Optimizing Tourism Data Extraction and Analysis: A Comprehensive Methodology

José Javier Galán-Hernández, Ramón Alberto Carrasco-González, and Gabriel Marín-Díaz

Abstract Objective: There are various sources that provide data related to tourism. However, at times, this data lacks structure or is found in sources that do not facilitate its easy, automatic, or unsupervised collection. In such situations, a methodology employing data science techniques offers a significant advantage to researchers. They can leverage the tools available through the proposed methodology to extract, process, and analyze information efficiently. While this methodology is applicable to various disciplines, this work presents a specific case focused on tourism in Spain. Methodology: Employing data science techniques like graph analysis and unsupervised machine learning, we collect and process data on tourists' origins and numbers in Spain, using Python, R, and VOSViewer. The analysis uncovers primary tourism sources and origin-country patterns. It delves deep into Andalusia due to its high tourist influx. Results: Our study reveals key Spanish tourism sources and visitor behavior patterns. Visual data illustrates tourist origins, visit numbers, and interactions. Additionally, Andalusia is thoroughly examined for visit counts and origin countries. Conclusions: Employing data science, our study yields insights into Spanish tourism, identifying core sources and understanding origin-country interactions. These findings inform strategic decisions and enhance Spain's tourism promotion and management.

Keywords Tourism · Data science · Vosviewer · Python · Methodology

J. J. Galán-Hernández (✉)
Departamento de Sistemas Informáticos y Computación, Facultad de Estudios Estadísticos, Universidad Complutense de Madrid, 28040 Madrid, Spain
e-mail: josejgal@ucm.es

R. A. Carrasco-González
Departamento de Marketing, Facultad de Estudios Estadísticos, Universidad Complutense de Madrid, 28040 Madrid, Spain

G. Marín-Díaz
Departamento de Sistemas Informáticos y Computación, Facultad de Estudios Estadísticos, Universidad Complutense de Madrid, 28040 Madrid, Spain

A. J. Guevara Plaza et al. (eds.), *Tourism and ICTs: Advances in Data Science, Artificial Intelligence and Sustainability*, Springer Proceedings in Business and Economics,
https://doi.org/10.1007/978-3-031-52607-7_4

1 Introduction

Tourism plays a fundamental role in Spain's economy (Gonzalez & Moral, 1996), making it one of the most visited countries in the world. Understanding the primary source of tourism and obtaining relevant visitor information is essential for driving the development and promotion of the sector. Fortunately, in the era of data science and advanced technologies, we have effective tools and methodologies to address this challenge (Medina-Munoz et al., 2013).

In this study, we employ a proprietary methodology applying data science techniques to analyze tourism in Spain. Using technologies like Python, R, and VOSViewer (Moral-Muñozet al., 2020), we gather, process, and visualize information about the origin and number of tourists visiting the country. This methodology allows us to draw significant conclusions about the main sources of tourism and understand patterns and interactions among the countries of origin of tourists.

In particular, this study focuses on the autonomous community of Andalusia, one of the most prominent regions in terms of tourist reception in Spain (Martínez & Nicolás, 2014). We examine in detail the number of visits received and the countries of origin of tourists in Andalusia. With this information, we aim to gain a better understanding of the flow of tourists in the region and make strategic decisions to boost its tourism development.

Data analysis and network visualization through graph theory play a key role in this study (Abbasi-Moud et al., 2021), allowing us to effectively represent and explore information about tourism in Spain. Through this data science methodology, we seek to obtain an evidence-based, in-depth view of tourism in Spain, with an emphasis on Andalusia, to contribute to sustainable development and informed decision-making in the tourism sector.

2 Methodology of Analysis

To achieve the introduced objectives, a proprietary methodology is carried out, see Fig. 1, based on an adaptation of the CRISP-DM methodology (Moine et al., 2011), widely used in the field of data mining (Burbano & Anderson, 2023). This methodology has been specifically adapted to address the study of tourists visiting Spain and has been enriched with advanced network visualization techniques using graph theory.

Each stage of the methodology is analyzed in the following sections.

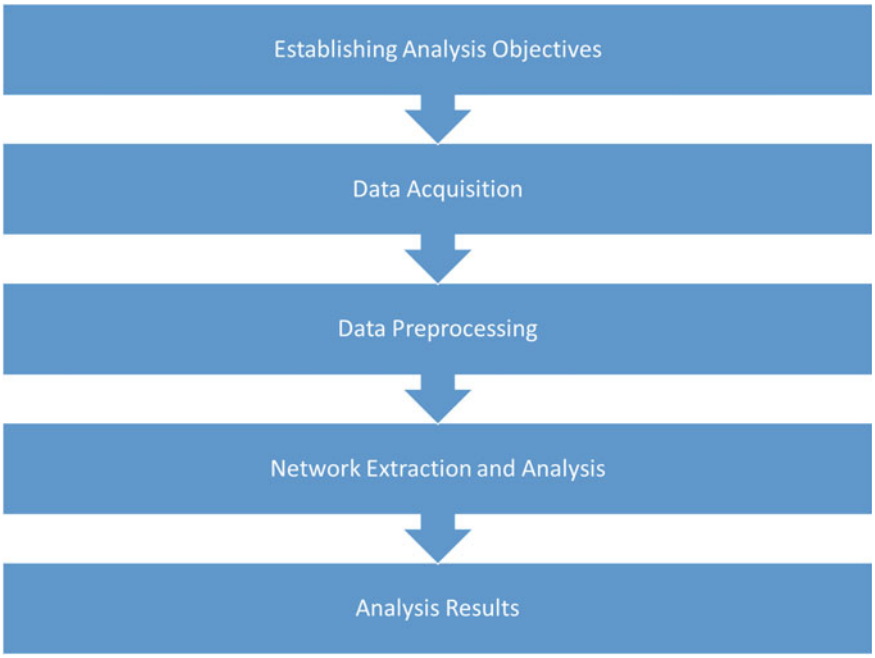

Fig. 1 Methodology used

2.1 Setting Analysis Objectives

Following the methodology outlined in Chap. 2, the objective is established, which is to collect data related to tourist visits to each of the autonomous communities of Spain by international tourists during the year 2022. This data is available on the official website https://www.dataestur.es/ (Diezma, 2021). Since the manual collection process would be extremely time-consuming, we have automated the data download and processing. Once the processed data is collected, we intend to create a network graph that allows us to perform a visual analysis of the flow of tourists to Spain, as well as a more specific analysis focused on the autonomous community of Andalusia.

2.2 Data Acquisition

As previously mentioned, the first objective is to obtain data from the official website https://www.dataestur.es/, which offers the option to consume its web service to obtain the desired information. However, this process involves numerous steps, as it is necessary to select each country for each autonomous community. With a total of 15 countries and 19 autonomous communities, there are a total of 285 possible

combinations. Performing this task manually would consume a significant amount of time and effort. For this reason, we have decided to use Python (Stančin & Jović, 2019) to automate the process and efficiently download a file with the information corresponding to each combination. Thanks to this automation, we can complete the task in approximately 2 h, unattended.

The automation of this process offers several significant advantages. Firstly, by using Python to automate the information download, we can save a considerable amount of time and effort. Given the large number of possible combinations (285 in this case), manually performing this task would be extremely laborious and error-prone.

Furthermore, using an automated approach ensures higher accuracy in data collection. By eliminating manual intervention, we reduce the possibility of human errors, providing greater reliability in the results obtained.

Another advantage is that automation allows us to carry out the task unattended. Once the process has been programmed in Python, we can run it and let it run in the background without requiring constant supervision. This allows us to use our time for other important activities while the process is ongoing.

In summary, the advantages of using automation with Python to download data from different combinations include time and effort savings, increased accuracy in data collection, and the ability to perform the task unattended.

2.3 Data Preprocessing

Once the data has been downloaded through the previous phase, we proceed to clean the information. In this process, we remove unnecessary columns for our analysis. Then, we use Python to merge all the files into a single Excel file.

This resulting Excel file serves as a consolidated information repository. While the Excel file is not a database itself, it can function as a large repository that stores relevant information for our study.

The use of Python allows us to automate this data cleaning and consolidation process efficiently. Thanks to Python's data manipulation capabilities, we can perform tasks such as removing unnecessary columns and combining files quickly and accurately.

In summary, after downloading the files corresponding to all possible combinations, we use Python (Sahoo et al., 2019) to clean and merge the information into a single Excel file. This Excel file acts as a centralized repository of information for our study, facilitating the analysis and processing of the collected data.

2.4 Network Extraction and Analysis

After preprocessing to obtain the relevant information, the next step is graph analysis. To do this, the processed data is exported to a format compatible with R, such as a CSV or Excel file.

Next, the powerful igraph library in R (Valdez, 2016) is used to build a graph that represents the connections between countries and autonomous communities of Spain in terms of tourist visits. This library allows for the straightforward creation of a directed or undirected graph and the assignment of attributes to nodes and edges based on the characteristics of interest.

The generated graph, see Fig. 2a, establishes a network between the countries visiting each autonomous community, providing a clear visualization of tourist relationships and flows. Each node in the graph represents a country or autonomous community, and the edges represent the connections between them based on the number of visits.

First, the graph is opened using the VOSViewer application (Van Eck & Waltman, 2011), a utility widely used in bibliometrics and data science, which allows for navigation of the network.

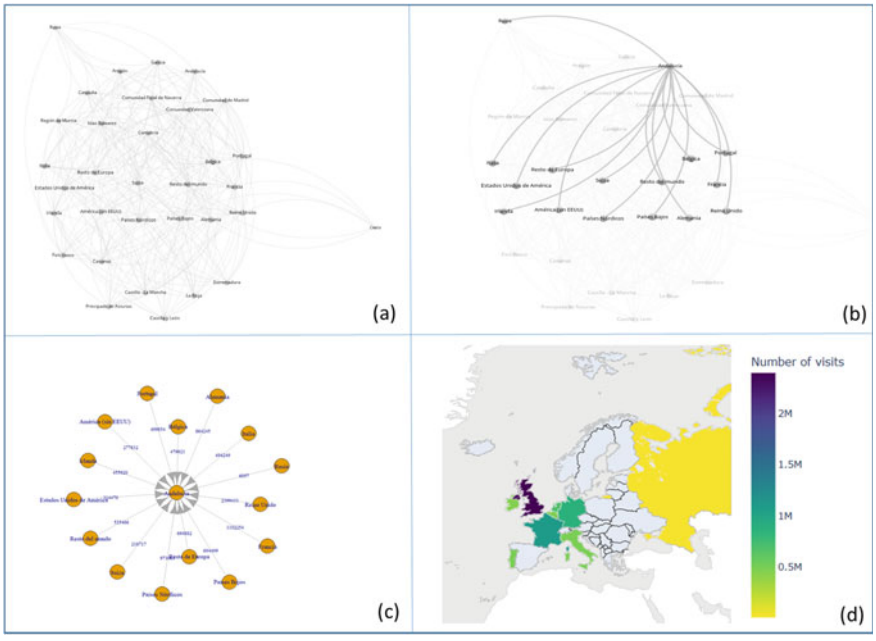

Fig. 2 Results obtained through the application of data science techniques: **a** general graph, **b** graph of Andalusia, **c** tourists heading to Andalusia, **d** number of European tourists who visited Andalusia, year 2022

Subsequently, we can focus on the specific case of Andalusia, see Fig. 2b, one of the autonomous communities that stands out in terms of receiving tourists from various countries.

Using exclusively graph theory, an analysis was conducted using the R language and the igraph library to visually represent the network of tourists visiting Andalusia, see Fig. 2c, focusing on the number of tourists per country. Using the igraph library in R, a network was constructed where each country was represented as a node (or vertex), and connections (or edges) were established between countries that had tourists visiting Andalusia in common. The strength of the connection between two countries was determined by the number of tourists they shared.

Finally, using the *pandas* and *Plotly* libraries in *Python*, a graduated color map of Europe is displayed, see Fig. 2d. You can choose the most significant countries without the need to select all of them. Visually, you can observe that darker blue represents countries with a higher number of tourists to Andalusia, while yellow represents those that provide fewer visitors to Andalusia.

2.5 Analysis Results

With the techniques performed in the previous chapter, it has been demonstrated that applying data science to the analysis of tourism-related data can be useful, providing automated and visually appealing results.

Focusing on Fig. 2d, with these results, it is possible to obtain the following analysis in a very intuitive way:

Key Tourism Sources

- United Kingdom: With a total of 2,399,433 visitors, the United Kingdom continues to be the primary source of tourism in Andalusia. This represents approximately 52% of the total visitors.
- France: France is the second most important source of tourism, with 1,102,254 visitors, equivalent to 24% of the total.
- Germany: Germany is another significant source with 864,245 visitors, accounting for approximately 19% of the total.
- Netherlands: The Netherlands contributes 694,499 visitors, which is about 15% of the total.
- Italy: Italy contributes 494,249 visitors, representing roughly 11% of the total.

Patterns and Interactions Among Origin Countries

- United Kingdom as a Key Market: The United Kingdom remains the largest and dominant market. Any tourism promotion and management strategy in Andalusia should continue to pay special attention to this market.
- Importance of European Markets: France, Germany, the Netherlands, and Italy are significant European markets and should be considered in tourism promotion and management strategies.

- Diversification Within Europe: While the United Kingdom is important, it is essential to diversify tourism sources within Europe. These European countries can be key targets for attracting more visitors.
- Emerging Markets: Russia, although representing a small proportion, may have potential for future growth. Specific strategies can be explored to attract more visitors from Russia.

Strategic Decision-Making

- Focus on the United Kingdom: Maintain and strengthen the relationship with the United Kingdom market through specific marketing and promotion campaigns.
- Key European Markets: Develop marketing strategies targeted at key European markets such as France, Germany, the Netherlands, and Italy.
- Promotion of Diversity: Highlight the diversity of experiences and attractions in Andalusia to attract visitors with different interests and preferences.
- Explore New Markets: Evaluate the feasibility of attracting more visitors from emerging markets, such as Russia, through campaigns and strategic alliances.
- Continuous Analysis: Continuously monitor tourism trends and adjust strategies to maintain competitiveness and attract visitors to Andalusia.

3 Benefits and Limitations of the Methodology Used

While there are increasingly more official data sources related to tourism, they are not always presented in a user-friendly, automatic, or unsupervised manner, and at times, they only provide data in a raw format. In such cases, having a reference methodology that equips researchers with the necessary tools to extract, process, and analyze the information can be highly beneficial.

The case presented in Sect. 2 serves as an example of data extraction from a recognized data source in the industry, which is used by many researchers without taking advantage of the proposed methodology. During this case, a data source was chosen from which data can be obtained in a user-friendly manner or via a web service. The latter method was chosen to demonstrate how information about tourist visits from a total of 15 countries and 19 autonomous communities as destinations, resulting in a total of 285 possible combinations, was successfully extracted. Manually downloading each of them would take approximately 5 min per download, totaling approximately 24 h of work and requiring constant attention. In contrast, using data science expertise within this methodology, the process only took 2 h and was entirely unsupervised.

In contrast to this methodology, a minimum level of data science knowledge and a basic understanding of programming languages are required to carry out data acquisition and processing properly. For this reason, it is likely that many researchers may choose to perform these tasks manually. However, this does not imply that this guide cannot be used, as the methodology itself serves as a guide in which each stage can be applied according to the available knowledge of each researcher.

4 Conclusions

Thanks to the methodology and visualization techniques used, it is demonstrated that a labor-intensive analysis, which would have initially required considerable effort and time, can be automated, allowing for the presentation of ready-made results for analysts to examine. This has made it possible to invest time in activities that truly add value. Thus, we can observe how a data science methodology and its associated techniques can be of great assistance in studying the tourism sector.

Through the use of data science and its techniques, the analysis process in the tourism sector has been simplified and expedited, freeing up resources and time for analysts to focus on interpreting results and generating relevant insights.

By employing graph theory, the R language, and the igraph library, we have successfully created an effective and easily interpretable visual representation of the tourist visits network to Andalusia. This tool provides valuable information for analysis and decision-making in the field of tourism, enabling us to better understand tourist flows and optimize promotion and tourism development strategies in the Andalusian region.

The proposed methodology has significant practical implications in the field of tourism. In destination management, it facilitates improved management of flows, marketing, seasonality, and predictions. For tourism businesses, it enhances decision-making in areas such as pricing, market segmentation, and service personalization. Furthermore, it underscores the growing need for data science training among the tourism sector workforce, preparing them for emerging industry challenges and promoting the acquisition of essential data analysis skills.

As indicated in the next section on future work, it will be interesting to delve deeper into the results obtained in the future. However, immediate conclusions can be drawn from the graphs obtained. During the year 2022, tourist flow between autonomous communities varied significantly. Russia and Ceuta were the territories examined with the lowest traffic. Andalusia stood out as a highly frequented destination within Spanish destinations, mainly by visitors from the United Kingdom and France, while receiving fewer visitors from Russia and Switzerland.

5 Future Work

As mentioned in the conclusions, the development of this work demonstrates how data science can assist in analyzing tourism-related data. The following are potential future works that can leverage the results:

- In-Depth Analysis of Graphs: In this future work, a comprehensive analysis of the graphs generated from tourism data would be conducted. This would involve examining the connections and relationships between different nodes, which could represent countries, regions, or specific tourist destinations. Graph analysis would help identify behavioral patterns, determine the relative importance of each

node, and gain a better understanding of interactions within the tourism system. This analysis would provide valuable insights for strategic planning, identifying opportunities, and decision-making in the tourism sector.

- Comparison with Other Communities: In this work, a comparison would be made between the results obtained for the autonomous community of Andalusia and those of other communities or tourist regions. This would help identify significant differences in terms of the number of visits, the origin of tourists, behavioral patterns, etc. These comparisons could help identify relative strengths and weaknesses of each community, as well as areas for improvement.
- Analyzing Why Some Countries Receive More Visitors: This work would focus on understanding the reasons why certain countries have a higher tourist presence in the study region. Factors such as geographical location, air connectivity, trade agreements, tourism promotion, safety perception, availability of tourism infrastructure, among others, could be analyzed. By examining these factors, a deeper understanding of the motivations and preferences of tourists from different countries could be obtained, allowing for the adaptation of marketing and promotion strategies to attract more visitors from other countries.
- Applying Findings to Tourists from Other Countries: In this future work, the goal would be to apply the knowledge gained from the previous analysis to attract tourists from other countries. This could involve adapting marketing and promotion strategies, improving the customer experience to meet the specific needs and preferences of tourists from those countries, and identifying opportunities for developing customized tourism products. By better understanding the factors that attract certain countries, the acquisition of international tourists could be optimized, increasing their satisfaction, which would, in turn, benefit the tourism industry as a whole.

References

Abbasi-Moud, Z., Vahdat-Nejad, H., & Sadri, J. (2021). Tourism recommendation system based on semantic clustering and sentiment analysis. *Expert Systems with Applications, 167*, 114324.

Burbano, C., & Anderson, K. (2023). Minería de Datos para mejorar los procesos de control de la demanda turística en el Ministerio de Turismo de la Provincia del Carchi en el año 2022. UPEC.

Diezma, F. R. (2021). Estadísticas turísticas: una herramienta clave para la planificación en el sector. *Indice: Revista de Estadística y Sociedad,* (81), 33–35.

Gonzalez, P., & Moral, P. (1996). Analysis of tourism trends in Spain. *Annals of Tourism Research, 23*(4), 739–754.

Martínez, E., & Nicolás, M. Á. (2014). The construction of tourist space by public administration and institutional communication: The image of the brand Andalucía as a tourist destination. *Journal of Promotion Management, 20*(2), 181–199.

Medina-Munoz, D. R., Medina-Muñoz, R. D., & Zuniga-Collazos, A. (2013). Tourism and innovation in China and Spain: A review of innovation research on tourism. *Tourism Economics, 19*(2), 319–337.

Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *Profesional de la Información, 29*(1).

Moine, J. M., Haedo, A. S., & Gordillo, S. E. (2011). Estudio comparativo de metodologías para minería de datos. In *XIII Workshop de Investigadores en Ciencias de la Computación.*

Stančin, I., & Jović, A. (2019, May). An overview and comparison of free Python libraries for data mining and big data analysis. In *2019 42nd International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 977–982). IEEE.

Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using Python. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8*(12), 2019.

Valdez, B. (2016). Análisis de grafos usando R e igraph. *Altamira, 1*(1), 1.

Van Eck, N. J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. arXiv:1109.2058.