Vilnius University Press

Content lists are available at Vilnius University Press

Transformations In Business & Economics
2026, Vol. 25 No 1 (67), pp.556-574.

ISSN 1648-4460; E-ISSN 2538-872X
https://doi.org/10.15388/Tibe.2026.25.1.27

Chapter: Understanding Transformative Shifts in Consumer Behaviour and Market Processes

CHARACTERIZING REGIONAL TOURISM IN ROMANIA THROUGH WEB-SCRAPED DATA AND MULTIVARIATE STATISTICAL ANALYSIS

Cristina Rodica Boboc
E-mail: cristina.boboc@csie.ase.ro
ORCID: https://orcid.org/0000-0002-7397-4821
Affiliation: Department of Statistics and Econometrics, Faculty of Economic Cybernetics, Statistics and Informatics, The Bucharest University of Economic Studies, Romania
ROR: https://ror.org/05a28rw58

Ana Maria Babaligea
E-mail: babaligeaana19@stud.ase.ro
ORCID: https://ORCID.org/0009-0003-5319-0995
Affiliation: Department of Statistics and Econometrics, Faculty of Economic Cybernetics, Statistics and Informatics, The Bucharest University of Economic Studies, Romania
ROR: https://ror.org/05a28rw58

Simona Ioana Ghita
E-mail: simona.ghita@csie.ase.ro
ORCID: https://orcid.org/0000-0002-5634-8570
Affiliation 1: Department of Statistics and Econometrics, Faculty of Economic Cybernetics, Statistics and Informatics, The Bucharest University of Economic Studies, Romania;
Affiliation 2: Institute of National Economy, Romania.
ROR: https://ror.org/05a28rw58

Claudiu Nicolae Ghinea
E-mail: ghineaclaudiu14@stud.ase.ro
ORCID: https://orcid.org/0009-0006-3645-1497
Affiliation: Doctoral School of Business Administration I, The Bucharest University of Economic Studies, Romania
ROR: https://ror.org/05a28rw58

Cristian Constantin Francu
E-mail: rancucristian23@stud.ase.ro
ORCID: https://orcid.org/0009-0001-4731-4299
Affiliation: Doctoral School of Business Administration I, The Bucharest University of Economic Studies, Romania
ROR: https://ror.org/05a28rw58

Annotation. The study investigates the tourism profile of Romania's counties by using automatically collected data from a major online travel platform. The analysis integrates variables related to the number of accommodation facilities, average prices, overall and category-specific ratings, number of reviews, and available amenities. To identify regional typologies and the determinant factors of tourism quality and development, cluster analysis and Principal Component Analysis (PCA) were applied. The results revealed four distinct clusters corresponding to different levels of development and attractiveness, as well as two principal components: a Tourism Offer Quality Component, reflecting visitors' satisfaction and perceptions of services, and a Tourism Development Component, associated with the scale and intensity of tourism activity. The study provides an integrated view of regional disparities, highlighting mature tourism destinations, counties with growth potential, and emerging areas, and offers recommendations for differentiated regional development and tourism promotion policies. The innovative contribution of this research lies in the use of alternative, web-scraped data that

Submitted on December 2024. Accepted on October 2025.

complement and enrich official statistics and data, offering timely, detailed insights, oriented towards the actual experience of tourists. Based on the findings, the study proposes recommendations for decision-makers, aiming to stimulate growth in emerging regions, improve service quality, and support balanced regional tourism development across Romania. By integrating digital data and multivariate methods, the research makes an original contribution to understanding the competitiveness and performance of regional tourism in Romania.

Keywords: web-scraping, regional tourism, multivariate analysis, cluster analysis, principal component analysis.

JEL classification: C38, L83, R11, Z32.

Introduction

The tourism industry generates significant economic and social benefits worldwide. It has the capacity to positively transform people's lives by stimulating economic growth and development, reducing poverty through job creation, and providing livelihoods for local communities. Tourism also promotes tolerance and peace through intercultural exchanges and mutual understanding between peoples and cultures (Sofronov, 2018).

According to the World Travel & Tourism Council, in 2019, before the COVID-19 pandemic, the tourism sector accounted for 10.5% of total employment (334 million jobs) and contributed 10.4% to global GDP (10.3 trillion USD). Recent WTTC research on the economic impact of tourism shows that in 2024, the travel and tourism sector contributed 10% to global GDP, amounting to 10.9 trillion USD. In 2024, tourism supported 357 million jobs globally (one in ten jobs). Domestic visitors spent 5.3 trillion USD (a 5.4% increase over 2023), while international visitor spending grew by 11.6%, reaching 1.9 trillion USD (World Travel & Tourism Council (WTTC), n.d.). A total of 1.4 billion international tourist arrivals were recorded in 2024, marking a return of international tourism compared to the pre-pandemic period (The World Tourism Organization (UNWTO), 2025).

The role of tourism in a country's economy and society depends on its level of development and tourism policy (Ștefan *et al.*, 2025). In Romania, tourism development is considered a strategic priority, given the country's substantial potential across multiple tourism segments (Constăngioară *et al.,* 2025). Effective destination planning and resource management are key challenges in leveraging this potential (Petroman, 2010).

The exploitation of tourism potential in Romania presents regional discrepancies, influenced both by the historical past of their development and by the level of the country's general infrastructure, thus generating difficulties in accessing highly attractive tourism resources. This situation has led to the development of some destinations to the detriment of others, with a more pronounced concentration, especially in the Southern Carpathians area, rich in mountain and spa tourist resorts of national and local interest and importance (Coroș, Negruța, 2014).

According to Hong (2008), the success of a tourist destination relies on how resources are used to attract visitors and to stand out in the competitive tourism market. Therefore, it is necessary to develop strategies for reallocating existing resources to change the comparative advantages of the destination compared to others.

National Institute of Statistics data show that in 2024, tourist arrivals in Romania's accommodation establishments increased by 4.5% compared to 2023, reaching 14.26 million arrivals (83.3% Romanian tourists). Growth was more pronounced among foreign tourists (+13.5% compared to 2023 vs. +2.9% compared to 2023 for domestic tourists). Overnight stays increased by 3.5% in 2024 compared to 2023 (National Institute of Statistics, 2024).

"*E-tourism*" is defined by Buhalis, Jun (2011) as both a phenomenon and a research field in which the adoption of digital technologies reshapes value chains, reflecting the digitalization of all processes in tourism, travel and hospitality. "E-tourism" and digitalization enable organizations to manage operations, conduct e-commerce, thus determining competitiveness by reorganizing internal processes, by developing transactions with trusted partners, but also by interacting with stakeholders.

The World Tourism Organization notes that with the rise of ICT worldwide, tourism was among the first industries to digitalize its business processes - bringing flight and hotel bookings online- and thus creating new business opportunities for competitiveness and sustainable industry growth. Big Data refers to large, complex and expanding datasets. In 2015, Narendra stated that Big Data can be interpreted as continuously growing data, which varies and creates new challenges in understanding it (Narendra, 2015). This phenomenon presents a certain dimension of data that is characterized by the four Vs: volume, velocity, variety and veracity (Stantic, Pokorny, 2014).

There are numerous data sources that measure flows or transactions in different statistical domains, and tourism statistics is among the first industries to allow innovation in sources and methods related to Big Data. The tourism sector tries to capture both physical and monetary flows of tourists using previously unavailable data sources and indicators provided by big data (Demunter, 2017). In 2014, Ahas *et al*. discussed the measurement of tourism using big data, the aim being to assess the possibility of using data to generate statistics on tourist inflows and outflows, but also to discover the strengths and weaknesses related to access, trust, cost and technological challenges in using such a new data source. Big Data is considered to be a solution to eliminate the shortcomings offered by official statistics (Demunter, 2017), but also a method to obtain accurate information about the real number of tourists, contributing to its sustainable development. Tokarchuk *et al*. (2021) highlighted their importance in research showing that residents' life satisfaction and tourists' positive experience are strongly influenced by tourism intensity, highlighting the need for precise monitoring of the actual flow of tourists.

In this context, such large volumes of unaggregated data can be processed quickly, efficiently and systematically through alternative methods, such as web-scraping, providing a detailed and up-to-date picture of tourist behavior and accommodation performance. Applied in the tourism sector, these techniques facilitate complex quantitative and qualitative analyses, supporting decision-makers in identifying trends, assessing service quality and optimizing tourism development and promotion strategies.

The main objectives of this study are: collecting and processing alternative data on the Romanian tourist offer through web scraping techniques, using updated online sources that are representative of the tourism market; building a unitary database that includes relevant indicators for analyzing the tourist offer: type of accommodation units, prices, available facilities, comfort level and tourist ratings; applying cluster analysis to group Romanian counties according to the structural similarities of the tourist offer, in order to identify territorial models and distinct regional profiles; using principal component analysis (PCA) to reduce dimensionality and highlight the main latent dimensions that describe the variability of the accommodation offer at the county level; interpreting the results in order to formulate conclusions

regarding the distribution, diversity and quality of tourist services, with an emphasis on regional differences, as well as to develop useful recommendations to substantiate decisions regarding the development and promotion of tourism at the regional level in Romania.

The paper is structured as follows: *the Introduction* presents the general context of the research, the motivation of the study and the objectives pursued, highlighting the importance of using web-scraped data in the analysis of regional tourism phenomena; *the Literature review* section offers a synthesis of the main works in the field of tourism and multivariate statistical analyses, by using alternative data sources; *the Data and Methodology* section describes the source of data automatically collected from tourism platforms, the variables analyzed and the stages of their processing, as well as the multivariate statistical analysis methods used; the *Results and Discussions* section is structured in two subsections: the first presents the results of the cluster analysis, through which the counties of Romania are grouped according to similarities regarding the structure and quality of the accommodation offer; the second details the results of the ACP analysis, highlighting the main dimensions that explain the interregional variation and the determining factors of tourism performance. Finally, *the Conclusions and Recommendations* chapter summarizes the main findings of the study, highlighting the theoretical and practical contributions of the research and formulating recommendations for the sustainable development and promotion of tourism at the regional level in Romania.

1. Literature Review

The growing digitalization of the tourism sector has fundamentally changed how destinations are monitored, evaluated and governed. Traditional tourism statistics - largely based on accommodation registers, border counts or surveys - are increasingly complemented by alternative data sources such as online booking platforms, mobile positioning, social media content or user-generated reviews. These new sources offer high temporal frequency, fine spatial granularity and rich qualitative information about tourists' perceptions and behaviours, which are crucial for understanding regional tourism dynamics and designing evidence-based development strategies.

1.1. Alternative Data Sources and Web-Scraped Information in Tourism

The emergence of digital technologies along the tourism value chain has generated large volumes of real-time data. Early academic contributions highlighted the web as a key information tool for tourism, facilitating the dissemination of information and enabling new forms of interaction between tourism suppliers and consumers (Corfu, Azevedo, 2007). More recent studies emphasize that digitalization is closely connected with sustainable tourism development. Săseanu *et al*. (2020) show that digital tools and data analytics support green tourism by optimizing resource use and strengthening competitiveness across European countries.

Web-scraping methods have become increasingly prominent in tourism research as a cost-effective way to collect micro-level data from booking platforms and review systems. Academic work demonstrates that web-scraped information - prices, availability, facilities, textual reviews - can be used for sentiment analysis, quality assessment, demand monitoring, and competitive positioning (Tanasă, 2024). Additional research finds that online reviews act as "open big data", capturing tourist satisfaction in real time and providing indicators for destination image and service quality (Trușescu, Nicolaie, 2024).

In Romania, Boboc *et al*. (2025) developed a web-scraped dataset to examine spatial patterns of accommodation supply, service quality and regional disparities. Their findings show that alternative data

can complement official statistics by providing granular, dynamic indicators that reveal differences in tourism performance across counties.

1.2. Regional Tourism Development and Spatial Disparities

Tourism development is influenced by the geographical distribution of resources, accessibility, infrastructure and local policies. Academic studies consistently highlight territorial disparities. Constantin *et al*. (2018) identify strong regional imbalances between Romania's well-developed tourism hubs and peripheral counties with underexploited potential.

A key contribution is Davidescu *et al*. (2018), who investigate the regional development of rural tourism using principal component analysis for mixed data and cluster analysis at the NUTS 3 level. Their results offer a clear segmentation of Romanian counties based on tourism supply and demand indicators and serve as a methodological and conceptual benchmark for the present study.

European research confirms that multidimensional statistical techniques - PCA, cluster analysis, factorial methods - are widely used to construct regional typologies and evaluate tourism development pathways. Zaharia (2024) applies PCA to Romanian tourism indicators and identifies latent components related to development, internationalization and performance dynamics between 2014 and 2022.

This article builds upon these contributions by incorporating web-scraped indicators - accommodation structure, facilities, prices, online reviews - providing a more comprehensive perspective on regional tourism development in Romania.

1.3. Multivariate and Multicriteria Approaches to Tourism Analysis

Multivariate and multicriteria techniques are standard tools in tourism research due to the complexity of the sector. PCA and cluster analysis are widely applied in destination evaluation, competitiveness studies and regional profiling. Ajčnerová *et al*. (2016) demonstrate how PCA and clustering can be combined to assess destination quality by reducing correlated indicators and grouping destinations into homogeneous categories.

In Romania, Davidescu *et al*. (2018) show that PCA and cluster analysis provide robust frameworks for identifying county-level patterns in rural tourism performance. Other scholars employ multicriteria approaches and competitiveness indices to compare Romania with peer destinations, such as Andronic (2025), who uses Competitive Importance-Performance Analysis (CIPA) to evaluate tourism strengths and weaknesses at national level.

Studies using web-scraped reviews increasingly apply sentiment analysis, PCA and clustering to understand perceived service quality and market segmentation (Tanasă, 2024). The analytical framework in the present study - hierarchical clustering, K-means and PCA - thus aligns with established methodological standards while innovating through the exclusive use of alternative data.

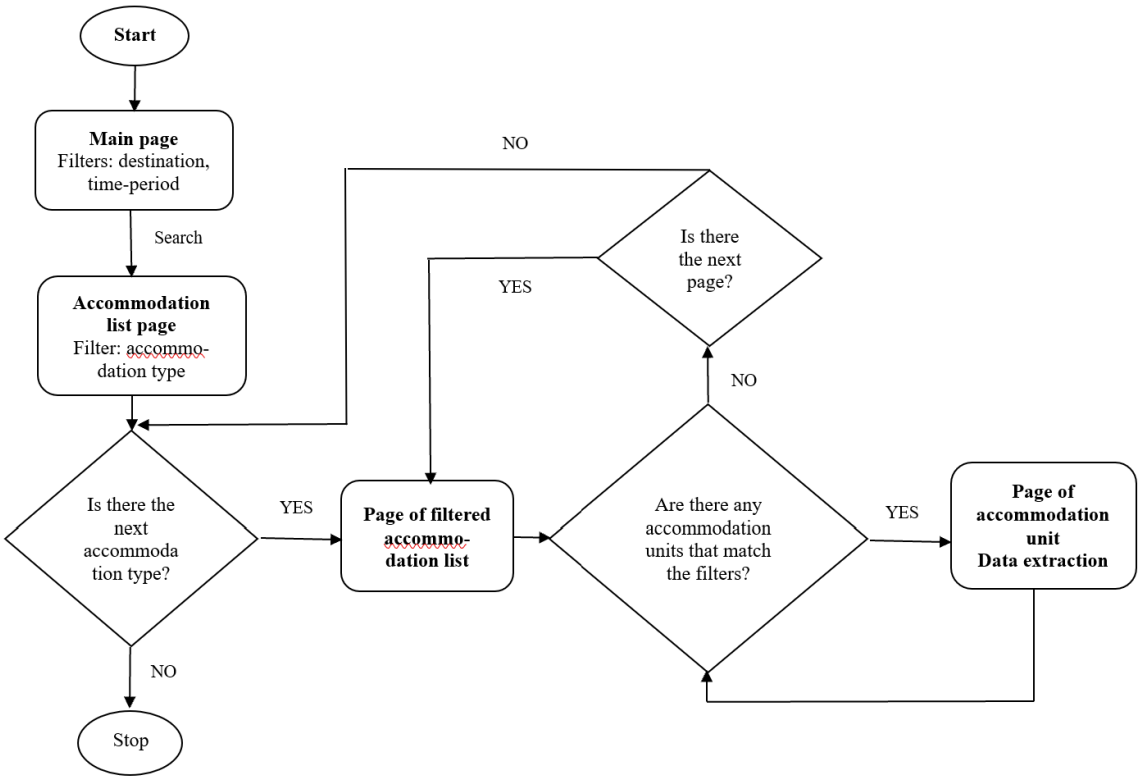
1.4. Data-driven Leadership and Evidence-Based Governance

Recent academic work emphasizes the role of data-driven decision-making (DDD) in tourism governance. ResearchGate-indexed studies on DMOs demonstrate that the adoption of data analytics and digital tools enhances organizational capacity to detect patterns, forecast demand and design evidence-based strategies. Academic literature conceptualizes data-driven leadership as the ability to transform complex datasets - including web-scraped indicators and multivariate analytical outputs - into actionable insights for strategic planning and regional development.

In this context, the analytical approach of the present article supports evidence-based tourism governance by offering county-level tourism profiles and identifying clusters with distinct levels of development and service quality. This contributes to academic discussions on data-driven leadership and provides practical foundations for differentiated regional policies in Romania.

2. Data and Methodology

In this research, a web scraping script was developed to obtain data that would allow characterizing the typology and quality of the Romanian tourism offer in a territorial profile. The program was created in the Python language in the PyCharm programming framework. The functioning and operation of the program is summarized in the diagram in *Figure 1*, which illustrates the sequence of data retrieval steps.



Source: Boboc *et al.,* 2025.

Figure 1. **Logical Workflow Diagram of the Web Scraping Program**

Agoda, part of Booking Holdings Inc., a globally recognized online travel services platform, was selected as the data source. The information collected referred to a one-night stay for two adults in one room, with the following attributes extracted: unit name, accommodation type, displayed rate, address, available facilities, overall score, detailed ratings for cleanliness, amenities, location, comfort, services, value for money, and total number of reviews.

By applying this algorithm to all counties in Romania, information was initially obtained on 7109 accommodation units, available between 1-2 July 2024, for two persons. The database was subsequently subjected to a cleaning and standardization process, the final dataset including 6888 valid observations.

The variables retained for subsequent analyses were: county, type of accommodation (*apartment, guesthouse/ bed and breakfast, hotel, hostel, private villa, camping, motel, complex, mansion, farm stay, inn*), price (in RON), address of the unit, facilities offered, overall rating and ratings for cleanliness, amenities, location, comfort, services and value for money, as well as the total number of reviews.

The data collected through the web scraping procedure, reflecting the structure and quality of the tourism offer at national level, formed the basis for performing a cluster analysis applied to Romania's counties. The purpose of this stage was to group the administrative-territorial units according to the similarities between the characteristics of the local tourism market - such as the typology of accommodation units, the price level, the facilities offered and the evaluations received from tourists. Through this classification method, the aim was to identify distinct regional profiles of Romanian tourism, reflecting the differences in the degree of development and tourist attractiveness between counties, thus representing a basis for further interpretations and analyses.

In the cluster analysis, Ward's hierarchical clustering method was applied to determine the number of clusters. Also known as the error sum of squares grouping method, it appeared in 1963 in a publication discussed by Ward, the basic idea being to group clusters where the loss of inertia between them is minimal. Ward's algorithm is an iterative process, minimizing the dispersion within the groups at each association, thus increasing the diversity between the groups (Murtagh and Legendre, 2014). Next, the K-Means algorithm is applied to verify the stability of the clusters obtained by Ward's algorithm, but also to identify the features of each group. The algorithm was developed in 1967 by James MacQueen and aims to classify individuals into a specified number of clusters so that the variability within the group is minimal, respectively, the variability between the groups is maximal.

In the second stage of the analytical approach, the set of variables - standardized - used in the cluster analysis was subjected to principal component analysis (PCA), in order to highlight the internal structure of the data and the relationships between the indicators describing the tourism offer. The application of the method allowed the dimensionality of the database to be reduced by transforming correlated variables - such as the type of accommodation units, the level of tariffs, the facilities offered and the evaluation scores - into a reduced number of independent components, which synthesize the essential information about the characteristics of the tourism market. Through this analysis, the main dimensions that explain the variation in the accommodation offer at the level of the Romanian counties were identified, facilitating a clearer interpretation of the regional differences and a statistical substantiation of the previously obtained groupings.

3. Results and Discussions

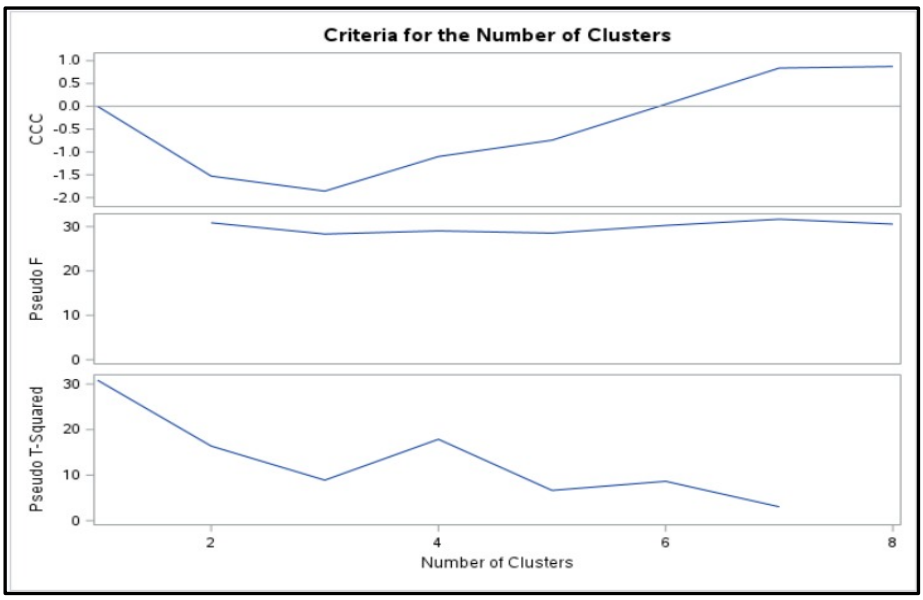
From the analysis of tourism data obtained through web-scraping on the online tourism platform considered, in terms of the distribution of the number of accommodation units in territorial profile, 10 counties in Romania have the lowest number of such units - under 38. These include two counties in Moldova (Botoșani and Vaslui), but also several counties in the south of the country (such as Olt, Ialomița, Călărași). At the opposite pole are the counties of Sibiu, Brașov, Prahova, Cluj, Timiș, Constanța, which present a high number of accommodation units, up to 800, indicating a developed tourist infrastructure, but also the Municipality of Bucharest with over 836 accommodation units. From the perspective of the average price of a night of accommodation, Călărași county has the lowest accommodation prices (under 230 lei), while Timiș and Alba counties record the highest prices (over 480 lei). The counties with a high number of reviews are Cluj, Timiș, Sibiu, Brașov, Prahova and Constanța, indicating an intense tourist activity and a closer interaction of tourists with travel and tourism websites. At the opposite pole - with a

low number of reviews left by tourists (under 10,000 reviews) - there are 17 counties (40% of the counties of Romania), located in the south of the country, in the south of Moldova or in the south-west of the country, areas characterized by a lack of tourist infrastructure and/or a lack of visibility and promotion (Boboc et.al, 2025).

3.1. Identifying Regional Patterns through Cluster Analysis of Tourism Web-scraped Data

Cluster analysis is a method of identifying groups as close to reality as possible, from a database, with little information about them (Anderberg, 1973). Thus, a way of organizing data with common elements into groups is achieved, which must be homogeneous, uniform inside, respectively heterogeneous, different outside. In the present research, the purpose of the developed cluster analysis is to segment the counties of Romania into homogeneous groups based on the main characteristics of the tourist accommodation offer, obtained following the application of the webscraping technique. By grouping counties according to similar profiles, the analysis allows the identification of regional patterns of tourism development and highlighting the differences between areas with high performance and those with insufficiently exploited potential.

Specifically, for the cluster analysis, the following clustering variables were taken into account for each county of Romania: the number of accommodation units, the average price of an accommodation night, the average general rating, but also for various services (cleanliness, amenities, location of the accommodation unit, degree of comfort, quality/price ratio, other services), the total number of reviews and the average number of facilities. In carrying out the cluster analysis, two complementary techniques were used in combination to increase the robustness and accuracy of the results: the Ward method, to identify the optimal grouping structure based on the distances between counties, and the K-Means method, to consolidate and validate the clusters obtained by dividing the counties into homogeneous groups.



Source: Created by the authors in SAS Studio

Figure 2. **CCC, Pseudo F and Pseudo T-Square Plots - Criteria for Choosing the Number of Clusters**

Following the application of Ward's method, the counties will be grouped into four clusters, chosen based on the three graphs below: the Cubic Clustering Criterion (CCC), Pseudo F, and Pseudo T-Square (*Figure 2*).

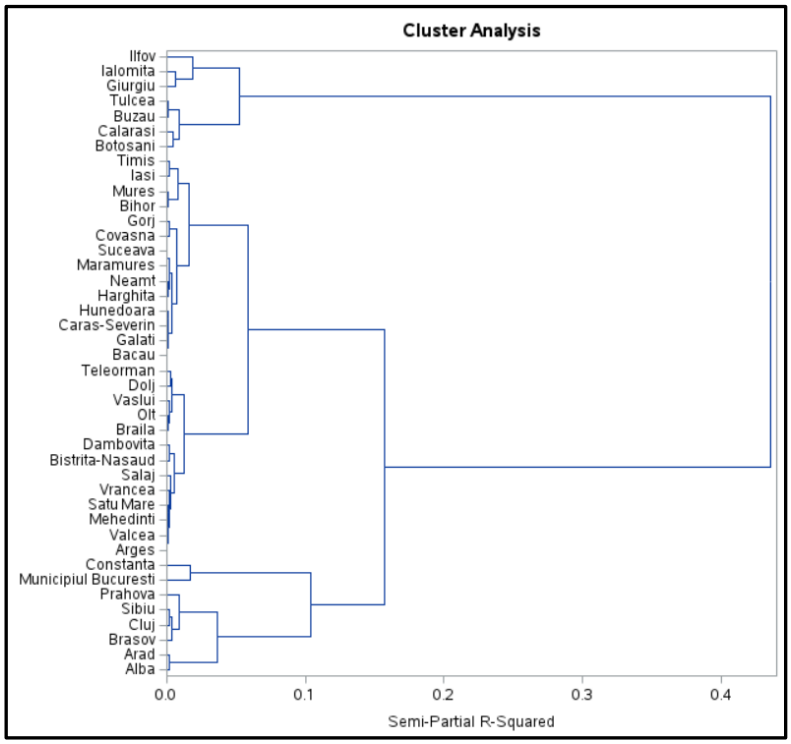
The K-Means algorithm grouped the counties into 4 clusters, so in the first cluster there are 2 counties, in the second cluster, 7 counties, in cluster 3, there are 24 counties, and in the last cluster 9 counties. Analyzing table 1, we can see the membership of the counties in the clusters. Thus, in the first cluster there are Constanța and Bucharest, in the second cluster there are Călărași, Buzău, Giurgiu, Ilfov, Ialomița, Botoșani and Tulcea. In cluster 3 there are 24 counties, including Mehedinți, Argeș, Gorj, Iași, Vâlcea, Galați, Dolj, Covasna, and in the last cluster there are Cluj, Mureș, Bihor, Brașov, Timiș, Alba, Arad, Sibiu and Prahova.

Table 1. Cluster composition

Cluster no.	Number of elements in cluster	Cluster composition (counties)
1	2	Constanța and București
2	7	Călărași, Buzău, Giurgiu, Ilfov, Ialomița, Botoșani and Tulcea
3	24	Iasi, Gorj, Covasna, Suceava, Maramures, Neamt, Harghita, Hunedoara, Caras-Severin, Galati, Bacau, Teleorman, Dolj, Vaslui, Olt, Braila, Dambovita, Bistrita-Nasaud, Salaj, Vrancea, Satu-Mare, Mehedinti, Valcea, Arges
4	9	Cluj, Mureș, Bihor, Brașov, Timiș, Alba, Arad, Sibiu and Prahova

Source: created by the authors.

The dendrogram in *Figure 3* illustrates the four clusters obtained, as well as the counties that are part of them.



Source: created by the authors in SAS Studio.

Figure 3. **Dendogram**

Figure 4 presents the characteristics of the four clusters obtained after applying the K-Means method, in terms of the number of observations (counties) included, the internal standard deviation, the maximum distance between an element and the cluster center, as well as the distance between the cluster centroids. The results indicate a high internal homogeneity (RMS Std Deviation between 0.1035 and 0.1950), with the counties in the third cluster presenting the highest degree of similarity in terms of the variables considered, while - at the opposite pole - Clusters 1 and 2 have a more pronounced internal variability, composed of counties with a more diverse tourist profile. The large distances between centroids (up to 1.3091) confirm the clear separation between the tourist profiles of the clusters.

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	2	0.1950	0.4574		4	1.3091
2	7	0.1875	0.8391		3	1.3037
3	24	0.1035	0.5373		4	0.7181
4	9	0.1466	0.7062		3	0.7181

Source: created by the authors in SAS Studio.

Figure 4. **Cluster Summary - Cluster Characteristics - K-Means Method**

Table 2 provides information on how the included variables contribute to the formation of clusters. Thus, for most variables, Within STD is significantly lower than Total STD, which confirms that the grouping resulted in homogeneous clusters. For most of the variables, the R-Square indicator has values above 0.7, suggesting that all these variables have a high power to segment clusters. The largest contributions to the formation of clusters are made by variables related to the quality of services (ratings on cleanliness, facilities, quality/price ratio, etc.) and volumetric indicators of the tourist market (total number of reviews and number of tourist accommodation units), while the average price of the stay and the average number of facilities contribute the least to the differentiation of these groups. Therefore, ratings related to services and the quality of offers are strongly discriminatory, which means that counties differ significantly in tourists' perception of the quality of services and offers and less in terms of prices and the number of facilities.

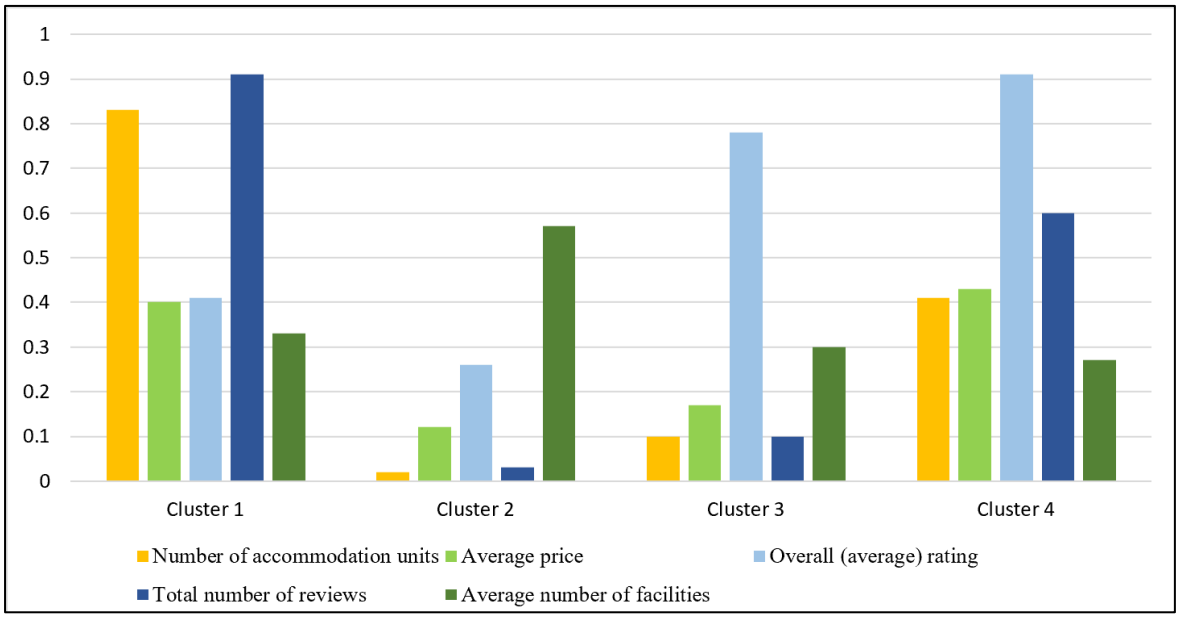
Table 2. Statistics for Variables - counties' contribution to cluster formation

Statistics for Variables				
Variables	**Total STD**	**Within STD**	**R-Square**	**R SQ/(1- R SQ)**
Number of accommodation units	0.22393	0.10485	0.796817	3.921673
Total number of reviews	0.29624	0.13585	0.805093	4.130652
Average number of facilities	0.18693	0.15975	0.323099	0.477321
Average cleanliness rating	0.24661	0.11669	0.792485	3.818918
Average amenities rating	0.23404	0.10993	0.795524	3.890557
Average location rating	0.23474	0.13365	0.699572	2.328582
Average comfort rating	0.23458	0.12999	0.715385	2.513517
Average services rating	0.22344	0.13178	0.677624	2.101969
Average value-for-money rating	0.24046	0.12383	0.754217	3.068624
average price	0.20321	0.16999	0.351422	0.541836
Overall (average) rating	0.25713	0.12702	0.773826	3.421370
OVER-ALL	0.23621	0.13251	0.708297	2.428147

Source: created by the authors in SAS Studio.

Based on the graph below (*Figure 5*), Cluster 1 is characterized by a large number of accommodations and reviews, with moderate prices and ratings in most categories. Bucharest and Constanța are part of this cluster, differing from the other counties by their popularity and wide availability, accessible offers for both locals and tourists, but the quality of services can vary. This group illustrates tourist centers with a diversified and accessible offer, where the high volume of demand and supply can lead to variations in the perceived quality of services.

Cluster 2 contains 7 counties, namely, Călărași, Buzău, Giurgiu, Ilfov, Ialomița, Botoșani and Tulcea, characterized by a very small number of accommodations, reviews and the lowest prices and ratings. Thus, these counties are perceived as less visited and known, offering economic options that do not excel in quality or popularity, but which may have potential for growth and development.



Source: Created by the authors in SAS Studio

Figure 5. **Cluster Means - Characterization of Clusters through the Distribution of Variables**

Cluster 3 contains the majority of counties, namely 24, among which Mehedinți, Argeș, Gorj, Iași, Vâlcea, Galați, Dolj, Covasna, stand out for a balance between the number of accommodation units, the number of reviews and relatively low prices, differentiating themselves from the other counties by their superior quality, highlighted by the high ratings in all categories. These counties thus offer a superior quality option, despite the moderate availability, at lower prices compared to other similar locations, attracting tourists eager for a satisfactory experience in various fields, from tourism to trade and services.

Cluster 4 includes the counties of Cluj, Mureș, Bihor, Brașov, Timiș, Alba, Arad, Sibiu and Prahova, which stand out for a considerable number of accommodation units and reviews, higher prices and the highest ratings in all categories. With a focus on excellence, these counties offer unique experiences, and the higher prices are justified by the high-quality services and products offered, with customer satisfaction being reflected in high ratings.

3.2. Identifying Key Dimensions of Regional Tourism through Principal Component Analysis

Principal component analysis is the general name for a technique that uses basic mathematical principles to simplify data sets and transform a number of correlated variables into a smaller number of variables called principal components, which retain the greatest variation in the data set. The main goal of this analysis is to identify trends, patterns, extreme values in the data and to reduce the size of the data set without losing too much information.

In the present research, the same variables were used to perform principal component analysis as in cluster analysis. Analyzing the correlation matrix, it is observed that all correlation coefficients are significant, with values greater than 0.6. All variables are positively correlated, most of which are strongly correlated, for example, the total number of reviews and the number of accommodations (correlation coefficient = 0.97), the average price and the average rating (correlation coefficient = 0.761). Thus, principal component analysis is useful in interpreting the links between the analyzed variables, the variables being sufficiently correlated to justify dimensionality reduction.

The eigenvalues in the matrix below (*Figure 6*) measure the amount of variance explained by each principal component. They decrease with the component index, with the first component having the maximum eigenvalue, the sum being equal to the number of initial variables, in this case 11. From the analysis of the eigenvalues it results that the set of variables can be divided into 2 principal components, the first value (9.22) explaining 9 variables, i.e. the 9 variables are replaced by a single one, reducing the size of the space greatly.

Eigenvalues of the Uncorrected Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	9.22225265	8.21289879	0.8384	0.8384
2	1.00935386	0.62267551	0.0918	0.9301
3	0.38667835		0.0352	0.9653

Source: Created by the authors in SAS Studio

Figure 6. **Eigenvalue Matrix**

Table 3. Eigenvectors

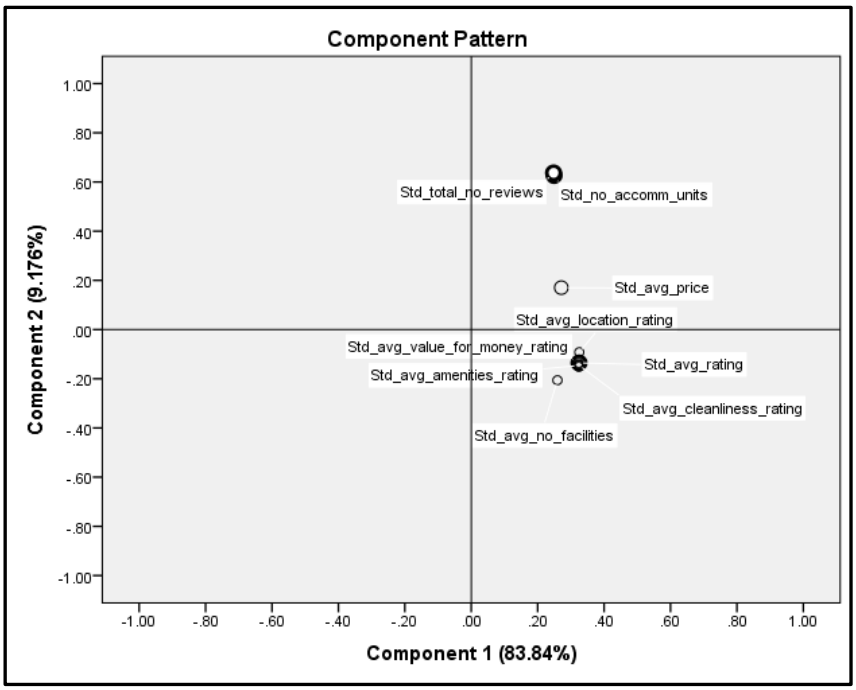
Eigenvectors				
		Prin1	Prin2	Prin3
Standardized_no_accomm_units	Number of accommodation units	0.249411	0.626311	-0.024108
Standardized_avg_price	Average price	0.271015	0.170122	0.287794
Standardized_avg_rating	Average rating	0.323924	-0.135418	-0.185068
Standardized_total_no_reviews	Total number of reviews	0.247930	0.637132	0.028538
Standardized_ avg_cleanliness_rating	Average cleanliness rating	0.323913	-0.144296	-0.133154
Standardized_avg_amenities_rating	Average amenities rating	0.323875	-0.146184	-0.146387
Standardized_avg_location_rating	Average location rating	0.325401	-0.092031	-0.103005
Standardized_avg_comfort_rating	Average comfort rating	0.323689	-0.151375	-0.138479
Standardized_avg_services_rating	Average services rating	0.323837	-0.142008	-0.105613
Standardized_avg_value-for-money rating	Average value-for-money rating	0.324392	-0.135392	-0.140327
Standardized_avg_no_facilities	Average number of facilities	0.259563	-0.206024	0.884159

Source: created by the authors in SAS Studio.

By projecting onto the first principal axis, 83.8% of the variability of the data set is preserved, and by projecting onto the plane determined by the two principal axes, 93% of the total variability is explained. Thus, since the first two principal components take over more than 80% of the variation of the initial data, the goal of dimensionality reduction is achieved.

Analyzing the eigenvector table (*Table 3*), we can identify the factors that determine each component. Thus, the first component (Prin1) is positively determined by the variables associated with the average rating for cleanliness, amenities, location, comfort, services, value for money. The second component (Prin2) is also positively determined by the number of accommodation units (0.626) and the total number of reviews (0.637).

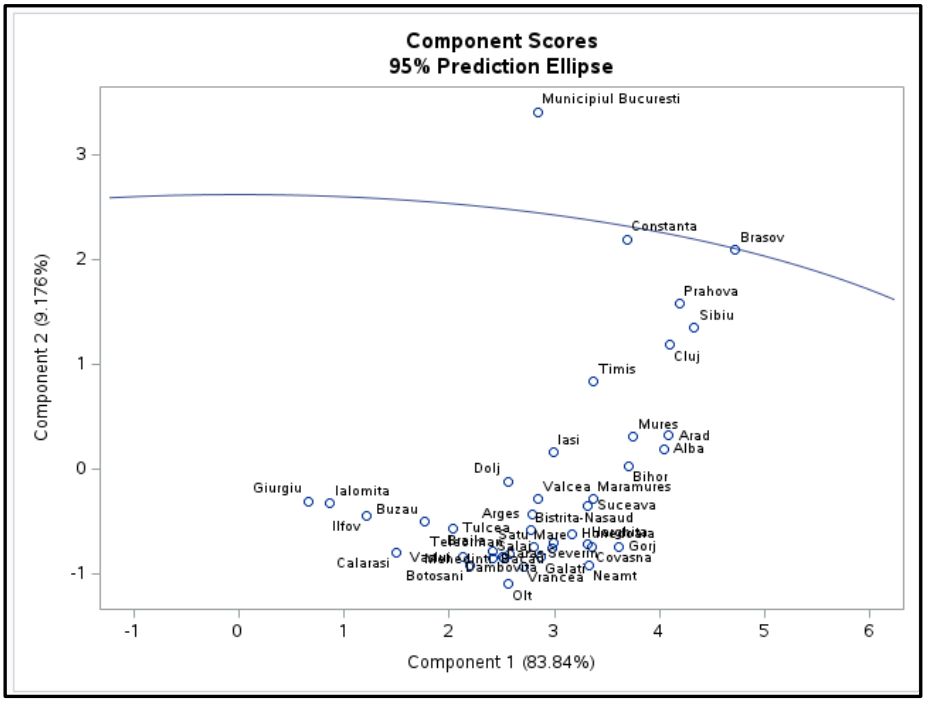
Regarding the relationship between the variables, it is observed that they are directly correlated for both components. The variables showing the number of reviews, the number of units and the average price are located close to each other, on the right side of the graph, indicating a strong positive correlation between them. In other words, they vary together, suggesting that counties with more accommodation units and more reviews also tend to have higher average prices (*Figure 7*).



Source: created by the authors in SAS Studio

Figure 7. **Graphical Representation of the Variables**

The variables describing the average rating and the average number of facilities offered by accommodation units are grouped in the lower part of the plot, being positively correlated with each other. Thus, if an accommodation unit has a good rating in one area, it is likely to have good evaluations in the other areas as well. In contrast, there is a negative correlation between the rating variables and the number of accommodation units, the number of reviews, and the average price, indicating that an increase in the number of units, in the number of reviews, or in the average price is associated with a decrease in the average ratings for cleanliness, amenities, location, comfort, services, value for money, and the average number of facilities, and vice versa.



Source: created by the authors in SAS Studio.

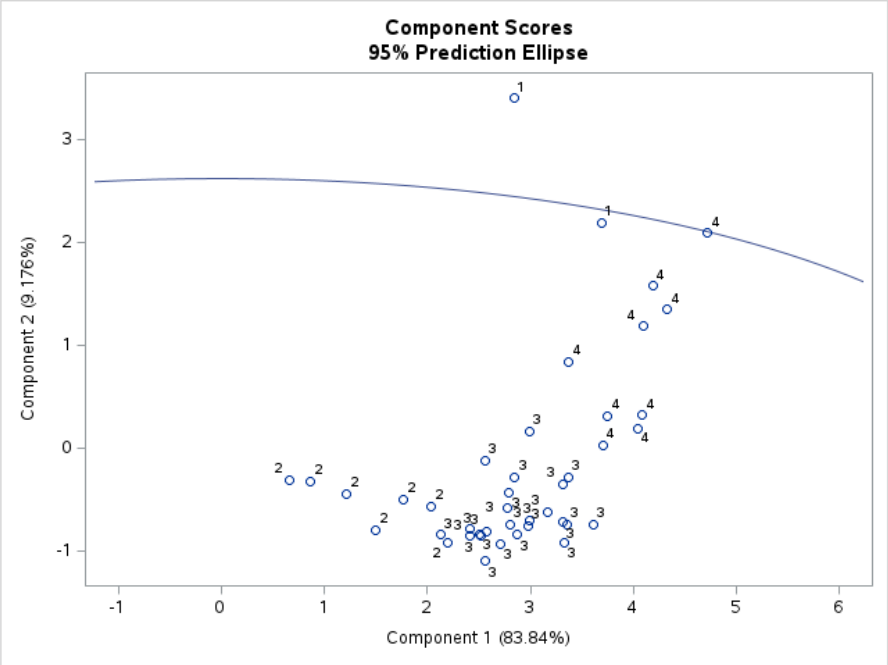
Figure 8. **Graphical Representation of Counties**

The purpose of this analysis is to highlight the disparities and similarities between the counties of Romania in terms of tourism development and the quality of tourism services, to identify counties with superior infrastructure and services, such as Bucharest, Constanța and Brașov, and to highlight counties that need improvements in these areas. Thus, the graph in Figure 8 illustrates the scores of the main components, using two main axes, called Tourism Development Index, on the horizontal axis (Component 1), respectively, Tourism Services Quality Index, on the vertical axis (Component 2).

It is observed that the Municipality of Bucharest is an outlier, being located at the top of the graph and characterized by a very high tourism development index, with a large number of accommodation units, many reviews and high prices, but also by a high tourism services quality index.

The counties of Constanța and Brașov indicate a high tourism development and a good quality of services, standing out with a well-developed tourism infrastructure due to their positioning. Constanța being a port city and a major tourist destination, and Brașov with its mountain and cultural attractions. The counties of Prahova, Sibiu, Cluj and Timiș present a high tourism development index, but also positive scores for Component 2, indicating a good quality of tourism services.

Most counties are concentrated in the central and left part of the graph. This suggests moderate characteristics for both axes, a moderate Tourism Development Index and a moderate Tourism Services Quality Index. Counties such as Giurgiu, Ialomița and Buzău are located in the lower part of the graph, indicating a lower level of both indices (*Figure 9*).



One observation from the graph above, included in Cluster 1, is an outlier, namely Bucharest.
Source: created by the authors in SAS Studio.

Figure 9. **Graphical Representation of the 4 Clusters**

Conclusions and Recommendations

In this paper, we have shown how the web scraping technique can serve as an alternative source of data for the tourism industry, becoming a valuable and up-to-date source, complementing and, in some cases, improving the information provided by official statistics.

The database used was created by a proprietary web scraping program written in Python, which collected data from an online booking site. Cluster analysis and principal component analysis contributed to achieving the objective of analyzing and segmenting the counties of Romania, depending on the type and quality of the tourism offer.

The study is notable for a high degree of methodological and application novelty, by using automatically collected data (web-scraped) from online tourism platforms, providing a current and detailed perspective on regional tourism in Romania. The originality of the research lies in the combined application of cluster analysis and principal component analysis to identify territorial typologies and the determinants of tourism quality and development. At the same time, by correlating objective indicators (number of units, prices) with subjective ones (ratings and reviews), the study proposes a complex approach, oriented towards the tourist experience, contributing to the foundation of differentiated strategies for tourism development and promotion at the regional level.

The cluster analysis highlighted significant differences in the perception of the quality of tourism services, by grouping the counties into four distinct clusters, where the strengths and weaknesses of the tourism offers at regional level were identified. Based on data collected through web scraping, this analysis allows for a rapid reaction to changes in consumer preferences and the adaptation of the tourism offer in a short time. Clusters based on different online ratings and reviews provided by tourists provide instant feedback

on customer satisfaction, allowing for rapid interventions by operators in the tourism industry. By carrying out this analysis, a high-quality tourism offer can be maintained and can help to continuously improve services in each county. Specifically, the counties included in Cluster 1 (Bucharest and Constanța) are distinguished by the very high number of accommodation units and reviews, as well as by moderate prices and ratings. Cluster 2 is made up of seven counties with low values for all the variables analyzed, reflecting regions with poorly developed tourism, characterized by a low number of accommodation units, low prices and lower ratings. These counties can be considered peripheral areas from a tourism point of view, but have growth potential through investment, promotion and diversification of the offer. Cluster 3, the most numerous (24 counties), is characterized by a balance between accessibility, quality and price. The counties in this group offer good quality accommodation services at moderate prices, being representative of the average Romanian tourism, which ensures tourist satisfaction without high costs. This cluster suggests a competitive segment of the market, capable of combining economic efficiency and quality of experience. Cluster 4 brings together the counties with the highest level of tourism development, such as Cluj, Brașov, Sibiu, Timiș or Prahova, which stand out for their extensive tourist infrastructure, high prices and excellent ratings. These destinations represent poles of excellence in Romanian tourism, where the higher price is compensated by the superior quality of services and customer satisfaction.

Principal component analysis helped simplify and understand the data by identifying two main components that influence both tourism development and the quality of tourism services, thus highlighting disparities and similarities between counties. A first component of the quality of the tourism offer integrates variables related to general and thematic evaluations of accommodation units (cleanliness, comfort, facilities, quality/price ratio, services, location), along with the average price. This reflects the general level of tourist satisfaction and the perceived quality of services, indicating counties that have a competitive tourism offer, oriented towards high standards and positive experiences. The second component, a component of tourism development, is mainly associated with the number of accommodation units and the total volume of reviews, representing the extensive dimension of tourism - respectively the degree of development and numerical attractiveness of destinations. This component differentiates counties with a more developed tourism infrastructure and a more active online presence from those in the early stages of tourism development. Together, the two components provide an integrated picture of regional tourism: the first captures the qualitative aspects of the tourism experience, and the second the quantitative aspects, related to the scale and dynamics of the offer. Their combined interpretation allows for a better understanding of the balance between infrastructure development and maintaining quality standards, essential information for guiding tourism development and promotion policies at county and national level.

This analysis also presents the essential aspects that require improvement through a more detailed assessment of the tourism experience, providing the necessary tools for making quick decisions in the tourism industry.

Thus, it is recommended that decision-makers take measures to stimulate investments in counties with poor tourism development (Cluster 2), through financing programs and public-private partnerships, in order to encourage infrastructure development, diversify the accommodation offer and create integrated tourism products (eco-tourism, cultural tourism, rural tourism, etc.).

Also, another direction in which action can be taken is to promote regional balance and inter-county cooperation. Thus, regions with consolidated tourism (Cluster 4) can become centers for mentoring and transferring good practices for emerging counties, through joint marketing and professional training

projects. Orienting public policies towards quality and sustainability can ensure the avoidance of oversaturation of established destinations, while development strategies can aim to increase the quality of services in areas with medium potential (Cluster 3), maintaining an optimal ratio between price and visitor satisfaction. Data shows that digital presence (number of reviews, ratings, etc.) plays an essential role in the attractiveness of tourist destinations. Digital marketing campaigns and collaboration with tourism platforms can help increase the notoriety of the counties in Cluster 2, helping to improve online visibility and promote lesser-known destinations. At the same time, regional tourism policies should be developed in a differentiated manner.

Since each cluster represents a different stage of tourism development, public strategies must be adapted to the specifics of each group, from consolidating excellence (Cluster 4) to stimulating emergence (Cluster 2).

The research can be extended, in the future, by using the web scraping method to other online booking platforms, as well as over a different time horizon, to obtain an even more comprehensive picture of the tourism sector. It would also be useful to integrate other data sources, such as social networks and reviews on travel platforms, to analyze tourists' perception in real time.

In conclusion, the results obtained confirm the usefulness of integrating web-scraped data and multivariate statistical methods in the analysis of regional tourism, offering a modern and applied perspective on territorial disparities and constituting a valuable tool for guiding policies for the development and sustainable promotion of tourism in Romania.

Literature

Ahas, R., Armoogum, J., Esko, S., Ilves, M., Karus, E., Madre. J.L., Nurmi, O., Potier, F., Schmücke, D., Sonntag, U., Tiru, M. (2014), *Feasibility study on the use of mobile positioning data for tourism statistics*, ISBN 978-92-79-39762-2, Luxembourg: Publications Office of the European Union.

Ajčnerová, I., Šácha, J., Ryglová, K., Žiaran, P. (2016), "Using the cluster analysis and the principal component analysis in evaluating the quality of a destination", *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, Vol. 64, No 2, pp.677-682.

Anderberg, M.R. (1973), *Cluster analysis for applications*, Academic Press.

Andronic, E. (2025). "Assessment of Romania's tourism competitiveness through Competitive Importance-Performance Analysis (CIPA)", *Administrative Sciences*, Vol. 15, No 9, p.358.

Boboc, C.R., Babaligea, A.M., Ghiță, S.I., Săseanu, A.S. (2025). "Leveraging web-scraping for tourism data analysis: A case study on Romania", in: Vasiliu, C., Dabija, D. C., Tziner, A., Pleșea, D., Dinu, V. (eds.), *BASIQ 2025 International Conference. New Trends in Sustainable Business and Consumption. Conference Proceedings, 26-28 June 2025, Oradea, Romania*, ASE Publishing, Bucharest, pp.174-181. ISSN 2457-483X. https://conference.ase.ro/papers/2025/Volum%202025.pdf.

Buhalis, D., Jun, S. H. (2011), *Contemporary Tourism Reviews: E-Tourism*. http://www.goodfellowpublishers.com.

Constantin, D.L., Popescu, C., Drăguşin, M. (2018), "A spatial analysis of tourism infrastructure in Romania", *Region*, Vol. 5, No 1, pp.1-20.

Corfu, A., Azevedo, S. (2007), "Some applications of the web as an information tool in the tourism industry", *Amfiteatru Economic,* Vol. IX(21), pp.20-25.

Coroş, M.M., Negruşa, A.L. (2014), "Contemporary approaches and challenges of sustainable tourism: analysis of the evolution and performance of the tourism offer in Romania and Transylvania", *Amfiteatru Economic*, Special Issue, Vol. 8.

Constăngioară, A., Ban, O., Ruge, P., Coita, D.C., Țarcă, N., Constăngioară, A.M. (2025), "Digital Transformation: the Impact of e-Commerce on Organisational Performance in Romanian Tourism and Accommodation SMEs", *Amfiteatru Economic*, Vol. 27, No Special Issue 19, pp.1404-1419. https://doi.org/10.24818/EA/2025/S19/11404.

Davidescu, A.A.M., Strat, V.A., Grosu, R.M., Zgură, I.D., Anagnoste, S. (2018), "The Regional Development of the Romanian Rural Tourism Sector", *Amfiteatru Economic*, Vol. 20, No Special No 12, pp.854-869. DOI:10.24818/EA/2018/S12/854.

Demunter, C. (2017), "Tourism statistics: early adopters of big data?", in *Session 5 - Producing Data on Sustainable Tourism,* ISBN 978-92-79-71899-1, Luxembourg: Publications Office of the European Union, pp.6-28.

Hong, W. C. (2008), *Competitiveness in the tourism sector. A comprehensive approach from economic and management points*, Physica-Verlag.

Murtagh, F., Legendre, P. (2014), "Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion?", *Journal of Classification*, Vol. 31, No 3, pp.274-295. https://doi.org/10.1007/s00357-014-9161-z.

Narendra, A.P. (2015). "Data besar, data analisis: dan pengembangan kompetensi pustakawan [Big data, data analyst, and improving the competence of librarians]", *Record and Library Journal,* Vol. 1(2), pp.83-93.

National Institute of Statistics (2024), "Informații statistice - seria Statistica de turism, nr. 4/2024: Activitatea de turism pe anul 2024", Bucharest. https://insse.ro/cms/sites/default/files/field/publicatii/seria_turism_in_anul_2024_0.pdf. [Statistical information – Tourism statistics series no. 4/2024: Tourism activity in 2024,

Petroman, I. (2010), *Managementul turismului cultural în Judeţul Timiş: politici de intervenţie*, Editura Eurostampa, Timişoara, [Cultural Tourism Management in Timiș County: Intervention Policies *in Romanian*].

Săseanu, A. S., Ghiță, S. I., Albăstroiu, I., Stoian, C.-A. (2020), "Aspects of digitalization and related impact on green tourism in European countries", *Information*, Vol. 11, No 11, p.507.

Sofronov, B. (2018), "The development of the travel and tourism industry in the world", *Annals of Spiru Haret University - Economic Series*, Vol. 18, No 4, pp.123-137. https://doi.org/10.26458/1848.

Stantic, B., Pokorny, J. (2014), "Opportunities in big data management and processing", *Frontiers in Artificial Intelligence and Applications*, Vol. 270.

Ștefan, S.C., Popa, S.C., Breazu, A., Mircioiu, C.E., Beldiman, C.M. (2025), "Unlocking Romania's Tourism Potential: a Data-Driven Analysis of Competitiveness Factors", *Amfiteatru Economic*, Vol. 27 No. Special Issue 19, pp.1292-1312. https://doi.org/10.24818/EA/2025/S19/1292.

Tanasă, A.M. (2024), "Web scraping and review analytics: extracting insights from online reviews", *Annals of Ovidius University of Constanța - Economic Sciences Series*, Vol. 24, No 1, pp.1-18.

The World Tourism Organization (UNWTO) (2025), *World Tourism Barometer*. https://www.untourism.int/un-tourism-world-tourism-barometer-data

Tokarchuk, O., Gabriele, R., Maurer, O. (2021), "Estimating tourism social carrying capacity", *Annals of Tourism Research*, Vol. 86. https://doi.org/10.1016/j.annals.2020.102971.

Trușescu, M.N., Nicolaie, D. (2024). "Sentiment Analysis: a Comparative Study of Online Booking Platforms Reviews for the Northern Oltenia Balneary Destination", *Balneo and PRM Research Journal*, Vol. 15, No 1, pp.651.

World Travel & Tourism Council (WTTC) (n.d.), *Travel & Tourism Economic Impact Research (EIR).* https://wttc.org/research/economic-impact.

Zaharia, R. M. (2024), "Analysis of the main components in Romanian tourism performance indicators", *Annals of Ovidius University of Constanța - Economic Sciences Series*, Vol. 24, No 2, pp.1-15.

REGIONINIO TURIZMO RUMUNIJOJE YPATUMAI: INTERNETO DUOMENŲ IR DAUGIAMATĖS STATISTINĖS ANALIZĖS TAIKYMAS

Cristina Rodica Boboc, Ana Maria Babaligea, Simona Ioana Ghiță, Claudiu Nicolae Ghinea, Cristian Constantin Francu

Santrauka. Straipsnyje nagrinėjamas Rumunijos apskričių turizmo profilis. Tam pasitelkti automatiškai surinkti duomenys iš didelės internetinės kelionių platformos. Į analizę integruoti kintamieji, susiję su apgyvendinimo įstaigų skaičiumi, vidutinėmis kainomis, bendrais ir konkrečioms kategorijoms skirtais įvertinimais, atsiliepimų skaičiumi ir turimais patogumais. Siekiant nustatyti regionines tipologijas ir lemiamus turizmo kokybės bei plėtros veiksnius, atlikta klasterinė analizė ir pagrindinių komponentų analizė (PCA). Rezultatai atskleidė keturis skirtingus klasterius, atitinkančius skirtingus išsivystymo ir patrauklumo lygius, taip pat du pagrindinius komponentus: turizmo pasiūlos kokybės komponentą, atspindintį lankytojų pasitenkinimą ir paslaugų suvokimą, ir turizmo plėtros komponentą, susijusį su turizmo veiklos mastu ir intensyvumu. Tyrime pateikiamas integruotas regioninių skirtumų vaizdas, akcentuojant brandžias turizmo vietas, apskritis su augimo potencialu bei besiformuojančias sritis. Taip pat suformuluotos rekomendacijos dėl diferencijuotos regioninės plėtros ir turizmo skatinimo politikos. Novatoriškas šio tyrimo indėlis susijęs su alternatyvių, iš interneto surinktų duomenų, kurie papildė ir praturtino oficialią statistiką ir duomenis, naudojimu. Aktualios ir laiku pateiktos išsamios įžvalgos, orientuotos į tikrąją turistų patirtį. Remiantis tyrimo išvadomis, sprendimų priėmėjams pateiktos rekomendacijos, kuriomis besivystančiuose regionuose siekiama skatinti augimą, gerinti paslaugų kokybę ir remti subalansuotą regioninio turizmo plėtrą visoje Rumunijoje. Integravus skaitmeninius duomenis ir daugiamatę analizę, tyrimas prisideda prie regioninio turizmo konkurencingumo ir veiklos rezultatų analizės Rumunijoje.

Reikšminiai žodžiai: internetinių duomenų rinkimas; regioninis turizmas; daugiamatė analizė; klasterinė analizė; pagrindinių komponentų analizė.