Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(12), 25–48 https://doi.org/10.59139/ws.2023.12.2
Statystyka w praktyce / Statistics in practice
Zgłoszony/submitted: 18.07.2023, poprawiony/revised: 3.11.2023, zaakceptowany/accepted: 20.11.2023

The evaluation of (big) data integration methods in tourism[1]

Marek Cierpiał-Wolan,[a] Galya Stateva[b]

Abstract. In view of many dynamic changes taking place in the modern world due to the COVID-19 pandemic, the migration crisis, armed conflicts, and other, it is a major challenge for official statistics to provide high-quality information, which should be available almost in real time. In this context, the integration of data from multiple sources, in particular big data, is a prerequisite. The main aim of the study discussed in the article is to characterise and evaluate the following selected methods of data integration in tourism statistics: Natural Language Processing (NLP), machine learning algorithm, i.e. *K*-Nearest Neighbours (*K*-NN) using TF-IDF and *N*-gram techniques, and Fuzzy Matching, belonging to the group of probabilistic methods.

In tourism surveys, data acquired using web scraping deserve special attention. For this reason, the analysed methods were used to combine data from booking portals (Booking.com, Hotels.com and Airbnb.com) with a tourism survey frame. The study is based on data regarding Poland and Bulgaria, downloaded between April and July 2023. An attempt was also made to answer the question of how the data obtained from web scraping of tourism portals improved the quality of the frame.

The study showed that Fuzzy Matching based on the Levenshtein algorithm combined with Vincenty's formula was the most effective among all the tested methods. In addition, as a result of data integration, it was possible to significantly improve the quality of the tourism survey frame in 2023 (an increase was observed in the number of new accommodation establishments in Poland by 1.1% and in Bulgaria by 1.4%).

Keywords: data integration methods, tourism survey frame, web scraping
JEL: C1, C81, Z32

Ocena metod integracji danych dotyczących turystyki z uwzględnieniem big data

Streszczenie. W obliczu wielu dynamicznych zmian zachodzących we współczesnym świecie, spowodowanych m.in. pandemią COVID-19, kryzysem migracyjnym i konfliktami zbrojnymi,

[1] Artykuł został opracowany na podstawie referatu wygłoszonego na konferencji *Metodologia Badań Statystycznych MET2023*, która odbyła się w dniach 3–5 lipca 2023 r. w Warszawie. / The article is based on a paper delivered at the *MET2023 Conference on Methodology of Statistical Research*, held on 3rd–5th July 2023 in Warsaw.

[a] Uniwersytet Rzeszowski, Kolegium Nauk Społecznych, Instytut Ekonomii i Finansów; Urząd Statystyczny w Rzeszowie, Polska / University of Rzeszów, College of Social Sciences, Institute of Economics and Finance; Statistical Office in Rzeszów, Poland. ORCID: https://orcid.org/0000-0003-2672-3234. Autor korespon-dencyjny / Corresponding author, e-mail: m.cierpial-wolan@stat.gov.pl.
[b] National Statistical Institute, Bulgaria. ORCID: https://orcid.org/0009-0005-0755-6970.
E-mail: gstateva@nsi.bg.

ogromnym wyzwaniem dla statystyki publicznej jest dostarczanie informacji dobrej jakości, które powinny być dostępne niemalże w czasie rzeczywistym. W tym kontekście warunkiem koniecznym jest integracja danych, w szczególności big data, pochodzących z wielu źródeł. Głównym celem badania omawianego w artykule jest charakterystyka i ocena wybranych metod integracji danych w statystyce w dziedzinie turystyki: przetwarzania języka naturalnego (Natural Language Processing – NLP), algorytmu uczenia maszynowego, tj. *K*-najbliższych sąsiadów (*K*-Nearest Neighbours – *K*-NN), z wykorzystaniem technik TF-IDF i *N*-gramów, oraz parowania rozmytego (Fuzzy Matching), należących do grupy metod probabilistycznych.

W badaniach dotyczących turystyki na szczególną uwagę zasługują dane uzyskiwane za pomocą web scrapingu. Z tego powodu analizowane metody wykorzystano do łączenia danych pochodzących z portali rezerwacyjnych (Booking.com, Hotels.com i Airbnb.com) z operatem do badań turystyki. Posłużono się danymi dotyczącymi Polski i Bułgarii, pobranymi w okresie od kwietnia do lipca 2023 r. Podjęto także próbę odpowiedzi na pytanie, jak dane uzyskane z web scrapingu wpłynęły na poprawę jakości operatu.

Z przeprowadzonego badania wynika, że najbardziej przydatne spośród testowanych metod jest parowanie rozmyte oparte na algorytmach Levenshteina i Vincenty'ego. Ponadto w wyniku integracji danych udało się znacząco poprawić jakość operatu do badań turystyki w 2023 r. (wzrost liczby nowych obiektów w Polsce o 1,1%, a w Bułgarii – o 1,4%).

Słowa kluczowe: metody integracji danych, operat do badań turystyki, web scraping

1. Introduction

The modern world is determined by many threats, both global and local. World economies are facing increasing problems such as instability caused by numerous armed conflicts, energy crises, and the unprecedented scale of global migration. Additionally, since the beginning of 2020, the world has been struggling with the COVID-19 pandemic, which continues to affect many areas of our lives. All these circumstances are causing various effects of a socio-economic nature, which impact several economy sectors, and the tourism industry in particular. This leads to the emergence of huge demand for tourism-related data.

To provide high-quality, real-time information, it is necessary to integrate data from various sources, i.e. administrative registers, databases using information from censuses and sample surveys, and, most importantly, big data. Developing innovative methods of data integration has therefore become an imperative for academia and official statistics.

Data sources used so far by official statistics to create frames for tourism surveys have proven insufficient. This is due to both the specifics of the tourism market (short-term and sometimes incidental activity) and the fact that some part of the tourism-related activity is hidden in the shadow economy. In this context, big data is becoming an indispensable source of data.

The main aim of the study discussed in the article is to characterise and evaluate the following selected methods of data integration in tourism statistics: Natural Language Processing (NLP), machine learning algorithm, i.e. *K*-Nearest Neighbours

(*K*-NN) using TF-IDF and *N*-gram techniques, and Fuzzy Matching, belonging to probabilistic methods. In addition, we evaluated the quality of the tourism survey frame by obtaining data from web scraping of tourism portals. Selected methods were used to combine data from booking portals (Booking.com, Hotels.com and Airbnb.com) with the tourism survey frame in Poland and Bulgaria in 2023.

2. Big data in tourism statistics

Nowadays, the term 'big data' is used to describe a way of acquiring knowledge and learning about reality that is possible thanks to new technologies creating and processing large data sets. New data sources combined with innovative methods of processing them (especially machine learning, etc.) provide, in many cases, the possibility of publishing information on socio-economic phenomena and processes in a real-time mode, as well as better quality forecasts.

There are many classifications of big data. We assume that these are data from the following sources:

1. social interactions, especially social networks;
2. data-processing systems directly or indirectly related to business performance;
3. systems of electronic devices that automatically exchange data without human intervention – Internet of Things (United Nations Department of Economic and Social Affairs Statistics Division, 2015).

Another classification defines big data as information derived from sensors and any records of activity from electronic devices, social networks, business transactions, digital files (web pages, audio recordings, videos, PDF files, etc.) and real-time transmissions.[2]

As regards tourism surveys, useful information is that obtained by means of web scraping, as well as data from mobile network operators, traffic sensors or payment card operators.

Using web scraping, a technique for automatic extraction of data from websites, one can obtain valuable information on tourism phenomena and processes. In this context, online accommodation booking portals are particularly useful, and it is very important to select appropriate platforms from which data will be drawn. Hence, it is necessary to get to know the domestic market of online booking and identify the information resources of these portals. Both platforms with international coverage (e.g. Booking.com or Hotels.com) and local ones (e.g. Pochivka.bg for Bulgaria and Nocowanie.pl for Poland) should be taken into consideration. Data from traffic sensors, the Automatic Number Plate Recognition System (ANPRS), mobile

[2] Read more at: United Nations Economic Commission for Europe (UNECE), n.d.

network operators and payment cards are also very useful in monitoring tourist traffic.

It is worth noting that in European Union countries, traffic sensor data have been used for many years. Combined with the data from the ANPRS or the smart city more generally, they have been playing an increasingly important role in tourism statistics. Data held by mobile network operators, on the other hand, are an important source of information on the population movements. However, it is important to point out that this type of data should be subjected to detailed processing, especially in order to separate information on traffic related to daily activities from data relevant to tourism statistics. Still, gaining access to mobile network operators' data is very complicated, both legally and technologically (e.g. in the case of border areas, devices repeatedly log in outside the home network). As regards travel expenses of residents and foreigners, payment card data is the key source of information, because it takes into account spatial aspects. Companies operating payment terminals and ATMs have data on transactions made in individual countries. However, it has to be remembered that in countries with large migration, payment cards may be held by foreigners, which makes it difficult to accurately estimate the scale of spending. Therefore, the use of data obtained through payment cards should be accompanied by the analysis of additional sources of information to properly determine the status of the cardholder. A detailed breakdown of expenses (e.g. for lodging, food, transportation, goods) can be performed using the MCC (Merchant Category Code) classification.

In general, big data can be used in both census and sample surveys in various ways, e.g. only for data validation, as complementary sources, or to replace existing surveys entirely. When using these sources, especially in cyclical surveys, a pre-requisite is to secure the continuity of access to data, both formally and regularly. This could be done by diversifying data sources, as it is always possible to loose access to one or another source. In this context, it is important to constantly develop methods of data imputation and calibration. Data integration systems must be resistant to the loss of access to different sources of information. In practice, for example, simulators that integrate different sources of information can be an alternative to mobile data.[3] For traffic data at the EU's internal borders, an alternative to information from traffic sensors or the ANPRS is to use other sources such as smart city systems in nearby cities, parking meters, drones or satellite imagery.

[3] An interesting proposal for such a simulator was created in the framework of *ESSnet Big Data II* project (Oancea et al., 2019).

Another interesting way to provide constant access to big data is to develop mobile applications that would provide information and services related to users' needs, which, in return, collect diagnostic information from mobile devices (e.g. a travel planning application providing information on transportation, lodging, meal planning, preferred kinds of entertainment, etc.).

An obvious prerequisite for the use of big data is the positive assessment of their quality. In 2009, Daas provided some guidelines for the evaluation of the quality of various data sources, pointing out that each 'dimension' of quality can be defined by several indicators (Daas et al., 2009, pp. 6–9). Maślankowski (2015, pp. 173–174) asserted that the following 'dimensions' of the quality of big data were crucial: unambiguity, objectivity (mapping and reduction errors), inclusion of a timestamp, granularity or degree of detail, presence of duplicate data, completeness, accessibility, precision, interpretability, integrity, and consistency.

3. Selected methods of combining data

The process of combining datasets can be implemented by means of many methods. In the literature there are at least a dozen different methods, not to mention their variants and modifications. However, all of them belong to one of the following four groups of methods (Asher et al., 2020):

- Deterministic Record Linkage;
- Probabilistic Record Linkage;
- Data Fusion;
- Statistical Matching.

In the Deterministic Record Linkage and Probabilistic Record Linkage methods, the combined sets must be large enough for the probability of an object from one set beign also in the other set to be large as well.

In deterministic methods, the linkage process is based on simple rules of the logical exact matching of variables – keys. Any deficiencies in the data increase the risk of error of type I and type II, i.e. false matches and missing matches of records that should be linked. In probabilistic methods, the estimation of the probabilities of a random match between two values of a given variable, assuming that the paired records do not belong to the same unit, and the probabilities of a random mismatch between the values of a given variable, assuming that the paired records belong to the same unit, are incorporated into the process of combining data sets. The Data Fusion and Statistical Matching methods are dedicated to the process of combining relatively small sets, i.e. those for which the chance of an entity occurring in both sets is negligible. In the first case, linkage does not occur, but a concatenation of sets (union of sets) is created, taking into account common variables. Missing data are

imputed, for example, by interpolation. Statistical Matching is procedurally more complex and involves both the micro and macro approaches. In the micro approach, combining sets is done by either concatenation or matching based on similarity. Missing values are then imputed and record weights are calibrated. In the macro approach, a synthetic set is not created. The relevant parameters are estimated taking into account the special role of the covariance matrix of the variables present in both sets.

In this paper, we present the following data linkage and deduplication methods: Natural Language Processing machine learning algorithm, i.e. *K*-Nearest Neighbours using TF-IDF and *N*-gram techniques, and Fuzzy Matching belonging to probabilistic methods.

3.1. Natural Language Processing data linkage method

The NLP method with the Faiss library developed by Facebook AI Research for deduplication candidate generation and rapid comparison involves the following steps: tokenisation, vectorisation, comparison and similarity assessment, and deduplication decision.

The first step in this method is to transform the texts in character variables into tokens and then into semantic vectors using the SentenceTransformer library. This allows the meaning of the text to be expressed in a numerical form, which is necessary for further analysis. To find potentially duplicated accommodation establishments, we generated them on the basis of the input data. For that, we used the Faiss library, which enabled us to efficiently search the vector space to find similar items. With this search structure, we could quickly find similar accommodation establishments. The deduplication process also focuses on evaluating the coincidence between the two strings to determine the degree of their similarity and deciding whether they could be considered duplicates. To do this, we used Euclidean metrics to calculate the degree of difference between the two vectorised text strings.

In the context of the NLP processing, the selection of appropriate libraries plays a key role, especially when talking about differences between national languages. For instance, for English, there is a wide range of libraries and models ready to use, which facilitates research and application work. However, when we move on to inflectional languages, e.g. Polish or Bulgarian, the situation is different. For Polish, libraries such as spaCy offer dedicated models, allowing a more precise processing of the language, taking into account its grammatical and lexical peculiarities. For Bulgarian, on the other hand, the availability of tools and models is so far limited (libraries such as bgNLP and transformers allow basic text processing operations in this language).

3.2. Machine learning algorithm – *K*-Nearest Neighbours using Term Frequency-Inverse Document Frequency and *N*-gram techniques

Data linkage and deduplication procedures can be carried out using machine learning algorithms such as Random Forest, *K*-Nearest Neighbours (*K*-NN), clustering algorithms or neural networks (Quinlan, 1983, pp. 463–482).

The *K*-NN algorithm is a popular algorithm in machine learning that can be used to find similarity between data. The method used for data integration involves the use of Term Frequency-Inverse Document Frequency (TF-IDF) and *N*-gram techniques combined with the *K*-NN algorithm.

TF-IDF is a technique used to assess the validity of terms or tokens (string) in a text in the context of the entire dataset. It works by assigning weights to words based on their frequency in the document (TF) and the inverse frequency in the entire document set (IDF). A high TF-IDF value indicates that the term is valid in the document, allowing unique features of the data to be identified. *N*-grams are sequences of *N* consecutive tokens in a text. For example, bigrams are sequences of two words, and trigrams of three. The use of *N*-grams allows contextual information to be taken into account in text analysis. This is useful in the deduplication process, where the structure and layout of data are important.

Thus, the linkage and deduplication process begins with tokenisation, which consists in dividing the text into strings of characters. The use of the TF-IDF technique allows a more accurate detection of unique sequences of words or phrases in the text, which can be characteristic of duplicate data. The use of TF-IDF also helps reduce errors due to the similarity of single words. The next step is the use of N-gram creation, which makes it possible to take the context into account in text analysis. This is especially important in the deduplication process, as it helps detect similarities between data that differ not only in individual words, but also in their arrangement in sentences or phrases.

The next step involves using the *K*-NN algorithm, which determines similarities between different data. The *K*-NN algorithm operates on a feature space, where features are the TF-IDF values of *N*-grams that were previously calculated. An important part of this algorithm is the calculation of distances, using the appropriate distance metric. The choice of a particular distance metric depends on the type of data and deduplication goals. It is worth noting that *K*-NN is an unsupervised algorithm, which means that it does not require prior classification or data labels. Therefore, we use it as a tool to determine which data are similar to the largest extent. After calculating the distance between data instances, a similarity threshold is defined.

3.3. Fuzzy Matching

The last applied method is Fuzzy Matching along with Vincenty's formula, which is used to calculate the exact geodetic distances between accommodation establishments.

The Fuzzy Matching method allows the comparison of textual data, such as names of accommodation establishments, taking into account spelling errors, typos or differences in format. This makes it possible to find potential matches between records that are not identical, but may represent the same establishments. The method is based on algorithms such as the Levenshtein or Jaro-Winkler distance.

The Levenshtein algorithm is a text-editing algorithm that calculates the minimum number of operations (insertions, deletions or substitutions of characters) necessary to transform one text into another. With this algorithm, we can find potential matches between records that are not identical, but are similar to such a degree that they can represent the same accommodation establishments. The Jaro-Winkler algorithm, on the other hand, is an extension of the algorithm used to calculate the 'Jaro distance', and takes values from the [0, 1] interval, where 1 means the texts are identical, and 0 means there is no similarity between them at all. The Jaro-Winkler algorithm additionally has a prefix scale, which gives a higher similarity score when the strings share a common prefix (Cierpiał-Wolan et al., 2022).

Since the combined data may contain geographical coordinates, we can additionally use Vincenty's formula to calculate the exact distances, which might be important in the process of deduplicating accommodation establishments. Unlike simpler approximations such as the Haversine formula, Vincenty's formula takes into account the fact that the Earth is not perfectly spherical, but has the shape of an ellipsoid.

Finally, we consider two criteria, i.e. the distances between the strings for the selected variables and the geodetic distances. Thus, if the established thresholds for both criteria are met, we link the accommodation establishments.

The above-described methods were used to combine data from web scraping booking portals (Booking.com, Hotels.com and Airbnb.com) with the tourism survey frame. The procedure for combining this type of data is usually a multi-step process. Christen (2012) proposed five stages to it: standardisation, indexing, comparison, combining and the evaluation of results.

It is also worth noting that an interesting solution related to combining and deduplicating data regarding accommodation establishments on scraped portals is the use of algorithms that compare images. This method is based on analysing the visual characteristics of the images that are assigned to each establishment. There are several algorithms that can be useful in this kind of deduplication process, such as

comparing the similarity of colour histograms, comparing visuals using feature descriptors, and digital fingerprints.

4. Data sources

Information on accommodation establishments advertised in Poland and Bulgaria was downloaded between April and July 2023 from three booking portals: Booking.com, Airbnb.com and Hotels.com (a portal equivalent to Expedia). It is worth noting here that the number of scraped variables varied from a portal to portal. The information obtained from booking portals can be used both to supplement the tourism survey frame with new establishments, and to improve the quality of the results of the survey on the number of tourists and nights spent. Particularly important in this context is the information on the rental offers of accommodation establishments belonging to the NACE group 55.2 (Holiday and other short-stay accommodation), which are often difficult to identify primarily due to some part of them operating within the shadow economy.

Among the scraped variables, there are those that are crucial in the process of combining data obtained by web scraping with the tourism survey frame. These are: the name of the establishment, its address including longitude and latitude coordinates, and data on the services offered, to determine the size of the accommodation establishment and its type according to the NACE classification.

It is worth mentioning here that a major problem with linking and further processing data is classification differences regarding the types of accommodation facilities. In international comparisons, we usually use the NACE classification, but it is not used by booking portals. Hence, we need to make the data comparable – therefore, when assigning types of establishments to this classification, double verification was used. For this reason, we both used the classification of establishments declared on booking portals and we adopted a machine learning method (a classification tree). The learning dataset consisted of accommodation establishments from the booking portals linked with those found in the tourism survey frame. The application of this data processing procedure resulted in a relatively high accuracy of the units' assignment to accommodation-related NACE classification groups.

Web scraping was performed on the basis of simulated user interaction with the site using screen scraping. The adoption of this solution enabled full interaction with selected web pages, based on a dynamic modification of the Document Object Model (DOM) tree, cascading style sheet (CSS) components and JavaScript. The used web scraping system was prepared in the Python programming language with good practices to minimise the burden on the scraped portals.

4.1. Data scraped from Booking.com

As a result of web scraping from Booking.com, information on 82,965 booking ads, which in practice means 8,884 accommodation establishments in Poland, was obtained. In Bulgaria, the procedure yielded 3,873 establishments. 33 variables were scraped from Booking.com: the web address on Booking.com, the shortened web address on the portal, the name of the establishment, its address (street, number, postal code, city), the accommodation type, room name, maximum number of guests, rental price, facility area, the quality rating, number of nights spent, number of adults, number of children, number of rooms, number of double beds, number of single beds, number of couches, number of views, ability to communicate in English, availability of a car park, restaurant, bar, availability of onsite breakfast, availability of the Internet, TV, facility service, air conditioning, laundry, spa, fitness facilities, pool, and the availablity of facilities for the disabled.

It is worth mentioning that it was not always possible to obtain a full set of variables (e.g. about 67% of owners did not specify the area of the offered facility). This is very important in the process of classifying establishments in accordance with the NACE. It should also be noted that the classification of accommodation establishments adopted by Booking.com, as well as by other booking platforms, often differs from the classification used in official statistics (e.g. guesthouses are classified as hotel or agritourism accommodation). Therefore, a preliminary classification is usually made on the basis of the characteristics of the establishment, thanks to which we can unambiguously determine its type (Cierpiał-Wolan & WPJ Team, 2020). An important role in the process of classifying an accommodation establishment into a specific type is played by information on the rental price or the services offered, such as bed-making, serving breakfast, or the available catering facilities. The aforementioned amenities are characteristic for hotels and similar establishments classified as NACE 55.1. The pre-classification process is particularly important for new accommodation establishments which are not included in the tourism survey frame.

4.2. Data scraped from Hotels.com

Data were downloaded from Hotels.com on 4,620 accommodation establishments in Poland and 532 in Bulgaria. Twelve variables were extracted, namely: the web address on Hotels.com, the shortened web address on the portal, the name of the establishment, its address, type, number of rooms, the availability of a car park, restaurant, bar, the availability of onsite breakfast, and the availability of room and laundry services.

In contrast to the data obtained from Booking.com, only 11 accommodation establishments (0.2%) did not match any classification type. It is important to note

that the Hotels.com portal almost always provides information on the number of rooms in the accommodation establishment, which makes it possible to determine its size (e.g. Poland divides accommodation establishments into those with up to 9 or 10 and more beds). Out of the total number of the yielded accommodation establishments, 60% were hotels, and apartments accounted for 24%.

4.3. Data scraped from Airbnb.com

Using the web scraping method on the Airbnb.com portal, it was possible to obtain data on 12,556 accommodation establishments in Poland and 4,174 in Bulgaria. Eight variables were scraped from the Airbnb.com portal, i.e. the web address on Airbnb.com, the shortened web address on the portal, the name of the establishment, its address and type, the maximum number of guests, the rental price, and the number of beds.

A distinctive feature of this portal is that among the listed variables, the address of the accommodation establishment is not directly available. The address data on the site is limited to descriptive information like 'a cottage near the beach' or 'a sunny apartment at the foot of the mountains'. The great majority of facilities listed on the portal were those with fewer than 10 beds (97.2%).

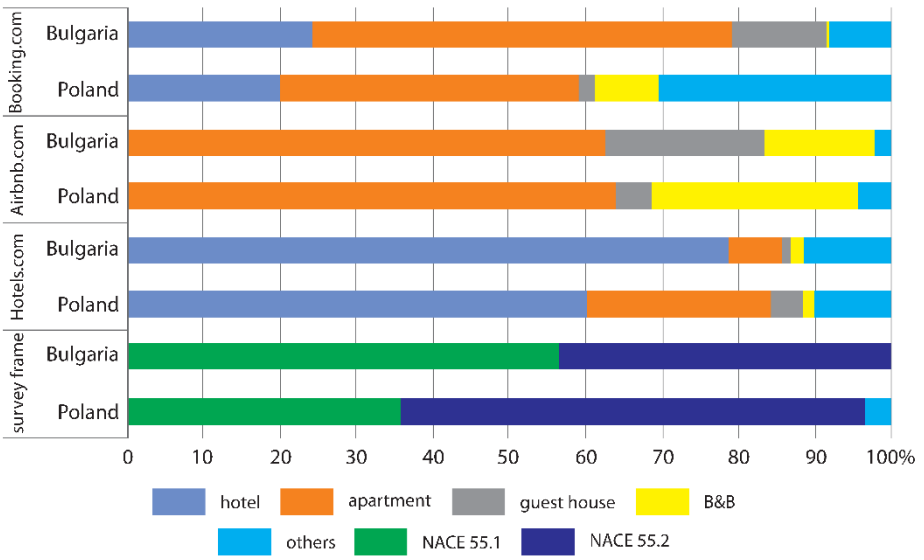
Compared to Booking.com, the vast majority of establishments on Airbnb.com (99.9%) were assigned a type. However, this classification contains generic names of the establishments that are not the same as those in the classification used in official statistics. Therefore, without additional information such as the area of the facility, the number of rooms or the available amenities (daily bed making, room cleaning and washing of sanitary facilities), it is not possible to classify an establishment into a particular type, which is labeled on the portal as e.g. a villa a loft or a condominium.

Figure 1 shows the structure of accommodation establishments on selected booking portals in Poland and Bulgaria by their generic name, as well as the classification of accommodation establishments in the tourism survey frame by the NACE classification in Poland and Bulgaria.

The structure of accommodation establishments scraped from Booking.com in Poland differs from such structure in Bulgaria. In the latter country, the vast majority of establishments are apartments (more than 50%), while the proportion of facilities classified as 'other' does not exceed 10%. In Poland, apartments also account for a considerable share of establishments (just under 40%), but the share of 'other' facilities is about 30%. This is due to the slightly different types accommodation on offer in these countries. In Poland, there are more establishments classified as other hotel establishments (which include, for example, B&Bs and

Aparthotels), while in Bulgaria, short-stay apartments rented by their owners prevail.

Figure 1. Structure of tourist accommodation in Poland and Bulgaria



Source: authors' work.

4.4. Tourism survey frame

The basis of our research was a tourism survey frame used in Poland and Bulgaria. In Poland it consisted of 13,804 establishments (with 10 and more beds), including 7,588 (55.0%) accommodation establishments classified as NACE 55.2. In Bulgaria, the tourism survey frame consisted of 4,031 establishments, of which 43.0% were classified as NACE 55.2.

From the tourism survey frame the following six variables were used: establishment name, establishment type, street, house number, postal code, city name.

5. Research results

The studies and analyses presented in this chapter were partially initiated in research projects[4] carried out by Statistics Poland and the National Statistical Institute in Bulgaria. Below, we present the results of the analyses for Poland.

[4] ESSnet Big Data II, Work Package WPJ (European Commission, n.d. b), ESSnet WIN, Work Package 3 – Use case 4 (European Commission, n.d. a).

In order to test the selected probabilistic methods of combining data, we used the tourism survey frame and the databases obtained by means of web scraping of the three booking portals mentiond before.

In addition, it was necessary to clean the web-scraped database (from typing errors, white spaces, HTML tags, etc.) in order to transform the unstructured data into a structured form corresponding to the data structure of the tourism survey frame. The cleaning process was performed in the Python programming language using the pandas, re, and the BeautifulSoup libraries. The cleaning process also used the web mapping and navigation service operated by HERE Technologies. Thanks to the aforementioned application, it was possible to carry out both the automatic parsing of address data into a common structure and the correction of language errors. The use of the HERE MAPS tool also made it possible to assign geographic coordinates to each accommodation establishment in the tourism survey frame and to do the same regarding the establishments in the web-scraped database (Cierpiał-Wolan et al., 2023).

All the above-described methods of data linkage and deduplication (NLP, *K*-NN, Fuzzy Matching) that we used generate linkage distances. To assess the usefulness of these methods, it was necessary to determine a distance threshold for the linkage. For this purpose, we used sensitivity (true positive rate) and specificity (true negative rate). In the case of data linkage, sensitivity is the probability of the correct match, while specificity is the probability of the correct non-match. Both terms are closely related to type I and type II errors. Since there is a trade-off between specificity and sensitivity, changing distance threshold always leads to improving one measure and worsening the latter. We evaluated various thresholds with the receiver operating characteristic (ROC) and found the optimal one by means of Youden's J statistic (Youden, 1950). With the optimal threshold, we generated a confusion matrix and derived a set of auxiliary statistics.

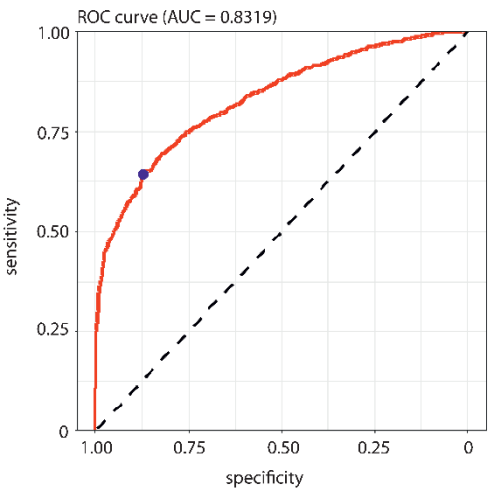
The ROC curve plots points of pairs – specificity and sensitivity – determined for a set of thresholds. When the curve is close to diagonal, this indicates that a given classifier is close to a random classifier. The better the classifier, the closer the curve to the top-left corner. Youden's J statistic is calculated as a sum of specificity and sensitivity minus one (Fawcett, 2006; Peirce, 1884; Powers, 2011).

5.1. Natural Language Processing

As a result of linking the establishments of the tourism survey frame with the establishments from web scraping by means of the NLP method (the main dataset was the tourism survey frame), 8,593 accommodation establishment connections were obtained. Then, based on the ROC curve, we determined the optimal threshold.

We examined a set of 2,314 thresholds ranging from 0.00 to 126.63. Figure 2 presents a ROC curve for the NLP method (solid red line), a ROC curve for the random classifier (dashed black line), and a pair of specificity and sensitivity for the optimal threshold derived from Youden's J statistic (blue point).

Figure 2. The ROC curve for the NLP method



Source: authors' work using R package.

The optimal distance threshold amounted to 15.159. For this threshold, specificity and sensitivity amounted to 0.8730 and 0.6455, respectively, while Youden's J statistic reached 0.5184.

Table 1. Results of the NLP method

NACE	Number of matched establishments	Number of perfect matches	Distance		
			mean	minimum	maximum
55.1	1,440	90	17.8091	0	119.4789
55.2	717	3	24.2153	0	126.6318
55.3	49	0	33.8532	13.1397	85.4241
55.9	187	0	22.4814	4.2541	82.6862

Source: authors' work using Python package.

The distance at 0 consisted of 93 establishments (see Table 1). Differences at low distance values (0.68 to about 10) were mainly related to typos, e.g. missing the '-' symbol (in the names of accomodation establishments, cities and streets), incomplete names of establishments or streets, or missing Polish letters. However, it

could be detected that the link in this case concerned the same accommodation establishments. In the 10–20 range of distance values, there may have already been differences in the names of accomodation establishments, streets or their numbers. However, in most cases, the linked records concerned the same establishments. There were also mismatches of completely different records. Starting from the distance of 16, only half of the establishments were matched correctly. Above the value of approximately 30 (455 establishments), almost all the establishments were incorrectly linked, usually only by a few common characters, such as a fragment of the postal code or the phrase 'hotel, apartment' in the name of the establishment.

The quality of matching can be checked by the correctly and incorrectly matched and mismatched establishments, preferably using the confusion matrix. Four situations may arise after matching accommodation establishments:

1. the establishments have been correctly linked (true positive, TP);
2. two establishments have been mistakenly linked due to the short distance between them and similar names (distance smaller than threshold) (false positive, FP);
3. two establishments have not been linked (correctly) due to the large distance between them and low similarity between the names (true negative, TN);
4. two establishments have not been linked, but should have been, because there was only a small discrepancy in the establishment names (false negative, FN).

There are several approaches to building a confusion matrix for a data-matching problem. One of them assumes that we find the best match for all establishments in a smaller dataset with respect to a given distance or similarity score. Applying the optimal matching threshold, we obtain predictions: the match and mismatch. After a manual review of the linked data, we obtain the true state of the match and mismatch. Finally, we build a confusion matrix based on a set of true and predicted labels (match and mismatch).

Table 2 presents a confusion matrix for the data linkage result for the optimal distance threshold.

Table 2. Confusion matrix for the NLP method

Actual	Predicted	
	match	non-match
Match	0.31 (TP)	0.17 (FN)
Non-match	0.07 (FP)	0.45 (TN)

Source: authors' work using R package.

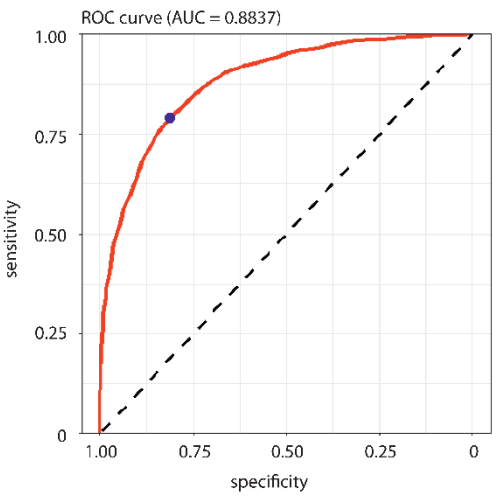
The accuracy amounted to 0.7622 with a 95% confidence interval (0.7446, 0.7792). The accuracy was tested against No Information Rate (NIR = 0.5132), and was significantly higher (p-value [Accuracy > NIR] < 0.0001).

5.2. Machine learning algorithm – *K*-Nearest Neighbours

Using the deduplication method based on machine learning algorithm, 3,157 establishments were combined. Similar to the NLP method, we set the optimal threshold.

We examined a set of 116 thresholds ranging from 0.00 to 1.29 (a large number of duplicated values of the metric). Figure 3, as in the case of NLP, presents ROC curves and a pair of specificity and sensitivity for the optimal threshold derived from Youden's J statistic (blue point).

Figure 3. ROC curve for the *K*-NN method



Source: authors' work using R package.

The optimal distance threshold amounted to 0.845. For this threshold, specificity and sensitivity reached 0.8121 and 0.7909, respectively, while Youden's J statistic totalled 0.6031.

Using the established value for the optimal distance threshold, data combining was performed.

The largest number of matches – 2,138 – was yielded for establishments belonging to the NACE group 55.1, and 54 of them linked perfectly (i.e. with zero distance). As regards accommodation establishments classified as NACE 55.2, 749 establishments were connected, of which three linked perfectly.

Table 3. Results of *K*-NN method

NACE	Number of matched establishments	Number of perfect matches	Distance		
			average	minimum	maximum
55.1	2,138	54	0.736908326	0	1.26
55.2	749	3	0.928958611	0	1.29
55.3	22	0	0.975454545	0.59	1.25
55.9	248	0	0.962540323	0.45	1.23

Source: authors' work using Python package.

A detailed analysis of the combined data showed that the distance measure ranged between 0 and 1.29. A distance of 0 covered 57 establishments. Differences at low distance values (0.14 to about 0.7) were related to incomplete names of accomodation establishments (e.g. lacking the owner's surname or abbreviation or company name), repetitions of keywords (e.g. 'hotel', 'apartment') and missing special characters. However, the links in almost all cases were true matches. Up to the distance level of 0.92, most establishment matches were correct; towards the end of this interval, differences in the names of accomodation establishments and numbers of buildings occured, but most street names and postal codes matched. Above the distance of 0.94, the vast majority of matches were incorrect, including street names or postal codes.

Table 4 presents a confusion matrix for data linkage result for the optimal distance threshold.

Table. 4. Confusion matrix for *K*-NN method

Actual	Predicted	
	match	non-match
Match	0.26 (TP)	0.07 (FN)
Non-match	0.13 (FP)	0.55 (TN)

Source: authors' work using R package.

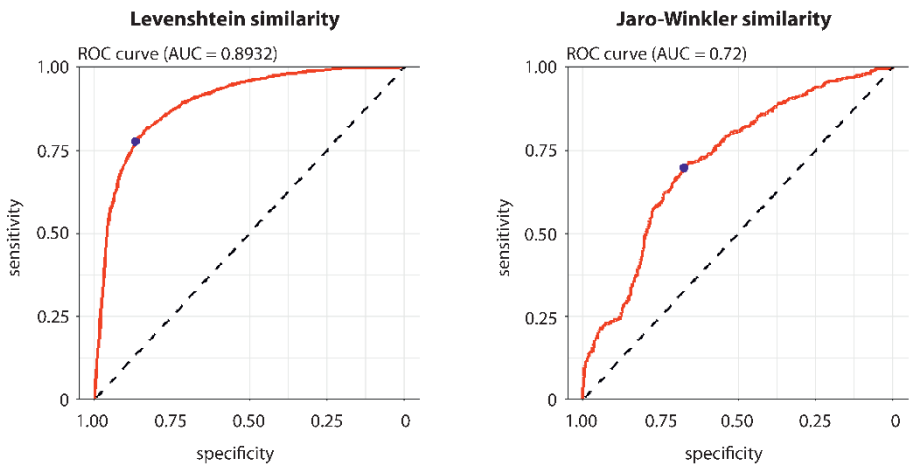
The accuracy amounted to 0.8052 with a 95% confidence interval (0.7969, 0.8134). The accuracy was tested against NIR (NIR = 0.6735). It was significantly higher than NIR (p-value [Accuracy > NIR] < 0.0001).

5.3. Fuzzy Matching and geolocation

The application of Fuzzy Matching and the use of geographic coordinates make it possible to obtain a set of linked accommodation establishments, containing two metrics for text strings (Levenshtein, Jaro-Winkler) and the geodetic distance between accommodation establishments in metres.

First, we checked which metric (edit distance) achieves better results. We examined a set of 1,977 thresholds ranging from 46 to 100 for the Levenshtein similarity and 1,843 thresholds ranging from 0.6 to 1 for the Jaro-Winkler similarity. Figure 4 presents the ROC curve for the applied method (solid red lines), the ROC curve for random classifier (dashed black lines), and a pair of specificity and sensitivity for the optimal threshold derived from Youden's J statistic (blue points).

Figure 4. ROC curve for the Levenshtein and Jaro-Winkler similarity



Source: authors' work using R package.

The optimal Levenshtein similarity threshold amounted to 83.5. For this threshold, specificity and sensitivity amounted to 0.8681 and 0.7757, respectively, while Youden's J statistic totalled 0.6438. The optimal Jaro-Winkler similarity threshold amounted to 0.73. For this threshold, specificity and sensitivity amounted to 0.6761 and 0.6975, respectively, while Youden's J statistic reached 0.3736. Table 5 presents a confusion matrix for the data linkage result for the optimal similarity threshold for the Levenshtein and Jaro-Winkler similarity.

Table 5. Confusion matrix for the Levenshtein and Jaro-Winkler similarity

Actual	Predicted	
	match	non-match
Levenshtein similarity		
Match	0.48 (TP)	0.07 (FN)
Non-match	0.04 (FP)	0.41 (TN)
Jaro-Winkler similarity		
Match	0.51 (TP)	0.25 (FN)
Non-match	0.07 (FP)	0.17 (TN)

Source: authors' work using R package.

For the two metrics tested, the best result was by far achieved by the Levenshtein algorithm, where approximately 48% of accomodation establishments were correctly matched. For 41% of the establishments from Booking.com, the accommodation establishments were not found in the tourism survey frame.

For the Levenshtein similarity, the accuracy amounted to 0.8912, with a 95%-confidence interval (0.8809, 0.9131). The accuracy was tested against NIR (NIR = 0.6355). It was significantly higher than NIR (p-value [Accuracy > NIR] < 0.0001). For the Jaro-Winkler similarity, the accuracy amounted to 0.6812, with a 95% confidence interval (0.6595, 0.7023). NIR (NIR = 0.7622) was higher than accuracy. It is worth noting that matching with the Jaro-Winkler similarity is near to a random classifier. The results of the analysis confirmed that combining Fuzzy Matching with the Levenshtein similarity is more effective than the Jaro-Winkler similarity.

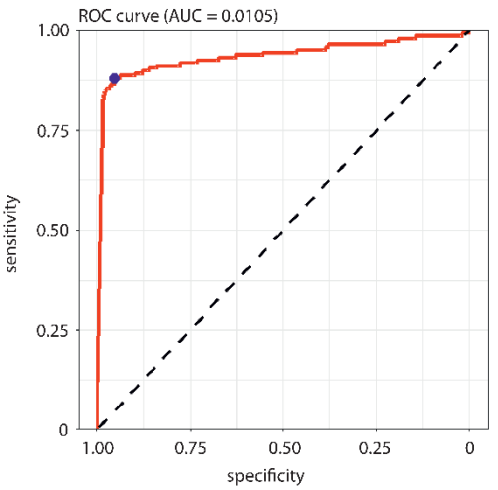
The Levenshtein similarity score ranged between 46 and 100. The similarity score of 100 covered 1,435 matches. Detailed analysis of the data showed that the differences in similarity scores in the range of 89–99 (924 establishments) were related to incomplete establishment names, repetition of keywords (e.g. 'hotel', 'apartment'), or lack of special characters. Linked records mostly involved the same establishments, but sometimes there were differences between the numbers of buildings or properties.

For the similarity score ranging from 85 to 88, the number of correctly and incorrectly matched establishments was similar. At the similarity score of 78, 80% of records were incorrectly matched. The reasons for these differences were the same as aforementioned. In addition, numerous discrepancies in the street and the name of establishments can be observed.

The detailed results of the algorithm using the Jaro-Winkler similarity were also analysed. The similarity score for addresseses using the Jaro-Winkler distance ranged between 60 and 100, and only five linked establishments reached the highest score (100). Similarity score values between 88 and 99 (41 links, including eight incorrectly linked) were the result of the incomplete name of the accommodation establishment (most often the lack of letters next to the building/apartment number, e.g. '9' where it should be '9a'), or the absence of individual special characters. Below the value of 88, correctly linked establishments totalled 404. However, their distribution was uneven.

To apply the geodesic distance between accommodation establishments, we also needed to determine the optimal threshold based on the ROC curve. For this purpose, we examined a set of 1,442 thresholds ranging from 0.006 to 697.3 km. Figure 5 presents a ROC curve for this method (solid red line), a ROC curve for a random classifier (dashed black line), and a pair of specificity and sensitivity for the optimal threshold derived from Youden's J statistic (blue point).

Figure 5. ROC curve for Vincenty's distance



Source: authors' work using R package.

The optimal Vincenty distance threshold amounted to 0.026 km. For this threshold, specificity and sensitivity totalled 0.9520 and 0.8795, respectively, while Youden's J statistic reached 0.832. Table 6 presents a confusion matrix for the data linkage result for the optimal similarity threshold.

Table 6. Confusion matrix for Vincenty distance

Actual	Predicted	
	match	non-match
Match	0.29 (TP)	0.02 (FN)
Non-match	0.02 (FP)	0.67 (TN)

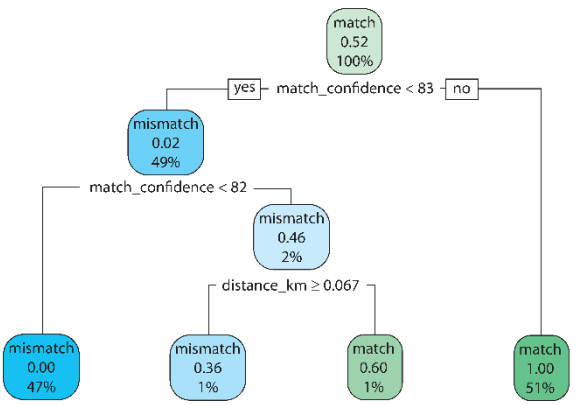
Source: authors' work using R package.

The accuracy amounted to 0.9612, with a 95% confidence interval (0.944, 0.9769). It was tested against NIR (NIR = 0.7622) and was significantly higher than NIR (p-value [Accuracy > NIR] < 0.0001).

As for the distance between accommodation establishments calculated using Vincenty's formula, all establishments within the distance up to 50 metres were linked correctly. In the distance range of 50–200 metres, the number of establishments linked correctly was comparable to the number of those linked incorrectly. A few establishments for which the distance was greater than 1 km were linked correctly, which was due to the specific notation of addresses obtained by web scraping.

Finally, we adopted a Fuzzy Matching method based on Levenshtein similarity and geolocation. Since we had two criteria to choose from, it was necessary to define decision rules for linking. The most intuitive solution seemed to be the conjunction of the two critical values. However, we did not want to rely on intuition, so we used a decision tree. The optimal complexity parameter value was determined at 0.016 by means of *k*-fold cross-validation. Figure 6 presents a decision tree using Levenshtein similarity and geolocation.

Figure 6. Decision tree for the Levenshtein similarity and geolocation



Source: authors' work using R package.

The model accuracy amounted to 0.9919, while specificity and sensitivity reached 0.9921 and 0.9917, respectively. Youden's J statistic totalled 0.9838.

Let us assume a match with a 100 similarity score and zero Vincenty distance is a perfect match. Table 7 presents the summary of the applied method.

Table 7. Results for the Levenshtein similarity and geolocation

NACE	Number of matched units	Number of perfect matches	Similarity score			Geodetic distance		
			mean	minimum	maximum	mean	minimum	maximum
Total	**3,170**	**1,400**	**94.13**	**82**	**100**	**0.63**	**0**	**711.86**
55.1	1,815	891	95.19	82	100	0.25	0	700.43
55.2	1,354	508	92.77	82	100	1.13	0	699.05
55.3	1	1	100.00	100	100	0	0	0

Source: authors' work using Python package.

For the NACE groups 55.1 and 55.2, the key criterion was a similarity score of at least 83 and less than 82. These two values separated 98% of cases. For values between 82 and 83, the geodetic distance was the deciding factor, which applied to about 2% of cases. The average value of the Vincenty distance was 0.250 metres for the NACE group 55.1 and over four times more, i.e. 1.13 meters, for the NACE group 55.2. Similarity score statistics were the same for both the above-mentioned NACE groups. In the case of the NACE group 55.3, only one establishment linked.

6. Conclusions

The growing demand for tourism information is caused both by external circumstances (the COVID-19 pandemic, large-scale migration, armed conflicts) and increasing expectations of tourists. In many countries, tourism is treated as a priority sector because of its role and the benefits it brings to the economy. The above circumstances as well as the dynamic development of new technologies and competition in information markets are forcing national statistical offices to carry out activities involving the continuous search for new sources of information and, above all, their integration into statistical databases and administrative records. Meeting these challenges is a highly complex phenomenon. Having reliable and real-time information on the tourist traffic, the average length of tourists' stay and the degree of their spending opens up new opportunities for effective tourism policies at the local, regional and international levels.

Data linkage and data deduplication from web scraping of tourism portals with the tourism survey frame are aimed at ensuring high-quality research results. Depending on the availability of data on websites, which also involves formal and legal considerations, different deduplication methods can be used. Each of these methods has its own strengths and weaknesses, which should be taken into account when choosing the right solution.

The article provides a detailed characterisation and evaluation of three data linkage and deduplication methods: NLP, *K*-NN and Fuzzy Matching. The use of NLP offers numerous benefits, such as scalability, flexibility and automation. Machine learning algorithms are also useful for data linkage and deduplication. When deciding which method to use, a combination of different algorithms for better results is worth considering (we, for example, considered the TF-IDF and *N*-gram techniques). A similar situation occurs with the Fuzzy Maching method, where, in addition, Vincenty's formula was used to calculate the exact geodetic distances between the establishments. It is also important to choose the appropriate distance metric, which affects the accuracy of the results.

The evaluation of the selected methods of linking and deduplicating the data was done using the confusion matrix, the ROC curve and Youden's J statistic. The best results were obtained by using the Fuzzy Matching method based on Levenshtein similarity combined with Vincenty's formula. It is worth noting that this method copes well with arbitrary notation of the names of establishments and can also be used to classify them.

The article analysed data from three booking portals, i.e. Booking.com, Hotels.com and Airbnb.com. It should be noted that these portals differ significantly in terms of the volume of information available. Therefore, Booking.com was chosen for the procedure of linking and deduplicating the data, due to its largest range of variables corresponding to the tourism survey frame. In this context, a very promising further research direction is the possibility of using algorithms that compare images. This way, it is possible to combine data from different portals more efficiently (photos become an additional key of correlation).

The presented research results are also important in the context of improving the quality of the tourism survey frame. It turns out that the use of web-scraped data resulted in an increase in the number of accommodation establishments classified as NACE 55.1 and NACE 55.2. In 2020, information was yielded on 151 new accommodation establishments in Poland and 56 in Bulgaria. These establishments accounted for 1.1% and 1.4% of the total number of accommodation establishments constituting the tourism survey frame in Poland and in Bulgaria, respectively. Most of them belonged to the NACE group 55.2 (64% of Poland-based units and 58% of the Bulgaria-based ones). Another large group were accommodation establishments belonging to the NACE group 55.1 (31% of Poland-based units and 26% of Bulgaria-based ones). They were mainly Aparthotels and B&B establishments, which, according to the adopted methodology, are classified as 'other hotel establishments'.

Currently in tourism statistics, information from booking portals is used for both data imputation and calibration. The process aimed at a full replacement of selected tourism surveys with information from online portals has already been launched. In this context, it is important to remember about prerequisites for the use of big data, namely a stable access to such data and a positive assessment of its quality. Combining new information with administrative registers and other sources of statistical data, by means of appropriate models, can lead to qualitatively new statistics at the micro-, meso- and macroscale.

References

Asher, J., Resnick, D., Brite, J., Brackbill, R., & Cone, J. (2020). An Introduction to Probabilistic Record Linkage with a Focus on Linkage Processing for WTC Registries. *International Journal of Environmental Research and Public Health*, *17*(18), 1–16. https://doi.org/10.3390/ijerph17186937.

Christen, P. (2012). *Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer. https://doi.org/10.1007/978-3-642-31164-2.

Cierpiał-Wolan, M., Truszyńska, A., Szlachta, P., Wnuk, Z., Sawicki, K., Oprych-Franków, D., Data, M., Ulma-Ciupak, B., Giełbaga, E., Wieczorek, G., Gumiński, M., & Mordan, P. (2022). *Feasibility project on digitalisation issues in national accounts*.

Cierpiał-Wolan, M., & WPJ Team. (2020). *Innovative Tourism Statistics Deliverable J2: Interim technical report showing the preliminary results and a general description of the methods used*. Eurostat, ESSnet Big Data II. https://ec.europa.eu/eurostat/cros/sites/default/files/WPJ_Deliverable_J2_Interim_technical_report_showing_the_preliminary_results_and_a_general_description_of_the_methods_used_2020_01_07.pdf.

Daas, P., Ossen, S., Vis-Visschers, R., & Arends-Tóth, J. (2009). *Checklist for the Quality evaluation of Administrative Data Sources* (CBS Discussion Paper No. 09042). https://ec.europa.eu/eurostat/documents/64157/4374310/45-Checklist-quality-evaluation-administrative-data-sources-2009.pdf/24ffb3dd-5509-4f7e-9683-4477be82ee60.

European Commission. (n.d. a). *Project Overview*. Retrieved July 8, 2023, from https://cros-legacy.ec.europa.eu/content/project-overview_en.

European Commission. (n.d. b). *WPJ Innovative tourism statistics*. Retrieved July 8, 2023, from https://cros-legacy.ec.europa.eu/content/WPJ_Innovative_tourism_statistics.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Maślankowski, J. (2015). Analiza jakości danych pozyskiwanych ze stron internetowych z wykorzystaniem rozwiązań Big Data. *Roczniki Kolegium Analiz Ekonomicznych SGH*, (38), 167–177. https://rocznikikae.sgh.waw.pl/p/roczniki_kae_z38_11.pdf.

Oancea, B., Necula, M., Salgado, D., Sanguiao, L., Barragán, S. (2019). *ESSnet Big Data II. Workpackage I: Mobile Network Data. Deliverable I.2 (Data Simulator). A simulator for network event data*. Eurostat.

Peirce, C. S. (1884). The Numerical Measure of the Success of Predictions. *Science*, *4*(93), 453–454. https://doi.org/10.1126/science.ns-4.93.453-a.

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. https://bioinfopublication.org/pages/article.php?id=BIA0001114.

Quinlan, R. (1983). Learning efficient classification procedures. In R. S. Michalski, J. G. Carbonell & T. M. Mitchell (Eds.), *Machine Learning. An Artificial Intelligence Approach* (pp. 463–482). Springer-Verlag. https://doi.org/10.1007/978-3-662-12405-5.

United Nations Department of Economic and Social Affairs Statistics Division. (2015). *Classification of Types of Big Data*. https://unstats.un.org/unsd/classifications/expertgroup/egm2015/ac289-26.PDF.

United Nations Economic Commission for Europe. (n.d.). *Unece Statswiki*. Retrieved July 8, 2023, from https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.