**Methods and analysis for combining web scraping data with data on tourist accommodations survey**

**Keywords:** Statistical matching, Jaro-Winkler, Arcgis, Experimental Statistics, Tourism Statistics, Accommodation establishments

# Introduction

This paper describes the work done by Italy for WPJ (Innovative Tourism Statistics) in the scope of the Essnet - Big Data II project, whose overall objective is to prepare the ESS for the integration of Big data sources into the production of official statistics. In this context, the acquisition of new sources of data is of particular importance, especially in the tourism sector. The statistical representation of tourism is particularly demanding, as it concerns the measurement of a phenomenon that does not involve a single sector, but a set of highly heterogeneous services. Tourism can be analyzed from both the demand and supply perspectives and is, therefore, already measured through different information sources. For many years, national statistical institutes have been looking at the Big Data available on the web as a relevant source for increasing and completing official tourism statistics. Portals offering tourist accommodations have a large amount of information that could be very useful if integrated into the current informative system on tourism. The feasibility of this integration is largely dependent on an appropriate matching strategy between the traditional and the new sources. This includes the choice of the method (data linkage, matching) and the evaluation of the results through a cross-checking of the sources, implementing an iterative, and complex process. Italy carried out the web scraping in 2019 and 2020 on the Booking.com portal, as it was verified that it contains more accommodation establishments on the Italian territory[[1]](#footnote-1). The high number of accommodations in the country required limiting the search to a region considered particularly interesting from a tourist point of view and which shows a high degree of reliability in the statistical coverage of the census survey: Emilia-Romagna (in the following ER). This paper illustrates the methods and the first results of the attempts of a combination of web scraping data and the administrative register of tourist accommodation establishments provided by the ER region.

# Methods

The register of the accommodation establishments of the ER region was one of the most suitable for the scope, having a good quality in terms of exhaustiveness and completeness of the information contained. Furthermore, there was a complete coincidence of its contents with the results of the Capacity survey. Additionally, ER is a relevant area of the country concerning tourism, both in terms of accommodations and tourist flows.

The acquisition of the regional registers of the accommodation establishments was possible only starting from the wave 2018 of the census survey on the Capacity of accommodation establishments, starting from the update of the National Statistical Program. In particular, this update included the mandatory transmission of microdata on the accommodations, namely the registers, by the Regions. This work was done with 2018 data (the update on 2019 is now in progress). The register contains 25,683 accommodations on 11,526 records: 15,950 of them are accommodation establishments while 9,733 are 'other private establishments' not disseminated by the National Statistical Institute. The accommodations represent 7.4% of Italian accommodations. Two different methods of matching were tried to compare this archive with the results of web scraping on the ER region: Jaro-Winkler distance and Arcgis Score.

## Matching with Jaro-Winkler distance

In this attempt of matching the tool Here Maps Geolocalization was used for the location of the scraped accommodations, and Arcgis was used for the location of accommodation establishments on the regional archive. Here Maps is a tool provided by the Polish colleagues of the project in the context of WPJ [1].

This tool expects a list of fields as input, namely the unit-id, the name, the city, the address, and the postal code and returns, among others, the state, the street, the house number, the latitude, and longitude unpacked on as many different information fields. The advantage of having more details allowed us to diversify the types of combinations and to be able to proceed for subsequent steps by loosening the constraints on the conditions to obtain a larger number of matches. Here maps location was successful in 86.6% of accommodation establishments, as the other cases were false negative (of which the majority was null). A unique code was necessary to identify each unit in the administrative data archive because the fiscal code sometimes was duplicated due to the presence of units belonging to the same owner; in other cases, it was absent. The same process was made on the scraped data.

The first comparison was an exact link between two strings on hotel name and postal code and linked only 607 accommodation establishments (5.3%). A more sophisticated attempt was tried using the Jaro-Winkler similarity [2], a string metric measuring an edit distance between two sequences. The Jaro measure is the weighted sum of the percentage of matched characters from each file and transposed characters. Winkler increased this measure for matching the initial words and then rescaled it by a piecewise function, whose intervals and weights depend on the type of string (first name, last name, street, and others). The Jaro-Winkler distance takes into account letter transpositions and the occurrence of a prefix. It uses a prefix *p,* which gives more favorable ratings to strings that match from the beginning for a set prefix length *l*.

## Matching with ArcGis score

In this attempt, Arcgis was used for the location of accommodation establishments in both sources. Pre-processing of the records for the standardization of text strings was done by isolating the name from the classification attributes (hotel, B&B, others) and the breakdown of the address to the highest possible level of detail (street, street number, postal code, and municipality), and also by eliminating duplicates (the same string of name-province, city, address). Then, the city and the postal code were extrapolated with a regular expression from the formatted address, obtained with ArcGIS API. The software returns the formatted address, the coordinates in terms of latitude-longitude, and a score of the calculated coordinates, useful by selecting only the coordinates with the highest score, considered the most reliable ones. Several attempts of linkage were made, by using the following variables:

* hotel name and municipality;
* formatted hotel name (upper case, removal of spaces, removal of words classifying the accommodation establishment like a hotel, room, camping, farmhouse) and municipality;
* formatted address (for multiples, it matches the accommodation establishment name with a tolerance of 1 character);
* latitude and longitude coordinates (even in this case, for multiples, it matches the accommodation establishment name with a tolerance of 1). In the latter attempt, a threshold of tolerance on distances was tried, namely 10 meters.

Every attempt was numbered and stored in a separate column of the database to trace it and identify the successful ones for each accommodation establishment. The candidate's score matching the address was evaluated, in a range of 0 to 100, in which 100 indicates the candidate is a perfect match.

## Cross-checking of the matching

Following the web scraping activity, many checks were made on the matchings. Some territories were not well-covered (under-coverage) by the web scraping results, so a geographical analysis was carried out on the areas covered by web scraping with GIS software. Moreover, an analysis by type of accommodation included and excluded from web scraping was carried out to detect if the type of accommodation was well matched. As a result, three lists of municipalities were defined, to allow IT experts to refine the criteria for carrying out web scraping:

* The first list of municipalities with low matching coverage, namely with the greater distances between the regional register (higher values) and the web scraping results (41 municipalities). In the list, there were provincial capitals and highly receptive municipalities.
* The second list with those absent in the web scraping results although they have accommodations in the regional register (165 municipalities). In this list, there were also some municipalities with many accommodations.
* The third list (with only one municipality) with the one present in web scraping and not in the regional register.

# Results

Table 1 shows the results of the two attempts of matching on 2018 ER data. The Jaro-Winkler similarity with the higher score (85) gave unsatisfactory results, with a match of only 242 accommodation establishments (2.0% of the total 11,526), whilst considering a lower score significantly increased the number of matchings (1,025, corresponding to 8.9%) but was still not a good result. One reason could be the different methods used to format and identify the address in the two sources, namely the use of Here Maps Geolocalization tool on the scraped accommodation establishments and Arcgis on the regional archive. Subsequent attempts will be carried out using both the addresses formatted with Arcgis and extrapolating all information on multiple fields to try different matches. The attempts made with the Arcgis score gave better results, both using the formatted addresses and the geographical latitude and longitude to pair accommodation establishments. Considering the attempt that maximizes the number of matchings (3,004), the percentage of 1-1 matching was about 26%. The challenge (and the issue) in both methods is finding the best matching score: if too high, namely more restrictive, the risk is the loss of information, but if it is too low, the risk is the comparison of the incorrect data. In particular, this can occur in case of less precise addresses, namely countryside areas where farmhouses have not a house number, so they need more approximation during the match. Also, the best threshold of tolerance on distances has to be identified. Thus, further analysis to count the false-negative and false-positive matches are in progress to evaluate the quality of matching.

**Table 1: Results of matching on 2018 Emilia-Romagna administrative archive**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Emilia-Romagna Archive** | **Booking.com on Emilia-Romagna** |
|  |  |  |
| Reference period | 2018 | 19.11.2019-03.12.2019 |
| Survey Population | 11,526 accommodations in 328 municipalities | 5,670 accommodations on 315 municipalities |
| 1-1 matching Hotel Name/Address |

|  |  |
| --- | --- |
| **Jaro-Winkler score** | **Matched Accommodations** |
| 8080 | 1025 |
| 8585 | 242 |

 |
| * 1. matching
 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Match nr.** | **Matching type** | **Score ArcGis** | **Matched Accommodations** |
| 1 | Linkage Hotel name and municipality | - | 648 |
| 2 | Linkage Hotel name formatted/municipality | - | 1,738 |
| 3 | Linkage formatted address  | 90 | 2,665 |
| 80 | 2,968 |
| 70 | 3,001 |
| 4 | Linkage Geo Latitude, Longitude distance 10 meters | 90 | 2,927 |
| 80 | 2,979 |
| 70 | 3,004 |

 |

# Conclusions

The building of a joint dataset containing accommodation establishments coming from the ER regional archive and Booking, made with the subsequent replication of the process on other regions, represents important indications to:

* define the degree of coverage of the Region's registers and of the Capacity survey (e.g. identifying accommodation establishments present only on the web and not in the registers). This process will be very important both concerning hotels but, above all, to the arising sector of the ‘other collective accommodations’ and in particular to ‘holiday dwellings’ (i.e. the so-called accommodations managed as enterprises) and ‘private accommodations’, to make statistics on these missing accommodations, in terms of type, location, etc.;
* analyze the quality of information on the accommodation establishments provided by the Regions; integrate the official information, which is already required by the EU Regulation on tourism statistics [3] and by the National Statistical Program by collecting new variables from the web.

## A brief evaluation of 2019 data

The analysis of the 2019 update of the ER regional archive is in progress and seems to be promising. The experience gathered in the exploration of 2018 data regarding the cleaning phase, the reduction of dimensionality, and the normalization phase will be crucial for improving the performance of the linking. As known, the cleaning phase is the hardest in the linkage process but it is needful to improve the quality of the datasets. The most significant evidence on 2019 data is on the unique hotel names. In the ER archive (12,877 units) and the Booking dataset (6,419 units), the unique hotel names are 11,135 and 5,775, corresponding to about 86.0% and 90.0% of the datasets, respectively. However, as mentioned before, this identifier used alone does not allow for an acceptable link. Other similarity measures between two records in different data sets are going to be evaluated in comparison with the ones existing in the literature and with other's country experience. The new archive is richer in terms of contents compared to the previous one, so new variables like mail address, telephone, and geographic coordinates could play a relevant role in the linkage process.

# References

1. WPJ [Deliverable J5: Final Report containing final results and a full description of the methodology used](https://ec.europa.eu/eurostat/cros/system/files/wpj_deliverable_j5_2020_11_17_final_report_containing_final_results_and_a_full_description_of_the_methodology_used_1.pdf), 17 November 2020
2. W. Winkler, String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage (1990)
3. Regulation (EU) no 692/2011 of the European Parliament and of the Council of 6 July 2011 concerning European statistics on tourism <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:192:0017:0032:EN:PDF>
1. <https://www.similarweb.com/top-websites/Italy/category/travel-and-tourism> [↑](#footnote-ref-1)