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

Work Package 3

New Use-cases

Deliverable 3.10: WP3 UC 4

Report on methods to be used for imputation of price data in price statistics.

Version, 2025-03-03

Prepared by:

UC coordinator(s):
Marek Cierpiał-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 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 web portals Expatistan.com and Trip.com. Data acquired from the aforementioned sources can be used when validating and imputing missing records in sample surveys of tourist travel and spending (demand side of tourism) conducted for balance of payments and tourism statistics.

2 Web scraping data collection and preliminary analysis

Information on the prices of various goods and services was obtained from Expatistan.com, Expatistan.com is an online portal that provides information on the cost of living in various cities and countries around the world. The platform is primarily aimed at people considering moving abroad, including expats, travelers and professionals working for multinational corporations. Expatistan.com's main goal is to allow users to compare the cost of living in different locations, which is particularly useful when planning travel-related changes.

The portal relies on user-provided data, which is regularly updated and verified, to provide reliable information on prices for food, lodging, transportation, health services and many other aspects of daily life. Expatistan.com also enables the generation of detailed comparison reports, which helps to better understand price differences between cities.

During the project, observations were conducted on 9 products and 5 services available in 26 Polish cities and 16 Bulgarian cities. Data were collected monthly on the 16th of each month using web scraping. For most cities, price information was successfully obtained, however, some data could not be retrieved due to challenges encountered during the web scraping process. The main information summarizes the overall dataset quality, indicating both the successfully acquired data and the gaps. The accompanying chart (Figure UC-4-PL-1) presents data for individual cities, while the table (Table UC-4-PL-1) details the missing data for each product.

During the project work, the observation covered 9 products and 5 services offered in 26 cities in Poland and 16 in Bulgaria. For most of the cities it was possible to obtain price information, unfortunately, due to the difficulties encountered during the web scraping process, some of the information was not retrieved. The main information on the acquired data (in purple) is shown in Figure UC-4-PL-1.

The nine products whose prices were analyzed during the project work include:

- Milk - 1 liter (1 qt.) of whole fat milk;
- Bread - Bread for 2 people for 1 day;
- Eggs - 12 eggs, large;
- Water - 2 liters of water;
- Cigarettes - 1 package of Marlboro cigarettes;
- Apples - 1 kg (2 lb.) of apples;
- Cappuccino - Cappuccino in expat area of the city;
- Gasoline - 1 liter (1/4 gallon) of gas.
- Transportation - Monthly ticket public transport;

The 5 services whose prices were analysed during the project work include:

- Taxi - Taxi trip on a business day, basic tariff, 8 km. (5 miles)
- Meals - Dinner for two at an Italian restaurant in the expat area including appetisers, main course, wine and dessert;
- Movie - 2 tickets to the movies;
- Meal - Combo meal in fast food restaurant (big mac meal or similar);
- Housing - Monthly rent for 85 m^2 (900 sqft) furnished accommodation in expensive area.

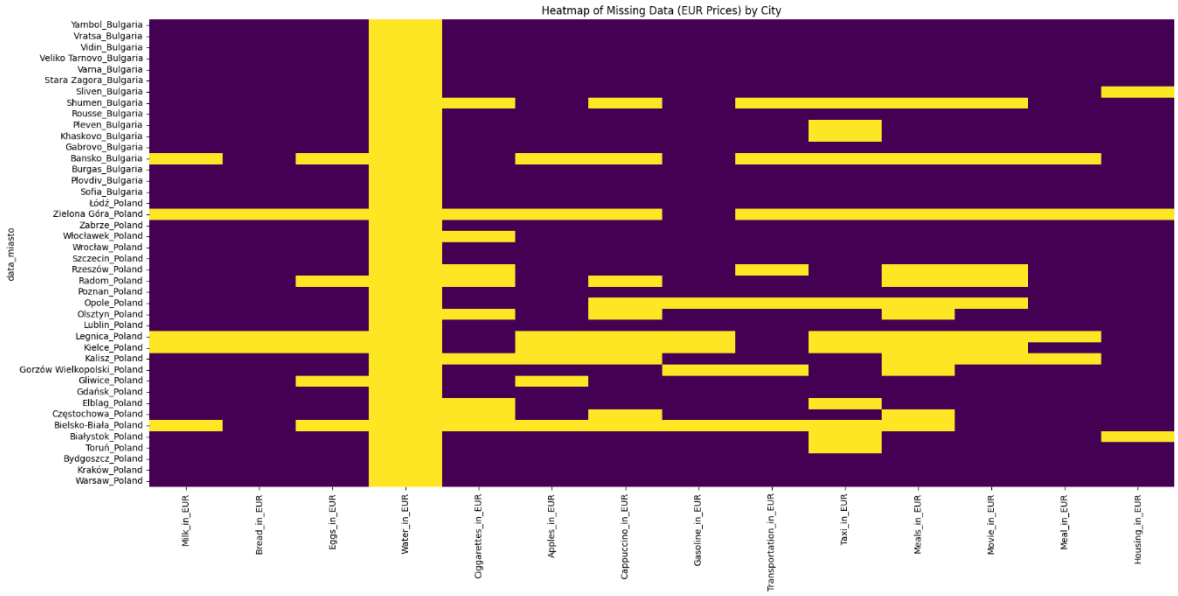

Figure UC-4-PL-1. Collected and missing data by city for one month

Table UC-4-PL-1. Missing values by variables for source *expatistan.com*

Columns	Count	Missing values	% of missing values
Apples	646	208	24.4
Bread	704	150	17.6
Cappuccino	545	309	36.2
Cigarettes	611	243	28.5
Eggs	657	197	23.1
Gasoline	683	171	20
Housing	659	195	22.8
Meal	675	179	21
Meals	553	301	35.2
Milk	685	169	19.8
Movie	592	262	30.7
Taxi	552	302	35.4
Transportation	618	236	27.6
Water	0	854	100

To understand the overall market trends, we analyzed the average prices of the observed products and services over the study period. It is important to note that these averages are computed without a weighting mechanism; therefore, high-value segments such as accommodation may disproportionately influence the overall figures. While this unweighted approach provides a clear snapshot of market dynamics and volatility, it also lays the groundwork for more detailed, weighted analyses in future studies.

The average price in Bulgaria remained relatively stable over the two-year period (Figure UC-4-PL-2), fluctuating between a low of EUR 52 in June 2023 and a high of EUR 63 in December 2022. After peaking at the end of 2022, the average price experienced a gradual decline through April 2023. From April 2023 to April 2024, the price remained stable, with a sharp increase in May 2024, reaching the December 2022 peak, and then maintaining a level slightly above EUR 60 in the subsequent months.

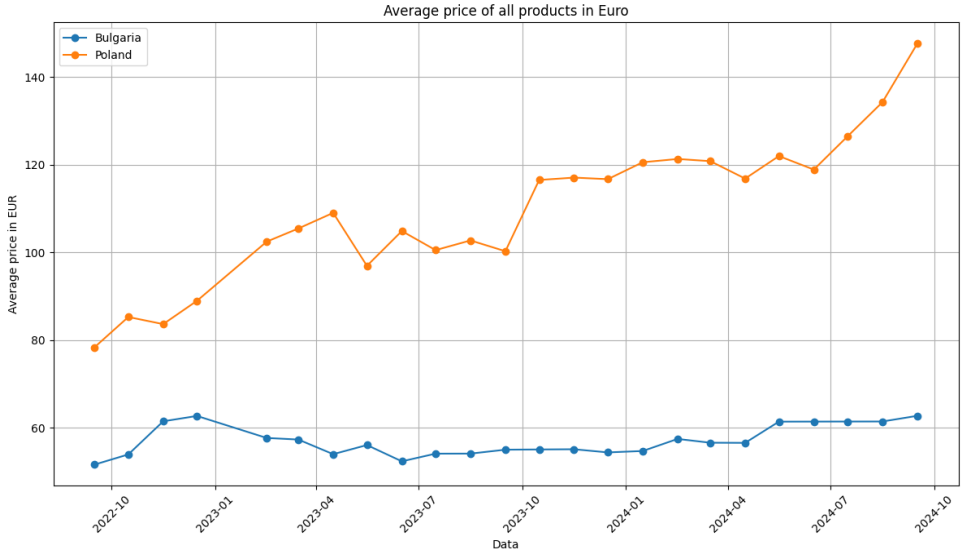

Figure UC-4-PL-2. Comparison of price changes in Bulgaria and Poland from 2022 to 2024

In contrast, the average price in Poland demonstrated a strong upward trend and significantly higher volatility. Starting from a low of EUR 79 in September 2022, prices rose steadily until reaching EUR 110 in April 2023. A decline followed in May 2023, with the average price stabilizing around EUR 100 for the next four months. Prices surged again in October 2023, remaining around EUR 120 until July 2024, after which they continued to rise, peaking at EUR 145 in September 2024 — an increase of over 80% compared to September 2022.

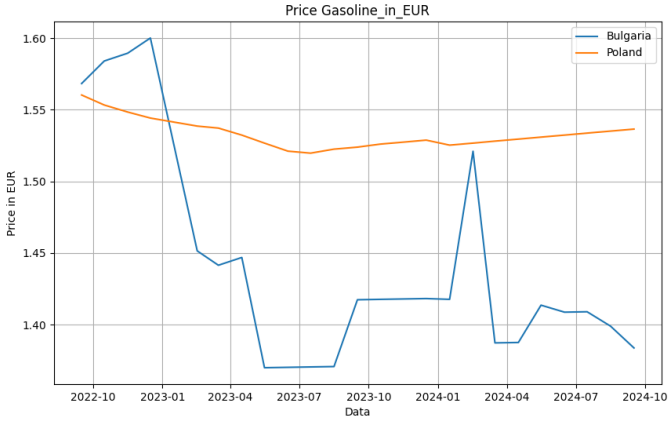

Figure UC-4-PL-3. Price of 1 litre (1/4 gallon) of gasoline in Bulgaria and Poland from 2022 to 2024

Considering the price of gasoline over the period under observation, it was found that in Poland the price of gasoline remained stable (around EUR 1.55 per litre), with a slight upward trend from mid-2023 (Figure UC-4-PL-3.). In Bulgaria, prices showed much more volatility. At the beginning of the period, the price of petrol was around EUR 1.60, before falling sharply towards the end of 2022. For most of 2023, prices remained at a lower level, around EUR 1.40, with periodic sharp increases and decreases. In 2024, price fluctuations were observed again, with short-lived increases. In Poland, compared to Bulgaria, petrol prices are more stable, which may be due to less fluctuation in the fuel market in that country.

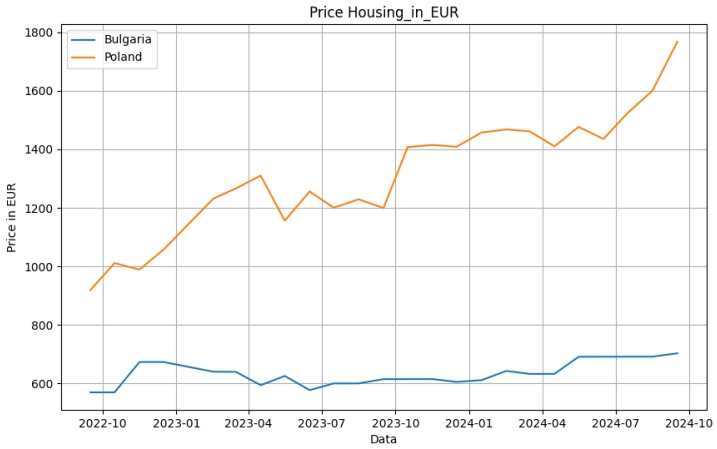

Figure UC-4-PL-4. Rent price of an 85 m^2 (900 sqft) furnished accommodation in an expensive area in Bulgaria and Poland from 2022 to 2024

In Poland and Bulgaria, based on the downloaded data, there is a clear difference in the dynamics of rental prices for an 85 m^2 flat. When analysing the collected data, it was found that rental prices in Poland have an upward trend (Figure UC-4-PL-4.). This increase is gradual, with some short-term fluctuations. At the beginning of the observation period, prices are around EUR 1,200, while in October 2024 they reach a level of around EUR 1,800. In Bulgaria, flat rental prices are much lower and more stable compared to Poland. At the beginning of 2023, the rental price of an 85 m^2 flat was slightly above EUR 500, and by the end of the study period the price had increased only slightly, reaching around EUR 600. In Bulgaria, flat prices remain relatively stable. This suggests that the property market in this country may be less dynamic or not as burdened by price pressures as the market in Poland.

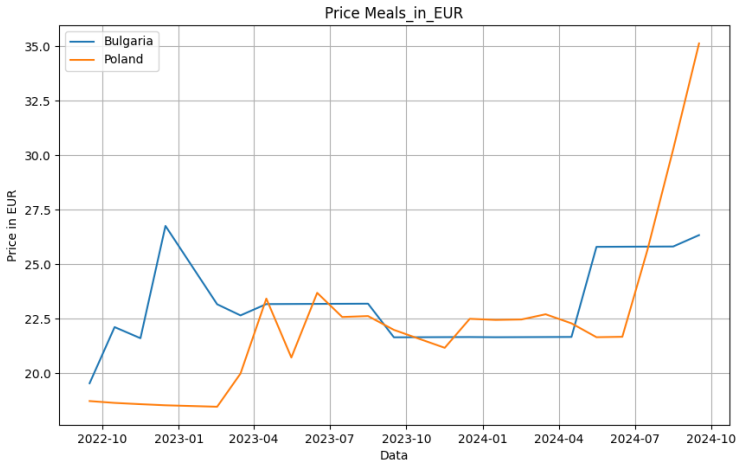

Figure UC-4-PL-5. Price of Dinner for two at an Italian restaurant in the expat area including appetisers, main course, wine and dessert in Bulgaria and Poland from 2022 to 2024

Figure UC-4-PL-5 shows a comparison of the prices of meals in an Italian restaurant in Bulgaria and Poland. Based on the analysis of the data, it was found that Bulgarian prices were quite volatile, especially in the first half of 2023. Initially, the price per meal was around EUR 22 and then, later on, it fluctuated between EUR 22 and EUR 27. In the second half of 2023, prices stabilised at around EUR 25 before rising to around EUR 27 by the end of 2024. In Poland, prices were initially lower, at around EUR 20, but by the middle of 2023 there was a rise. By the end of 2024, the purchase price of a meal for two in an Italian restaurant had risen to EUR 35, which was a significant increase compared to earlier values. In both Poland and Bulgaria, there was a difference in the growth rate of meal prices, with Poland experiencing a significantly stronger price increase in the final part of the analysed period.

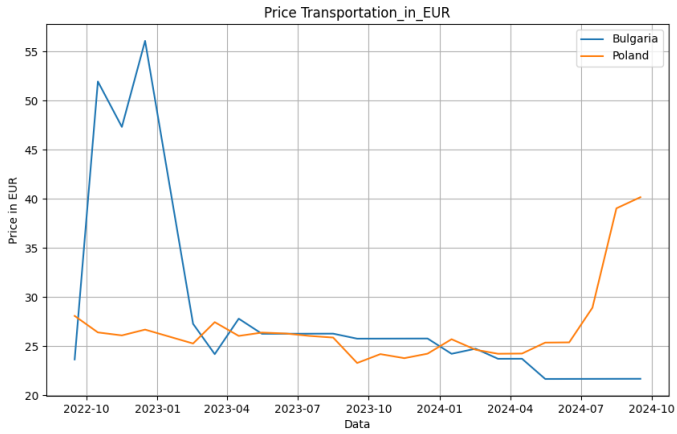

Figure UC-4-PL-6. Price of a monthly ticket for local transport in Bulgaria and Poland from 2022 to 2024

When analysing the price level of a monthly ticket for local transport in Bulgaria and Poland, it was found that its price in Bulgaria showed an increase at the beginning of the analysed time period, reaching a peak between November and December 2022 at around EUR 55 (Figure UC-4-PL-6). However, there was a drastic drop in the transport price to around EUR 25 in early 2023. In the following months, the price level of the monthly ticket remained stable, with minor fluctuations, until October 2024, when it fell again, this time slightly (to EUR 22). The purchase price of a monthly ticket in Poland during the period under review, unlike

in Bulgaria, was characterised by greater stability overall. At the end of 2022, it was around EUR 27, with only slight fluctuations in the following months, remaining mostly between EUR 25 and EUR 27. In the second half of 2024, there was a significant price increase. In October, the cost of buying a monthly ticket for local transport in Poland was EUR 40. The overall trend shows that Bulgaria experienced more volatility in transportation prices early in the period, with a sharp spike and subsequent decline, while Poland maintained more stability, with a noticeable rise in prices only in the latter part of 2024.

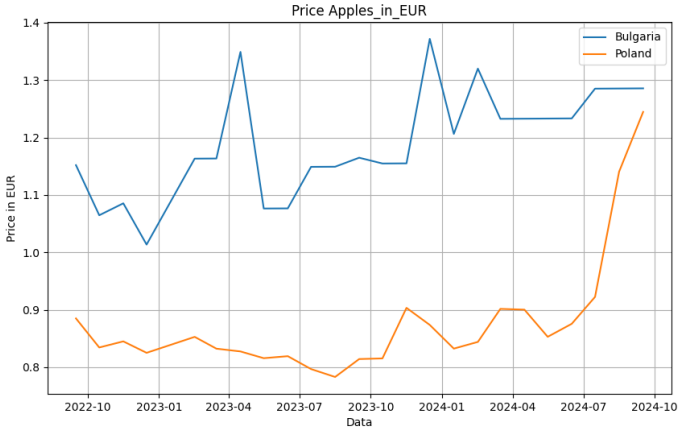

Figure UC-4-PL-7. Price of 1 kg (2 lb.) of apples in Bulgaria and Poland from 2022 to 2024

The price level of 1 kg of apples in Poland and Bulgaria shows a large variation. In Bulgaria, their price was higher throughout the analysed period (Figure UC-4-PL-7). It was also characterised by high variability. The highest price was recorded in mid-2023, when the purchase cost of 1 kg of apples was more than EUR 1.3. Prices in Poland were significantly lower. In the first meis months of the research period, they oscillated between EUR 0.8 and EUR 0.9, followed by a sharp increase (to EUR 1.2) in July 2024.

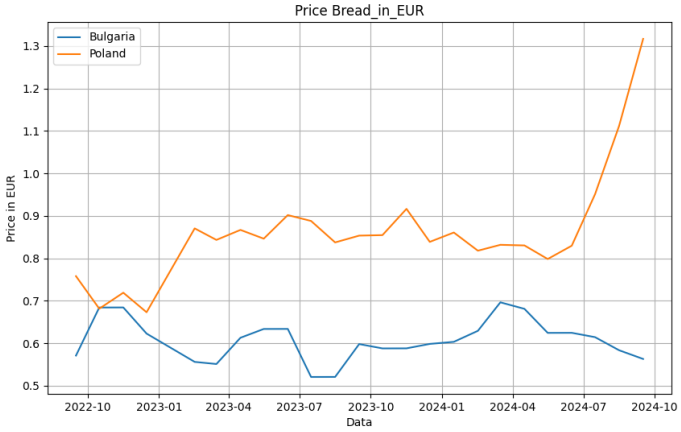

Figure UC-4-PL-8. Price of bread for 2 people for 1 day in Bulgaria and Poland from 2022 to 2024

Poland was characterised by higher bread prices for 2 people throughout the study period (Figure UC-4-PL-8). Prices oscillated around EUR 0.7-0.9 until mid-2024. Then, from mid-2024, they increased sharply reaching EUR 1.3 by the end of the research period. In Bulgaria, bread prices were significantly lower, oscillating between EUR 0.5-0.7.

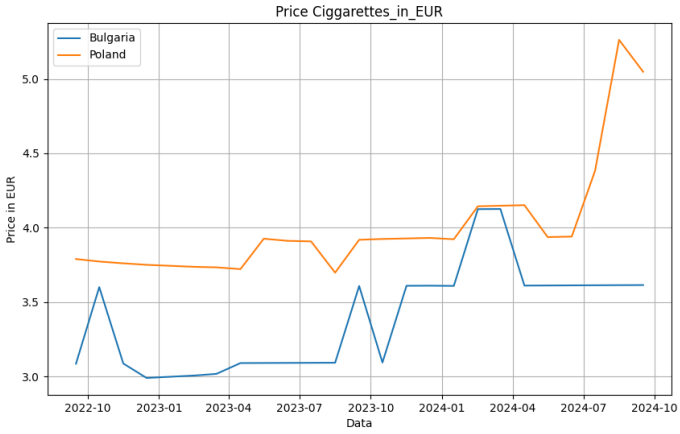

Figure UC-4-PL-9. Price of a pack of cigarettes in Bulgaria and Poland from 2022 to 2024

The price of cigarettes in Bulgaria fluctuated between EUR 3 and EUR 4 throughout the observed period (Figure UC-4-PL-9). After an initial rise and subsequent dip around early 2023, the price stabilized around EUR 3.0, experiencing a few minor fluctuations from mid-2023 to mid-2024. The price remained flat at approximately EUR 3.5 from April 2024 onwards. In Poland the price showed a generally upward trend, starting slightly below EUR 4 in late 2022. After a small decline during early 2023, prices rose moderately over the next several months. A significant increase was observed after April 2024, where prices sharply escalated, reaching a peak above EUR 5 in August 2024.

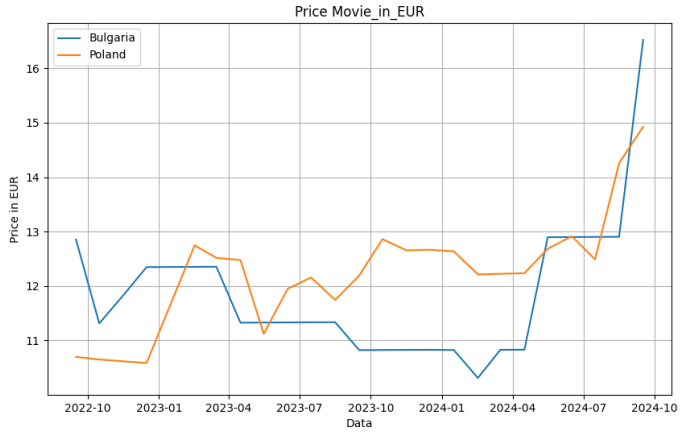

Figure UC-4-PL-10. Price of movie tickets in Bulgaria and Poland from 2022 to 2024

In both countries presented (Figure UC-4-PL-10) the price of movie tickets experienced significant changes throughout the observed period, with a sharp increase in 2024. In Bulgaria the price continuously dropped from EUR 13 in September 2022 to just below EUR 11 in October 2023, at which point it remained steady until April 2024. In May 2024 a significant increase of 2 euro was recorded, followed by an even larger increase in September, where it reached a high of over EUR 16. In Poland prices started slightly below EUR 11, staying steady until December and rising sharply between January and March 2023. Throughout late 2023 and early 2024 the price fluctuated around EUR 12, then surged in the third quarter of 2024 reaching EUR 15.

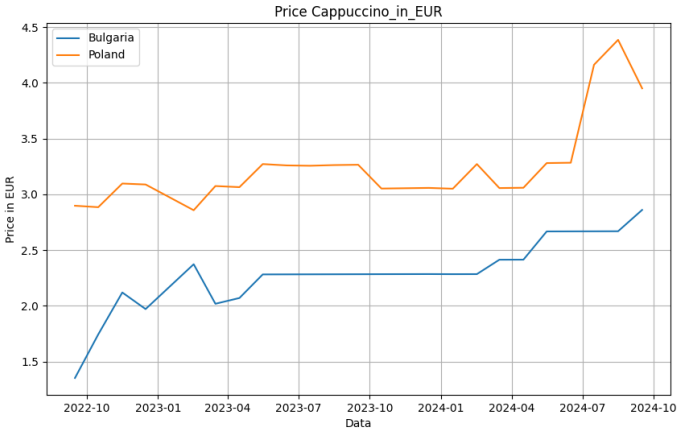

Figure UC-4-PL-11. Price of a cappuccino in expat area of a city in Bulgaria and Poland from 2022 to 2024

The price of a cappuccino in Poland and Bulgaria saw similar trends throughout the reference period (Figure UC-4-PL-11), with an initial increase in price followed by a period of stability, with another increase at the end of the period. In Bulgaria the initial price of EUR 1.5 rose quickly to a level of EUR 2.3 in February 2023. After a brief dip it remained steady until February 2024, from which point forward it increased gradually reaching the high of EUR 28 in September 2024. In Poland the fluctuations in price were more frequent but generally small in scale. The price increased between September 2022 and May 2023, then for the rest of 2023 and early months of 2024 it remained above EUR 3. A surge in price was seen in July 2024 with a peak of almost EUR 4.5 in August, followed by a small decrease in September.

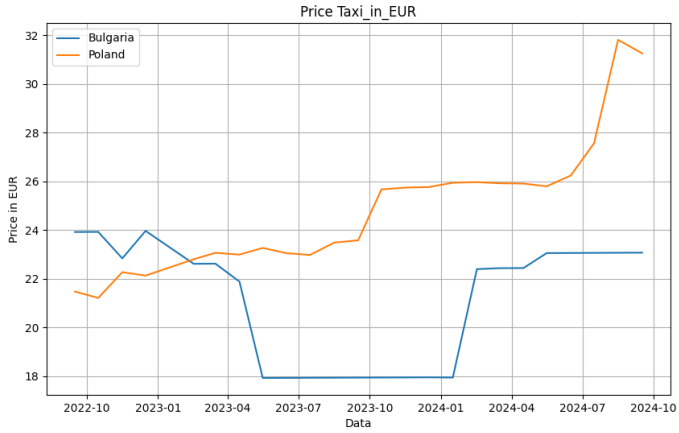

Figure UC-4-PL-12. Price of Taxi trip on a business day, basic tariff, 8 km. (5 miles) in Bulgaria and Poland from 2022 to 2024

During the research period, the price level of taxi travel in Poland, was characterised by a gradual increase (Figure UC-4-PL-12). At the beginning of the research period, the price of travel was found to be less than EUR 22. In May 2024, there was a sharp increase in the price level, which reached a maximum around September 2024. At that time, it was noted that a taxi fare was around EUR 32. In Bulgaria, taxi fares were significantly lower. At the end of 2022, the fare was around EUR 24. It declined in the following months to reach EUR 18 per fare in May/June 2023. From January to March, an increase in taxi fares was recorded. A stablisation of their level was observed between February and October 2024 when prices oscillated between EUR 22 and EUR 24 per course.

3 Trips and trip metrics

During the progression of the project work web scraping was performed on Trip.com, where information regarding passenger flights from Warsaw and Sofia to the following cities was collected: Athens, London, Madrid, Paris, Rome, Warsaw (or Sofia), and Vienna. Flight data was gathered every 6 days. For each flight, data were collected one month prior to the scheduled departure, and this process was repeated again one week before the departure date. This approach was chosen due to the repetitive nature of flight schedules and in adherence to best practices in web scraping.

The web scraping process extracted information on 15,997 flights. Considering the exact aircraft model, it was found that the Airbus A319 was the most frequent on the indicated destinations (15.5% of the total flights), while slightly less frequently used than the Embraer 195 (13.8%) and the Airbus A320 (12.9%). For 1284 (8.6%) flights, the aircraft model could not be determined. Details of the number of flights are shown in table UC-4-PL-2.

Table UC-4-PL-2. Number of flights by aircraft model

Aircraft	Number of flights	% of flights
Airbus A318	62	0.39
Airbus A319	2 822	15.52
Airbus A320	1 922	12.93
Airbus A320-212	1 633	11.21
Airbus A321	1 307	9.48
Airbus A350 XWB	96	0.43
Boeing 737	101	0.43
Boeing 737-700	104	0.43
Boeing 737-800	277	1.72
Boeing 737MAX8	213	1.29
Bombardier Regional Jet 1000	160	1.29
Bombardier Regional Jet 900	247	2.16
de Havilland DHC-8-400 Dash 8/8Q	871	5.17
Embraer 170	299	2.16
Embraer 175	863	6.03
Embraer 190	872	6.90
Embraer 195	2 864	13.79
Not determined	1 284	8.63

Of the total flights taken, only 7.4% were direct flights. Flights with 1 connecting flight accounted for 73.4% of the total, while 19.2% were flights with two connecting flights (Table UC-4-PL-3).

Table UC-4-PL-3. Number of direct flights and connecting flights

Stops	Number of flights
0	1 184
1	11 742
2	3 071

The majority of ticket (economy class, 1 person, one way ticket) prices to the indicated destinations were higher one week before departure than one month before. Considering the average ticket price, it turned out that only for departures from Warsaw to Madrid and Sofia was it higher one month before departure (UC-4-PL-4).

Table UC-4-PL-4. Price comparison from web scraping one month before departure and one week before departure

Destination	Avg. price month before departure	Avg. price week before departure
Athens	1 178.32	1 278.58
London	1 001.67	1 375.05
Madrid	1 466.67	1 588.36
Paris	1 214.73	1 299.96
Rome	1 728.22	1 993.55
Sofia	1 186.13	1 088.66
Vienna	1 156.43	1 162.00

In a collaborative effort between Poland and Bulgaria, a framework has been developed to capture key variables pertaining to various sectors within the tourism industry. New indicators will improve the quality of research related to both domestic and international travel:

- Average price per flight to selected countries per quarter/month (flight booking portals)

$$average\ price\ _{country,period} = \frac{\sum_{i=1}^{n_{country}} price_i}{n_{country}}$$

where $n_{country}$ represents the total number of flights during specified period to the selected country and i represents the price of each individual flight to the selected country during the specified period.

The average price per flight to selected countries per quarter/month is a crucial metric that provides insights into the affordability and demand for air travel to specific destinations within a given period. By calculating the average price of flights to selected countries, stakeholders can understand fluctuations in airfare prices, identify trends in travel demand, and make informed decisions regarding travel planning and marketing strategies.

- Number of flight offers to individual countries (flight booking portals)

$$Number\ of\ flight\ offers_{country} = \sum_{i=1}^{n} offers_i$$

where n represents the number of flight offers data points available and i represents the number of flight offers for each data point.

The number of flight offers to individual countries is a metric that quantifies the availability and accessibility of air travel options to specific destinations. It provides insights into the demand for air travel, route network coverage, and the competitiveness of airlines in serving various markets. By analyzing the number of flight offers, stakeholders can identify emerging travel trends, assess market demand, and make informed decisions regarding route planning, marketing strategies, and pricing.

- Average cost of a trip booked through a Travel Agent (travel agencies portals)
-

$$average\ cost\ of\ trip = \frac{\sum_{i=1}^{n} cost\ of\ trip\ _i}{n}$$

where n the total number of trips booked through a travel agent and i represents cost of each individual trip booked through a travel agent.

The average cost of a trip booked through a travel agent to a particular country is a metric that quantifies the overall expenditure incurred by travelers when arranging their trips through travel agencies. It encompasses various components such as transportation, accommodation, meals, activities, and additional services arranged by the travel agent. By analyzing the cost of trips, stakeholders can assess travel affordability, compare destination competitiveness, and understand traveler spending patterns.

- Average Price per category and country (cost of living portals)

$$average\ price\ _{category,country} = \frac{\sum_{i=1}^{n} price\ _i}{n}$$

where n represents the total number of data points analyzed for a specific category and country.

This metric encompasses various aspects of the cost of living within different countries, including average prices of local transport, fuel, cigarettes, alcohol, and meals. By analyzing these cost components, insights into the overall cost of living and affordability in different destinations can be obtained.

- Cost of Living Index per Quarter/Month (cost of living portals)

$$cost\ of\ living\ index = \left(\frac{\sum_{i=1}^{n} P_i * Q_i}{\sum_{i=1}^{n} P_{0i} * Q_i}\right) \times 100$$

where P_i is the price of the basket of goods and services in the current period or location, P_{0i} is the price of the basket of goods and services in the base period or location and Q_i is the quantity of each item in the basket of goods and services.

The Cost of Living Index per quarter/month is a pivotal metric that quantifies the relative cost of living in different countries or regions during specific periods. It encompasses various factors such as housing, transportation, food, healthcare, and other essential expenses. By analyzing the Cost of Living Index, stakeholders can compare living expenses between different destinations, assess purchasing power parity, and understand the economic dynamics of various regions.

4 Survey data

The primary source of information on the participation of Polish residents in travel is the household survey: 'Participation of Polish residents in travel' (SAMPLE SURVEY).

The aim of the survey is to determine the scale of participation of Polish residents in travel, the characteristics of domestic trips with at least one overnight stay and trips abroad both with and without an overnight stay. Expenditure related to trips is also examined. The survey on participation of Poles (residents) in travel is a sample survey conducted in households, on a quarterly basis (in the month following the quarter). The unit of observation in the survey is members of single or multi-person households in drawn dwellings. Information obtained in the survey concerns journeys of Polish residents - domestic with accommodation and foreign with and without accommodation and included:

- characteristics of the trip (e.g. start and end date of the trip, main aim of the trip, places visited, number of nights, type of accommodation, how the trip was organised, how services were booked),
- trip-related expenditure (on accommodation, meals, transport, shopping, cultural and leisure services, etc.).

Travel-related expenditure includes amounts for the purchase of services and consumer goods (including durable consumer goods and high-value items) incurred before and during the trip in cash and non-cash form (e.g. payment card, bank transfer) directly by the travellers (household members) as well as financed or reimbursed by the workplace or other persons or institutions. All travel-related expenditure is covered, even if the services were booked and paid for before the trip or if the actual payment was made after the trip. In contrast, they do not include expenditure on the purchase of goods for resale.

The analysis draws on 31,263 survey responses collected throughout the study period, from September 2022 to September 2024. Of these responses, 86% pertain to domestic trips within Poland, while the remaining responses concern travel to Bulgaria.

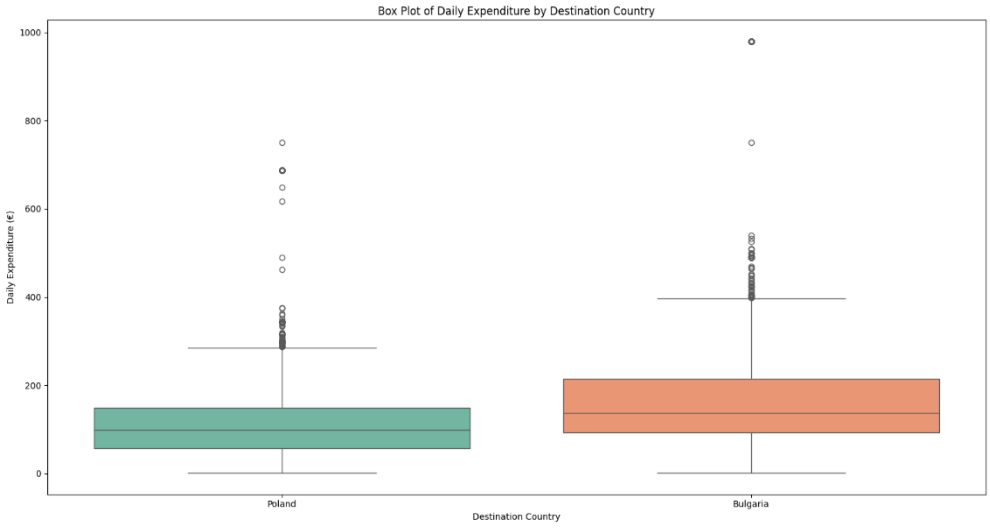

Figure UC-4-PL-13 Daily expenditure in EUR of Polish tourists on travel in Poland and Bulgaria in 2022-2024

The box plot (Figure UC-4-PL-13) shows the daily spending of Polish tourists (EUR) by destination country. On the horizontal axis (X) there are two destination countries Poland and Bulgaria The vertical axis (Y) shows

the daily expenditure in euros, with a range from 10 to 1,000 euros. Based on the results obtained, the median daily travel expenditure of Polish tourists was found to be slightly higher than in Bulgaria (about EUR 200, for Poland about EUR 150). The scatter of the data (values between the first and third quartile) is larger for Bulgaria than for Poland, implying greater variability in expenditure The average travel expenditure in Poland is EUR 237, while in Bulgaria it is EUR 278. The whiskers of the graph suggest maximum values, with relatively similar values in both cases. Outliers above EUR 400 per day are noted in both countries. In Poland, they are more concentrated in the EUR 400 to EUR 800 range, while in Bulgaria some values reach above the EUR 800 mark.

5 Methodology

Accurate and timely statistics are crucial for understanding the dynamics of tourism demand, travel behavior, and associated expenditures. However, challenges arise when relying on sample surveys that may be outdated or lack consistency across different regions. A significant issue is that travel records from sample surveys can be several years old, leading to potentially misleading statistics in a rapidly changing tourism market. Moreover, different countries or regions may have varying depths of data collection, resulting in inconsistencies in data quality and gaps in information. These disparities pose challenges to creating a standardized view of tourism demand and expenditure.

To address the issue of outdated survey data, the use of web scraping techniques has been incorporated into the data collection process. Web scraping provides access to the most current data on airfare prices, and living expenses. By integrating this real-time data, the imputation process becomes more dynamic and reflective of the current market environment. This approach significantly reduces the impact of outdated information, providing a more accurate depiction of tourism demand and expenditure patterns.

To further enhance data accuracy, the imputation process involves identifying "Golden Records" within the available survey datasets. These records are deemed the most reliable and are used as benchmarks for data imputation. By selecting Golden Records, we ensure that only high-quality, accurate information forms the basis for filling gaps or updating outdated data. This careful selection minimizes errors and aligns the imputation with real-world dynamics.

Once the Golden Records are identified, they serve as the foundation for updating or completing missing or outdated data in the dataset. This step is crucial for providing a comprehensive and reliable view of tourism demand and expenditure patterns. A methodical approach to data filling ensures that data gaps do not lead to misleading conclusions. This comprehensive dataset allows for more robust analysis of trends, such as travel frequency, destination choice, and expenditure levels, reflecting the latest market conditions.

Ensuring that the updated data remains consistent and accurate involves a thorough validation and correction process. This step includes additional checks and adjustments to the imputed data, ensuring it aligns with both historical trends and current market dynamics. Continuous validation is essential to maintain data integrity, making the statistics more reliable for policy formulation, economic forecasting, and strategic planning by tourism-related stakeholders.

The imputation process is continuously refined to incorporate the most up-to-date information available, mitigating issues related to outdated survey data and inconsistencies. By combining Golden Records, web scraping and validation techniques the imputation strategy becomes a new tool for generating reliable tourism demand and expenditure statistics.

6 Results

The data obtained by web scraping on the prices of passenger flights and individual goods and services allowed the validation and imputation of micro-data (using the developed indicators) in the sample survey on the expenditure of Polish tourists in Bulgaria and Poland on international air transport and the purchase of food commodities in connection with travel in the period from September 2022 to September 2024.

The analysis adheres to a strict methodology that includes a verification step using "Golden Records." These records serve as high-quality data entries that best represent the travel expenditures and are instrumental in validating the survey results. Based on this methodology, the distribution of Golden Records across destination countries is as follows:

- Poland: 10,179 Golden Records
- Bulgaria: 1,860 Golden Records

This approach ensures the reliability of findings and provides an accurate reflection of travel expenditure patterns in both Poland and Bulgaria.

The results of the experimental work carried out showed that after the implementation of web scraping data in the above-mentioned study, there was an increase in the number of tourists' expenditures in particular research periods due to changes in travel related expenses. In the case of the travels of Polish tourists to Bulgaria, the lowest price increase translating into the value of goods and services purchased was recorded in September 2022 when it was 3.0%. The highest increase in prices affecting the results of the SAMPLE SURVEY survey was shown in May and September 2023 (in both cases by approximately 9.2%) (Figure UC-4-PL-14).

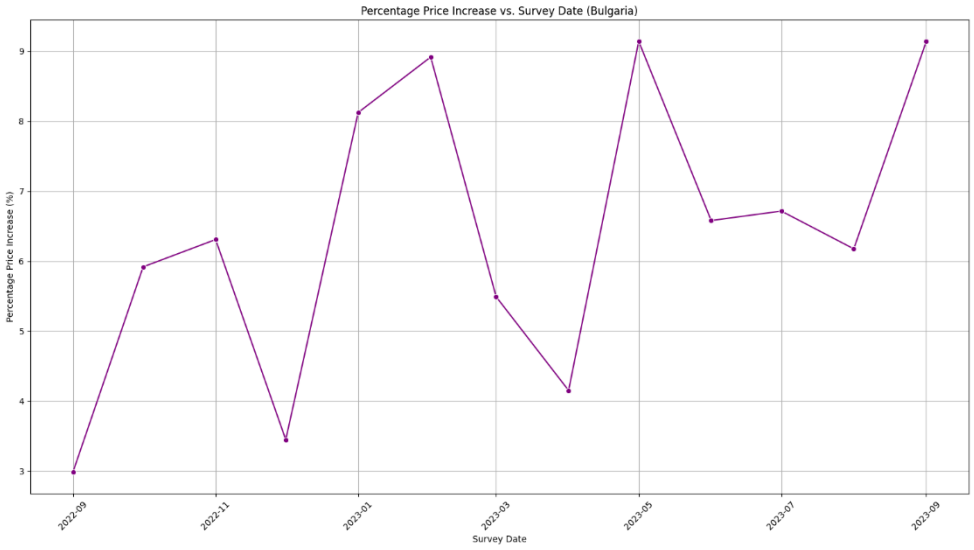

Figure UC-4-PL-14. Monthly increase of domestic travel expenditures based on web scraping data (Bulgaria) sample survey on the expenditure of Bulgarian tourists

The figure illustrates monthly estimations of domestic travel expenditures in Poland, derived from data collected via web scraping, covering the period from September 2022 to September 2023. Initially, in September 2022, the estimated expenditures show a significant increase of approximately 9%. This is followed by a sharp decline in November 2022, suggesting a potential drop in demand or cost reductions

in the domestic travel market. Throughout the beginning of 2023, fluctuations in the data are observed, with smaller peaks in early 2023 indicating periodic increases in expenditure. A notable rise occurs in May 2023, where the estimated expenditures reach their highest point, nearly 13%. This peak might be attributed to seasonal factors, such as increased travel during the spring and early summer months, combined with price hikes due to heightened demand for travel-related services. After this peak, the data indicates a gradual decline in travel expenditures through the summer months, culminating in a sharp drop by September 2023.

The use of web scraping techniques ensures a large volume of data, contributing to the accuracy and reliability of these expenditure estimates.

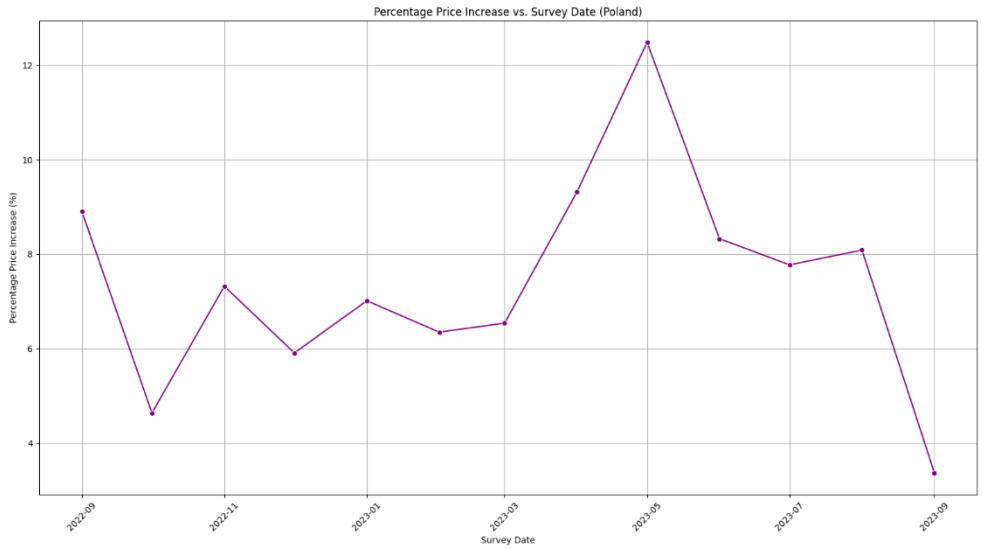

Figure UC-4-PL-15. Monthly increase of domestic travel expenditures based on web scraping data (Poland) and sample survey on the expenditure of Polish tourists

When comparing the results of the experimental study, in which the developed indicators were used, with the results of the sample survey study in relation to the whole year, an increase in the spending of Polish tourists on travel to both Bulgaria and Poland was found. On trips to Bulgaria in the sample survey, Polish tourists spent an average of EUR 921.06, while after applying the developed indicators an average of EUR 979.54 was obtained. The Percentage Change in Average Expenditure was 6.35%.

In the sample survey study, Polish tourists spent an average of EUR 640.80 on travel in Poland. Based on the results of the experimental work carried out, it was found that the average expenditure of travellers increased to EUR 687.71. This means that a percentage change in average expenditure of 7.32% was recorded.

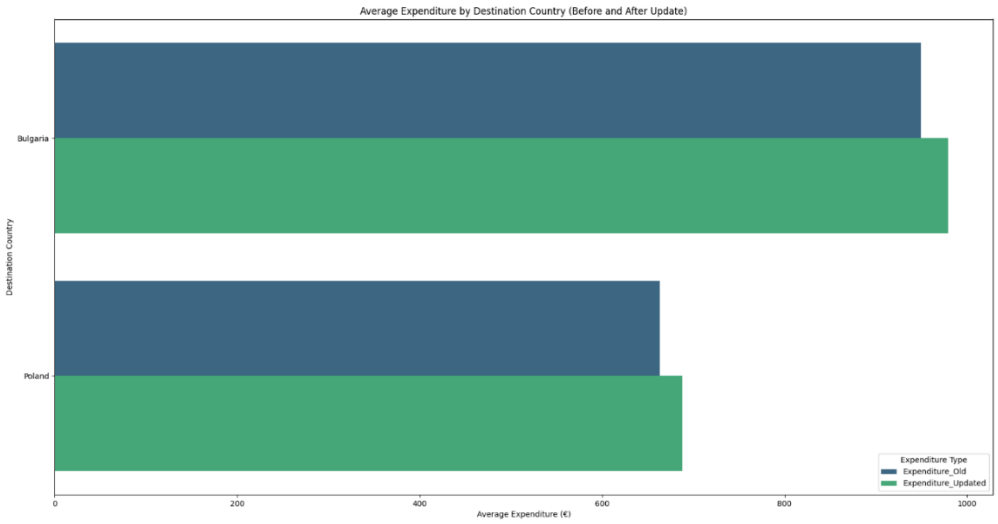

Figure UC-4-PL-16. Average spending of Polish tourists on domestic travel and travel to Bulgaria

7 Conclusions

The research demonstrates potential in utilizing web scraping techniques to estimate travel-related expenditures. While the methodology is still under refinement, the initial findings are promising. Key conclusions from the study include:

- Using the web scraping method, data can be obtained from online platforms concerning prices for products, services, and air transportation within a specified timeframe, providing a rich source of real-time pricing information.
- Frequent changes in online platform structures necessitate continuous monitoring and regular adjustments to the web scraping codes, requiring significant time and developer resources to maintain the functionality of data extraction processes.
- The obtained data on flight prices and other travel-related goods and services can be integrated into sample surveys on travel and tourism expenditures in European Union countries. This data can be utilized for validating individual data entries (e.g., price comparisons) and imputing missing data due to non-responses from survey participants.
- The use of price indicators developed from web scraping data has led to an adjustment in the reported expenditure levels in the survey on “Participation of residents of Poland in trips.” After updating outdated prices, the average reported expenditures for travel to Bulgaria increased by approximately 6.35%, while the average expenditures for travel within Poland rose by approximately 7.32%.

While the methodological work is still ongoing, these early results indicate the value web scraping offers in enhancing the accuracy of travel expenditure survey.