Leveraging Web-scraping for Tourism Data Analysis: A Case Study on Romania

Cristina Rodica Boboc[1], Ana Maria Babaligea[2], Simona Ioana Ghiță[3] and Andreea Simona Săseanu[4]
[1)2)3)4)] *Bucharest University of Economic Studies, Bucharest, Romania*
[1)3)] *Institute of National Economy, Bucharest, Romania*

E-mail: cristina.boboc@csie.ase.ro; E-mail: simona.ghita@csie.ase.ro;
E-mail: babaligeaana19@stud.ase.ro; E-mail: andreea.saseanu@com.ase.ro

Please cite this paper as:
Boboc,C.R., Babaligea, A.N., Ghita, S.I. and Saseanu, A.S.,2025. Leveraging Web-scraping for Tourism Data Analysis: A Case Study on Romania. In: C. Vasiliu, D.C. Dabija, A. Tziner, D. Pleșea, V. Dinu eds. 2025. *11th BASIQ International Conference on New Trends in Sustainable Business and Consumption.* Oradea, Romania, 26-28 June 2025. Bucharest: Editura ASE, pp. 174-181
DOI: 10.24818/BASIQ/2025/11/020

Abstract

In the digitalization era, the availability and accessibility of data have experienced a significant increase, opening up new opportunities for the analysis of the tourism sector—an area where the primary source of data for performance evaluation has traditionally been official statistics. This paper aims to investigate the potential of using an alternative data source, namely web scraping, by emphasizing the additional insights and advantages that this method can offer in comparison to conventional statistical data. The information was collected through web scraping from an online tourism booking platform, using a specially developed program written in the Python programming language. This data is employed to conduct an in-depth analysis both at the national level and from a territorial perspective, examining the types and quality of Romania's tourism supply. Thus, the analysis focuses on the number and types of accommodation establishments, the average price per room for a one-night stay during the peak summer season, as well as on the reviews provided by tourists for the accommodation units. Based on the results of this analysis, a set of recommendations is formulated to support the enhancement of Romania's tourism supply. Authorities in less developed tourist areas can boost investment through incentives and promote diverse, authentic accommodations and tourism types, while improving visibility and modernizing existing facilities.

Keywords

Web-scraping, Tourism, Python, Tourism supply, Accommodation establishment.

DOI: 10.24818/BASIQ/2025/11/020

Introduction

Tourism is the largest global industry, generating billions of dollars in revenue and creating millions of jobs annually (UNWTO, 2024). In 2005, Buhalis and O'Connor stated that Information and Communication Technology (ICT) had globally transformed tourism, reshaping its structure and introducing a wide range of opportunities and challenges within the tourism industry. In the digitalization era, the availability and accessibility of data have significantly increased, offering new opportunities for analyzing the tourism sector. The primary source of data for evaluating tourism performance has traditionally been official statistics, which provide indicators such as the number of tourist arrivals, length of stay, and tourist expenditures. However, these data have certain limitations in terms of update frequency, granularity and accuracy. Nowadays, data is becoming increasingly valuable in a highly competitive world, where companies strive to derive insights through data-driven models to gain an advantage. According to Carrigan, Green and Rahman-Davies (2021), each person on the planet generates approximately 1.7 MB of data per second. As an alternative to direct data collection, Big Data techniques have emerged.

This paper aims to explore the use of an alternative data source, specifically web scraping, in order to highlight the additional benefits this method offers compared to official statistics. Web scraping involves

the automatic extraction of data from various websites, allowing for the collection of large volumes of information in real time. Zhao (2017) defines web scraping as a method of extracting data from the World Wide Web (WWW) and storing it in a file or database for subsequent analysis, emphasizing its efficiency as a data collection technique, particularly given the vast amount of information generated online. An important aspect of the tourism sector that is not captured by official statistics is the quality of the tourism offer. Through web scraping, it is possible to extract tourist reviews, ratings, and comments from various online platforms, thereby gaining a clearer understanding of the quality of tourism services and customer satisfaction. By analyzing tourist feedback, it becomes possible to identify the strengths and weaknesses of various destinations and tourism services, as well as to gather insights on customer satisfaction levels, accommodation quality, cleanliness, staff hospitality, the diversity and quality of activities offered, and any challenges encountered. The use of this alternative data source brings significant benefits to stakeholders in the tourism sector. Public authorities can use such information to develop more effective strategies and make informed decisions. Tourism industry operators—such as hotels, restaurants, and travel agencies—can use the data to improve their services and better respond to tourist needs and expectations. In addition, researchers and analysts can benefit from an extensive and detailed database for conducting complex studies and analyses of the tourism sector.

Therefore, the objectives of this study are: (a) to use the web scraping technique as an alternative source of data, highlighting its advantages over official statistics and enabling the calculation of various tourism industry indicators; and (b) to use the information obtained through web scraping to analyze, both at national and territorial (county) levels, the type and quality of Romania's tourism supply. In order to achieve these objectives, a web scraping program was developed using the Python programming language, designed to extract various types of information from an online booking website.

This study is structured as follows: it begins with an introduction, followed by a review of the main findings from the relevant literature, emphasizing the importance of using alternative data sources in the tourism industry obtained through modern, intelligent methods such as web scraping (Section 1). Section 2 presents the dataset used in the analysis, as well as the main data processing methods, Section 3 details and discusses the main research findings, while the last part of the study provides the conclusions drawn from these results.

1. Technology and innovation in the tourism industry. A review of the literature on the use of Web Scraping in tourism

An overview of the development of research over the past two decades in information technology and tourism is presented by Xiang (2018), who distinguishes two distinct eras that reflect the way technology has transformed society and the economy:

- The Digitization Era (1997–2006) was characterized by the emergence and growth of the Internet as a commercial tool. During this time, numerous technical terms such as website, browser, email, laptop, desktop, mobile phone, and e-commerce became part of common vocabulary.
- The Acceleration Era (2007–2016) marked the proliferation of technologies such as Wi-Fi, search engines, tablets, sensors, as well as the emergence of machine learning and artificial intelligence. Users' access to online information, the spread of social networks, and other collaborative tools redefined the Internet from a publishing platform to one of participation and social interaction.

Today, tourists have access to a wide array of gadgets, which they use to generate and contribute vast amounts of data. These include web analytics from tourism platforms, app log data from hotel services, call center histories, traffic volume at destinations, tourism service sales records, search engine query volumes, social media tagging, location and GPS data, media files, and more. All of these can serve as potential indicators for tourists' preferences, motivations, planning behaviors, and actual travel experiences (Pan, 2015). An alternative data source employed in this study is a data mining technique applied to websites—web scraping. Data mining refers to the process of extracting useful information from large datasets, uncovering hidden relationships and patterns within the data (Adeniyi et al., 2016). Web scraping provides a significant advantage, being both cost-effective and efficient. It reduces the need to purchase data from external sources, allowing companies to extract and analyze valuable information from the massive volume of data available online. These processed data can be used to anticipate future tourist behaviors and to develop strategies based on past experiences.

Numerous studies highlight the use of web scraping as an alternative source in the tourism industry to identify non-traditional indicators, different from those found in official statistics—such as indicators measuring customer satisfaction, service quality, and experiential aspects based on tourist reviews and

comments from various travel websites. Web scraping can also reveal destination search trends on booking platforms, provide real-time average price indicators, and track price fluctuations based on time of year or specific events. Barcaroli (2015) published a study outlining three potential scenarios for using Big Data in official statistics, in response to growing societal demands for broader and higher-quality information: a) using Big Data for data collection without altering traditional survey methodologies (e.g., web scraping and text mining); b) combining Big Data with statistical surveys to achieve faster and more accurate estimates; c) fully replacing traditional surveys with Big Data. O'Reilly (2007) utilized web scraping to define relevant tourism indicators that allow companies to obtain valuable insights. Data were extracted from TripAdvisor to collect user-generated content related to the city of Minas Gerais, Brazil. The study revealed patterns in tourist behavior in the area and validated the findings by comparing them with observable tourism trends in the region. Choong & Tunku (2019) developed an automated web scraping tool for the Malaysian tourism sector, using the Python programming language in combination with Selenium and BeautifulSoup. They concluded that there is an abundance of tourism-related data published online that remains underutilized and not leveraged to its full potential. A 2020 study by Adhinugroho et al. describes the "Development of Online Travel Web Scraping for Tourism Statistics in Indonesia", using web scraping techniques on online travel agency websites such as Agoda and Pegipegi. They generated numerous graphs and tables, ultimately concluding that the data extracted closely matched official statistics. The study emphasized that web scraping is a highly effective and powerful technique capable of yielding valuable results.

2. Data and methodology

A web scraping program is typically implemented in two main steps: the first involves creating an HTTP request, which can take the form of a URL containing a GET query in order to obtain resources from a website. In the second step, the request is received and processed, the requested information is retrieved from the website, and then sent back to the web scraping program (Zhao, 2017). In this study, a web scraping program was developed in accordance with these two fundamental steps. In the first part, the HTTP request was executed using Selenium, which automatically launched a temporary browser session to access the online booking website. In the second part, the relevant information was extracted and analyzed. For further analysis, the data were exported to a .csv file using the Python programming language and the PyCharm integrated development environment (IDE). Several Python libraries and packages were used in the development of the program. The main library was Selenium, for browser automation. To build and control the browser, the selenium.webdriver module was used. For automatic browser configuration and management (specifically Google Chrome), the package webdriver_manager.chrome was employed. For data extraction from web pages, selenium.webdriver.common.by was utilized, which allows element selection by different methods such as By.ID, By.CLASS_NAME, and By.XPATH. During the scraping process, a waiting mechanism was needed to ensure that elements were fully loaded or visible before extraction. For this purpose, selenium.webdriver.support.expected_conditions was applied. Once the data were extracted, the **Pandas** library was used to organize the data into a table format and to export it as a .csv file. For the data extraction phase, the selected platform was Agoda, an online booking website that is part of Booking Holdings Inc., the global leader in online travel and related services. The information retrieved referred to accommodation units based on the following search criteria: one night of stay, two adults, one room. The extracted elements included: Name of the accommodation unit, type of accommodation, price, address, available facilities, overall rating, ratings for cleanliness, amenities, location, comfort, services, and value for money, number of reviews. To facilitate a better understanding of the web scraping program, the application steps were illustrated in the diagram shown in Figure no. 1. The program begins on the Main Page, where the destination and period are entered. It then proceeds to the Accommodation Listings Page, where each accommodation type is selected in turn. The program checks whether there is a next accommodation type available. If yes, it loads the Filtered Listings Page, where it verifies whether any units match the filters. If matching units are found, the program opens a new page—the Accommodation Unit Page—for each listing, from which the relevant data are extracted. If no matching units remain, the program checks for the existence of a next page. If a next page is available, the extraction process repeats until no further pages are found. Then, the program proceeds to the next accommodation type. The process ends when no further accommodation types are available (Figure no.1). Following the execution of the web scraping program for the counties of Romania, information was obtained for 7,109 available accommodation units for the period of July 1st – July 2nd, 2024, for 2 persons. The database required additional processing for data cleaning, such as removing duplicate data, handling missing values, eliminating outliers, standardizing filters, as well as text cleaning by removing special characters and unnecessary spaces. After the data cleaning process, 6,888 observations remained, with accommodation units containing missing values being eliminated. The main variables considered are: County; Type of accommodation (11 types of accommodation units are present in the database, namely, apartment,

guesthouse/bed and breakfast, hotel, hostel, private villa, camping, motel, complex, mansion, farm stay, inn); Price of the accommodation unit – expressed in RON; Address of the accommodation unit; Facilities offered; Overall rating; Rating score; Number of reviews; Rating for cleanliness, facilities, location, comfort, services, and value for money.

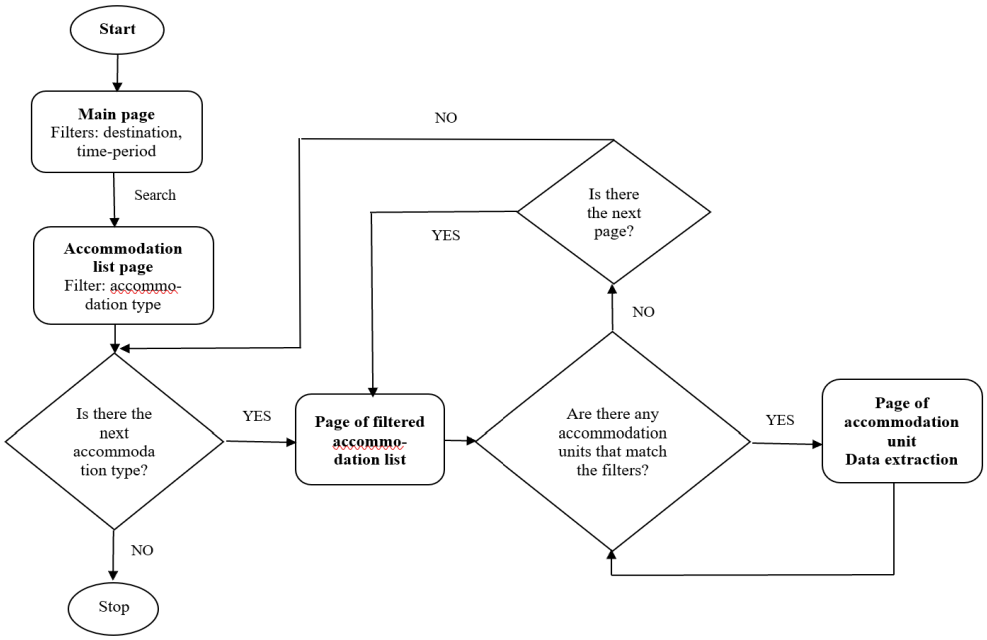

Figure no. 1. The logical diagram of the web scraping program
Source: Created by the authors, based on the program in PyCharm

3. Analysis of the data obtained through web scraping

Most of the accommodation units in the database are apartments, representing 53.38% of the total. This type of accommodation offers tourists flexibility and privacy. Guesthouses, specifically Bed & Breakfast, make up 25.28% of the total, providing a personalized atmosphere in a family setting. Hotels account for 20.09% of the total accommodation units, serving as an option for both leisure and business tourism and providing comfort and full services. Other types of accommodation, such as hostels, private villas, campsites, caravan sites, motels, resorts, and farm stay have a very low presence. Figure no. 2 illustrates the average price based on the type of accommodation, highlighting significant differences between the accommodation options. The most affordable options are hostels and motels, with average prices of 187.04 lei and 192.92 lei, respectively, offering basic conditions at low costs. Guesthouses and campsites have average prices of 303.42 lei and 290 lei, being accessible for tourists seeking an authentic experience at affordable prices.

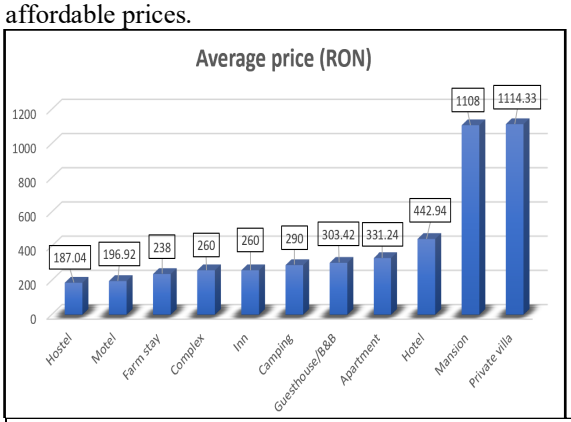

Figure no. 2. Average price (RON) by accommodation type

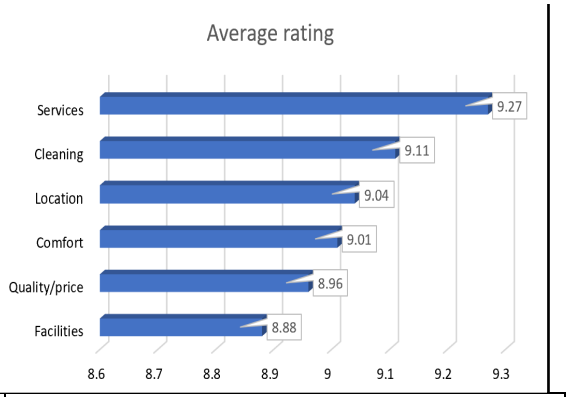

Figure no. 3. The average rating by rating criteria

Source: Created by the authors

The types of accommodation that reflect exclusivity are private villas and mansions, with very high average prices, exceeding 1,100 lei. Hotels and apartments also represent high-end options, with average prices of 443.94 lei and 331.24 lei, respectively (Figure no. 2). Figure no. 3 shows the aspects most appreciated by tourists regarding the quality of the tourism offer. The highest average rating is 9.27, given for various services such as safety, accessibility, and staff attitude. Cleanliness is also highly appreciated, with an average rating of 9.11, suggesting that the accommodations maintain high hygiene standards. Location and comfort have close average ratings of 9.04 and 9.01, respectively, while the lowest average rating of 8.88 is given for facilities, indicating room for improvement. Tourists consider the prices reasonable in relation to the quality offered, with an average rating of 8.96. The facilities provided by the accommodations are also very important. The most commonly encountered ones are parking, Wi-Fi, air conditioning, and pet acceptance. Tourists place great importance on the comfort and accessibility of services, so most accommodations include family rooms, private bathrooms, laundry services, smoking areas, and security.

For a better understanding of the data obtained through web scraping, the GeoDa program was used to represent the spatial distribution of the total number of accommodation units for each county, the number of reviews, the share of the most popular types of accommodation, as well as the average price. Two types of maps were used: Natural Breaks Map and Standard Deviation Map.

The number of accommodation establishments for each county in Romania is shown in Figure no. 4. It is observed that 10 counties – including Vaslui, Botosani, Olt, Ialomita, and Calarasi (with the lightest shade) – have the lowest number of accommodation units (less than 38 accommodation units). With a slightly darker shade, 11 counties are colored, which have between 38 and 112 accommodation establishments, such as Satu Mare, Neamt, Bacau, Covasna, Galati, Mehedinti, Buzau, and Tulcea. In the counties of Gorj, Dolj, Valcea, Arges, Arad, and Hunedoara, there are between 112 and 191 accommodation establishments. These counties are dispersed in various parts of the country, indicating regions where the promotion of tourism development can be improved. The counties of Sibiu, Brasov, Prahova, Cluj, Timis, and Constanta have a high number of accommodation establishments, ranging from 555 to 836, indicating a developed tourist infrastructure. The Municipality of Bucharest, with over 836 accommodation establishments, represents an outlier due to its complexity, reflecting the size, large population, and high density, compared to other counties (Figure no. 4).

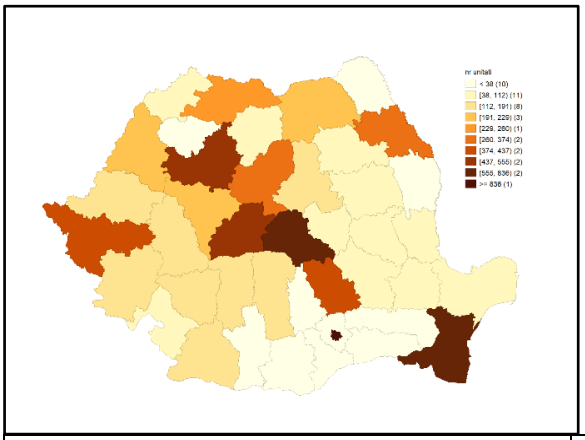

Figure no. 4. Number of accommodation establishments, by counties

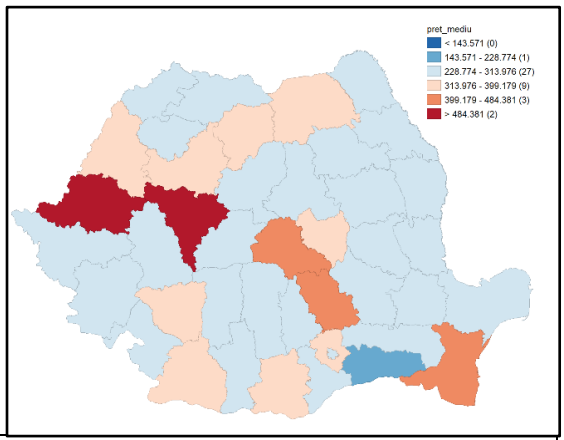

Figure no. 5. Average accommodation price, by counties

Source: Created by the authors in GeoDa

Price of accommodation establishments: In Figure no. 5, a standard deviation map of the average price for each county is represented, with the overall average being 313.97 lei. It is observed that Calarași County has an average price below the general average, at one standard deviation. The counties of Bihor, Cluj, Suceava, Gorj, Dolj, Ilfov, and Teleorman have average prices between 313.97 lei and 399.17 lei, more than one standard deviation above the national average. Three counties, namely Brasov, Prahova, and Constanta, are at two standard deviations from the mean, with prices ranging between 399.17 lei and 484 lei. The counties of Timis and Alba are considered outliers, as they are more than two standard deviations above the average (Figure no. 5).

Type of accommodation establishments: The three maps shown in Figure no. 6 illustrate the share of hotels, guesthouses, and apartments in the total accommodation establishments in each county. It is observed that in the counties of Gorj, Caras-Severin, Sibiu, and Vaslui, the share of hotels is below 10%, the share of guesthouses is over 50% in three of the four counties, while in Sibiu County, apartments

dominate with more than 60%. In Botosani County, hotels are predominant with over 53%, followed by guesthouses at 30%, and apartments represent the smallest share. In the southeastern part of the country, the share of hotels is over 40%, while the counties with the highest shares of apartments are Timis, Arad, Brasov, and Galati. Thus, it is observed that hotels are more concentrated in counties with tourist attractions or large urban centers, guesthouses prevail in rural counties, and apartments are more commonly found in counties with a high population density (Figure no. 6).

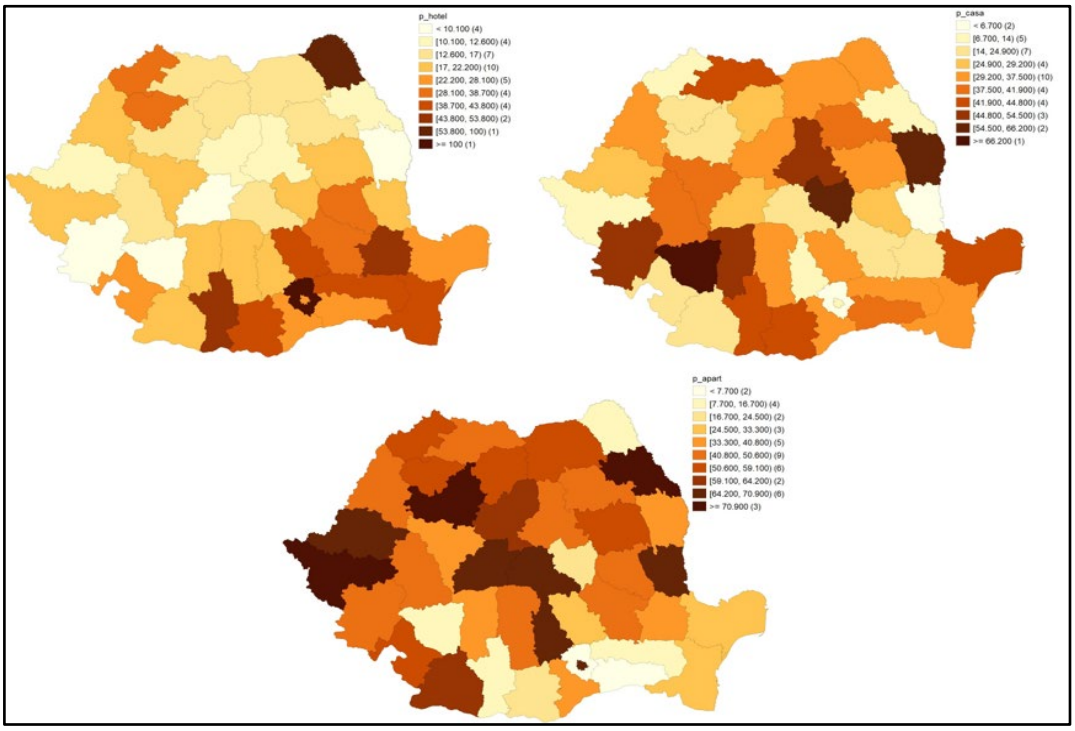

Figure no. 6. The share of hotels, guesthouses and apartments in the total accommodation establishments in each county

Source: Created by the authors in GeoDa

Reviews of accommodation establishments: To better understand the areas where tourists interact the most with online booking sites, as well as the most popular tourist regions, a map has been created to represent the number of reviews in each county. Thus, counties with a large number of reviews include Cluj, Timis, Sibiu, Brasov, Prahova, and Constanta, indicating intense tourist activity and good quality of accommodation services, as it is reasonable to expect that tourists are more likely to leave reviews in places where they had significant experiences. The map also shows 17 counties with fewer than 9,547 reviews, including Botosani, Salaj, Vaslui, Covasna, Vrancea, Buzau, Olt, Mehedinti, and Ialomita. These counties may indicate a lack of tourism infrastructure or insufficient visibility and promotion (Figure no. 7).

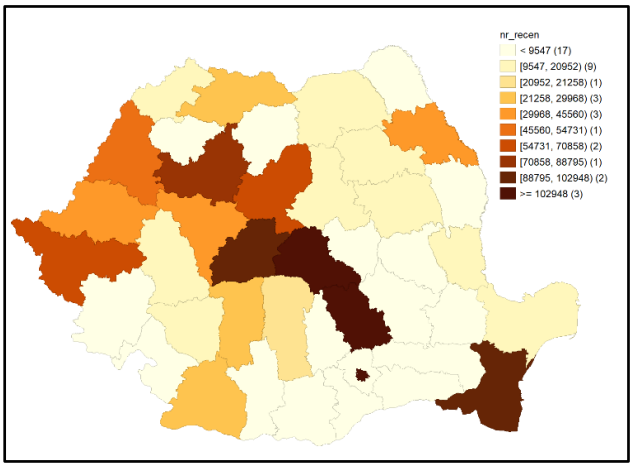

Figure no. 7. The number of reviews, by counties

Source: Created by the authors in GeoDa

Conclusions

This paper highlights how web scraping can serve as an alternative data source for the tourism industry, becoming a valuable and up-to-date resource that complements and, in some cases, enhances the information provided by official statistics.

From the analysis conducted on the data obtained through web scraping, regarding the tourism offer in Romania, it resulted that apartments represent more than half (53.38%) of the total accommodation units analyzed, indicating a clear preference for flexible, modern, and intimate lodging, especially among tourists who desire independence during their stay. The accommodation offer is diverse in terms of prices: hostels and motels are the most affordable (under 200 RON/night), preferred by budget-conscious travelers, while at the other end of the spectrum, private villas and mansions exceed 1,100 RON/night, indicating premium, exclusive options. The ratings given by tourists to accommodation establishments reflect high overall satisfaction, with tourists particularly appreciating safety, cleanliness, and the attitude of the staff (associated with scores above 9), and less so the facilities (scores around 8.8). This suggests an investment opportunity for improving equipment and facilities offered. Additionally, tourists most frequently seek facilities adapted to modern needs, such as Wi-Fi, air conditioning, parking, and pet-friendly options. Regarding the analysis of the tourism offer at the territorial level, the counties of Sibiu, Brasov, Prahova, Cluj, Timis, Constanta, as well as the Municipality of Bucharest, present the highest number of accommodation establishments, and these are also the areas where tourists have provided the most reviews for the accommodation establishments, as well as the areas with the highest average price for accommodation services (Brasov, Prahova, and Constanta). In these counties, tourists interact the most with online booking sites, constituting the most popular tourist destinations. Seven counties in the south of Romania and two in Moldova have a very low number of accommodation establishments (under 38), and in these areas, tourists have left a lower number of reviews.

The analyses conducted highlight that data obtained through web scraping can provide unique perspectives on the quality of tourism offers and consumer preferences. This information is crucial for public authorities, tourism operators, and researchers, aiding in informed decision-making and the development of effective strategies to increase the competitiveness and attractiveness of tourist destinations in Romania. Thus, in counties with a low number of accommodation establishments, local authorities can stimulate investment by offering fiscal incentives, grants, or support programs for the development of guesthouses, bed-and-breakfasts, or campsites. They can also promote rural tourism, eco-tourism, and cultural tourism projects to attract investment and capitalize on local heritage. Furthermore, the lower rating given to the facilities of accommodation units indicates the need for modernization, for which co-financing or favorable credit programs can be launched for upgrading accommodation spaces. Apartments dominate the market, but authorities can encourage the development of alternative units (glamping, traditional cottages, agritourism, etc.) to offer authentic experiences and attract various tourist segments. Areas with few reviews can be integrated into digital promotion campaigns, collaborations with tourism influencers, actively listed on booking platforms, or can have local events, festivals, and thematic routes developed to increase visibility and attractiveness. This study can be expanded by using web scraping on other online booking platforms to gain an even more comprehensive view of the tourism sector. Additionally, integrating other data sources, such as social media and reviews on travel platforms, would be useful for analyzing tourist perceptions in real-time. In conclusion, using web scraping to collect and analyze data from the tourism sector provides a detailed and up-to-date picture of the market, complementing and enhancing official statistics.

References

Adeniyi, D.A., Wei, Z. and Yongquan, Y., 2016. Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing and Informatics*, 12(1), pp. 90–108. https://doi.org/10.1016/j.aci.2014.10.001

Adhinugroho, Y., Putra, A. P., Luqman, M., Ermawan, G. Y., Takdir, Mariyah, S. and Pramana, S., 2020. Development of online travel web scraping for tourism statistics in indonesia. *Information Research*, 25(4), paper 885. https://doi.org/10.47989/irpaper885

Barcaroli, G., 2015. Use of Big Data in official statistics. Conference Paper, https://www.researchgate.net/publication/281244223

Buhalis, D. and O'Connor, P., 2005. Information communication technology revolutionizing tourism. *Tourism Recreation Research*, 30(3), pp. 7–16. https://doi.org/10.1080/02508281.2005.11081482

Carrigan, C., Green, M. W. and Rahman-Davies, A., 2021. The revolution will not be supervised: Consent and open secrets in data science. *Big Data & Society*, 8(2). https://doi.org/10.1177/20539517211035673

Choong, W. J., 2019. An automated web scraping tool for Malaysia tourism. Diss, UTAR. http://eprints.utar.edu.my/3493/

O'reilly, S., 2007. Nominative Fair Use and Internet Aggregators: Copyright and Trademark Challenges Posed by Bots, Web Crawlers and Screen-Scraping Technologies, *Loyola Consumer Law Review,* 19(3)*, Article 4.* http://lawecommons.luc.edu/lclr/vol19/iss3/4

Pan, B., 2015. E-Tourism. In: J. Jafari, H.Xiao, ed. 2015. *Encyclopedia of Tourism*. New York: Springer.

The World Tourism Organization (UNWTO), 2024. *World Tourism Barometer*, [online] Available at: <www.unwto.org/market-intelligence> [Accessed 7 February 2024].

Xiang, Z., 2018. From digitization to the age of acceleration: On information technology and tourism. *Tourism Management Perspectives*, 25, pp.147–150. https://doi.org/10.1016/j.tmp.2017.11.023

Zhao, B., 2017. Web Scraping. In: Schintler, L., McNeely, C., ed. 2017. *Encyclopedia of Big Data*. Springer, Cham, pp. 1-3. https://doi.org/10.1007/978-3-319-32001-4_483-1