ISTAT Farm Register: Data Collection by Using Web Scraping for Agritourism Farms

Giulio Barcaroli*, Daniela Fusco*, Paola Giordano*, Massimo Greco*, Valerio Moretti*, Paolo Righi*, Marco Scarnò**
(*) Italian National Institute of Statistics, via C. Balbo 16 00184 Rome, Italy
(**) CINECA, Headquarters of Rome, via dei Tizi 6/B 00185 Rome, Italy
Correspondence author: Massimo Greco (e-mail address: msmagrec@istat.it)

DOI: 10.1481/icasVII.2016.f29d

Abstract

The Farm Register is a key element for the Agricultural Statistical System. Agritourism Farms (AFs) represent a small sub-population of the units included in the Farm Register (around 20,000 out of 1.7 million in 2013), but their number is increasing along the time and acquiring importance from an economic point of view. Given the tendency of using the World Wide Web to substitute the traditional way of acquiring information, ISTAT is now experimenting the possibility to collect such information directly from the sparse and unstructured information in the Internet, belonging to the vast category of Big Data, by means of a web scraping technique. A specific scraping application is developed for one of the most important hubs (TripAdvisor) and an another one for scraping individual websites. The text collected in this way requires a specific processing step finalised to extract and structure the information of interest. At the end of the process, the obtained information is used not only to update the existing information available on the Farm Register, but also to enrich it, permitting the production and the periodical dissemination of statistics related to the activities and to the services offered by the AFs, at a minimum cost. This strategy permits also to check the information (regarding newly born farms or ceased ones) stored in the Register, pertaining to the coverage of the frame. From a statistical point of view, such derived information represent a new methodological framework, that requires the evaluation of specific quality indicators.

Keywords: Farm Register, Agritourism, Web Scraping, Internet as a Data source, Big Data

1. Introduction

Agritourism Farms (AFs) are farms where the agricultural activities are integrated with the touristic ones. They are in a growing trend in Europe at least since the 1970s, particularly in Italy. Now they are more and more popular in other parts of the world as well, especially Australia, Asia and North America. A reason of such increase could be identified by observing the decline in agricultural and other forms of rural employment in many countries, that created a need for a diversified range of rural businesses. In most cases, agritourism has become an important element of development in rural areas. The initial concentration of agritourism in Italy was in small mountains and hill farms. Today the offer is more sparse, and is characterized by structures ranging from simple family farms to luxurious estates. In 2010, with 19,973 operators and about 200,000 beds available, there were more than two million guests, out of which 50% foreigners, who spent in the farm an average of 4.5 nights (Santucci, 2013).

The Farm Register maintained by the Italian National Institute of Statistics (ISTAT) is a key element for the Agricultural Statistical System. Moreover, it represents the frame for selecting samples for any survey in this sector. Such frame is built by integrating ten different administrative and statistical sources. Agritourism farms (AFs) represent a small sub-population of the units included in the Farm Register (around 20,000 out of 1.7 million in 2013), but their number is increasing along the time and acquiring importance from an economic point of view. The main source of information for these farms is represented by the administrative data of the authorised Agritourism. A dedicated survey is carried out by ISTAT, on a yearly basis and with standard questionnaires, to collect some useful information about their characteristics. Given the tendency of using the World Wide Web to substitute the traditional way of acquiring information, ISTAT decided to experiment the possibility to collect such information through the Internet, accordingly to a strategy typical of the most recent Big Data paradigm. It has to be noted that, despite the objects of the analysis (the AFs) are limited, the quantity of information that can be derived for them from the web is huge and changes rapidly over the time. This is in line with the Gartner's definition of Big Data, that considers "*high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing*"[1]. Moreover, it has to be observed that the information obtained under such paradigm is sparse and unstructured, but its proper use could permit the pursuing of the following objectives:

- improve the quality and the completeness of the administrative information already existing, that consists of *identification* (ID fiscal code of the holder, name, address, telephone, etc.), *agricultural production* (Total Area, Utilized Agricultural Area, main crops, livestock, labor force, geo-localization) and *touristic* (rooms, places, food service, other activities) data.
- collect new data in the topic as prices, wi-fi facilities, swimming-pool, organic and quality production, email address, website, horse riding, etc..
- implement a statistical system that can be updated more frequently and at a minimum cost.

The general idea of such work is to obtain structured information useful to update and enrich the Farm Register. From a statistical point of view, the use of the Big Data paradigm in this context represents a new methodological framework that should be evaluated in order to verify its efficiency and its quality. To this purpose, the administrative database (the *master dataset*) represents the benchmark according to which it will be possible to introduce specific indicators able to verify the coherence of the derived information.

Such paradigm implies the use of a new technique to collect the information: the "web scraping". But, also, it is necessary to define an integrated strategy in order to correctly link each AF referenced in the *master dat set* to the information collected on the web.

[1] http://www.gartner.com/it-glossary/big-data/

2. Web scraping and the integrated strategy

Web scraping is a computer software technique for extracting information from the web. Its objective is the collection and transformation of unstructured data (typically represented by a document written in HTML language) in more structured data. It is implemented by automatic procedures that permit to access the whole content of one or more websites, or that can select specific parts from these, by simulating a human behaviour when browsing for a specific purpose (Barcaroli *et al*, 2015a).

The use of this technique is more and more considered in the official statistics community: see Hoekstra *et al* (2012) and Ten Bosh and Windmeijer (2014).

In general, each website is characterized by a specific, and often unique, structure; an automatic approach can hardly be conceived to ensure the correct recognition of the relevant information in correspondence with all this diversity. Some facilities in this sense could derive if a special schema is applied to mark up the elements in the web pages; such approach is often used by the major search engines (that could be intended as special scrapers) to help them in identifying the target information. It has to be noted that specific standards are already defined (for instance, the definitions introduced by the Schema.org community), though not commonly used by the webmasters.

When the structure of the website is not known a priori, a phase of post-processing of the scraped texts is required, in order to gather the desired information, possibly by making use of text mining and machine learning techniques (Barcaroli *et al*, 2015b). But there are also situations (usually limited to a small number of websites) in which the structure is well defined and the information can be easily extracted with personalised applications; in such cases the results are usually more reliable (Polidoro *et al*, 2015).

In this study the two approaches have been integrated, by considering, from one side, all websites pertaining to AFs in the *master dataset* (each other different in structure and content) and, from the other side, a specific "hub" website (TripAdvisor), i.e. an aggregator of information for travellers that have large collections of pages generated dynamically from an underlying structure, like a database. This last reference gave us the possibility to retrieve data encoded into pages characterized by a common structure from which the content can be translated into the original relational form.

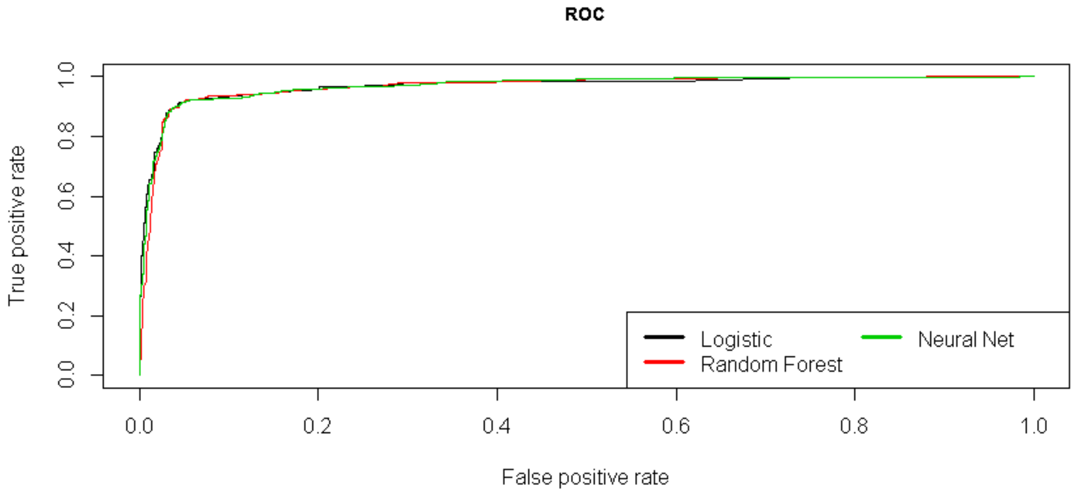
The main issue in this study refers to the proper identification of the links between the AFs as in the master frame (i.e. the database of all Agritourism Farms, originated by the legal obligation to register this activity) and their web sites or their reference in the hub website. To this purpose, the only information that can be used is the denomination and the address; unfortunately, these are often not correctly or uniquely indicated. For instance, the owner's personal information could be used in one source and the denomination of the farm in another source; the address could be different or differently written (by means of abbreviations). Moreover, there could be an implicit difference in the two sources, because in the master frame the information depends on the peculiar administrative purposes, while in the other the aim is purely informative and promotional.

To solve such issue, we used a "natural linker", able to treat also incomplete information. The perfect candidate for this scope was a "web search engine", to which specific queries (each one having in input the denomination and the address of a given AF) were submitted. The result of each query consists of different URLs, from which we should identify those referred to the specific website or those pointing to the hub. This second situation was easier to be solved, due to the fact that the presence of the text "tripadvisor.com" in one of the URLs represents the pointer to the information of the AF collected in the hub itself.

In order to identify the correct website for each AF (if existing and included in those rsulting from the query), a specific procedure was developed in which both the URL text and its content are examined: for the first the similarity between the URL and the denomination of the AF is evaluated, while as for the content, the presence of the phone number and of the address in one of the page of the related website is verified.

This led us to obtain a vector of variables containing different values resulting from the above comparisons for each one of the 10 URLs obtained by the search engine for a given AF.. A set of models (logistic, random forests, neural networks) have been fitted in order to determine for each AF the most probable corresponding URL among those found by the search engine. The training set is the one of those AFs for which the URL is known as it is indicated in the *master dataset*. The three models perform nearly the same (as illustrated in Figure 1), and the logistic was chosen because the computed score obtained by applying this model can be interpreted as a probability.

Figure 1 – *Performance of the models in determining the true URL for each AF*



Once obtained in this way the list of the URLs pertaining to the AFs contained in the *master dataset*, the contained texts were collected and analysed. In particular, we looked for specific terms in the pages of the website (single or combinations) to verify if the AF has a “restaurant”, a “swimming pool”, a “wifi” and if it sells products derived from the proper activities.

Obviously, the correct determination of the presence of one of the above characteristics is not precise because it could be affected by the use of different terms, by terms written in different languages or with different lemmas, etc. More precise could be, instead, the results deriving from the characteristics extracted from the hub; in this case, in fact, the structure of the page permits to better recognize the different terms that are located in some “standard position”.

Referring to the hub website (TripAdvisor), we were able to look for more additional information, like the possibility to host pets, to play specific sports (golf, tennis), to have a baby sitter service, to be near the sea, etc. Moreover from this hub it is possible to obtain information on the number of rooms and of the prices of each AF.

To implement this strategy we used the software ADaMSoft[2], an Open Source general-purpose software written in Java for data management, data analysis, ETL, etc., that integrates, between others, methods and libraries to parse HTML pages or to interpret the results of queries submitted to a web search engine. The results of these scraping procedures consist of a dataset in which the elements of the web pages are stored in its rows, while their different characteristics in proper columns. Such representation permits to preserve the logical structure of the page and to identify the hierarchy of the elements (i.e. the Document Object Model). As an example, it is possible to consider that a column in the output data set contains the reference URL of the page; then, for each

[2] http://adamsoft.sourceforge.net/

element of the page a record is created, characterized by a column that specifies its type (for instance BODY, INPUT, DIV, TEXT, IMG, etc.), a column with the associated text. It has to be noted that the procedures are executed in a multithread approach: this means that more than one website can be scraped in parallel. The number of threads depends on the amount of available memory: obviously, this has an impact on the time needed to scrape all the websites, while the final performance depends on their complexity (i.e. the number of pages and their response time). The limit to the full scrape of a website derives from the possibility to be blocked by the server due to a "high traffic" concentrated in a small amount of time. In some cases (for instance when referring to TripAdvisor) we divided the queries in blocks of one hundred and we set the maximum waiting time to 30 minutes.

The administrative database of the authorized AFs (the *master dataset*) includes 18,818 farms. As explained above, the first step of our selected strategy was the execution of a query in specific web search engine (Microsoft Bing). Through the query, 17,081 AFs (92%) have been identified in the web (namely with at least one URL found). The reason of the 1,737 AFs not found are various: different reference time, different farm denomination in the two sources, not existence of the AF in the web, errors in the denominations and addresses, etc

For those for which the query was successful, by applying the logistic model we were able to determine 3,289 websites (20%) specific of the AF; 5,525 (32%), instead, were those AFs for which a link in the TripAdvisor hub. The number of AFs for which were found both an entry in TripAdvisor and a specific website were 1,935.

Table 1 summarizes these results.

Table 1 – *main results obtained after the execution of the query related to the search of each AF*

Result of the query	**Frequency**
Not found by the search engine	1,737
Found by the search engine but the retrieved URLs are not referred neither to Tripadvisor nor as a correct link to the AF specific website	10,202
URLs recognized as correct specific websites of AFs not found in TripAdvisor	1,354
AF found in Tripadvisor, without a recognized specific website	3,590
AFs with both recognized specific websites and found in TripAdvisor	1,935
Total	**18,818**

The total number of AFs for which it is possible to gather information by the Internet is therefore 6,879 (36.6% of the total).

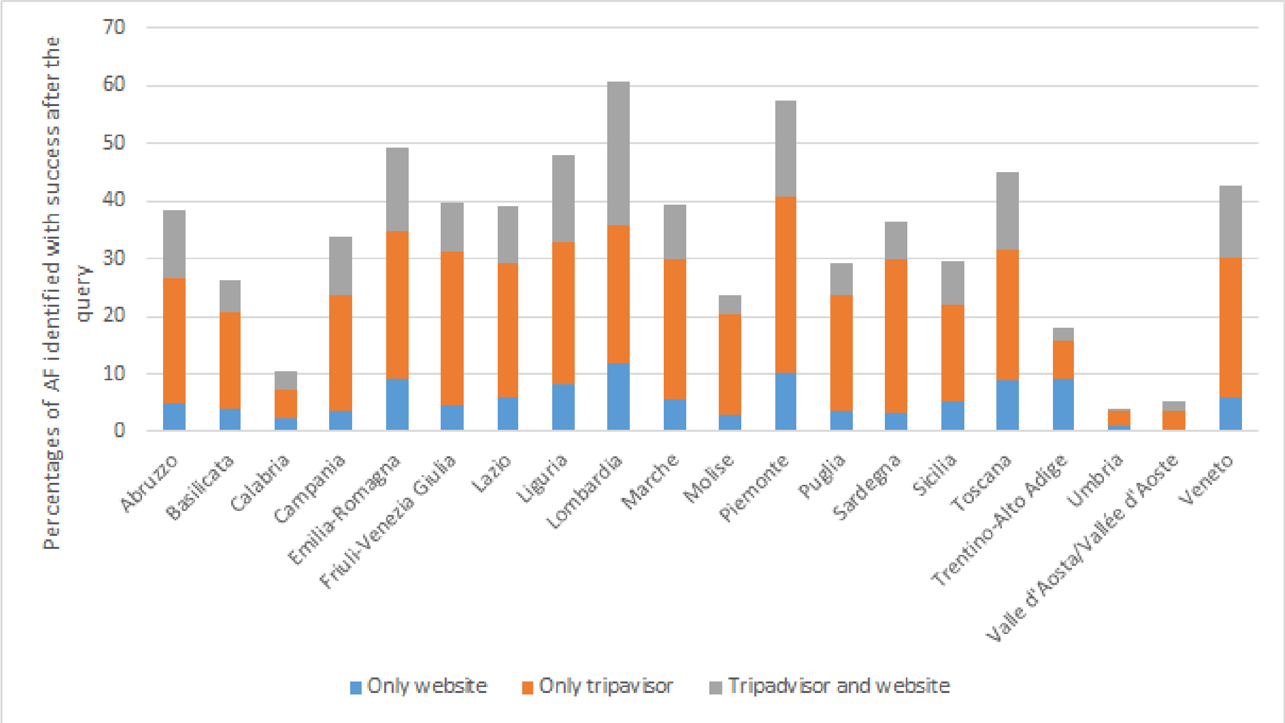
3. Analysis of the results

It is of the utmost importance to assess the representativeness of the subset of AFs whose information can be obtained after the execution of the query illustrated in the previous paragraph.

At geographical level, the coverage rates show a high variability among the Italian Nuts2 areas (Figure 2).

Figure 2 shows that the AFs identified by our strategy are not equally distributed among the Nuts2; the higher percentages (more than the 40%) are in the biggest Regions of the North and Center of Italy (Lombardia, Piemonte, Emilia Romagna, Liguria, Toscana and Veneto, while in other regions (Calabria, Umbria, Valle d'Aosta) it falls under 10%: it is therefore not possible to consider the subset of AFs found as representative by a geographical point of view.

Figure 2 - *Coverage Rate of AFs by Region*



To make proper inference we assume that the identification probabilities achieved by the retrieval procedure are uniform conditionally to the region and change among regions (that is, each AF has the same probability to be found on web given the region). If this assumption holds at least approximately we could solve the coverage issue by making use of proper weights to be assigned to each AF. Such calibration factors can be obtained by the inverse of the coverage rate in each region. This is possible because the number of AFs in the regions is known and certain.

Let us consider some important variables contained in the master DB, as the *availability of restaurant (yes/no)*. This information is contained in the *master dataset*: it is possible to verify its coherence with reference with the corresponding information found (i) only in the hub, (ii) only in specific websites, (iii) in both hub and specific websites. To this purpose let us consider the "confusion matrix" (Table 2), that cross-classifies the presence/absence of such facility with respect to the different sources. It is possible to evaluate the effectiveness of the method by considering the "accuracy rate", i.e. the ratio between the concordances (frequencies on the main diagonal) and the total number of units involved .

Table 2 – *Confusion matrix related to the availability of restaurant as in the master data set and as recognized by referring to the AFs identified only in the hub, only by considering their web sites or jointly*

Presence of restaurant as in the master dataset	**Information derived only from y the hub**		**Information derived from the specific websites**		**Information derived both from the hub and from specific websites**	
	No	**Yes**	**No**	**Yes**	**No**	**Yes**
No	1667	640	537	475	325	812
Yes	741	542	148	194	141	657
Rate of success	*61.5% (100*2209/3590)*		*50.6% (100*685/1354)*		*50.7% (100*982/1935)*	

Table 2 shows the concordances and discordances between administrative data and web data and it opens questions on the quality both the master dataset and the database obtained by the web-scraping procedure. Hereinafter, we assume the discordances between web and administrative data depend on real discrepancies between the information (and not depending on errors due to the data collection process). So how to explain these differences? In this phase of the study we give some plausible justifications that agree with the evidences given in Table 3:

i) the *master dataset* is likely to suffer of a general problem of information updating. Some structural information (as the *availability of rooms*) is correctly indicated at the moment of registration of a new AF, while some other characteristics (as the *availability of restaurant*, or the *sales of products*, or others), that may be offered at a later period, are not added in the *master dataset*, but only in the hubs and/or in the AF websites. This can explain why the differences for *availability of rooms* are small in the four columns, while are relevant for *availability of restaurant*, and highly relevant for *sales of products*;

ii) the difference among columns (2), (3) and (4) should be given by the coverage effect on the master dataset. For *availability of rooms*, the frequencies are quite stable and the performance of the weighting procedure could indicate to better investigate the procedure;

iii) the discrepancy related to the coverage is more relevant for *availability of restaurant*. The difference between 27.1% and 33.9% depends on the coverage. The remaining differences depend on the values of the variables;

iv) for *sales of products* the coverage effect is well defined by the difference between column (1) and (2) even though the frequencies of columns (3) and (4) show a relevant problem related to the quality of the variable, presumably in the master dataset.

Table 3 – *Percentages of AFs with a restaurant, that sell products and that have rooms, as in the master or as identified after the web scraping*

Variables	**Percentages of positive values (=yes)**			
	(1) **Total in the master**	**(2)** **In the common subset with value given by the master**	**(3)** **In the common subset with value obtained by web scraping**	**(4)** **In the common subset with obtained by the web scraping and after weighting**
Availability of restaurant	27.1%	33.9%	38.5%	35.9%
Sales of products	7.1%	4.4%	21.8%	18.6%
Availability of rooms	34.1%	35.1%	33.5%	32.3%

This analysis suggests the importance to collect information from the web even though it is already available in the *master dataset*. This information can be used for validating or updating it, for imputing missing values and for identifying outliers.

But one of the objectives of the strategy here proposed is also to collect additional information. Considering only the population of agritourisms identified by our strategy (6,879), *presence of swimming pools* is in 18% of cases (unweighted) and 17.1% (weighted). It is also possible to note that *wifi facilities* are in 20.6% (unweighted) and in 19.4% (weighted).

The implemented strategy would permit also to identify the prices of offered services, as related information is available in 41% of cases. Other information can be obtained with respect to the *possibility to host animals*, to offer some *sports facilities*, to *accept credit cards*, etc.

4. Conclusions and future work

This study proposes an approach based on collecting data directly from the Internet with the aim of improving the quality and the completeness of the administrative information on the Italian Agritourism Farms already available, and to produce statistics regarding structural data.

Up to 36.6% of the official AFs can be accessed in the Internet because of the availability of their websites and this number is likely to increase in the time with the spread of Internet and its potentiality in the touristic field.

Information collected from the Internet can be used in different ways, in order to:

- provide final estimations for some structural variables;
- identify outliers with respect to existing variables of the master frame;
- impute missing values in the existing variables of the master frame;
- add new variables to the master frame.

In order to assess the quality of the information obtained with this new strategy, a further step will be to provide a direct comparison of the different values obtained by the different sources, individuating also the true values: to this purpose, a sample of AFs will be selected, the corresponding websites will be manually accessed and inspected.

This manual procedure will allow to discard or confirm two strong assumptions we made when analysing the results: the adopted procedure (i) leads to identify the real value of the variable of interest on the web without error and (ii) the information on the web is updated.

Finally, new developments will regard:

- the improvement of the web scraping technique for increasing the coverage of AFs;
- the possibility to produce prices indices concerning agritourism.

REFERENCES

Barcaroli G., Scannapieco M., Scarnò M., Summa D. (2015a). Using Internet as a Data Source for Official Statistics: a Comparative Analysis of Web Scraping Technologies. In NTTS 2015 https://ec.europa.eu/eurostat/cros/system/files/Barcaroli-etal_WebScraping_Final_unblinded.pdf

Barcaroli G., Nurra A., Salamone S., Scannapieco M., Scarnò M., Summa D. (2015b). Internet as Data Source in the Istat Survey on ICT in Enterprises. Austrian Journal of Statistics Vol 44, n. 2 April 2015

Polidoro F., Giannini R., Lo Conte R., Mosca S., Rossetti F (2015). Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. Statistical Journal of the IAOS 31 (2015) 165–176

Santucci F.M. (2013). Agritourism for Rural Development in Italy, Evolution, Situation and Perspectives. British Journal of Economics, Management & Trade, 3(3), 186-200

Ten Bosh O., Windmeijer D. (2014). On the Use of Internet Robots for Official Statistics (2014). In MSIS-2014. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2014/Topic_3_NL.pdf

Hoekstra R., Ten Bosh O ., Harteveld F. (2012). Automated data collection from web sources for official statistics: First experiences." Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, 28(3-4), 2012.