manually detected and web scraped and then linked to the product codes coming from phase 4. A comparison between February and March 2013

Products	Manually detected lists of prices February 2013	Web scraped lists of prices March 2013
Cordless or wired telephones	195	185
Mobile phones	102	111
Smartphones	174	171
PC desktop	102	83
PC peripherals: monitors	142	310
PC notebook	328	433
PC peripherals: printers	383	421
PC Tablet	100	87
Digital Cameras	392	322
Video cameras	103	83
Total	**2021**	**2206**

Source: Istat

In terms of time saving, the following tables (from 5 to 7) resumes the achievements of the projects and quantifies an estimation of the gain obtained (in table 5 and 6 the on line shops for which web scraping macros have been currently used are reduced from nine of the beginning of the experimentation to six, for ordinary turnover of data collection units; from January 2014, the on line shops for which web scraping techniques are 7 as it is illustrated in table 8).

In table 5 it is showed 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).

In table 6 the current (monthly) workload is compared between semi-automatic detection ("copy and paste") and web scraping data collection.

Finally in table 7, it is carried out, on annual basis, the comparison, in terms of workload, between semi-automatic data detection techniques and 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 resumed as follows: on annual basis the workload necessary to manage the survey is reduced from about 23 working days to 16 working days. It means that the adoption of web scraping techniques for this sub sample of shops has allowed save more than 30% of time, increasing the amount of elementary quotes usable for the index compilation through the automatic link via SAS procedure.

Table 5. 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	5x12	60	24	60	5x23	115	
www.ekey.it	6	11	5x11	55	22	120	5x21	105	
www.keyteckpoint.it	9	16	5x16	90	16	30	5x15	75	
www.misco.it	10	11	5x11	55	22	45	5x21	105	
www.pmistore.it	7	10	5x10	50	10	30	5x9	45	
www.softprice.it	10	22	5x22	110	46	45	5x45	225	
www.syspack.it	8	14	5x14	70	14	30	5x13	65	
Total time			**8 hours**			**6 hours**	**12 hours**		**8 hours**

Source: Istat

Table 6. 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**	**1 hour 40 minutes**	**4 hours 30**

Source: Istat

Table 7. Annual working hours for half of shops sample for data collection of prices of consumer electronics products. Comparison between semi-automatic detection and web scraping data collection

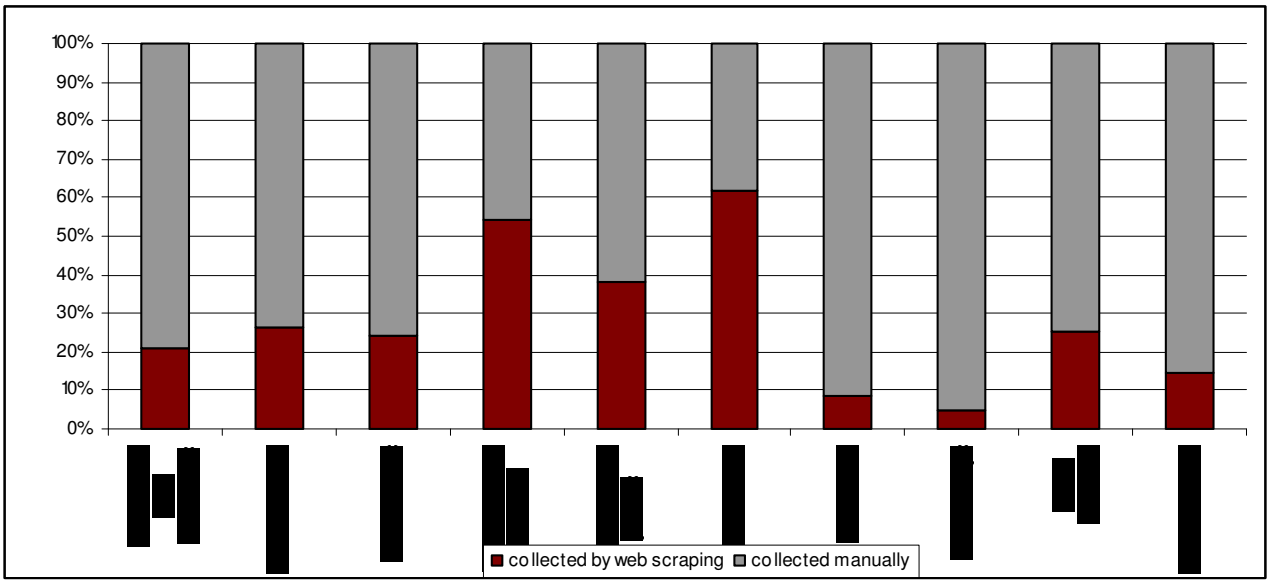
	Manual detection	Web scraping
Starting workload (annual changing base)	0	34
Current maintenance	0	12
Current data collection	173	74
Total working hours	173	120

Source: Istat

The results resumed before, pushed Istat to adopt, since March 2013, in the current survey, for all the shops for which it was possible to implement web scraping techniques, a procedure that implements the use of these techniques for the collection of prices: navigation of the sites and collection of information (model number, brand description and price) are automatically recorded through macros produced by iMacros for the on line shops for which it was previously adopted the semi-automatic techniques.

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

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



Source: Istat

Table 8 illustrates, with reference to the month of January, the current situation of the survey in 2014 for different products of consumer electronics for which web scraping techniques are currently used for price data detection (on line shops involved are 7 whereas those ones for which manual detection goes on are 9). In the first column of data it is reported the number of models selected in the sample in phase 4. In the second column of data the amount of elementary price quotes scraped is showed. In the third column of data it is displayed 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).

Table 8. Sample of models, price quotes scraped and price quotes collected for consumer electronics products survey for Italian CPI/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

From the results obtained, it is clear how many are the potentialities of web scraping techniques 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 CPI/HICP compilation. At the same time crucial issues emerge: it is possible to use the "big data" scraped for statistical purposes, enhancing the capability of the survey on consumer prices to cover the reference universe consumer prices proposed via web? 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 final comments this topic will be discussed in the light of the results of preliminary testing of web scraping techniques on airfares data collection.

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

3.1 Survey on airfares

The specific survey concerning airfares is carried out centrally by Istat. This choice was adopted since long time and it is due both to practical reasons (data collection of airfares in the field is highly inefficient and, if centrally conducted, it is possible to optimize it, exploiting Internet potentialities) and to 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 in or leaving from Italian airports, not considering charter flights and taking into account only holiday/leisure travels on both traditional (TCs) and low cost carriers (LCCs).

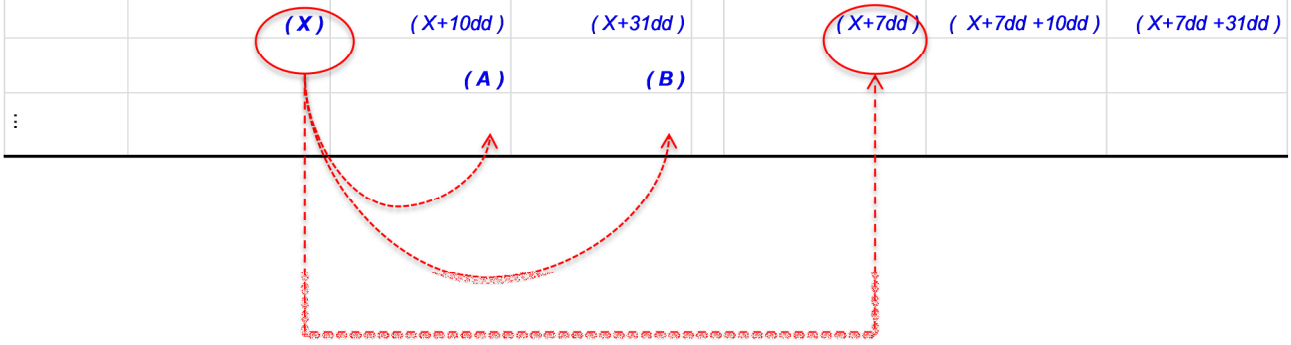
The COICOP class (passenger transport by air, weight on the total HICP basket of products equal to 0.85% in 2013) is articulated in three main consumption segments, for which a specific index is compiled monthly: Domestic flights, European flights, Intercontinental flights. The three consumption segments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

In 2013 the final sample size consisted of 208 routes (from/to16 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+10days and (B) = X+1month, considering 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+7days, of each month,). Table 9 shows an example of calendar for airfares data collection carried out twice a month.

Table 9. Example of airfares data collection calendar for Italian CPI/HICP compilation. 2013

Month	COLLECTION 1	Departure A	Departure B	COLLECTION 2	Departure C	Departure D
⋮						
November	5-Nov-13	15-Nov-13	6-Dec-13	12-Nov-13	22-Nov-13	13-Dec-13
	(X)	(X+10dd)	(X+31dd)	(X+7dd)	(X+7dd +10dd)	(X+7dd +31dd)
		(A)	(B)			
⋮						



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 are collected. More than 960 elementary price quotes were registered monthly, which correspond 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.

Istat resources involved in airfares data collection process are two persons, for 15 hours monthly each, distributed along three days.

3.2 Experimental results obtained testing web scraping techniques for airfares survey.

In the second half of 2013 and the beginning of 2014, the activity has been focused on the study and development of procedures for applying web scraping to record prices of air transport services.

The aim of testing web scraping techniques on airfares is twofold: verifying the possibility of improving the efficiency of the survey (analogously to consumer electronics products) and evaluating the chance of extending data collection to further dates (two and three months of "purchasing advance") with respect to those ones ordinary scheduled (monthly or twice a month with departure dates ten days and one month later), exploiting the potentialities of web scraping procedures. The latter aim is crucial for a survey for which sampling the time of purchasing is so important for the measurement objective (inflation), taken for granted that a change in the time distance between the purchase of the ticket and the departure date should affect in a relevant way the price paid for the selected flight.

Characteristics and peculiarities of the survey on airfares of passenger transport (as described in paragraph 3.1) lead to a very detailed online purchasing simulation regarding dates, airports, airline companies and cost definition in order to reproduce a specific purchasing behavior. Therefore, the activity of testing web scraping techniques on airfares data collection has required not only developing and assembling scraping macros but also implementing a multitude of logic controls, relying on the usage of the powerful Scripting Interface that makes possible an interchange of communication between iMacros (the software chosen to develop and test web scraping procedures – see paragraph 4) and every Windows Scripting or programming language used on the involved web sites.

First of all, the following low cost airline companies have been scraped: 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 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, presented CAPTCHA (**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part"), that is a type of challenge-response test used to determine whether the web user i**s** human or not, that stopped us from developing a scraping macro at all[3]; Meridiana

[3] CAPTCHA has recently been removed from Ryanair website.

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

Finally, attention was concentrated on EasyJet and the macros developed have provided excellent results in correctly replicating manual data collection; however, improvements in terms of time saving have been quite 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 to the limited amount of elementary quotes involved (60) that does not allow to have a meaningful measure of time saving deriving from the adoption of web scraping techniques as a powerful tool to acquire big amount of elementary data in an efficient way.

On the other side, web scraping techniques for airfares offered by traditional airlines companies have been applied on the web agency Opodo (www.opodo.it). In this case, an amount of about 160 monthly price quotes was involved. Also for Opodo, the results of the macro have been evaluated analyzing both the improvement in terms of efficiency and the coherence with the data manually downloaded. For what concerns the improvements in terms of efficiency, the results obtained are more meaningful than those ones obtained with EasyJet macro. 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. But also 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 total time necessary for automatic detection of prices is not so different with respect to the manual detection; and time to update the macro is also needed. But it has to be considered that, if the Opodo macro works correctly and only little check activity is needed, then the two hours' time of manual work is saved and could be dedicated to other phases of the production process or to improve quality and coverage of the survey.

The main issues emerged with Opodo macro regard the recognition and exclusion of LCC's (as it is carried out with manual data collection). The prices automatically scraped in some cases differ from those manually detected because they are referred to different airline companies: 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 is collected done by a human collector following the rules stated for the survey).

In conclusion, 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 web data capturing of airfares for the current survey on consumer prices is still on the way. If for EasyJet, it is possible to schedule a coming switch to the use of the web scraping macro for the current monthly data collection, further improvements and "fine tuning" (to exactly replicate current manual data collection) are necessary for Opodo and, presumably, additional efforts will be needed to develop and maintain specific macros to apply web scraping techniques to the other web agencies and LCC's web sites. Moreover, a sizeable reengineering (i.e. interfacing iMacros with Oracle data base) and reorganization (moving human resources from data collection to data check and analysis) of the production process will be needed when, starting from Easyjet web site, Istat will adopt a widespread use of web scraping macros in the survey on airfares.

Anyway, as far as fully exploiting the potentialities of the use of automatic web data capturing procedures for statistical aims is concerned, a crucial issue emerges from another point of view: is it enough to consider this tool to make the current surveys more efficient without questioning about the possibility of changing them? Is it necessary to revise sampling designs of the surveys in order to fully take advantage of the big amount of data potentially available using these innovative techniques of data collection? In paragraph 5 some comments about these general topics are proposed.

4. IT choices adopted to implement web scraping procedures

As it was described in the previous paragraphs, prices data collection on internet was carried out, above all, through "copy and paste" operations, that are, generally speaking, inefficient and costly in terms of human resources involved, even though this technique appears to be the only practical method to be used when websites are setup with barriers and machine automation cannot be enabled.

Websites involved in testing web scraping techniques for collecting data for Italian survey on consumer prices, do not present heavy restrictions to web automation technologies. In this context the use of web scraping software represents a clear improvement from various points of view with respect to the "copy and paste" technique. Web scraping software automatically retrieves and makes recognizable the information off the web page, writing it in local database/data-store/files.

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 stands for Hyper-Text Query Language (HTQL) and it is a simple language for querying and transformation of HTML Web documents. It is possible to apply HTQL to both XML and plain text documents and it is possible to use it to extract HTML contents from Web pages, to build table structures from a Web page, to modify automatically HTML pages.

IRobotSoft is a visual Web automation and Web scraping software using HTQL. The leading edge of IRobotSoft software lies in its completely automated Web robot generation technology. By observing some user Web explorations from an embedded Internet Explorer, IRobotSoft is able to generate a Web robot that does what the user intends to do. IRobotSoft has an internal scheduler so that users can set up it to run each robot in a particular time or frequency. IRobotSoft targets at common Web users who have very limited programming skills. However, its powerful data manipulation language supports complex Web computations needed by skilled programmers. IRobotSoft's Web robot engine uses sophisticated artificial intelligent and machine learning techniques for Web robot learning, and offers a complete computational platform for Web data manipulation. IRobotSoft is the best for market researchers who need to collect frequently market data from public Web domain, for example, realtors who collect house information in a certain area, and researchers who continuously track certain topics on the Web.

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 & XML files, databases, or any other source. The choice of iMacros as software to be used to develop and test web scraping procedures for consumer price data collection, it is based on the following main reasons:

- It is a product that allows speeding 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) that provide code examples and problems that have already been addressed and solved by others, helping in speeding up the development of the macro.
- 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.

Testing web scraping techniques has been started relying on the base version of iMacros and 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.

After a short period of time of usage this approach has shown, with evident good results, some important advantages and also disadvantages.

The main advantages are:

- ✓ Easy maintenance due to modularity that helps the identification of problems when they occur.
- ✓ In all cases in which problems reside into pointing macro, there is no need of IT specialist support in maintenance, because also the collector can regenerate easily pointing macro.

The main disadvantages are:

- ✓ Lower usability, because collectors are forced to use two macros instead of one.
- ✓ More time necessary to execute the complete activity of web scraping, due to the idle time between pointing macro and web scraping macro.

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.

5. Possible future developments and conclusive remarks

Developing and testing web scraping procedures to collect data for the Italian consumer price survey have confirmed the enormous potentialities of the use of automatic detection of prices (and of related information) on web and at the same time have highlighted some very important challenges that are in front of the statisticians in terms of use of “big data” for statistical purposes.

If the results obtained affect, above all, the dimension of the “efficiency” of statistical production process, the challenges for the future and the open questions regard the adoption of web

scraping techniques to gather big amount of data useful to better estimate inflation. 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. Moreover MIT project stresses how the access to this big amount of data has become “easier” in the last years (in particular for what concerns consumer prices), posing further challenge to the official statistics whose “monopolistic position” with respect to data capturing is no longer the same as some years ago.

Concerning the dimension of efficiency of the production process, the tests developed, with regards to consumer electronics and airfares, have provided clear evidences of the important improvements that it is possible to achieve through the adoption of web scraping techniques. For the time being, these improvements are clear when data collection is carried out on a few websites with a big amount of information. The situation appears to be partially different if it is necessary to collect few prices on 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 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.

Concerning the challenges, what has emerged in testing web scraping for inflation estimate purposes is the issue of the present survey design facing the potential availability of big amount of data not usable or only partially usable, for the time being, within the present schemes of sampling. The possibility of fully exploiting the potentiality of web scraping (and of the “big data” virtually available) proposes a reasoning about the necessity to discuss (and revise?) how a statistical survey, as that one on consumer prices, has been conceived and organized until now. In the field of consumer price statistics this issue seems to pose more pressure on the official statistics that, in the last years, is operating in a different situation, with respect to the past, for what concerns the accessibility of the elementary data.

References

[1] DGINS (2013) Scheveningen Memorandum: Big Data and Official Statistics

[2] United Nations (2013) Big data and modernization of statistical systems. Report of the Secretary-General

[3] United Nations (2012) Big Data for Development: Challenges & Opportunities