ESSnet Trusted Smart Statistics – Web Intelligence Network
Grant Agreement Number: 101035829 — 2020-PL-SmartStat

Work Package 3

New Use-cases

Deliverable 3.9: WP3 UC 4

Report on methods for analysing hotel price data and computing various indices of interest

Version, 2025-03-03

Prepared by:

UC4 coordinator(s):
Marek Cierpial-Wolan – (GUS, Poland); M.Cierpial-Wolan@stat.gov.pl

Contributors:
Łukasz Zadorożny – (GUS, Poland)
Szlachta Piotr – (GUS, Poland)
Galya Stateva – (BNSI, Bulgaria)
Kostadin Georgiev – (BNSI, Bulgaria)

This document was funded by the European Union.

Table of Content

1. Introduction

The objective of ESSnet WIN Work Package 3, use case 4 – „Experimental indices in tourism statistics (UC4) is to develop experimental indicators based on data collected through web scraping from online platforms for the purpose of conducting statistical research in the field of tourism. The information for this report was obtained through web scraping from the Booking.com portal. Data acquired from the aforementioned source can be valuable for analyzing both the accommodation base in tourism (supply side of tourism) as well as for studying tourists' travel patterns and expenditures (demand side of tourism) for balance of payments purposes.

2. Web scraping

In order to collect data from various online portals, a proprietary web scraping concept developed by Statistics Poland was implemented, based on two separately executed scripts (Figure UC-4-PL-1). The first script was run twice a week: on Mondays (collecting data for the following Monday regarding a single night stay) and on Thursdays (gathering information for the upcoming weekend covering two nights). This script collected only basic data about the properties available in the portal's offers, such as the URL, the name of the property, and the price per night. The second script was run once a month, on the first day of each month, to obtain more detailed information about the properties listed on the offer subpages. For this purpose, the URLs acquired during the operation of the first script were used. This method ensured the minimization of the load on the websites from which the information was retrieved.

The entire web scraping process included, among others, the identification of websites, an analysis of their structure, the examination of legal aspects related to conducting web scraping, as well as the implementation and storage of extracted data in a temporary database. The retrieved information was characterized by an unstructured format, making it unsuitable for immediate analysis. Therefore, after each data retrieval, the information underwent processes of cleaning and normalization. Only after these steps were completed were the formatted data loaded into the repository for further analysis.

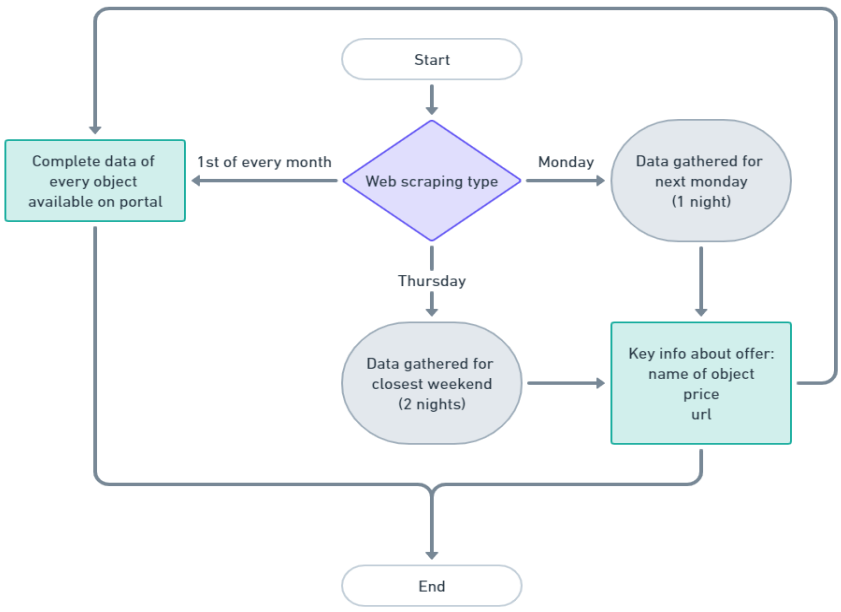

Figure UC-4-PL-1. Solution for data collection for booking portal

Online booking platforms such as Booking.com, Airbnb, Expedia, and Hotels.com serve as key sources of information on the availability and pricing of accommodations worldwide. Therefore, web scraping appears to be an excellent technique for extracting data from these platforms. However, over the past two years, these booking platforms have implemented a series of changes that have significantly impacted the effectiveness of web scraping. The main challenges and changes introduced on these platforms, along with their impact on the automated data collection process, include:

1. Tightening of anti-Scraping policies on booking platforms

- In the past two years, booking platforms have intensified their efforts to protect against web scraping by implementing advanced technologies and strategies aimed at detecting and blocking automated data extraction:
- IP Blocking and Bot Detection Systems: Most booking platforms use systems to analyze traffic patterns on their websites to detect and block IP addresses that generate an unusually high number of requests. Such measures compel scrapers to use proxy networks and IP rotation, which increase both the costs and the risks associated with data acquisition.
- CAPTCHA and reCAPTCHA Technologies: Many booking platforms, such as Booking.com and Expedia, have implemented CAPTCHA systems that appear after a high volume of requests or unusual user activities, significantly complicating the automation of scraping processes. Circumventing these mechanisms often requires advanced solutions, such as image recognition or human intervention.
- Limitation of Queries per User: To prevent large-scale data acquisition, booking platforms frequently restrict the number of queries sent from a single account or user session. This may necessitate the creation of multiple accounts or sessions, further complicating the scraping process.

2. Frequent changes in website structure and user interface

Booking platforms regularly update their websites, making changes to page layout, user interface, and code structure, which pose significant challenges for scrapers:

- Seasonal and promotional visual changes: Platforms often conduct seasonal updates to their websites, adapting the appearance and layout to match tourism seasons, promotions, or special events. These changes can involve alterations in the HTML structure, repositioning of key data, or modifications in the way information is presented, requiring scrapers to continuously update their tools.
- Changes in the presentation of search results: To enhance user experience and minimize the impact of scraping, platforms such as Booking.com and Airbnb experiment with different methods of displaying search results. These may include dynamic content loading, variable sorting, and new filters, making systematic data collection more difficult and necessitating more sophisticated adaptive methods from scrapers.
- Dynamic content loading and use of AJAX technology: Techniques such as AJAX and other dynamic loading mechanisms are increasingly used by booking platforms to load content, such as prices and availability, only upon request. This complicates traditional scraping, which relies on extracting complete HTML pages, and necessitates the use of more advanced technologies, such as browser emulation or JavaScript rendering.

3. Limiting the number of results displayed to users

Another strategy employed by booking platforms to protect their data resources involves limiting the number of results presented in response to a user's query:

Limiting the number of search result pages: Platforms often restrict the number of results that can be displayed at once for a specific query. For instance, a user may be able to view only a predetermined number of result pages, forcing scrapers to execute multiple queries, which increases the risk of detection by anti-scraping systems.

Pagination and infinite scrolling: Platforms such as Airbnb use infinite scrolling instead of traditional pagination, meaning that subsequent results are loaded only as the user scrolls down the page. This approach complicates scraping scripts, which need to emulate user behavior, increasing both the complexity and the time required to retrieve data.

The changes implemented by booking platforms over the past two years have significantly complicated web scraping. The tightening of anti-scraping policies, frequent changes in website structure, and limitations on the number of results displayed are just a few of the measures that require constant monitoring and code updates. This process can be highly time-consuming, ranging from minor code modifications to the need for a complete re-analysis of the portal and the development of new scraping scripts.

3. Web scraping results

One of the main platforms from which web scraping was conducted between 2022 and 2024 was the international website Booking.com, which serves as an online accommodation reservation portal. Data was collected from this platform regarding the number of listed accommodation properties in Poland and Bulgaria, including the names of the properties, their exact locations (street, building number, postal code), rental prices (for 2 adults), and the amenities offered at each property. An example of the locations of the various variables involved in web scraping on Booking.com is presented in Figure 1. The choice of Booking.com was based on user rankings, indicating that this platform is the most frequently used by tourists for accommodation reservations.

Information about accommodation properties was gathered for the entire country in both Poland and Bulgaria. In Poland, data was collected across 16 voivodeships, while in Bulgaria it was obtained from 28 districts. The assignment of accommodation properties to specific locations was made possible by acquiring information regarding their addresses. Each year, during the analysis of the collected datasets, it was noted that a portion of the addresses (the correlation between postal codes and specific towns) listed on the platform was incorrect. However, the number of inaccuracies was minimal. In Poland, the number of properties with incorrectly assigned postal codes accounted for approximately 1.6% of the total properties listed on the portal, while in Bulgaria, it was around 1.3%.

There were also concerns regarding the classification of properties into specific NACE categories. The classifications presented on Booking.com often did not correspond to the classifications used in official statistics. Furthermore, the names of the properties published on the website were sometimes misleading; for example, a property might be labeled as a hotel (the term included in the property name) when, in fact, it was a different type of lodging (e.g., a motel, guesthouse) or classified under NACE category 55.2 (Holiday and other short-stay accommodation). However, these discrepancies did not pose significant issues during the project work. For classifying properties into specific NACE categories, a proprietary decision tree method was employed.

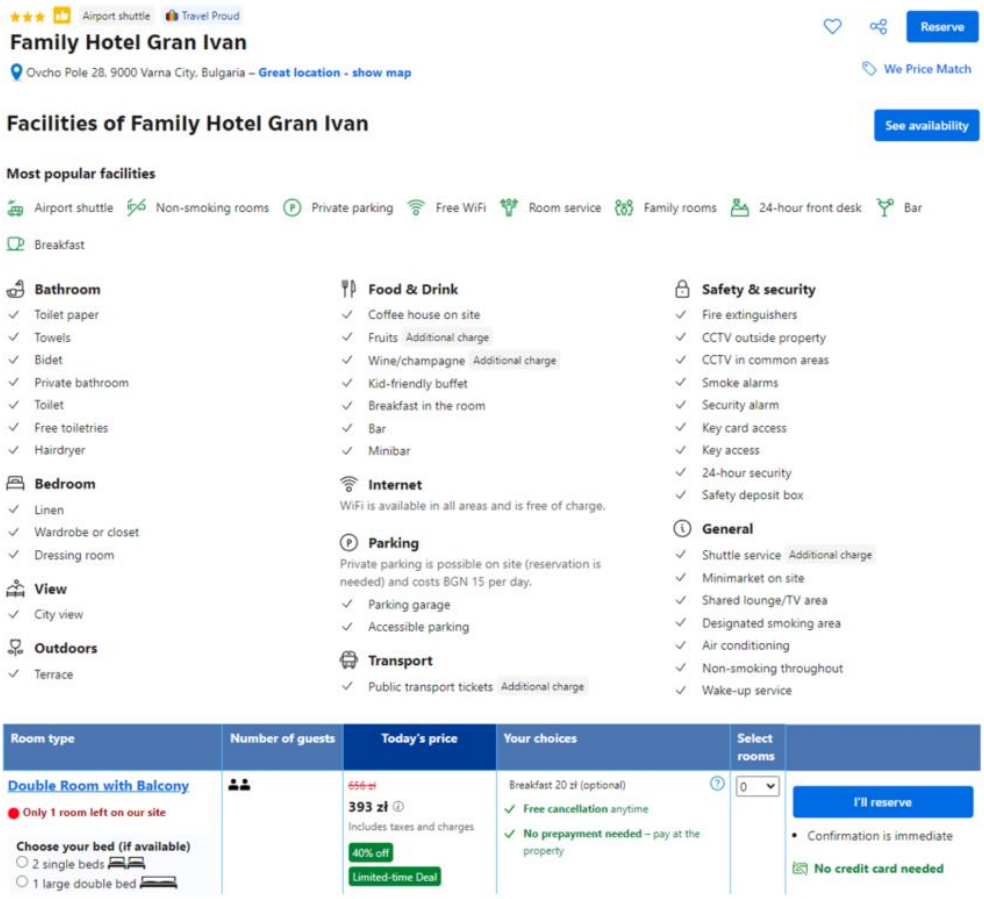

Figure UC-4-PL-1. Example of Variables Collected from Booking.com Using Web Scraping

In 2022, web scraping was conducted 122 times on the Booking.com platform. The results are presented in the table (UC-4-PL-1).

Table UC-4-PL-1. Web scraping - basic information on variables

variable	Total objects scraped	Unique objects scraped	No of missing values [qty]	No of missing values [%]
web_scraping_date	1 164 430	122	-	-
region	1 164 430	44	-	-
url	1 164 430	25 020	-	-
object_name	1 164 430	25 020	-	-
object_type	1 164 430	28	-	-
price	1 152 926	1 899	11 414	0,9

During the analysis of the acquired dataset, a total of 25,020 unique accommodation properties were recorded (from both Poland and Bulgaria), classified into 28 categories according to the classification used by the online portal. Almost all web-scraped variables were successfully collected, with the exception of data concerning rental prices, where the missing values were estimated at 0.98%.

In 2023, web scraping was conducted 118 times on the Booking.com platform (as shown in Table UC-4-PL-2). The dataset revealed a total of 24,144 unique accommodation properties, belonging to 29 categories. It was also noted that there were missing values for the price variable, amounting to 2.3%.

Table UC-4-PL-2. Web scraping - basic information on variables

variable	Total objects scraped	Unique objects scraped	No of missing values [qty]	No of missing values [%]
web_scraping_date	1 099 715	118	-	-
region	1 099 715	44	-	-
url	1 099 715	24 144	-	-
object_name	1 099 715	24 144	-	-
object_type	1 099 715	29	-	-
price	1 027 588	1821	11 005	2,3

In 2024, web scraping was conducted 125 times on the Booking.com platform (as shown in Table UC-4-PL-3). A total of 24,144 unique accommodation properties were recorded, which were classified into 32 categories by Booking.com, including capsule hotels, chalets, and even love hotels. For the price variable, it was noted that there were missing values, accounting for approximately 1.2% of the data.

Table UC-4-PL-3. Web scraping - basic information on variables

variable	Total objects scraped	Unique objects scraped	No of missing values [qty]	No of missing values [%]
web_scraping_date	1 200 499	125	-	-
region	1 200 499	44	-	-
url	1 200 499	24 815	-	-
object_name	1 200 499	24 815	-	-
object_type	1 200 499	32	-	-
price	1 231 856	1994	12 258	1,2

4. Data linkage

Extracting data from portals can provide new indicators on its own, but to fully unlock its potential, it needs to be combined with statistical sources. Due to the fact that the classification of accommodation facilities on booking platforms is not as precise as in the registries used for statistical surveys, comparing the average rental prices of accommodations between various categories may lead to errors, such as underestimations or overestimations of rental prices. According to the NACE classification, accommodation facilities are categorized into three groups: 55.10 (Hotels and similar accommodation), 55.20 (Holiday and other short-stay accommodation), and 55.30 (Camping grounds, recreational vehicle parks, and trailer parks). Accommodation facilities participating in statistical surveys are grouped based on the aforementioned classification. However, on platforms such as Booking.com, Hotels.com, and Tripadvisor.com, grouping is often determined by the owner's declaration.

As a result, facilities classified under NACE 55.2, where rental prices are typically slightly lower (Figure UC-4-PL-3), may be mistakenly categorized as hotels (55.10) since the term "hotel" is frequently included in the names of 55.20 facilities. Conversely, facilities such as guesthouses or motels, which belong to the hotel category (55.1), may be incorrectly described and treated as 55.2 facilities. These discrepancies complicate

data analysis from these platforms and hinder the drawing of accurate conclusions regarding accommodation rental prices.
Integrating data obtained from booking platforms at the individual facility level with registry data used for statistical surveys of accommodation facilities effectively addresses this issue, resulting in more accurate estimaes of accommodation rental prices.

In Poland, the process of merging registry data with data from booking platforms serves an additional important function. It facilitates the identification of new accommodation facilities that are not yet included in the registry used for the accommodation base survey. Based on the results of this integration, new facilities are added to the registry and informed about their reporting obligations to public statistics. This solution increases the number of facilities participating in the survey, thereby improving the quality of its results.

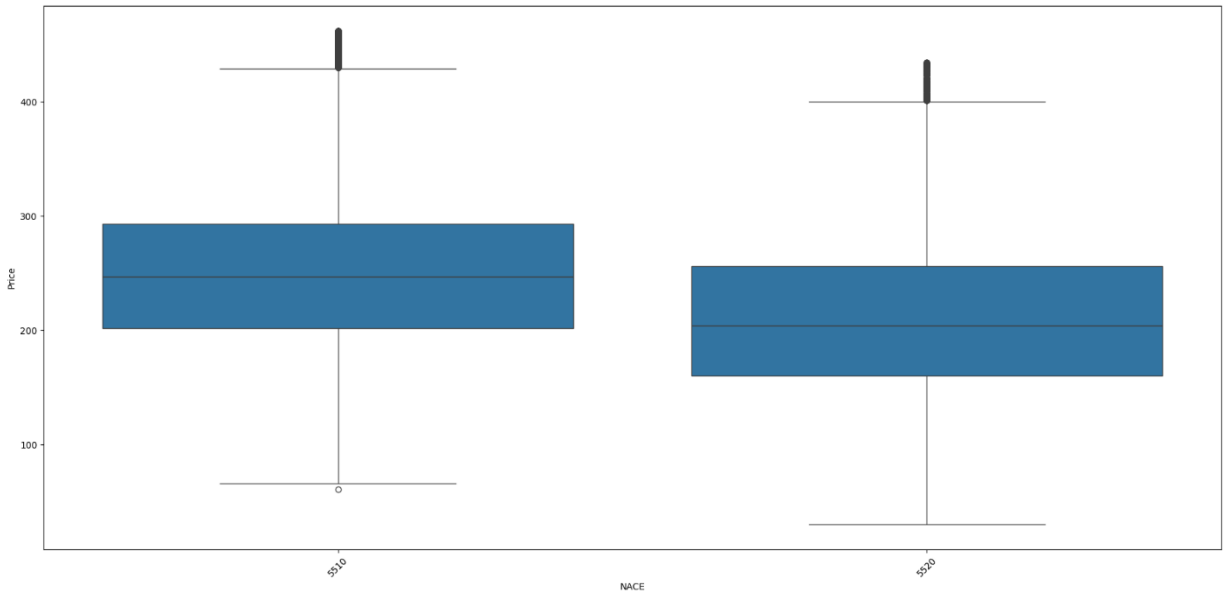

Figure UC-4-PL-3. Range of rental prices for Hotels and similar accommodation (55.1) and Holiday and other short-stay accommodation (55.2) according to information collected from booking platforms.

In this context, two primary data matching approaches can be identified:

1. Deterministic matching: This method identifies the properties within datasets and seeks exact matches. It operates on a binary principle of complete agreement or disagreement, offering limited flexibility.
2. Probabilistic matching: This approach calculates the likelihood that records match, providing a probability score to indicate whether the records correspond or not. One specific form of probabilistic matching is fuzzy matching.

Errors, such as incorrect formatting or spelling mistakes, frequently occur in property names or addresses on booking platforms. Deterministic matching can only confirm if an address is precisely identical, making it less effective at handling variations caused by formatting or spelling discrepancies. In contrast, probabilistic linking methods, like fuzzy matching, are better suited for identifying and correcting these inconsistencies.

Fuzzy logic effectively addresses many issues that deterministic methods cannot manage, including:

- Assessing the probability of a complete match, which facilitates more accurate data comparison.
- Accommodating minor typos and formatting inconsistencies by analyzing the similarity based on the number of differing characters.
- Handling unique or non-standard address elements. With fuzzy logic, it is possible to establish rules for standardizing addresses, allowing for corrections to street names and other components for more precise alignment.

Several techniques can be utilized for fuzzy matching. In this project, two different methods were employed and compared. The first method involved the FuzzyMatcher library. This tool merges data from different sources that may not perfectly align. FuzzyMatcher utilizes fuzzy string matching techniques to compare and assess the similarity of text strings. Techniques such as Levenshtein Distance, Ratio, and Partial Ratio can be applied within this library.

The second method tested was the Python Record Linkage Toolkit, which is designed for connecting data records and identifying duplicates. The toolkit offers several features, including:

- The ability to define linkage types based on data categories.
- Optimized solutions for faster data linking.
- Data ranking capabilities for linkage results.
- Support for multiple string similarity algorithms.
- Options for both supervised and unsupervised learning models.

Different criteria can be used to compare columns, such as searching for exact matches in standardized fields like postal codes or city names. Moreover, the toolkit allows for customized string comparisons by setting specific thresholds and algorithms. Additional comparisons can incorporate custom elements, such as numeric values, dates, or geographical data.

Both libraries provide solutions for data linking, especially when dealing with addresses. To assess the similarity between addresses and object names, each library relies on string comparison methods. In the FuzzyMatcher library, Levenshtein Distance was applied, whereas the Record Linkage Toolkit enables the assignment of various methods to different columns. In this project, both Levenshtein Distance and Jaro-Winkler Distance were tested.

The final accuracy depends heavily on the quality of the data and the chosen comparison techniques. In cases where there were significant variations in object names, additional tools like geocoding were employed to refine the results. A comparison of the two libraries' results showed that the record linkage toolkit outperformed FuzzyMatcher by 9%. Incorporating object distance thresholds further increased the accuracy, achieving a 12% improvement over the results obtained using only FuzzyMatcher.

Fuzzy logic is generally more effective than traditional address matching due to its greater flexibility, but it is not without its limitations. Several key points should be considered when applying fuzzy logic techniques: Although fuzzy logic can estimate the probability of a match, certain spelling and formatting discrepancies might still go undetected. It does not fully address the issue of data deduplication, as there remain obstacles to ensuring reliable deduplication. Limited additional address information for validating matches can result in errors, even with specific probability estimates. Despite overcoming many challenges that direct address matching cannot manage, fuzzy logic is not a flawless solution. Nevertheless, its numerous benefits and the ease of creating generalizable approaches make it a practical choice for integrating data from web scraping with statistical sources.

5. Accommodation establishment metrics

In the field of tourism statistics, understanding the landscape of accommodation establishments is crucial for understanding traveler preferences, market segmentation and infrastructure development.

Although the metrics bear names similar to those used in official statistics, their values may differ due to inherent biases and errors in web-scraped data. These differences arise from factors such as platform-specific coverage, exclusion of smaller establishments, and potential validity errors. For instance, the classification of establishments or the granularity of data might vary depending on the data source.

It is crucial to emphasize these distinctions to avoid misinterpretation and to align expectations when comparing web-derived metrics to traditional statistical indicators.

Cooperation between Poland and Bulgaria has led to the identification and establishment of a set of variables that cover the diversity of accommodation establishments and facilities:

- Total number of Establishments by Type
  For each accommodation type, the total number of establishments based on the NACE 55 categories is computed as:

$$Total\ number\ of\ Establishments_t = \sum_{j=1}^{n_t} \mathrm{E}_{t,j}$$

where n_t represents the number of NACE 55 categories for type t and $E_{t,j}$ denotes the number of establishments in the jth NACE 55 category within that type.

This metric quantifies the distribution of accommodation establishments based on the NACE 55 classification categories, encompassing diverse segments such as hotels, motels, and holiday centers. Understanding the distribution of establishments within each category offers insights into market segmentation and business dynamics.

- Total number of Establishments in the District by type

$$Total\ number\ of\ Establishments_d = \sum_{t=1}^{T} E_{d,t}$$

where T represents the total number of accommodation types and $E_{d,t}$ is the number of establishments of type t in district d.

This metric delves into the geographical distribution of accommodation establishments within provinces or districts. By analyzing the concentration of establishments in different districts, trends in tourism infrastructure development and demand patterns can be identified.

- Number of Accommodation Establishments in the Region Broken Down by the Number of Beds/Rooms
  There are two approaches to aggregating the capacity of accommodation establishments within a region/district:
  a) Aggregation at the Establishment Level:

$$Total\ number\ of\ beds/rooms_d = \sum_{i=1}^{E_d} b_i$$

where E_d is the total number of establishments in the region/district and b_i denotes the number of beds/rooms in the *i*th establishment..

b) Aggregation by Type:

$$Total\ number\ of\ beds/rooms_d = \sum_{t=1}^{T_d} (E_t \times b_t)$$

where T_d represents the total number of accommodation types in region/district, E_t is the number of establishments of type *t*, and b_t is the average number of beds/rooms for establishments of type *t*.

This metric provides detailed insights into the capacity and scale of accommodation providers within a region/district, categorized by the number of beds or rooms available. Understanding the distribution of beds/rooms facilitates strategic planning and resource allocation.

- Average Price per Overnight Stay at a Facility in the District

$$average\ price\ per\ overnight\ stay_d = \frac{\sum_{i=1}^{n} price\ per\ night_i}{n}$$

where n represents the total number of accommodation establishments and i represents each accommodation establishment.

This metric analyzes the pricing dynamics of accommodation establishments within specific regions. Calculating the average price per overnight stay provides insights into market competitiveness and concentration of tourist areas. The developed indicators may be used for tourism accommodation base studies to facilitate comparisons and validations with officially existing statistics, as auxiliary data, or for purposes related to building statistical registers containing information about accommodation facilities.

6. Results

Based on data obtained through web scraping from online portals, indicators regarding the average rental prices of accommodations in Poland and Bulgaria have been developed. In both countries, rental prices for accommodation in tourist lodging establishments exhibited noticeable seasonality (as shown in Table UC-4-PL-4). The highest average rental price for accommodation in Poland was recorded in July, followed by a slightly lower average in August. This trend is associated with increased tourist traffic and the peak of the tourist season occurring during these months.

According to official statistics, the majority of tourists in Poland visit accommodation properties between May and September. During this period, over 50% of all tourists utilize lodging services, with approximately 25% of the total using accommodation from July to August.

Table UC-4-PL-4. Average rental price of accommodation in Poland by month

Year	Month	Objects in Poland	Avg. Price in Poland
2022	January	8968	257,25

2023		8276	259,11
2024		8435	258,67
2022	February	9507	260,85
2023		9233	259,39
2024		9401	261,99
2022	March	9174	299,68
2023		8993	302,74
2024		9055	302,65
2022	April	9903	297,22
2023		10017	300,15
2024		9986	303,39
2022	May	10442	305,27
2023		10687	314,45
2024		10613	319,78
2022	June	10897	309,21
2023		10721	310,02
2024		10777	309,23
2022	July	6935	407,40
2023		6886	411,97
2024		6913	412,99
2022	August	8585	366,36
2023		8499	376,28
2024		8406	378,65
2022	September	9241	355,93
2023		9490	365,03
2022	October	11490	313,83
2023		11573	311,24
2022	November	11864	325,02
2023		10451	325,30
2022	December	10446	299,43
2023		10396	300,44

An opposite trend can be observed regarding the number of accommodation properties listing their offerings on the platform. There were significantly more rental offers available on Booking.com outside the summer season. This may indicate a high occupancy rate of lodging establishments during July and August, leading to a reduction in the number of available accommodations and, consequently, fewer listings on the portal.

The increased demand for rental accommodations in the summer months, coupled with a limited number of available places, may also encourage property owners to raise their rental prices. Between 2022 and 2024, a noticeable increase in rental prices for accommodation in lodging establishments was recorded. A particularly significant rise was observed in December 2024, with the average rental price increasing by 25.87 euros compared to data from December 2022.

In Bulgaria, similar to Poland, the highest average rental price for accommodation was also recorded in July (as shown in Table UC-4-PL-5). However, a different trend was noted regarding the number of accommodation properties listing their offerings on Booking.com. Due to strong tourist interest in renting properties during the summer season, the available number of accommodations in August and September was the lowest for the entire year.

Table UC-4-PL-5. Average rental price of accommodation in Bulgaria by month

Year	Month	Objects in Bulgaria	Avg. Price in Bulgaria
2022	January	3712	256,71
2023		3702	252,44
2024		3749	257,83
2022	February	4995	262,27
2023		4888	272,00
2024		5017	269,74
2022	March	4808	306,05
2023		4812	341,86
2024		4885	335,47
2022	April	4441	307,74
2023		4397	305,61
2024		4429	314,66
2022	May	4609	312,89
2023		4579	330,20
2024		4675	331,88
2022	June	3544	310,29
2023		3506	310,88
2024		3521	315,42
2022	July	1946	372,15
2023		1933	394,55
2024		1863	390,87
2022	August	2631	366,96
2023		2645	376,06
2024		2600	380,01
2022	September	2824	357,41
2023		2910	321,77
2022	October	3434	319,44
2023		3454	320,90
2022	November	3587	322,15
2023		3299	317,46
2022	December	2971	312,54
2023		3045	308,71

Both countries show growth in the number of tourist objects and average prices from 2022 to 2024, reflecting a recovery or expansion phase post-pandemic

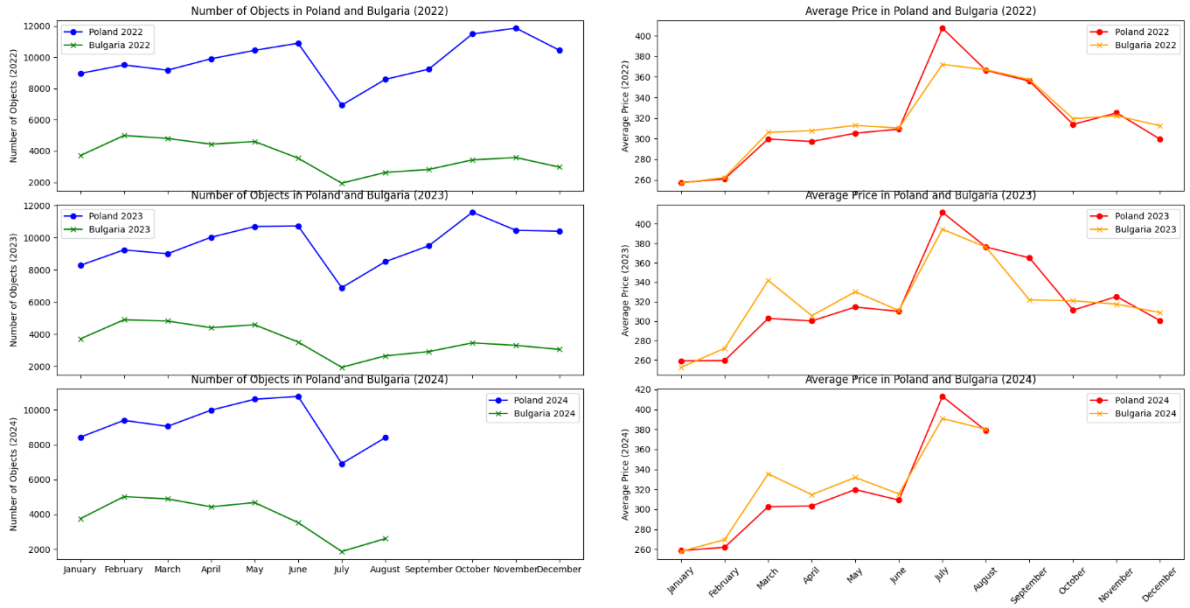

Figure UC-4-PL-4. Number of establishments and average rental price of accommodation in Poland and Bulgaria by month

7. Utilization of data from booking portals in statistical surveys in field of tourism

Data from booking portals, combined with appropriate analytical tools, provide detailed and valuable information that can support statistical research in the field of tourism. This pertains to both studies conducted within the framework of demand-side monitoring (research on travel and spending by domestic and international tourists) and supply-side monitoring (accommodation base) in tourism.

Information regarding rental prices for accommodation can primarily be used for imputing missing records in sample travel surveys conducted in European Union countries. Price imputation can be utilized, among other things, for:

determining the attractiveness index of tourist destinations:

- Higher accommodation prices may indicate more attractive and popular destinations, which could correlate with other variables such as the number of tourists, length of stay, or choice of accommodation type.
- Using prices for imputation can assist in estimating missing data regarding the number of tourists, their characteristics (e.g., higher income groups), and preferences concerning destinations.

Modeling the relationship between price and other variables:

- Prices can be used as independent variables in predictive models (e.g., regression) to impute missing values for other variables, such as the total cost of travel.
- The average rental price of accommodation in a given city allows for the imputation of missing records in sample travel surveys. It also enables the validation of expenditures reported by respondents related to accommodation rentals during the database verification stage.

Market segmentation:

- Based on prices, market segmentation can be conducted, allowing for the imputation of missing data tailored to specific tourist segments (e.g., luxury hotels for wealthier tourists vs. budget hostels for economical travelers).
- In this context, prices can serve as proxies for characteristics that may be missing, such as the age of tourists or their preferred travel style (budget vs. luxury).

Acquiring data on accommodation rental prices in specific countries/regions/cities can be an essential element for those conducting sample studies. In travel survey forms (in Poland, these include: Participation of Polish residents in travel, Non-residents' travel to Poland, and Vehicle and pedestrian traffic at Poland's borders with EU countries (PDP)), questions regarding the costs incurred for accommodation rental during travel are directed to respondents. Since both sample studies (residents and non-residents) are conducted quarterly, the rental amounts provided by respondents may sometimes deviate from the actual expenses incurred. This is related, among other factors, to the timing of the interview with the respondent, which may take place about three months after their travel. The long period between the trip and the survey can distort the estimated information reported by the respondent to the interviewer.

Having information on accommodation prices for specific groups of people (e.g., 1 adult, 2 adults, 2 adults + 1 child) during a given period (month, quarter, etc.) and for a specified tourist destination, along with travel survey data on the actual number of participants, allows for estimating the actual cost of accommodation rental. This enables the validation of the accuracy of the information provided by respondents during the record verification stage when checking the database.

Information on accommodation prices can also serve an important function in cases where the respondent may not know the answer to the question (for example, when the interview is conducted with someone who participated in the trip but did not make the payment for accommodation rental). This information gap can be filled through imputation by replacing missing data with values derived from the knowledge of all price values in the given tourist destination obtained from web scraping datasets.

In imputing records related to tourists' expenses on accommodation rental, one can use the median of the prices present in the web scraping dataset. The median is less susceptible to the influence of outliers (very high or very low prices) that can significantly distort the mean. In the context of accommodation prices, outlier values can occur quite frequently (e.g., luxury apartments or very cheap hostels), while the median provides a more representative measure of typical prices. The median also better reflects the average value that a respondent would encounter in the visited region, which may be more in line with reality than the mean, especially in the case of non-standard distributions (e.g., with a long "tail"). Additionally, in tourism, the price range can vary significantly depending on various factors such as seasonality, type of accommodation, etc. Taking this into account, the median better represents the central tendency in such a diverse dataset.

8. Image deduplication – out of box approach

This process is essential for determining the number of accommodation facilities that appear exclusively on the second of the mentioned platforms. AirBnB, unlike Booking.com, is a platform where rental offers predominantly pertain to properties classified under NACE group 55.2 (Holiday and other short-stay accommodation). This category typically includes private properties, often operating in the informal rental sector (apartments managed by private individuals), which is a particularly significant issue in the context of accommodation base studies. Unfortunately, the specific characteristics of this platform—namely, the lack of published data regarding the exact address and often even the precise name of the property—make it impossible, under the currently employed methodology, to link this data to statistical research registers

and, consequently, to determine the number of such units. Therefore, to effectively address the issue of duplication, a non-standard or innovative approach is required, enabling reliable matching of information about the same properties regardless of the descriptions available. One such solution could involve using photographs of accommodation facilities, which, despite being published across different booking platforms, exhibit a significant number of common elements.

In this chapter, we present an innovative approach to image deduplication, utilizing visual feature matching algorithms such as SIFT (Scale-Invariant Feature Transform). By analyzing the images themselves, without relying on metadata or descriptions, it becomes possible to effectively identify the same properties across different platforms. This approach is particularly useful in situations where textual data is insufficient or inconsistent, and the goal is to create a coherent and accurate tourism database.

Methodology

To address the issue of image deduplication using an out-of-the-box approach, the following steps were taken:

- **Data preparation**: A set of images representing tourist properties from various platforms was collected. Properties were labeled with symbols, where the same property from different platforms was assigned the same symbol, but with a different number representing the platform (e.g., A1 and A2 refer to the same property from different platforms).
- **SIFT feature extraction**: For each image, key points and descriptors were extracted using the SIFT (Scale-Invariant Feature Transform) algorithm, which is robust to changes in scale, rotation, and lighting conditions.
- **Image comparison**: Pairwise comparisons of images were performed by calculating the number of "good" matches between key points. The match percentage for each pair was determined by dividing the number of good matches by the total number of matches.
- **Establishing a matching threshold**: Based on the analysis of the results, a threshold of 20% was set as the minimum percentage of shared key points between two images to consider them as representing the same property.

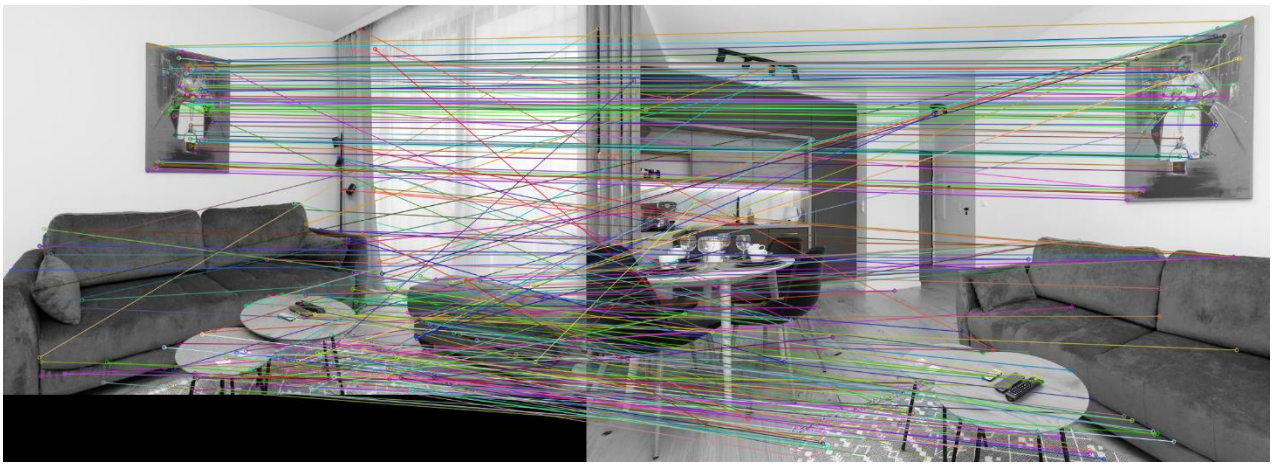

Figure UC-4-PL-5. Photo comparison of accommodation establishments.

Results

The applied out-of-the-box approach achieved an accuracy rate of 75%, indicating that the majority of image pairs representing the same object were correctly identified using the established 20% threshold for shared key points.

- **Effectiveness of the SIFT algorithm:** The SIFT algorithm successfully identified pairs of images of the same object, even when there were variations in perspective or lighting. This confirms its usefulness in the non-traditional approach to deduplication. The matching threshold was determined based on the distribution of match percentages across different image pairs. Pairs depicting the same object typically exceeded the 20% match threshold, whereas pairs of different objects exhibited lower percentages. This threshold allowed for a balanced trade-off between the sensitivity and specificity of the method.
- **Challenges and limitations:** In some cases, the method failed to correctly identify image pairs representing the same object. This could be attributed to significant differences in image quality, angle of capture, or the presence of noise and distortions. These factors can impact the effectiveness of the algorithm in detecting sufficient common key points.

The proposed out-of-the-box approach to image deduplication, utilizing the SIFT algorithm, presents an effective solution to the problem of object duplication in tourism databases. Setting the matching threshold at 20% shared key points proved to be successful, achieving 75% correct identifications. Due to its independence from textual descriptions, this method is particularly valuable in integrating data from multiple sources and can significantly improve the quality of information in the tourism sector. Combining the SIFT algorithm with other techniques, such as metadata analysis or location-based information, could further enhance the results.

9. Web scraping: solution or problem?

Web scraping has become one of the most widely used methods by statisticians across Europe. As one of the simplest ways to access new data sources, it holds particular importance in the tourism sector, where services are primarily directed towards travelers. Although identifying sources that can improve research quality seems straightforward, various challenges emerge at each stage of the web scraping process.

Modern websites are often characterized by advanced designs and the use of cutting-edge technologies, many of which require JavaScript execution. Basic scraping libraries frequently struggle with content loading, and while more advanced libraries can handle interactive elements, they are typically slower and may require additional solutions. Increasingly, even basic information—such as the exact address of a property or a specific flight number—requires additional user interaction, complicating the scraping process and eliminating simple, quick solutions.

Large platforms like Booking.com and Trip.com dynamically define classes and identifiers in their HTML structures. Such dynamic generation requires a deeper understanding of the website's architecture and makes scraping solutions more vulnerable to changes in the portal structures. A growing trend among booking sites is to completely alter their page structure depending on the season, often necessitating a complete rewrite of web scraping code.

The taxonomy of property types on reservation platforms frequently differs from the classifications used in statistical research, complicating data analysis. Identifying properties—usually hotels—that offer individual apartments as separate accommodation options adds another layer of complexity.

Efforts to extract data from travel agency websites and cruise portals have encountered specific challenges. Attempts to gather data on selected variables from sites such as Alexandertour.com and Bohemia.bg were unsuccessful for several reasons:

- **Lack of standardized variables:** There is no uniform representation of variables like "Offer price (2 adults + 1 child)" or "Types of services included in the trip price" across different data sources and advertisements.
- **Inconsistent HTML structures:** Advertisements lack consistent HTML structures, making it difficult to locate and extract necessary information.
- **Variability in information presentation:** Differences in how information is presented in advertisements hinder data collection and classification.
- **Absence of universal scraping tools:** There are no universal scripts suitable for data sources from travel agencies, complicating the data collection process.

Overcoming these challenges is crucial for the effective use of web scraping techniques in statistical research within the tourism sector. The development of advanced algorithms and methodologies may help navigate these obstacles and expand the possibilities for data collection. Close collaboration with website developers to create standardized data formats and integrate APIs could provide easier access to travel agency information. Additionally, the development of methodologies for calculating indicators and their use in producing experimental statistics plays a vital role in improving tourism research.

Overcoming these challenges is crucial for the effective use of web scraping techniques in statistical research within the tourism sector. The development of advanced algorithms and methodologies may help navigate these obstacles and expand the possibilities for data collection.

Close collaboration with website developers to create standardized data formats and integrate APIs could provide easier access to travel agency information. In addition to these measures, negotiating access to databases of booking platforms presents a promising alternative.

Establishing formal collaborations with platform owners could provide standardized and consistent data sources, bypassing many of the technical and legal complexities associated with scraping. Such partnerships would allow statistical offices to access anonymized datasets that comply with privacy regulations while ensuring data reliability and consistency. Furthermore, platform owners could benefit from enhancing their corporate social responsibility profiles by contributing to public good initiatives. These collaborative efforts, along with the development of methodologies for calculating indicators and their use in producing experimental statistics, play a vital role in improving tourism research.

10. Conclusions

- Using the web scraping method, it is possible to obtain data on accommodation rental prices for specific tourist destinations and during a specified period of tim.
- Reservation platforms implement solutions on their websites that hinder the process of web scraping. As a result, the user is forced to continuously monitor the structure of the platforms and update the web scraping codes accordingly.
- Utilizing data obtained from reservation portals through web scraping, indicators regarding the average rental prices of accommodations in Poland and Bulgaria have been developed. The rental prices exhibited distinct seasonality.

- The average monthly rental prices of accommodations in Poland were generally higher than those in Bulgaria.
- Information on accommodation rental prices in specific tourist destinations is useful for validating and imputing missing records in sample survey results on travel and tourist expenditures conducted by national statistical offices within the European Union.
- Comparing images across platforms that offer accommodation bookings can assist in the deduplication of accommodation property datasets. This is particularly important in cases where data from the platforms do not include information about the property's address.