The Role and Use of Big Data in Spatial Analysis of Touristification – Example of the Historic Core of Dubrovnik

Dino Bečić[1,*]

[1] Faculty of Science, University of Zagreb, Zagreb, Croatia, dbecic.geog@pmf.hr
* corresponding author

doi: 10.5281/zenodo.11584606

Abstract: Big data is a term used to define large volumes of data which are generated, collected, and analysed. It can be structured, unstructured, or semi-structured. Also, the come from various sources such as sensors, mobile devices, satellite imagery, and internet, including social networks that generate big amounts of data daily. Big data represent a significant potential data source that can be utilized in various analyses, both qualitative and quantitative. The process analysed in this study is touristification, the process by which destinations become increasingly oriented towards tourism. Since touristification is a highly complex process, this study will not cover all its elements but will focus solely on analysing specific social media data. In the context of researching the touristification of the Historic Core of Dubrovnik, social media data analytics can provide additional insights into the characteristics of the process. Big data were collected from the Booking.com platform, one of the most visited tourism platforms. This platform allows users to search, view, and book touristic accommodation. These accommodations, along with their descriptive data and location information, can be observed and analysed as elements of touristification. The research involved data collection, analysis, and visualization, through the R programming language.

Keywords: touristification; R; spatial analysis.

1 Introduction

In the era of big data, many scientific disciplines are leveraging extensive datasets for complex analyses and data-driven decisions. The interdisciplinary field of "data science" encompasses principles, algorithms, and processes to extract useful patterns from large datasets (Kelleher and Brendan 2018). Big data represents a paradigm shift in data management, characterized by the "3Vs": volume, velocity, and variety. The fusion of data science and Geographic Information Systems (GIS) has revolutionized spatial analysis, enabling access to vast amounts of geospatial information (Oliviera et al. 2024). Spatial data management in GIS involves data science techniques like spatial indexing, geocoding, and spatial queries for efficient geographic data processing. This area is often seen as a subset of data science, termed spatial (geographic) data science (Brunsdon and Comber 2019, Rey et al. 2023). Traditional GIS applications and software, initially designed for smaller datasets, now incorporate big data technologies to handle large spatial datasets from sources like satellite imagery and social media. Programming languages and open-source tools such as R and Python have become popular in both data science and GIS for their flexibility, scalability, and cost-effectiveness, facilitating advanced spatial analysis (Brunsdon and Comber 2019, Oliviera et al. 2024). A notable application of these technologies is analysing touristification, a process transforming a destination to focus primarily on tourism, impacting socio-cultural, ecological, and economic dimensions. Touristification often leads to the commercialization of residential spaces and displacement of residents, driven by platforms like AirBnB and Booking.com. This growth in tourist accommodation affects central urban areas, impacting housing prices and residents' quality of life. Cities like Lisbon, Malaga, Barcelona, and Madrid have seen rapid increases in short-term rentals, leading to significant social, economic, and urban transformations (García Bujalance et al. 2019, Antunes et al. 2021, Ardura et al. 2021, Porfido et al. 2023). This paper demonstrates the integration of geographic data analytics methods in collecting and processing spatial data to manage the touristification process, highlighting the importance of geographic data analytics as an enhancement of traditional GIS techniques.

2 Materials and methods

Considering social media and internet platforms in general, which contain vast amounts of data, including spatial data, they offer opportunities for spatial analysis. As mentioned, one aspect of touristification is tourist accommodation. This study will demonstrate data collection from the Booking.com online platform, followed by processing, analysis, and visualization. The selected online platform is Booking.com. By retrieving data from this platform and analyzing it, certain spatial patterns can be observed. The chosen area, or tourist destination for this research, is Dubrovnik, specifically the Historic Core of Dubrovnik. Booking.com is one of the most visited online platforms for browsing, booking, and reviewing accommodation units. According to another internet portal, Statista.com, the number of monthly visits to the Booking.com website amounts to 556.1 million. The portal contains numerous data about accommodation units, including nightly rates, location, ratings, reviews, number of accommodation units, number of beds, and many others. Such information can serve various analyses, including spatial ones, given that platforms naturally possess addresses of accommodation units. For research purposes, the following data about accommodation units have been selected: name, address, unit rating, number of reviews, indication of whether the property is managed by a private individual or a company, and number of accommodation units. Such data are not structured for spatial analyses, and they need to be collected and processed to be suitable for spatial analyses. Getting such type of data, available on internet portals, can be achieved

using a method called web scraping. Web scraping is the process of extracting data from websites. This technique involves using software tools to access web pages, extract desired data, and store it in a structured format for further analysis or use. This technique is crucial for obtaining geographic information and other attributes from web content. The steps include identifying the target website, reviewing the structure of the page, writing a scraper that involves setting up the environment, accessing the website, parsing HTML, handling dynamic content, extracting data, cleaning, and storing the data (Brenning and Henn 2023).

Key concepts of web scraping include automation tools such as Selenium, BeautifulSoup, Scrapy, rvest, which are used through programming languages like R or Python, as well as standalone programs and platforms like Octoparse, etc. They are all designed to automate the process of accessing web pages and extracting data. Websites are primarily composed of HTML. Data extraction involves parsing HTML to identify and retrieve relevant information. Selectors, such as CSS selectors and XPath, allow for precise determination of parts of the website from which data is extracted. After extraction, the data is stored in a structured format to facilitate easier access and analysis (Brenning and Henn 2023). For this research, a code was written in the R programming language, which includes accessing the Booking.com page for the location "Historic Core of Dubrovnik," where all accommodation units in that area are listed, using web scraping packages. Then it includes saving all links for each accommodation unit separately followed by accessing each accommodation unit by opening the link and retrieving desired data and storing it for each unit. Then workflow includes - Processing attributes and structuring. After that we do spatial analysis and geovisualization.

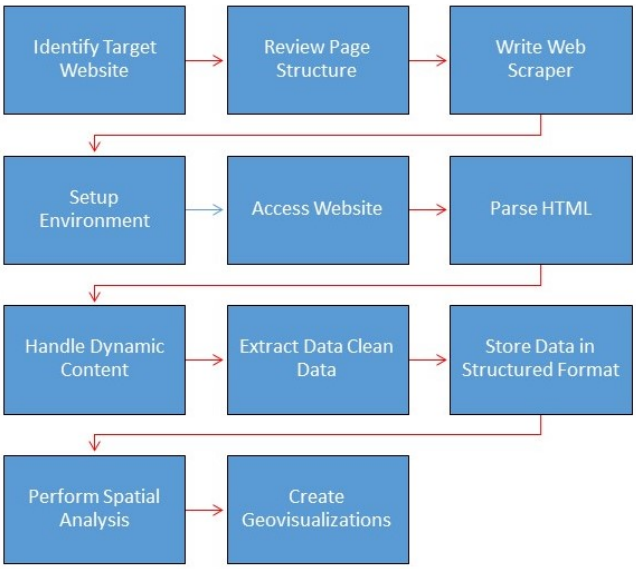

Figure 1. Flowchart.

This generated workflow used packages, or libraries, or extensions, which enabled such an approach to research, from data retrieval, storage, processing, and finally visualization, including RSelenium, rvest, sf, ggplot2, dplyr. Through the written code, each accommodation unit was accessed separately, automatically, and the necessary data was retrieved from it, which was later structured into a new format suitable for analysis (Figure 1).

3 Results

Data from the Booking.com platform about accommodation units in the Historic Core of Dubrovnik were collected on March 31, 2024. On that day, there were 550 advertised tourist accommodations on the platform. Using web scraping tools within the R language, data for all 550 advertised accommodations were gathered. After data processing, the processed table contained information about the name of the property, property rating, number of reviews, coordinates, number of accommodation units, and whether the property is managed by a physical or private individual. The processed data, in this format, within the RStudio application, can be analysed qualitatively, quantitatively, and spatially.

As mentioned, web scraping collected data for a total of 550 advertised listings. The average rating for all properties is a high 8.91, with the highest rating being 10, while one property has an average rating of 4.2. Regarding the management of properties, 299 are managed by individuals, while 251 are managed by companies. The average rating for properties managed by individuals is 9, while the rating for properties managed by companies is 8.79. The maximum number of accommodation units at one location is 16, while the minimum is 1. The average number of accommodation units per property is 1.78. As mentioned, there is one property with a rating in the category up to 5, while in the rating range from 9 to 10, there are as many as 245 properties, which makes up almost half of the total number of advertised properties (Figure 2).

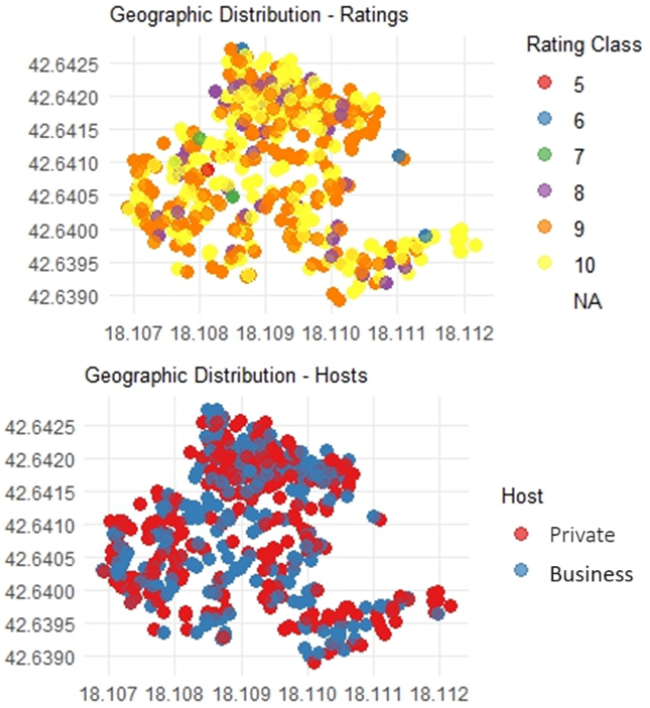

Figure 2. Geovisulalization of geographic distributions.

In a geographic context, since the data includes coordinates, it is possible to create spatial clusters using the k-means method. K-means is a data clustering

algorithm used to identify similar groups of data, or clusters. The algorithm starts by randomly selecting initial cluster centers, or centroids. Then, each data point is assigned to the cluster whose centroid is closest to it. After that, new cluster centers are calculated by computing the mean of all points in each cluster. These steps are repeated iteratively until the centroids stabilize. The result is the grouping of data into a certain number of clusters, where points within the same cluster are as similar as possible, while points from different clusters are as different as possible. In this specific analysis, the variable used for data clustering is location. The number of clusters was defined using the elbow method, which resulted in three clusters. The elbow method in clustering is a technique used to determine the optimal number of clusters in a dataset. It involves plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. The elbow point represents the number of clusters where adding another cluster does not significantly improve the model's performance. The variable used for data clustering is location. Specifically, the clustering algorithm utilized the latitude and longitude coordinates of the data points to form clusters. No other variables, such as rating, were used in the clustering process. Overall, the k-means clustering with three clusters based on geographic coordinates provides a meaningful way to segment the data into distinct geographic regions. Adjustments to the number of clusters can be made depending on the desired granularity of the analysis (Figure 3).

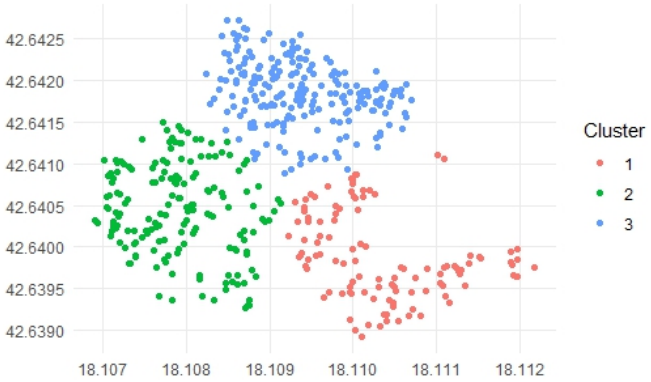

Figure 3. K-means clusters.

4 Discussion

In the context of analyzing touristification in the Historic Core of Dubrovnik, the analysis of collected data suggests that the destination is facing a pronounced phenomenon of touristification that may have negative implications. The high number of 550 advertised accommodation units indicates an intense pressure of tourism on this area. The increase in the number of accommodation units, especially owned by individuals, can lead to issues such as excessive resource consumption, infrastructure overload, and loss of authenticity of the local community. This aligns with findings from other historic cities experiencing similar trends (Curto et al. 2022, Chamizo-Nieto et al. 2023). The spatial clusters identified areas particularly affected by tourism Spatial analysis through the k-means method can identify areas that are particularly under pressure from tourism, serving as a basis for planning measures to protect and preserve natural and cultural resources. The study demonstrates data collection from Booking.com, using web scraping to retrieve information about tourist accommodations in Dubrovnik. Data on property names, addresses, ratings, reviews, and management types were collected and processed using R packages such as RSelenium, rvest, sf, ggplot2, and dplyr. Spatial clustering was performed using the k-means method to identify geographic patterns.

5 Conclusions

Web scraping serves as a method for collecting data that can contribute to big data analysis, including in a geographic context. Web scraping can gather large amounts of data from multiple web pages or entire websites, significantly contributing to the volume of big data. Automated tools can collect data at high speed, providing a continuous flow of information, which aligns with the characteristic of velocity in big data. Although web scraping itself is not synonymous with big data, it is a critical component for gathering diverse and large datasets needed for big data analysis. In a geographic context, web scraping provides additional opportunities for data collection. Programming languages like R, used in this research, enable the structuring and management of data, as well as their analysis and visualization. Considering touristification as the observed process, the collected data on the locations of accommodation units provided additional insight into their spatial distribution within the Historic Core of Dubrovnik. Additionally, although this study involved spatial analysis of accommodation unit locations, they were not presented and analyzed through GIS applications like ESRI ArcGIS PRO, or QGIS; instead, geographic data science methods were used with R programming language. This approach showcases the potential for enhancement, as spatial analysis in GIS integrates principles of big data analysis and advanced statistics into a geographic context. Integrating these interdisciplinary disciplines allows for a deeper understanding of spatial phenomena through complex models and analytical techniques. The usage of geographic data science provides a comprehensive approach to spatial data analysis, which is crucial for addressing complex spatial problems, such as the component of touristification analyzed in this study, namely the location of accommodation units.

6 References

Antunes, G., Ferreira, J., 2021. Short-term rentals: how much is too much – spatial patterns in Portugal and Lisbon. Tourism and hospitality management 27 (3), 581-603.

Ardura Urquiaga, A., Lorente-Riverola, I., Ruiz Sanchez, J., 2020. Platform-mediated short-term rentals and gentrification in Madrid. Urban Studies 57 (15), 3095-3115.

Brenning, A., Henn, S., 2023. Web scraping: a promising tool for geographic data acquisition. arXiv Preprint 2305.19893v1.

Brunsdon, C., Comber, L., 2019. An Introduction to R for Spatial Analysis and Mapping, second ed. SAGE Publications Ltd., London.

Chamizo-Nieto, F.J., Nebot-Gómez de Salazar, N., Rosa-Jiménez, C., Reyes-Corredera, S., 2023. Touristification and conflicts of interest in cruise destinations: the case of main cultural tourism cities on the Spanish mediterranean coast. Sustainability 15 (8), 6403.

Curto, R.A., Rubino, I., Verderosa, A., 2022. Investigating Airbnb evolution in an urban tourism context: the application of mathematical modelling and spatial analysis. Current Issues in Tourism 25 (10), 1666-1681.

García Bujalance, S., Barrera-Fernández, D., Scalici, M., 2019. Touristification in historic cities. Reflections on Malaga. Revista de Turismo Contemporâneo 7 (1), 93-115.

Goodchild, M.F., 2016. GIS in the Era of Big Data. Cybergeo: European Journal of Geography. n. pag.

Kelleher, J.D., Brendan, T., 2018. Data Science. The MIT Press, Cambridge.

Lestegás, I., Seixas, J. and Lois-González, R.-C., 2019. Commodifying Lisbon: A Study on the Spatial Concentration of Short-Term Rentals. Social Sciences 8 (2), 33.

Oliveira, A., Fachada, N., Carvalho, J., 2024. Data Science for Geographic Information Systems. arXiv Preprint 2404.03754.

Ojeda, A.B., Kieffer, M., 2020. Touristification. Empty concept or element of analysis in tourism geography? Geoforum 115, 143-145.

Porfido, E., Tomàs, M., Marull, J., 2023. A new urban diagnostics approach for measuring touristification: The case of the Metropolitan Area of Barcelona. Journal of Urban Management 12 (3), 195-207.

Rey, S., Arribas-Bel, D., John Wolf, L., 2023. Geographic Data Science With Python, first ed. Chapman and Hall/CRC, New York.

Simas, T.B., Oliveira, S.A.L.C.D., Cano-Hila, A.B., 2021. Tourismophobia or touristification? An analysis of the impacts of tourism in Poblenou, Barcelona. Ambiente Construído 21 (3), 117-131.