Data Descriptor

Datasets on Tourism Services in an Emerging Destination: The Case of Riohacha-Colombia

Andres Solano-Barliza [*] , Aida Valls , Antonio Moreno , José Escorcia-Gutierrez , Melisa Acosta-Coll [*]

Data Descriptor

Datasets on Tourism Services in an Emerging Destination: The Case of Riohacha-Colombia

Andres Solano-Barliza [1],*, Aida Valls [2], Antonio Moreno [2], José Escorcia-Gutierrez [3] and Melisa Acosta-Coll [3],*

[1] Faculty of Engineering, Universidad de La Guajira, 44001, Riohacha-La Guajira, Colombia

[2] Departament d'Enginyeria Informàtica i Matemàtiques, Escola Tècnica Superior d'Enginyeria, Universitat Rovira i Virgili, 43007, Tarragona, Spain

[3] Department of Computational Science and Electronics, Universidad de la Costa, CUC, Barranquilla, 080002, Barranquilla, Colombia

* Correspondence: andresolano@uniguajira.edu.co; macosta10@cuc.edu.co

Abstract

In the context of emerging tourist destinations, the availability of specific and detailed datasets for tourist services, such as restaurants and hotels, is essential for the development of smart and personalised solutions. This study presents the design and construction of two datasets aimed at the tourism sector, optimised for scenarios with limited information availability. The datasets include attributes relevant to restaurants, such as types of food, vegetarian options, payment methods, home delivery services and proximity to key points of interest; as well as for hotels, incorporating services offered, languages available and precise geographical details. A structured methodology is described for the systematic collection of data from recognised platforms such as TripAdvisor and Booking.com, as well as from sources related to emerging tourist destinations. As a result, 18 records were compiled for restaurants and 46 for hotels, constituting a valuable resource for the development and evaluation of multi-criteria tourist recommendation systems. The construction of these datasets facilitates future research on intelligent systems and provides a replicable framework for the generation of comprehensive datasets covering various tourism activities and services.

1. Summary

In emerging tourism destinations, Arnegger (2016) [3], one of the main limitations in the design and development of effective technological solutions, especially in the integration of artificial intelligence [4] and analytics in multi-criteria decision making [5], is the lack of adequate, organized and reliable data [6]. This weakness stems from the lack of a clear policy on data management and processing by government agencies in these areas. Without well-defined guidelines, it becomes a challenge to obtain concrete and accurate data, which is necessary for the development of advanced technological solutions that can boost the tourism sector [7].

This scenario becomes a challenge when it comes to integrating technological solutions such as recommender systems, which can help to improve tourism competitiveness indicators in tourist destinations, such as the supply of products and services, as well as aspects such as tourism security [8,9]. Therefore, it is necessary to build a dataset that integrates data related to tourism services, such as hotels and restaurants. To achieve this, it is necessary to develop a methodology for data selection

and to integrate reliable data sources, such as tourism portals like booking.com and TripAdvisor, which provide up-to-date and relevant information on the destination's services [10,11]. The construction of such a dataset will not only facilitate multi-criteria analysis, but will also allow decision makers to identify areas for improvement and develop effective strategies to increase the competitiveness of the destination.

We present two datasets, one for restaurants and one for hotels. The Restaurants Riohacha dataset contains 18 records with detailed information on restaurants, including type of food, vegetarian options, payment methods, delivery services, distances to key points, category, rating, geographical coordinates and website. On the other hand, the Hotels Riohacha dataset contains 46 records with information on hotels, such as sustainable travel programmers, credit card acceptance, security, accommodation services, breakfast included, languages, overall rating, price category, distances to major points of interest, geographical coordinates and website. Both datasets were collected in the emerging tourism destination of Riohacha. Although they are small datasets, they serve as a valuable input that provides a path for other destinations to develop similar datasets, which are fundamental to building intelligent recommendation systems and improving the tourism experience and competitiveness of the destination [12].

To build the dataset, a methodology was followed that included validation of criteria by tourism sector experts, selection of reliable data sources, use of Web Scrapy to extract information from tourism portals and use of Google Maps to obtain distances between key points in the destination. In addition, data types were transformed to ensure they were in a format suitable for analysis. This process ensures the quality and relevance of the data collected, facilitating its use in advanced analyses to improve tourism competitiveness. With the dataset, we have developed proposals for multi-criteria recommendation systems, which become highly effective and necessary methods when destinations do not have large amounts of data. In this case, as a result of our research, we have used the proposed datasets for multi-criteria recommendation systems for hotels [13] and restaurants [14].

2. Data Description

The dataset focuses on tourism services in the emerging destination of Riohacha, the capital of the department of La Guajira in Colombia. La Guajira, located in the north of South America and bordering Venezuela, covers most of the La Guajira peninsula facing the Caribbean Sea, see geographical location in Figure 1. This department has enormous tourism potential due to its natural, cultural and ethnic diversity, although it has faced several challenges in recent years.
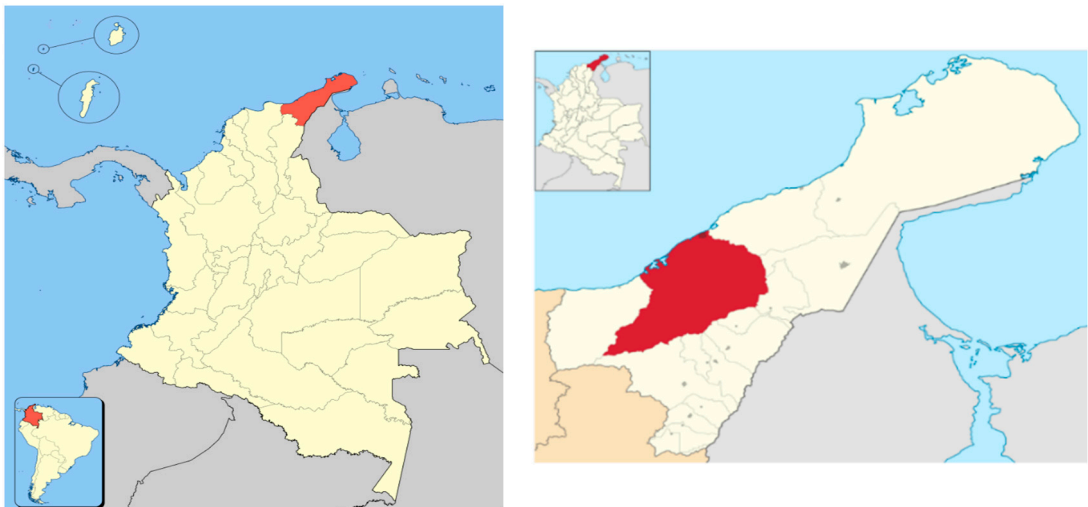

Figure 1. Location of Riohacha. Note: Source: wikipedia.org.

The hotel dataset shown in Table 1 contains relevant information on various aspects of tourism establishments in an emerging destination.

Table 1. Dataset hotels Riohacha.

No	Name variable	Description	Type of data
1	Hotel name	The name of the hotel.	String
2	Sustainable travel program	Indicates if the hotel participates in any sustainable travel programs, promoting environmentally friendly and responsible practices.	Binary
3	Accepts credit cards	Information on whether the hotel accepts credit cards as a method of payment.	Binary
4	Security Outside camera	Indicates whether the hotel has security cameras outside the establishment.	Binary
5	Includes breakfast	Information on whether breakfast is included in the room rate.	Binary
6	Language English:	Specifies whether the hotel offers services in English.	Binary
7	Global Hotel Score:	The overall rating of the hotel, based on guest reviews and opinions.	Float
8	Category (Price)	The hotel category according to its price range (it may be classified as budget, mid-range, or luxury).	Float
9	Distance to Beach:	The distance from the hotel to the beach.	Float
10	Distance to Historic Center:	The distance from the hotel to the historic center of the city or tourist area.	Float
11	Distance to Pier Police Tourism:	The distance to the pier or dock related to tourist or police activities.	Float
12	Distance to Clinic:	The distance from the hotel to the clinic or healthcare center.	Float
13	Geographical Latitude	The geographical latitude of the hotel's location.	Float
14	Website	The official website of the hotel, where users can find more information or make reservations.	String (URL)

On the other hand, the restaurant dataset contains detailed information on various restaurants in a tourist destination. The data included are shown in Table 2.

Table 2. Dataset Restaurant Riohacha.

No	Name variable	Description	Type of data
1	Restaurant name	The name of the restaurant.	String
2	Type Meals	The type of meals or cuisine offered by the restaurant (e.g., Italian, Mexican, Asian, seafood, etc.).	String
3	Vegetarian	Indicates whether the restaurant offers vegetarian options on its menu.	Binary
4	Payment Cards:	Information about whether the restaurant accepts payment cards as a method of payment.	Binary
5	Delivery	Specifies whether the restaurant offers home delivery service.	Binary
6	Distance to Beach:	The distance from the hotel to the beach.	Float
7	Distance to Historic Centre:	The distance from the hotel to the historic centre of the city or tourist area.	Float
8	Distance to Pier Police Tourism:	The distance to the pier or dock related to tourist or police activities.	Float
9	Category (Price)	The category of the restaurant according to its price range, type of food, or level of service (e.g., budget, mid-range, or luxury).	Float

10	Restaurant Score:	The overall rating of the restaurant based on customer reviews and opinions.	Float
11	Geographical Latitude	The geographical latitude of the restaurant's location.	Float
12	Website	The official website of the restaurant, where users can find more information, view the menu, or make reservations.	String (URL)

These datasets are useful for multi-criteria work, allowing the accessibility and quality of tourism services to be assessed according to a number of variables, such as proximity to points of interest, amenities offered and guest ratings.

Figure 1. Word cloud of restaurant food types.

The word cloud shows that the most common types of food in restaurants in this destination include a variety of culinary influences. The top 10 most common words are: Southern (8), American (6), Colombian (5), Caribbean (4), Fusion (4), Seafood (4), Fast (3), American (3), Colombian (3) and Latin (3). This diversity reflects the cultural and gastronomic richness of the region and offers visitors a wide range of options to enjoy.

3. Methods

A structured methodology was developed for the construction of the dataset for an emerging tourism destination such as Riohacha, with the aim of ensuring the quality, reliability and accuracy of the data, see Figure 2. This methodology ensured that the information collected was accurate and useful for the construction of the tourism services dataset. The steps taken to create the dataset are detailed below.

1. Validation of variables by tourism sector experts

2. Selection of data sources for dataset construction

3. Use of web scraping to extract data from tourism platforms

4. Use of Google Maps to collect distances to key destination points

5. Data type transformation

Figure 2. Structured methodology to build the dataset.

3.1. Validation of Variables by Experts in the Tourism Sector

In this step, relevant attributes or variables were identified from a review of scientific articles related to tourism, focusing on elements that could be incorporated to build useful data for intelligent systems for the hotel and restaurant sector. A survey was carried out with six destination experts, specifically professors from the Tourism program at the University of La Guajira, to validate that the variables selected were the most appropriate. These variables included aspects related to hotels and restaurants, distances and geolocation, destination context and elements related to tourism safety. In addition, the experts were asked to suggest other possible variables that they considered relevant.

3.2. Selection of Data Sources to Build the Dataset

Given the lack of organized data for the destination, and considering the validation of the proposed variables by tourism experts, sources such as Booking.com were selected to obtain data on hotels and TripAdvisor for restaurants. In addition, an institutional source was chosen that provides a competent body to validate the legality of service providers, which in the case of Colombia is the Chamber of Commerce, see figure.

Figure 3. data sources to build the dataset.

3.3. Use of Web Scraping to Extract Data from Tourism Platforms

Once the variables to be extracted from the websites of the selected tourism portals were identified, the scraping technique was used for academic or research purposes to collect the necessary data from the tourism services registered in Riohacha. This process facilitated the collection of information from both hotels and restaurants. Table 3 presents information on the data extracted from hotels in Riohacha from booking.com using the web scraping technique. Includes details such as hotel name, participation in sustainable travel program, credit card acceptance, presence of security cameras, breakfast included, availability of English-speaking staff, overall rating, price category and the hotel's website. There is also a column indicating the type of data originally extracted from the portal.

Table 3. Riohacha hotels data extracted from booking.com.

No	Name variable	Original data type
1	Hotel name	String
2	Sustainable travel program	Boolean (True/False)
3	Accepts credit cards	Boolean (True/False)
4	Security Outside camera	Boolean (True/False)
5	Includes breakfast	Boolean (True/False)
6	Language English:	Boolean (True/False)
7	Global Hotel Score:	Float
8	Category (Price)	Float
9	Website	String (URL)

Table 4 presents information on the extracted data on restaurants in Riohacha, obtained using TripAdvisor's web scraping technique. This table includes variables such as restaurant name, type of

food, vegetarian options, payment card acceptance, delivery service, price category, restaurant rating and website. There is also a column indicating the type of data originally extracted from the portal.

Table 4. Riohacha restaurant data extracted from TripAdvisor.

No	Name variable	Original data type
1	Restaurant name	String
2	Type Meals	String
3	Vegetarian	Boolean (True/False)
4	Payment Cards:	Boolean (True/False)
5	Delivery	Boolean (True/False)
6	Category (Price)	Float
7	Restaurant Score:	Float
8	Website	String (URL)

In addition, a manual check was carried out to collate and select those establishments with data on the portals and that had an active registration with the government body responsible for verifying the legality of the provider, in order to verify that they were duly legalized. This step was necessary to ensure the reliability and security of the establishments included.

3.4. Use of Google Maps to Collect Distances to Key Points at the Destination

A manual collection of geospatial data was carried out by analyzing optimized routes using the Google Maps routing algorithm. This provided the shortest distance (in kilometers) between two points, relying on the intelligence of Google's routing system. Table 5 presents data on the location of hotels in Riohacha, highlighting their proximity to key points of interest. Variables include distance to the beach, the historic center, Pier/Police Tourism and the clinic, all measured in kilometers. This data, extracted from Google Maps, shows the location of hotels in relation to the city's main attractions and services. Furthermore, Table 5 presents data on the location of Riohacha restaurants, highlighting their proximity to the main points of interest. Variables include distance to the beach, to the historic center and to the pier/tourist police, all measured in kilometers.

Table 5. Riohacha hotels/restaurants data extracted from google maps.

No	Name variable	Type of data	Units of measurement
	Hotels		
1	Distance to Beach	Float	Kilometres
2	Distance to Historic Centre	Float	Kilometres
3	Distance to Pier/Police Tourism	Float	Kilometres
4	Distance to Clinic	Float	Kilometres
	Restaurants		
1	Distance to Beach	Float	Kilometres
2	Distance to Historic Centre	Float	Kilometres
3	Distance to Pier Police Tourism	Float	Kilometres

Table 6 presents data on the location of Riohacha restaurants, highlighting their proximity to the main points of interest. Variables include distance to the beach, to the historic center and to the pier/tourist police, all measured in kilometers.

3.4. Transformation of Data Types

Considering the research interests, different data types in the dataset were transformed from Boolean to binary and distances from kilometers to meters to adapt the data to the needs of the analysis. The transformed data types are shown below. Table 6 presents the transformation from Boolean to binary data, carried out to satisfy the research interests where 1 is true and 0 is false. This

transformation allows a more accurate and structured analysis of the characteristics of the hotels and restaurants in Riohacha.

Table 6. Riohacha hotels/restaurant data Transformed for dataset.

No	Name variable	Original data type	Transformed data
	Hotels		
1	Sustainable travel program	Boolean (True/False)	Binary
2	Accepts credit cards	Boolean (True/False)	Binary
3	Security Outside camera	Boolean (True/False)	Binary
4	Includes breakfast	Boolean (True/False)	Binary
5	Language English:	Boolean (True/False)	Binary
	Restaurants		
1	Vegetarian	Boolean (True/False)	Binary
2	Payment Cards:	Boolean (True/False)	Binary
3	Delivery	Boolean (True/False)	Binary

Table 7 shows the unit conversion of the distance variables for hotels and restaurants from kilometers to meters. This conversion was done by multiplying the values in kilometers by 1,000, which allows a more detailed and accurate analysis of distances within the dataset.

Table 7. Riohacha hotels/restaurant units of measurement changed for dataset.

No	Name variable	Units of measurement original	Units of measurement changed
	Hotels		
1	Distance to Beach	Kilometres	Meters
2	Distance to Historic Centre	Kilometres	Meters
3	Distance to Pier/Police Tourism	Kilometres	Meters
4	Distance to Clinic	Kilometres	Meters
	Restaurants		
1	Distance to Beach	Kilometres	Meters
2	Distance to Historic Centre	Kilometres	Meters
3	Distance to Pier Police Tourism	Kilometres	Meters

4. Discussion

In order to create a dataset that meets the specific requirements of an emerging tourism destination, it is essential that contextually relevant variables are carefully selected and validated. In the case of Riohacha, for instance, significant factors included the variety of food available, vegetarian options, payment methods, delivery services and proximity to attractions [13,14]. These factors are indicative of the needs of both tourists and local businesses. This methodological approach guarantees that the collated information is both representative and useful for understanding the particularities of the destination. The generation of personalised recommendations that align with users' expectations and enhance the region's competitiveness in the tourism sector is thus facilitated [12].

From a technical standpoint, it is imperative that the dataset fulfils quality, structure and format criteria that facilitate its effective integration into multi-criteria recommendation systems. This encompasses the utilisation of web scraping techniques to procure updated data from reliable sources, the employment of expert validation to ensure the accuracy and relevance of the information, and the incorporation of geographic data to enhance the analysis. Furthermore, the judicious transformation of variables facilitates computational processing and enhances the system's capacity to manage multiple criteria effectively [14]. Consequently, the dataset not only reflects the

characteristics of the destination, but also meets the technical requirements necessary to develop robust and scalable technological solutions in the tourism sector.

5. Conclusion

The construction of the dataset followed a rigorous process that included the validation of variables by experts in the tourism sector, the selection of reliable sources such as Booking.com and TripAdvisor, the use of web scraping techniques and the incorporation of geographic data using Google Maps, thus ensuring the quality and relevance of the information collected for advanced analysis.

The design of the datasets meets the detailed and structured technical requirements necessary for the development of multi-criteria recommendation systems, especially in emerging destinations where data availability is limited, facilitating decision-making and improving tourism competitiveness.

The usefulness of the dataset lies in its ability to integrate multiple relevant variables (services, location, payment options, among others), which contributes to the design of recommendation systems that enhance tourism supply and promote sustainable development in emerging destinations such as Riohacha.

References

1. Solano-Barliza, A.; Valls, A.; Moreno, A.; Acosta-Coll, M.; Escorcia-Gutierrez, J.; De-La-Hoz-Franco, E. Dataset_Jimataa_hotels_Riohacha 2025.
2. Solano-Barliza, A.; Valls, A.; Moreno, A.; Acosta-Coll, M.; Escorcia-Gutierrez, J.; De-La-Hoz-Franco, E. Dataset_Jimataa_Restaurants_Riohacha 2025.
3. Arnegger, J.; Herz, M. Economic and Destination Image Impacts of Mega-Events in Emerging Tourist Destinations. *J. Destin. Mark. Manag.* **2016**, *5*, 76–85, doi:https://doi.org/10.1016/j.jdmm.2015.11.007.
4. Sarkar, J.L.; Majumder, A.; Panigrahi, C.R.; Roy, S.; Pati, B. Tourism Recommendation System: A Survey and Future Research Directions. *Multimed. Tools Appl.* **2023**, *82*, 8983–9027, doi:10.1007/s11042-022-12167-w.
5. Monti, D.; Rizzo, G.; Morisio, M. A Systematic Literature Review of Multicriteria Recommender Systems. *Artif. Intell. Rev.* **2021**, *54*, 427–468, doi:10.1007/s10462-020-09851-4.
6. Solano-Barliza, A.; Arregocés-Julio, I.; Aarón-Gonzalvez, M.; Zamora-Musa, R.; De-La-Hoz-Franco, E.; Escorcia-Gutierrez, J.; Acosta-Coll, M. Recommender Systems Applied to the Tourism Industry: A Literature Review. *Cogent Bus. Manag.* **2024**, *11*, 2367088, doi:10.1080/23311975.2024.2367088.

7. Lee, G.-C. A Data-Driven Approach to Tourism Demand Forecasting: Integrating Web Search Data into a SARIMAX Model. *Data* **2025**, *10*, doi:10.3390/data10050073.

8. Solano-Barliza, A.; Acosta-Coll, M.; Escorcia-Gutierrez, J.; De-La-Hoz-Franco, E.; Arregocés-Julio, I. Hybrid Recommender System Model for Tourism Industry Competitiveness Increment. In Proceedings of the Computer Information Systems and Industrial Management; Saeed, K., Dvorský, J., Nishiuchi, N., Fukumoto, M., Eds.; Springer Nature Switzerland: Cham, 2023; pp. 209–222.

9. Borràs, J.; Moreno, A.; Valls, A. Intelligent Tourism Recommender Systems: A Survey. *Expert Syst. Appl.* **2014**, *41*, 7370–7389, doi:https://doi.org/10.1016/j.eswa.2014.06.007.

10. Silaa, V.; Masui, F.; Ptaszynski, M. A Method of Supplementing Reviews to Less-Known Tourist Spots Using Geotagged Tweets. *Appl. Sci.* **2022**, *12*, doi:10.3390/app12052321.

11. Asani, E.; Vahdat-Nejad, H.; Sadri, J. Restaurant Recommender System Based on Sentiment Analysis. *Mach. Learn. with Appl.* **2021**, *6*, 100114, doi:10.1016/j.mlwa.2021.100114.

12. Solano-Barliza, A.; Acosta-Coll, M.; Escorcia-Gutierrez, J.; De-La-Hoz-Franco, E.; Arregocés-Julio, I. Hybrid Recommender System Model for Tourism Industry Competitiveness Increment BT - Computer Information Systems and Industrial Management.; Saeed, K., Dvorský, J., Nishiuchi, N., Fukumoto, M., Eds.; Springer Nature Switzerland: Cham, 2023; pp. 209–222.

13. Solano-barliza, A.; Valls, A.; Moreno, A.; Dujmovic, J.; Solano-barliza, A.; De-la-hoz-franco, E.; Escorcia-gutierrez, J.; De-la-hoz-franco, E. Personalized Hotel Recommender System Based on Graded Logic Personalized Hotel Recommender System Based on Graded Logic with Asymmetric Criteria with Asymmetric Criteria. *Procedia Comput. Sci.* **2024**, *246*, 2864–2873, doi:10.1016/j.procs.2024.09.385.

14. Solano-Barliza, A.; Valls, A.; Acosta-Coll, M.; Moreno, A.; Escorcia-Gutierrez, J.; De-La-Hoz-Franco, E.; Arregoces-Julio, I. Enhancing Fair Tourism Opportunities in Emerging Destinations by Means of Multi-Criteria Recommender Systems: The Case of Restaurants in Riohacha, Colombia. *Int. J. Comput. Intell. Syst.* **2024**, *17*, doi:10.1007/s44196-024-00700-8.