SOSYOLOJİ ARAŞTIRMALARI DERGİSİ
JOURNAL OF SOCIOLOGICAL RESEARCH

Geliş Tarihi / Submited: 17.07.2023
Kabul Tarihi / Accepted: 23.09.2023

Cilt / Volume 26 Sayı / Issue 2 (Ekim/2023 – October/2023) Araştırma Makalesi / Research Article

ISSN/e-ISSN: 2148-9947

DIGITAL TRACKS BEYOND BORDERS: A SYSTEMATIC REVIEW ON THE MIGRATION CRISIS

Beyza Yılmaz[1], Emre ÖZCAN [2]

Abstract

This study aimed to systematically examine the studies conducted and published on immigrants, asylum seekers, and refugees by using big data written in English. Articles were searched on Scholar, The Web of Science, ProQuest, Science Direct, PubMed and Scopus databases. The concept set centered around the concepts of immigration and big data was used in the surveys. In accordance with the PRISMA protocol principles, 49 articles were examined according to the inclusion and exclusion criteria among 258 articles obtained from the relevant databases until the end of December 2022. The reviewed articles were categorized under the headings of "topics examined", "dataset", "analyses", "software used" and "key findings". The studies provide indications on how to obtain information about this population, which is difficult to reach group especially due to its massiveness, using big data tools. In the findings, it has been seen that studies based on big data on immigrants, asylum seekers and refugees contribute to facilitating the integration of these groups into the target country. Also, it has been revealed that these studies may lead to undesirable results in terms of violating the confidentiality of research groups, producing labeling, and increasing surveillance for these groups. In addition to these, it has been found that these studies have methodological handicaps in terms of representativeness, accuracy, excessive homogenization, and easy generalization. It is thought that the findings of the study will shed light on the international migration and refugee policies to be carried out using big data analysis tools.

Keywords: Big data, migration, systematic review.

[1] Arş., Gör., Başkent Üniversitesi, Sosyal Hizmet Bölümü, beyzayilmaz@baskent.edu.tr, ORCID: 0000-0002-6963-2036
[2] Dr. Öğr. Üyesi, Başkent Üniversitesi, Sosyal Hizmet Bölümü, eozcan@baskent.edu.tr, ORCID: 0000-0002-0877-2457

SINIRLARIN ÖTESİNDEKİ DİJİTAL İZLER: GÖÇ KRİZİ ÜZERİNE SİSTEMATİK BİR DERLEME

Öz

Bu çalışma, uluslararası literatürde İngilizce olarak kaleme alınmış, büyük veri analiz araçları kullanılarak göçmen, sığınmacı ve mültecilerle ilgili yapılmış ve yayınlanmış çalışmaların gözden geçirilmesini ve çalışmalardan elde edilen verilerin sistematik bir biçimde incelenmesini amaçlamıştır. Makaleler Scholar, The Web of Science, ProQuest, Science Direct, PubMed ve Scopus veritabanları üzerinden taranmıştır. Taramalarda göçmen ve büyük veri kavramları etrafında yararlanılan kavram seti kullanılmıştır. PRISMA protokol ilkelerine uygun olarak 2022 yılı Aralık ayı sonuna kadar ilgili veritabanlarından elde edilen 258 makale arasından dahil etme ve hariç tutma kriterlerine göre 49 makale incelenmiştir. Taranan makaleler "ele alınan konular", "veri seti", "analizler", "kullanılan yazılım" ve "başlıca bulgular" başlıkları altında incelenerek kategorileştirilmiştir. Araştırmalar, büyük veri araçlarının kullanılması yoluyla özellikle kitleselliği nedeniyle erişilmesi zor bir grup olan bu nüfus hakkında nasıl daha kolay bilgi elde edilebileceğine dair göstergeler sunmaktadır. Bulgularda göçmen, sığınmacı ve mültecilerle ilgili büyük veriye dayalı çalışmaların, bu grupların hedef ülkeye entegrasyonunu kolaylaştırma noktasında katkı sağladığı görülmüştür. Ayrıca bu çalışmaların bu gruplar açısından araştırma gruplarının gizliliğinin ihlal edilmesi, etiketlemeyi üretmesi, gözetimi artırması bağlamında sakıncalı sonuçlar doğurabileceği ortaya konulmuştur. Bunlara ek olarak bu araştırmaların temsil edilebilirlik, doğruluk oranı, aşırı homojenleştirme ve kolay yoldan genelleştirme gibi hususlarda metodolojik handikaplar taşıdığı bulgulanmıştır. Araştırmanın bulgularının büyük veri analiz araçları kullanılarak gerçekleştirilecek uluslararası göç ve mülteci politikalarına ilişkin ışık tutacağı düşünülmektedir.

Anahtar Sözcükler: Büyük veri, göç, sistematik derleme.

INTRODUCTION:

The substantial technological developments experienced since the 1980s have resulted in one of the major breaking points in societal-historical transformation. This transformation described by conceptualizations such as "information society" (Masuda, 1991), "information age" (Castells, 1996), "digital society" (Fukuyama, 2018), and "post-industrial society" (Bell, 1973) has swiftly traversed the social arena. Information and communication technologies have become a part of daily life and led to radical changes in the lifestyles of individuals and communities. Users of information produce data every second by using multiple devices, record events, and can become the object of data production themselves. In addition to individuals, it is seen that organizations and governments with almost all their units and departments have become both the subject and object of this production. The increase of data at such an extent has created the concept of "big data" in the literature. The term big data was first used by Michael Cox and David Ellsworth in the study titled "Application-Controlled Demand Paging for Out-of-Core Visualization" in the Proceedings of the 8th Conference on Visualization held in 1997 to emphasize that some datasets could not be stored anymore in computer storage or external hard drives due to their sheer size (Cox & Ellsworth, 1997). Francis X. Diebold (2016) on the other hand, stated that the term was first used by Kohn Mashey in 1988 in his presentation titled "Big Data and the Next Wave of InfraStress."

Although the concept of big data started to be discussed in the late 1990s, its widespread use in the literature corresponds to the late 2000s. There is no clear consensus on what elements such as processing capacity, speed, access to information, and type of the data stored should be considered regarding the definition of big data. However, to summarize, big data is a concept that defines heterogeneous data at different volumes which are not possible to process through traditional database techniques and comprise various digital contents (Gahi et al., 2016). Beyer and Laney (2012) defined big data as "information entities with high volume, high speed, and great variety that have innovative forms of data processing to promote insight and decision-making and require cost-effectiveness." Davenport defined the concept as "data that are too big to be stored in one server (more than 100 terabytes in size), not structured in lines and columns, or continuously flowing in a way not to fit in a stagnant data warehouse."

Big data is a phenomenon that has been commonly used in many areas of both the natural sciences and the social sciences for some time. Despite its handicaps such as serving the interests of certain groups, neglecting certain masses, maintaining inequality, labeling, extending surveillance, and violating confidentiality, big data has been widely used especially in the social sciences in recent years due to reasons such as the ability of users to make generalizations in a simple way through homogenization, providing the opportunity to have easy access to greater volumes, opening the door to more advanced statistical techniques, and establishing connections between datasets (Yilmaz and Ozcan, 2022). The most attractive aspect of big data is the opportunity it provides to have access to data at great volumes which are difficult to reach quickly and easily.

Considering that one of the groups that are difficult to reach is asylum seekers and immigrants, who have gained these statuses as a result of recent mass migrations that emerged as a global crisis, it might be claimed that data on these individuals are within the scope of big data. These data cover many indicators that accelerate their integration into the host country, but they also lead to many adversities that label them, keep them under check, and risk the safety of their lives. On the other hand, it is indispensable for states to have big data that are reliable, qualified, and obtained timely on asylum seekers and immigrants to effectively manage such crises – and to prevent

decision-makers from making ill-advised decisions and society from having misled perceptions (Atar, 2021: 158). However, it should be underlined that despite limitations regarding the sharing of big data on refugees and immigrants, there has been an increase in recent years in the number of academic studies that investigate the mobility of these groups and their integration with the host society through big data analysis. Many researchers utilize datasets that include information obtained from mobile phone usage, social media, and other sources, especially public institutions, to perform analyses and make estimations regarding socioeconomic and cultural issues as much as the big data allow (Korkmaz, 2021: 241). Although these studies in which the quality and representation of datasets are important aims at supporting the integration of asylum seekers and immigrants and offer solutions to and inform states, they should be structured meticulously and focus on the principle of "do no harm", and they should also be sensitive not only in the analysis of the data but also in terms of the source, avoid political manipulations, not risk the safety of the relevant groups by violating their confidentiality, not reproduce discriminating factors, and in short, they should not ignore the rights and freedoms of these groups stemming from international law (Korkmaz, 2021, pp. 243-44).

In this context, this study aimed to review studies conducted and published on immigrants, asylum seekers, and refugees and written in English in the international literature by using big data analysis tools and/or datasets and systematically analyze the data obtained in these studies. In line with this main objective, the study sought to answer the following research questions:

- What issues/variables are addressed in studies on immigrants, asylum seekers, and refugees analyzed using big data tools?
- What types of datasets are used in studies on immigrants, asylum seekers, and refugees in which analyzed using big data tools?
- What statistical analysis methods and analysis software are used in studies on immigrants, asylum seekers, and refugees analyzed using big data tools?
- What are the main findings of studies on immigrants, asylum seekers, and refugees analyzed using big data tools?
- What are the main handicaps of studies on immigrants, asylum seekers, and refugees in which analyses were performed using big data tools?

1. MATERIALS AND METHODS

1.1 Research Strategy

In this study, articles using big data and big data analysis tools on immigrants, asylum seekers, and refugees were screened on the Scholar, Web of Science, ProQuest, Science Direct, PubMed, and Scopus databases without making any distinctions such as domestic-international migration, voluntary-forced migration, registered-unregistered immigration, or source country-destination country. In this context, to analyze the phenomenon of migration, key phrases such as "refugee[s]", "asylum seeker[s]", "immigrant[s]", "migrant[s]", "migration[s]", "displacement[s]", "mass movement[s]", and "immigration[s]" were used, and to include big data studies, the key phrases of "big data", "machine learning", and "data analytics" were employed. Each of these concepts were matched one by one, and different combinations were created. The review was performed according to the appropriate search methods in each database (e.g., "machine learning AND immigration", "(big data) AND (migration)", "data analytics, refugee", etc.).

In the review process on each database, first of all, the titles of the articles were examined, and they were saved in an Excel file. Articles that did not include at least one keyphrase from the identified immigration and big data keyphrase groups were excluded from the study. After the selected databases were reviewed, the titles of the collected articles were examined, and duplicate articles were removed. Inquiries were discussed with the team members. The articles were uploaded to Mendeley Desktop, Version 1.19.8.2008∓2020 Mendeley Ltd., and they were evaluated in terms of singularization, review, full text, and relevance. After this, the abstracts of the articles in the Excel file were analyzed, and the articles which were decided to be unsuitable for the purpose of the study were discussed and removed from the analyses by the researchers. Following the analysis of the abstracts, the full texts of the remaining articles were evaluated, and the articles whose full text could not be accessed and those which did not fit the purpose of the study were excluded from the study. Finally, the remaining articles were presented in a table which included “issues discussed”, “dataset”, “analyses”, “software used”, and “main findings.”

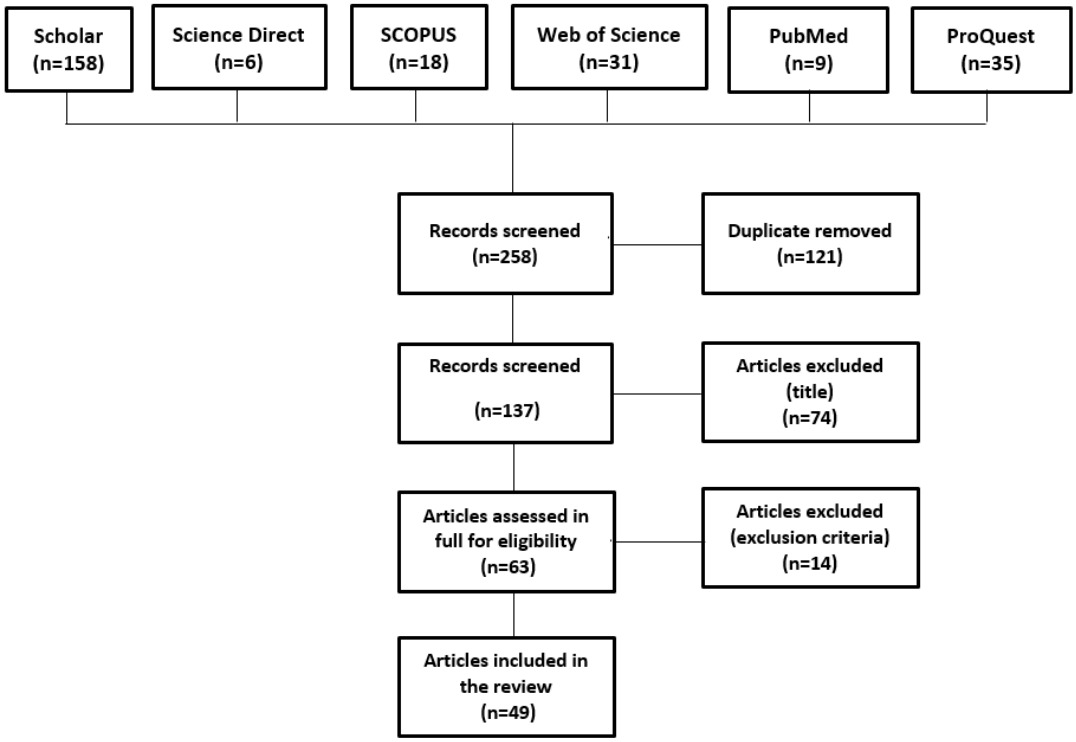
1.2 Inclusion and Exclusion Criteria

The inclusion criteria were determined as being a research article and being written in English. All studies published in the relevant databases by the end of December 2022 were included in the study. The exclusion criteria were specified as big data research articles not related to immigrant, asylum seeker, or refugee groups, not including the terms among the search criteria in the title, and publications other than research articles such as conference texts and book chapters.

1.3 Article Selection

The research articles were selected according to the PRISMA checklist. In total, 258 studies were found. The same studies which were found in different databases (121 articles) were removed. The abstracts of the remaining 137 studies were read, and 74 studies which were non-scientific or written in languages other than English were excluded. Then, the remaining 63 articles were subjected to full-text analysis. Finally, 19 articles were removed as their full texts were inaccessible, or they did not meet the purpose of the study. Thus, 49 articles were included in the analyses.

Figure 1. Articles according to PRISMA protocol principles.



2. RESULTS

Within the scope of the study, 49 articles were analyzed. The included articles were analyzed and classified under the categories of "issues discussed and main findings", "dataset", "analyses", and "software". Information about the findings is presented in Table 1.

Table 1. Research results and selection of studies reporting the use of big data in studies on immigrants, asylum seekers and refugees.

	Reference	Issues discussed	Dataset	Analyses	Software used	Main findings
1	Ahmed et al., (2020)	Machine learning was used to predict the landslide risk of the camp infrastructure based on the geospatial features of Bangladesh.	Geospatial features (elevation, lithology, Normalized Difference Vegetation Index (NDVI), Topographic Wetness Index (TWI) data) data were used.	Logistic Regression (LR), Multi-Layer Perceptron (MLP), Gradient Boosted Trees (GBT), Random Forest (RF)	Python, JavaScript	Normalized Difference Vegetation Index (NDVI) values in the study area are low due to deforestation and/or damage to vegetation in areas where camp

						infrastructure is available. For example, toilets were installed in places with the highest landslide risk rate (landslide tendency) with 40.23%. This was followed by hand pump tube wells with 37.49%.
2	Ahmed et al., (2019)	Population estimation approaches of Rohingya refugees in the south eastern border region of Bangladesh were investigated.	Data on the living conditions and needs of Rohingya immigrants residing in Bangladesh in 2018 provided by the International Organization for Migration (IOM) and U.S. Landsat-8 satellite images provided by the Geological Survey were used.	Linear Regression, LASSO Regression, Linear Regression with Elastic Net Regularization (Elastic Net), k-Nearest Neighbor (KNN), Decision Tree, Support Vector Regression (SVR), Ada Boost, Gradient Boosting, Extra Tree, RF	Python, Tensor Flow, Google Colab	It has been demonstrated that the data found in camp blocks and satellite images can be successfully used in the population estimation task. The Gradient Boosting Model is a suitable choice for both target-hunting and satellite image target-hunting approaches. In addition, it was concluded that artificial neural

						networks are a very good choice for data-driven approach and the most accurate option for satellite image-based approach to population estimation.
3	Aslan et al., (2019)	The comments made under the news titled Alan Kurdi and Drowning Refugees in Muğla on the websites of the five best-selling newspapers were examined.	The comments in the newspapers Liberty, Morning and latest news.com, which are selected from the top 25 websites clicked on Alexa.com in Turkey, were studied.	Perception Analysis Model (PAM)	Unspecified	It was found that almost every commentator reflected a negative opinion in 735 comments examined. 21.4% of commenters blame people with whom they differ on many issues, as well as political parties. While 62.4% of the comments point out that the Syrian refugee crisis is an internal problem,

						4.35% think that it is an out-group hatred. Those in the second group question why they used Turkey as a bridge, why they fled from Turkey afterwards, why they cooperated with the imperialist powers and why they did not seek asylum in other countries.
4	Augsburger & Elbert, (2017)	It was carried out to examine whether there was an increase in risky behaviors among IDPs who have settled in Germany.	The Adverse Childhood Experiences scale, an abbreviated version of the ViVo checklist, the abbreviated event list of the Post Traumatic Diagnosis Scale, the Post Traumatic Stress Symptoms Interview, the Patient Health Questionnaire, and the Risk-Taking Behavior BART were administered to 56 participants from various countries who settled in Germany.	Stochastic gradient boosting machine	R	All participants experienced at least one traumatic event. Childhood maltreatment was experienced by 94% of participants. Physical abuse is the most common. Emotional abuse comes next.

5	Aydemir, (2022)	Machine learning models were compared to predict both the number of immigrants and the income distribution of immigrants.	Data from the World Bank 2010 Economic Development Indicators and data on the number of immigrants between 1960 and 2010 obtained from the "Immigrant Quantity" were used.	Support Vector Machines (SVM), Naive Bayes (NB), LR, KNN, RF, Xgboost algorithms	Python	By using the features in the data, 86.04% success rate for LR, 83.72% for SVM, 83.72% for KNN and 69.76% for NB were obtained. In this application, it is seen that the highest success is achieved through the LR algorithm. In another application to estimate the number of immigrants, 98.37% success rates were obtained with XGBoost and 96.42% with RF.
6	Azizi et al., (2021)	By revealing the number of immigrants from Mexico to the United States, pre-immigration and post-immigration variables were discussed in order to predict how many people will immigrate and	The "life" dataset from the "Mexican Immigrant Project" provided by Princeton University was used. In this dataset, there are variables such as labor force information, immigration information,	RF	Unspecified	It has been observed that Mexican immigrants mainly immigrate to the United States for the duration of their migration and then decide depending

		how long they will stay in the United States.	marriage history, and number of children of 25,298 people.			on variables such as status and employment.
7	Baird et al., (2022)	Children's drawings were examined with traditional econometric and machine learning tools to learn about the mental health of children in migration and disaster environments.	Drawings of 2480 Syrian refugee children aged 5-12 in Jordan were digitized and analyzed. In addition, data were obtained from the self-portraits of children.	LASSO, K-fold cross-validation	Unspecified	The results show that children's drawings can be used as a diagnostic tool in crisis environments. The use of a political slogan or image in the drawings is the strongest indicator of psychological distress in a child refugee. Features such as sketchy drawings, shadowy drawing of the face or body, drawing in monochrome, lack of detail, incomplete drawing of the nose or mouth are presented as important indicators of their likelihood of being exposed to violence in the past.

8	Bertsimas & Fazel-Zarandi, (2021)	In order to design an immigration policy that prevents crime in the United States, it was aimed to develop an algorithm to predict the risk of retrial of non-citizens convicted of various crimes.	From the inception of "Safe Communities" in October 2008 to November 2015, 904,896 detainees and 3,640,599 crimes were examined among detention orders across the United States.	Unspecified	Unspecified	The algorithm was found to reduce crime by 25% and was successful in reducing recidivism rates for inmates in both the federal and state prison systems.
9	Best et al., (2022)	The migration situation in Bangladesh, which was considered one of the most vulnerable countries to climate change, has been handled with machine learning.	The migration background, employment and financial situation of more than 1695 people belonging to 3000 households were discussed through the survey conducted by the Bangladesh Environment and Migration Survey (BEMS) in 2014.	RF, Survival analysis, Regression	R	Cox proportional hazard models revealed that the number of members in a household, the first year of residence, and the household goods owned by the household are important in terms of economic status. The results show that the total number of members of a household has a negative effect on migration, and migration

						mobility increases with the increase in the number of non-workers in the household.
10	Carammia et al., (2022)	It was desired to develop an early warning and forecasting system regarding migration mobility in EU countries.	Global Events, Language and Tone Database (GDELT) project data, geolocated events and internet searches in countries of origin, Frontex, EASO data were used.	DynENet, Vector Autoregression (VAR), LASSO, ARIMA, RF	Google Trends	The model offers a highly effective system for early warning and prediction of migration.
11	Chang, (2018)	It was aimed to apply big data technologies on Hakka genealogy to investigate migration patterns of Hakka ethnic group in Taiwan between 1954 and 2014.	It was studied with 4,492 Hakka pedigree data from FamilySearch genealogy collection service.	Unspecified	GOHAKKA, JavaScript, Timemap, Google Map	A very detailed analysis of the genealogy is provided for users, enabling the use of big data at different stages of the research process.
12	Chen et al., (2022)	Urban economic resilience was discussed in the context of Baidu migration in China.	The China City Statistical Yearbook, which targets 287 cities in China, and the provinces and cities statistical yearbook are the main data sources in this study. The population mobility index data is taken	Spatial regression model	Python	When the population mobility index increases by one unit, the urban economic flexibility index increases by 0.36-0.56%. Innovation input has a positive

			from Baidu Migration Big Data on the Baidu migration web page.			mediating effect on urban economic flexibility and population mobility.
13	Choi et al., (2020)	The levels of psychological distress of Korean immigrants in the United States during the COVID-19 pandemic were addressed by the prediction of discrimination, coping mechanisms, and socio-demographic factors.	A survey was conducted with 790 people selected by purposive sampling among Korean immigrants over the age of 18 living in 42 states in the United States. Data were collected online between 24 May and 14 June 2020 through a questionnaire consisting of seven scales.	Artificial neural network (ANN), descriptive analysis	SPSS	The ANN model examined the predictability of three types of discrimination and three types of coping mechanisms on participants' psychological distress. Psychological resilience of a person was found to be the most important determinant. Daily discrimination experiences are the second most important predictor variable. Individuals' age, gender, education level, race and ethnic identity showed less importance in predicting the level of psychologic

						al distress. Finally, an individual's income level, employment status and marital status were found to have the least predictive features in the ANN model.
14	Emami et al., (2020)	Four machine learning models (BRT, MARS, MDA, and RF) were tested to predict collective motion sensitivity in Chaharmahal and Bakhtiari Provinces in the Southwest of Iran.	Mass movement events reported by the Forest, Range and Basin Organization between 1977 and 2019 -three types of mass movement, including debris flow, landslide, and rockfall- were collected and used aerial photographs based on extensive field surveys.	Boosted regression trees, Mixture discriminate analysis, RF, Multivariate adaptive regression splines	R, ARCGIS	According to the LASSO algorithm, altitude and slope are the most influential factors. Also, distance to roads and faults is the next priority. Random Forest was found to be the most accurate model for mass movement susceptibility mapping.
15	Fernández-Martínez et al., (2021)	The prevalence of communicable diseases of immigrants in Spain was discussed.	All immigrant patients in the Tropical Medicine Unit were selected. During the first visit, all patients were asked to complete a questionnaire	Fisher's exact test, T Test, Mann Whitney U, kNN, Leave-one-out cross-validation	PSPP	Screening was found in 566 patients (74.5%). The most commonly diagnosed diseases are intestinal parasites,

			containing demographic variables. A detailed medical history and physical examination were also performed at this time. A standard screening protocol was administered to all patients, regardless of the presence, absence, or type of symptoms. In addition, data from the country's official civil registry were also used.			followed by syphilis, HIV infection, chronic HBV, filariasis, malaria, chronic HCV and Chagas disease. Syphilis and HIV+ patients are mostly women from Central Africa.
16	Gao et al., (2022)	High-speed rail HSR presentations in China are examined to provide new causal evidence about the impact of improved transportation on urban tourism.	Tencent big dataset collected from smartphone location information was used on one-day migration of Chinese cities from 2015 to 2019.	DID method, ordinary least squares (OLS) method	Python	Through daily panel data from Tencent migration big data, it is revealed that transportation improvements are facilitating tourism economies among Chinese cities.
17	Garha & Domingo, (2019)	The ethno-linguistic diversity of the Indian diaspora population with different socio-demographic	Facebook data of individuals born in India and living abroad (13-65 and over) were extracted. According to Facebook data,	Unspecified	Unspecified	60.7% of the total diaspora population is in Asia, 20.9% in North America, 10.5% in

		characteristics such as age, gender, location, language, ethnicity and citizenship was investigated through the Facebook advertising platform.	the Indian diaspora in 2017 consisted of 12.8 million people living in 150 countries around the world.			Europe, 2.5% in Africa and 0.8% It is based in Latin America. In 2017, Bangladesh, Pakistan and Indonesia had the highest proportion of young adults (13-24 years old) among the population of Indian descent. The United States, United Kingdom, and Canada have the highest proportions of older people (50 and over). The largest proportion of the working population (ages 25-49) is settled in Singapore, Saudi Arabia and Kuwait.
18	Giang et al., (2022)	Simulated experiments were examined to estimate labor exports in Taiwan,	It uses a database of Vietnamese labor migration to Korea, Japan, and Taiwan from 1992 to	Back-propagation Neural Network (BPNN), Random Forest Regression (RFR), KNN	Python	The accuracy levels of the three prediction models were

		Japan, and Korea.	2020, obtained by the Overseas Ministry of Labor.			evaluated. BPNN findings were found to be the closest when compared to actual data in Taiwan, Korea and Japan. It was found that kNN algorithms reached the second most accurate level in Taiwan and Japan, while kNN reached the lowest level in Korea over the years from 1994 to 2016.
19	Havas et al., (2021)	Focused on analyzing refugee movements from the Near East to Central Europe in 2015 and 2016.	Two Twitter datasets on refugees called Geo-Tweets and General-Tweets, UNHCR dataset, Harvard CGA Geotweet Archive were used.	Convolutional Neural Network (CNN), hotspot analysis, ARIMA stochastic model	DBpedia Spotlight Web Application, and Stanford CoreNLP	At the beginning of 2015, hotspot maps show that refugees gather mostly at the border of Turkey and Greece to begin their journey. In autumn 2015, there is more concentration towards the borders of North Macedonia, Serbia,

						Hungary, Austria and Germany, which are the hot spots.
20	Huang & Shao, (2022)	Focused on the application of big data statistics to solve the construction of the Arab migration and entrepreneurs hip data system.	Pseudo data on Arab immigration and international trade from 2015 to 2020 are used.	Artificial servo cluster algorithm	Unspecified	In the context of big data, the pseudo-modeling method of the intelligent iterative pseudo-servo cluster algorithm is presented.
21	Juric, (2022a)	It was conducted to test the usefulness of Google Trends indices to predict forced migration from Ukraine to the EU (particularly Germany) and to obtain demographic information from social networks on the age and gender structure of refugees.	Google Trend data from February 24, 2021 to February 24, 2022 and UNHCR statistics are used.	Linear regression	Google Trends	The fastest growing Google search terms in Ukraine (December 7, 2021 - March 7, 2022) are shown to be Western Union, asylum, refugee and Schengen, excluding the term "border". The most frequent search in Poland since the outbreak of the Russia-Ukraine war is "Border Crossing + Germany". All search queries with

						an indication about migration planning show a positive linear relationship between the Google index and data from official statistics.
22	Juric, (2022b)	The focus point was on monitoring conditions that indicate the intention of refugees from Ukraine to move to Germany.	Since the outbreak of the war in Ukraine, Facebook, Instagram and Youtube data have been collected about users in Ukraine's neighboring countries and Germany. State-level estimates of Ukrainian refugees were also obtained from the Meta Business Suite database and the FB Marketing API.	Unspecified	YouTube, Facebook, Instagram, Twitter	The results show that the increase in Facebook and Instagram index frequency is associated with increased immigration from Ukraine. It shows that the interest in Germany, especially since the age of 23, and the curiosity of the Ukrainian Facebook and Instagram users in learning German has

						increased rapidly.
23	Jurić, (2022c)	It was carried out to obtain sociodemographic information about Ukrainian refugees through Google Trends and social media applications.	Between February 1 and March 11, 2022, a query was made on Google Trends for the terms "граница" and "кордону" searched on Google. Results have been compared with UNHCR's official statistics. Also, YouTube, Instagram, Facebook and Twitter networks were examined.	Unspecified	Google Trends	As of the beginning of the war in Ukraine, the number of shares of Facebook and Instagram users in Poland, Slovakia, Hungary, Moldova, Romania and Germany has increased rapidly. Women aged 25-44 made up the majority of Facebook and Instagram users among Ukrainian refugees in Poland, Slovakia, Hungary, Moldova, Romania and Germany.
24	Katsikouli et al., (2022)	Summaries of asylum case decisions in Denmark were examined.	Decision excerpts from asylum appeal cases published by the Danish Refugee Appeals Board are used. The dataset is from the Flygtningenævnets (FLN)	Decision Tree, RF, SVM, Logistic Regression, Neural Networks, Naive Bayes classifier	Python	The results showed that it is possible to have information with potentially predictive properties about the outcome of

			Nævnsdatabase 2020 repository, which became publicly available on October 3, 2020.			any asylum case. When the Random Forest Classifier is used, it has been seen that a lot of information about the asylum seeker is used to predict whether the case will be overturned by the Refugee Appeals Board.
25	Khangahi & Kiani (2021)	A machine learning-based model is presented by exemplifying the Urmia Lake case study in Iran.	Population data of Urmia and five other cities for the last 40 years are used.	Random Forest, KNN and Support Vector Machine	Unspecified	According to the results obtained, people in the relevant region had to migrate due to unemployment and economic difficulties, air pollution, the emergence of various diseases, and decrease in drinking water. It has been determined that the SVM method gives better results than other algorithms.

						The SVM-based model showed the best performanc e with an increase rate of 88%.
26	Kılıç, vd. (2019)	It was carried out in order to shed light on the lives of refugees in Turkey from various aspects.	Data collected from 992,457 Türk Telekom customers, of which 184,949 are tagged as refugees and 807,508 as Turkish citizens, provided by D4R Challenge in 2017 were used. In addition, geospatial data sets, population data from the Turkish Statistical Institute (TUIK) and various statistics, coordinates of houses and workplaces for rent and sale in various neighborhoods in Istanbul via Hürriyet Emlak and the results of the Address-Based Population Registration System for the year 2017, in which the education levels of the residents on the basis of the	K-means (SK-means) algorithm, network analysis	R, Python, ArcMap, Google Places API, QGIS, MySQL, NoSQl and Pajek	The top 10 cities with the highest refugee traffic are Istanbul, Gaziantep, Ankara, Izmir, Mersin, Adana, Hatay, Antalya, Şanlıurfa and Kocaeli. Wealthy refugees visit summer locations as often as non-refugees. The analysis showed that medium- and high-status refugees use a very large area, travel long distances, and have a regular pattern in their movements, while low-status refugees seem to be

			neighborhood are given, were used.			trapped in a small neighborhood, meaning it is unlikely to travel.
27	Kiossou, vd. (2020)	An artificial neural network (ANN) model has been established to gain deeper insights into how migration is affected by its drivers.	Two different datasets were used, consisting of features and flows from 1960 to 2000 and features and flows from 2001 to 2010.	Random Forests, "extreme" gradient boosting regression (XGBoost) model, artificial neural network (ANN) model	Unspecified	In the model, it has been seen that the farther the two countries are from each other, the lower the relevant migrant flow. The effect of droughts in the target countries on the flow of incoming migrants was examined. It has been revealed that destination countries tend to receive fewer migrants during drought years, mainly due to worse employment opportunities for low-skilled labor.
28	Krupenkin & Rothschild (2019)	In the 2016 US election process, it has	CNN, MSNBC and Fox News	Structural topic models	Unspecified	Immigration news volume on

		been tried to deal with whether the coverage of the media regarding immigration has changed before the campaign, during the campaign and after Donald Trump was inaugurated.	transcripts were used.			Fox News nearly doubled after the election compared to others. There has also been a dramatic increase in the use of the crime framework in immigration news. After Trump was inaugurated, there has been a significant decline in media debates on comprehens ive immigration reform, and instead a significant increase in discussions of legal immigration restrictions and sanctuary cities. Media discussions on immigration policy seem to have shifted to the rightward.
29	Lai, & Pan, (2020)	The structure of China's urban population flow network	Travel information data from 346 cities from Tencent	Complex network analysis method	Gephi, SPSS	The urban network is more densely distributed

		was examined.	Migration is used.			in the east than in the west. The direction of population mobility before the festival is symmetrical with the direction of the population flow in the middle of the festival and after the festival. Cities with high administrative levels are the main population distribution centers.
30	Lee & Song (2022)	Using a simulation study, it was investigated whether the negative relationship between the academic performance of both native students and rural immigrant students could be weakened when discrimination decreases.	In the 2013-2014 school year, the first wave of China Education Panel Survey, a nationally representative survey providing detailed information about students, families and schools in grades 7 and 9, was used.	Wasserstein generative adversarial network (WGAN), bootstrap method	Unspecified	The findings show that discrimination is negatively related to the academic performance of both immigrants and natives. The simulation study reveals that in a hypothetical scenario where discrimination is mitigated, discrimination is no longer

						negatively associated with students' academic performanc e.
31	Lu, (2022)	It has been tried to determine whether skilled immigrants can replace natives in the labor market in the USA.	US Census and American Community Survey (ACS) data were used..	LASSO, regression	Unspecified	The PDS Lasso estimate gives robust and precise estimates of substitution and categoricall y rejects that immigrants can replace natives. Based on these improved substitution estimates, the wage effects of immigration are simulated with confidence.
32	Martey & Armah (2021)	It examines the impact of international migration on household expenditures, labor outcomes and poverty on left-behind household members in Ghana.	The Ghana Living Standard Survey and World Bank employment data were used, which include data on household demographics, migration, income and expenditure, agricultural production, and household investment decisions.	LASSO , Double Selection	Unspecified	The results show that internationa l migration increases household spending, largely driven by investments relative to spending on consumer goods. In addition, the transfer of remittances by

						immigrants to lagging households results in reduced weekly working hours in the labor market and modestly reduces household poverty.
33	Micevska, (2021)	It was carried out in order to understand which factors play an important role in forced migration abroad in African countries.	A panel dataset containing data on forced migration from 45 African countries covering the years 1999–2017 was used. In this context, UNHCR Population Statistics (data on refugee stocks and asylum applications), Political terror index, Armed Conflict Location and Event Dataset, population data obtained from the World Development Indicators database provided by the World Bank, and measurements of climate conditions are included.	Random Forest, LASSO	Unspecified	The importance of the changing nature of the conflict in Africa has been demonstrated. It was emphasized that internet access rate is an important factor in explaining asylum applications. In Nigeria and Cote d'Ivoire, insurgency appears to play a major role in migration, while in Guinea, Nigeria, Gambia and Mali, population appears to be a major driver for migration. Algeria and

						the Democratic Republic of Congo are important sources of asylum applicants due to factors not included in the model.
34	Molina, vd. (2022)	It was carried out to determine the weather variables that best predicted the migration decisions of 54,986 people of Mexican origin between 1989 and 2016.	Data from the Mexican Migration Project (MMP), which included the biographies of 100,572 people between 1980 and 2016, were drawn from 161 communities located in major migrant-sending regions in 21 states of Mexico. MMP data were combined with daily gridded forecasts of weather data from 1980 to 2016 from the Oak Ridge National Laboratory Distributed Active Archive Center, one of the NASA Earth Observation System Data and Information System data centers.	Random Forest, Logistic Regression	Python	The model that includes all weather indicators outperforms the base model with only individual, household, and community-level features. Parametric models have not lagged far behind more complex methods such as random forests in predicting the migration process. The analysis identifies the most predictive weather features, but does not identify the direction of the effects

						on migration.
35	Nair, vd. (2019)	MM4SIGHT, a machine learning system that provides predictions about mixed migration, defined as refugees fleeing persecution and conflict in Ethiopia, victims of trafficking, and migrants seeking better lives and opportunities, was analyzed.	Net migration figures from the United Nations Department of Economic and Social Affairs (UNDESA), estimates of the United Nations High Commissioner for Refugees (UNHCR) regarding refugees, asylum seekers, other persons concerned and returnees, Danish Refugee Council survey estimates including unofficial counts of migrants moving to Saudi Arabia and data from 4MI's migration clusters and a number of institutional data providers such as the World Bank, UNHCR are used. Spatially, only data on Ethiopia was tested, followed by a dataset enriched with 21 other countries from Sub-Saharan Africa.	Cross-correlation, a gradient boosting ensemble (xgboost), random forest, a linear regression, and a support vector regression.	Unspecified	Error rates are shown so that annual forecasts for migration flows from Ethiopia to the six countries are within a few thousand persons per year for most destinations . Under the current modeling framework, causality between macro indicators and the resulting mixed migration flows has not been inferred.

36	Olberg, & Seuken, (2022)	For a project initiated by the Swiss State Secretariat for Migration for the resettlement of refugees, mechanisms that address families' preferences for resettlement locations were examined.	Historic resettlement data	Unspecified	Ubuntu 18.04	Two mechanisms have been proposed, constrained random serial dictatorship mechanism (CRSD) and constrained rank value mechanism (CRV). It shows that the CRV is generally superior in terms of family well-being where families have a complete order of preference over the locations to be placed. Preliminary simulations of the data show that both mechanisms can significantly improve family well-being with only a small loss in the overall employment rate of refugees.
37	Quinn, vd. (2018)	The settlements of refugees and IDPs in various locations in Africa and the	Remote sensing data from satellites was used.	Mask-RCNN model	ResNet101. ImageNet	In the analyzes, it has been shown that it is possible to detect a

		Middle East were analyzed.				large part of the structures in the examined settlements. However, it has been revealed that there are still significant differences in the characteristics of the images collected from different satellite sensors, geographical regions and types of settlements.
38	Rahaman, vd. (2022)	The relationship between the growth trend of Rohingya refugee camp settlements and deforestation in the Cox's Bazar region is discussed.	Landsat 5 and Landsat 8 satellite images of 1990, 2013 and 2020 were collected. Control was provided with Google Earth historical images.	maximum likelihood classification, support vector machine, random forest and artificial neural network	ArcGIS, Neural Net at ENVI	Shallow vegetation has increased over time, indicating forest areas close to the settlement camps. The greatest loss has occurred in deep forest areas, where both the waterbody and pasture land have been significantly reduced.
39	Raman, Vera, & Manna (2022)	Nearly 6 million immigration	The publicly available dataset of the	Random forest classifier, linear support vector	Python	While represented asylum

		court proceedings and 228 cases in the United States were analyzed.	US Executive Office of Immigration Review (EOIR), which includes information on litigation, asylum seeker representation, and child applicants, and Transactional Records Access Clearinghouse (TRAC) website resources, which contain detailed biographies of immigration judges were used These are limited to cases filed up to January 2022 after the implementation of the 1980 Refugee Act.	classifier, regression		seekers can get asylum, unrepresented asylum seekers cannot apply for asylum and are more often detained. The model found that the less partisan cases were associated with the detention at the time of the verdict of an asylum seeker from a Latin American country who was tried in a state that had voted Republican in the previous presidential election. The results show that asylum grants are associated with high partisanship and rejections are associated with low partisanship . Consistency scores of female

						judges were found to be significantly lower than male judges.
40	Ran, vd. (2022)	The effect of migration in China on online social behavior and the mediating effect of immigrants' characteristics are discussed.	We studied users' data from a large four-month dataset containing 2.29 million records from online social networks (OSN), one of the largest online social networks in China. The dataset also includes the location based on the IP address the users are logged into.	Propensity scores matching technique and difference-in-differences analysis (PSM-DID)	SQL, R	The findings reveal that migration affects individuals to communicate with more friends and receive more social support from different people, but their communication power decreases, which may be limited by cognitive and time resources. Migration increases the number of people added as friends, but reduces the number of messages. The characteristics of migrants, including gender, age and degree, play a moderating role.

41	Ren & Bloemraad, (2022)	Machine learning was used to identify immigrant-oriented nonprofits.	A dataset of immigrant-oriented organizations (data collected by GuideStar) and a second dataset of non-immigrant-oriented organizations (986 nonprofits from the National Center for Charitable Statistics (NCCS) dataset) were used. In addition, the ICEP dataset from the Immigrant Civic Engagement Project, the NY dataset extracted from the "A Guide to Community-Based Organizations for Immigrants" file produced by the Department of Education of New York was used.	Natural language processing, machine learning techniques	Python	Simpler NLP models are poor at categorizing and distinguishing between immigrant and non-immigrant nonprofits. Also, dictionary-based methods that use preset keywords for country name, nationalities, and migrant-specific words have been shown to perform better. The machine learning categorization strategy works much better in both validation datasets, accurately identifying 90 percent and 67 percent of immigrant-oriented nonprofits.

42	Ruhnke, vd. (2022)	A new method of attribution of legal status for the health of the undocumented population in the USA is discussed by comparing various analyzes.	The NHIS, a stratified random sample of the largest health survey of the US population, was used. Variables such as years in the USA, age, education attainment, poverty status, Medicaid coverage, household size, spousal citizenship, region of birth, marital status, difficulties speaking English, number of children, employment status, race, and Hispanic ethnicity were discussed.	Logistic regression models, Random Forest algorithm	R	The results suggest that undocumented immigrants experience a health advantage compared to the U.S.-born population. Undocumented respondents were significantly more likely to report their health as excellent than their US-born and documented immigrants, and less likely to describe their health as fair or poor.
43	Ruhnke, Wilson, & Stimpson (2022)	A new machine learning method has been developed to give legal status to immigrants.	Nationally representative survey data from the Survey of Income and Program Participation (SIPP) and the National Health Interview Survey (NHIS) were used.	KNN classifier, RF, Logistic regression	R	The results show that using machine learning to determine the legal status of immigrants, especially the Random Forest Algorithm, is more accurate in identifying undocumented immigrants and

						minimizes bias. The undocumented population in the NHIS was found to be, on average, younger, more likely to be Hispanic, and of lower socio-economic status than those with documents.
44	Simionescu, (2021)	The situation of Italian immigrants since 2008, when the global financial crisis began, has been analyzed using microeconomic, macroeconomic and big data.	Microeconomic data and surveys (including questions such as sociodemographic information, labor force) provided by the National Institute of Statistics (Istat) in Italy were used. In addition, exports by Eurostat, persons at risk of poverty or social exclusion, employment in knowledge-intensive activities, etc. data is provided.	Regression model	Unspecified	In the Italian labor market, women are more likely to be employed than men. The economic crisis in Italy affected the agriculture and construction sectors and brought about a decrease in the number of men. Immigrants who graduated from post-secondary education levels were less likely to be employed than those with higher

						education or lower. Immigrants from higher-income households prefer to wait longer until they find a better-paying job.
45	Szocska, vd. (2021)	It was carried out with the aim of monitoring mobility changes during the COVID-19 outbreak in Hungary.	Call Detail Records (CDR) for the period between February 1, 2020 and May 20, 2020 were used. Data were also compared to Google's Covid-19 Community Mobility Reports.	Unspecified	Microsoft Power BI Pro	While a significant decrease in movement is seen on weekends, the decrease is most marked on Sundays. At the end of March, people who had a weekend house near Lake Balaton (one of the most popular local tourist destinations in summer) decided to move in large numbers from Budapest and other cities in order to stay in quarantine there.

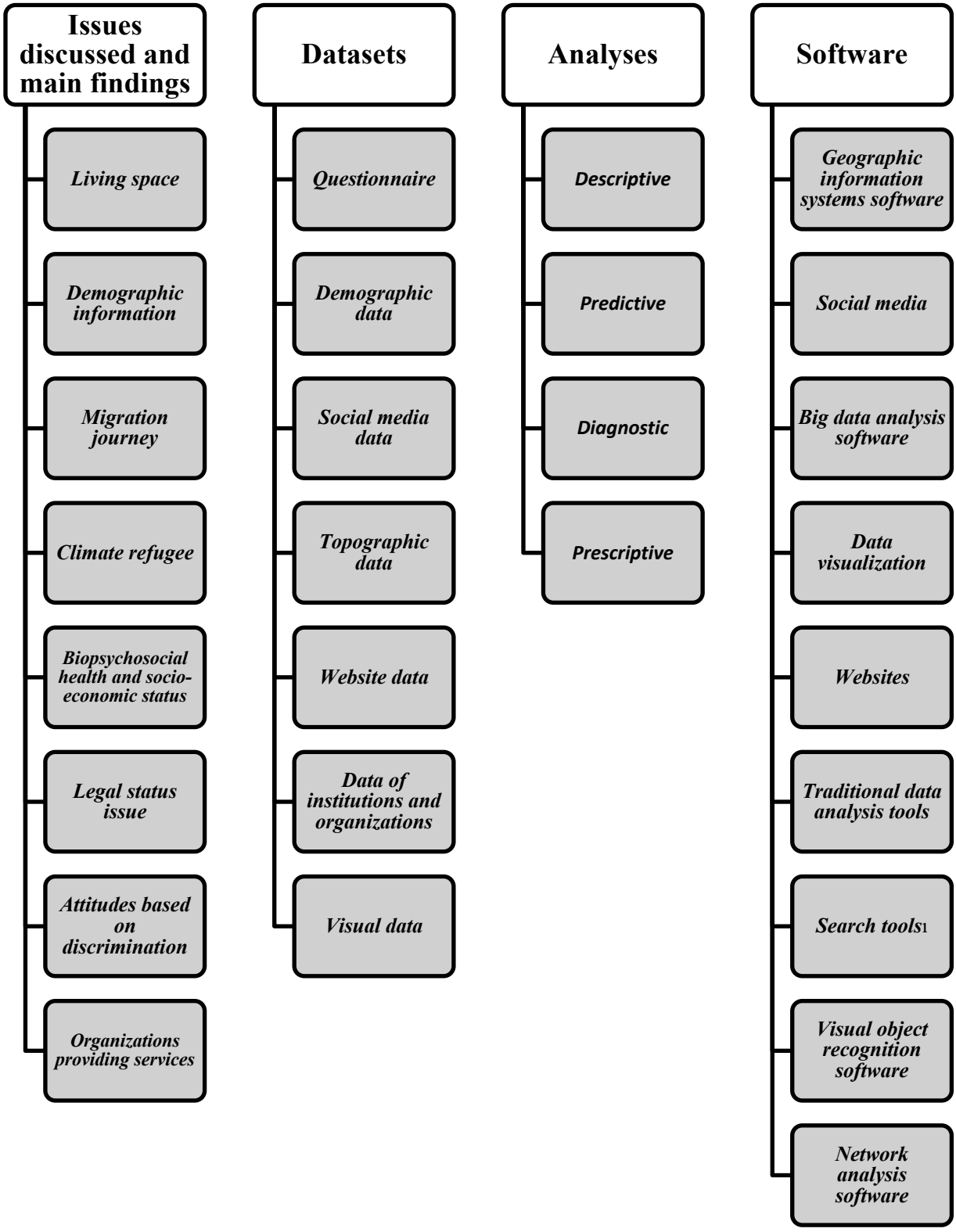
46	Weber, (2020)	It was carried out with the aim of developing a model that predicts both domestic and international migration for two demographic groups: youth aged 18-24 and families aged 30-49 including underage children.	Data on economic, demographic data and other characteristics such as distances to large cities or universities were collected for approximately 3,000 German municipalities.	Linear regression, random forest, extreme gradient boosted tree, deep neural network	R, TensorFlow	While it can predict the future net migration of young people aged 18 to 24 quite well, family migration is not as well predicted. When a high proportion of foreigners coincides with a rural setting, these rural municipalities with large numbers of foreigners actually attract slightly fewer young people than urban and wealthy municipalities with few foreigners. Families leave areas with high child poverty and high unemployment.
47	Wilson, vd. (2020)	An alternative methodological approach is presented to examine the differences in healthcare use and	2016-2017 Medical Expenditure Panel Survey (MEPS), is a large-scale, nationally-represented in-	Random forest classifier	STATA	Immigrants have been found to have significantly lower healthcare expenditure

		expenditure between unauthorized and authorized immigrants and US-born individuals.	person survey administered by the Agency for Health Care Research and Quality data were used.			s than US-born individuals. It has been estimated that around half (47.1%) of unauthorized immigrants are uninsured; this is significantly higher than the rates of authorized immigrants (15.9%) and US-born individuals (6.0%).
48	Xiao, vd. (2021)	The research was conducted to better characterize the population distribution pattern that takes place during the Spring Festival travel season.	Population movement data collected by a mobile application belonging to Tencent, China's largest social media company, was used.	spatial analysis methods and regionalization models	Unspecified	The population flow during the Spring Festival Transport in China is mainly concentrated east of the Hu Huanyong Line, which has played a relatively stable role in describing the differences in population flow in China. Net population exit areas are mainly concentrate

						d in the easternmost developed regions, some state capitals and some western regions.
49	Zhou, (2021)	It was conducted to investigate the discrimination experienced by rural immigrants in urban China at work.	2010 census data, data of residences and workplaces of 296,796 rural migrants and 581,731 residents from other groups obtained from cellular network data, and Open data from the Chinese Family Panel Studies (CFPS) were used.	Linear regression analysis, ANOVA	Unspecified	Results show that rural migrants experience higher levels of discriminati on in many areas than other groups. Also, rural migrants living in outer suburban areas are more likely to be isolated from local residents in residential neighborho ods and workplaces. Among all service sector employees, those working in suburban areas may experience lower work-place exposure.

These categories were divided into subcategories. These subcategories are presented in Figure 2.

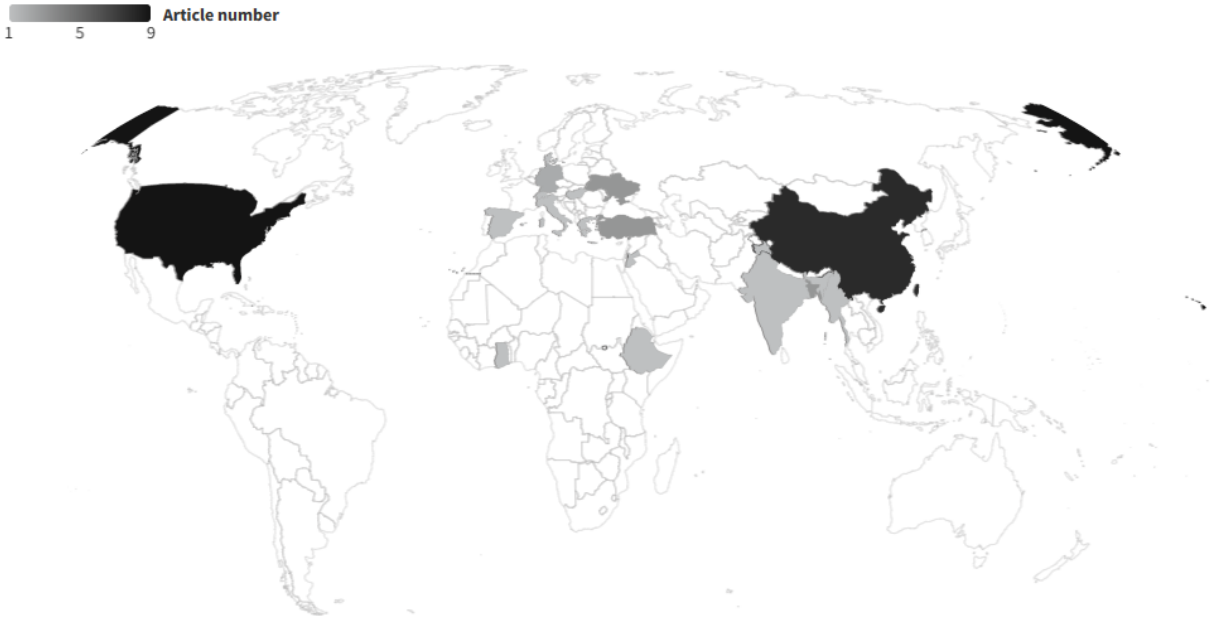
Figure 2. Categories and subcategories determined for research.



2.1 Issues Discussed and Main Findings

The issues discussed in the studies that were included in the analyses using big data analysis tools regarding the migration experiences of immigrants, asylum seekers, and refugees and the countries where these were experienced were visualized and highlighted on a world map. In this context, it has been observed that studies have been conducted mainly with migrants and refugees in the USA and China, while research has focused on asylum seekers in countries such as Turkey and Ukraine.

Figure 3. Distribution map of articles on immigrants, asylum seekers and refugees by country



2.1.1 Living Spaces

In the included studies, the living spaces of refugees and immigrants were examined, and it was observed that these spaces were separated as on-camp and off-camp. For instance, in a study that focused on the on-camp lives of these groups and examined the infrastructures of the camps, the Normalized Difference Vegetation Index – an index which analyzes and reveals the intensity of the vegetation in a certain place and whether it is healthy or not – was found to be low (Ahmed et al., 2020). In another study, similarly, deforestation and significant decreases in water resources and pasture lands were determined in these camps (Rahaman et al., 2022). In a study conducted by Quinn et al. (2018), it was reported that it was possible to determine a great majority of the structures in the encampments through big data analysis, but there were still significant differences in the properties of the images collected from geographical regions and settlement types using different satellite sensors.

2.1.2 Demographic Data

Analyses in the included studies were mostly performed by making estimations regarding the demographic data of immigrants, asylum seekers, and refugees, and models were established. For example, in one study, the ethno-religious diversity of an Indian diaspora population was

investigated, and it was seen that much of the total diaspora population lived in Asia (Garha and Domingo, 2019). In another study, by using social network data to obtain the demographic information of asylum seekers and refugees in the case of forced migration from Ukraine to the EU, it was revealed that Facebook and Instagram users among the Ukrainian asylum seekers and refugees in Poland, Slovakia, Hungary, Moldova, Romania, and Germany were mostly individuals in the age range of 25-44 years (Juric, 2022a; 2022c). In the study conducted by Ahmed et al. (2019), it was concluded that the best option for population estimation was the approach based on satellite images. Aydemir (2022) showed that the XGBoost algorithm displayed a high level of success in estimating the number of immigrants. In another study conducted by Ruhnke et al. (2022), it was found that the Random Forest Algorithm was a healthier tool in identifying undocumented immigrants.

2.1.3 Immigration Journey

Another issue which was frequently examined in the included studies was the immigration journey of immigrants, asylum seekers, and refugees. This category covered their pre-immigration decisions, experiences during immigration, post-immigration process, and those left behind. Kiossou et al. (2018) examined how the immigration process was influenced by driving forces, and it was found that the more distant the two countries were from one another, the lower the rate of the relevant immigration flow was. Analyses were performed to estimate the national and international immigration mobility of families who had children aged between 18 and 24 years, family members between 3 and 49 years, and minor children in Germany, and in the model that was established, it was demonstrated that families left regions where there were high rates of child poverty and unemployment (Weber, 2020).

In a study in which immigration movements from the Near East towards Central Europe between 2015 and 2016 were examined, the hot spots were mostly detected either on the routes of the refugees and immigrants towards Central Europe or the Western border of Turkey where most refugees and immigrants started their journey (Havas et al., 2021). Analyses revealed that refugees and immigrants with moderate and high socioeconomic status used a wide area, they travelled long distances, and those with low socioeconomic status seemed to be trapped in small neighbourhoods (Kilic et al., 2019).

It can be stated that the use of technological tools in obtaining information about immigration processes has become widespread. Those who migrate from Ukraine search terms such as Western Union, immigration, immigrant, and Schengen on Google, and they research videos about places to immigrate, especially Germany (Juric, 2022a; 2022b; 2022c). It was seen that all these Google searches, which include an indicator for planning immigration, were in parallel with the data obtained from official statistics (Juric, 2022a). In a study in which the factors that played a significant role in international forced immigration in African countries were examined, it was emphasized that internet usage rates were an important factor in explaining immigration applications (Micevska, 2021).

Technology use also has an important role not only in the pre-immigration period but also in the integration process following immigration. Ran et al. (2022) analyzed the effects of immigration in China on online social behaviors and revealed that immigration increased the number of individuals added as friends on social media, but it reduced the number of messages.

Various algorithms and models are used in the analysis of the immigration processes of immigrants, asylum seekers, and refugees. Molina et al. (2022) demonstrated that the model which included weather condition indicators performed better than the basic model which involved only the properties at the individual, household, and community levels. Emami et al. (2020) performed various analyses to examine the mobility in the southwest of Iran and found that according to the LASSO algorithm, elevation, slope, and distance to roads and fault lines were effective factors. Here, the Random Forest Model was found to be the best model for mass movement sensitivity mapping. Olberg and Seuken (2022) examined the mechanisms which dealt with the settlement preferences of families regarding the relocation of immigrants and refugees in Switzerland and demonstrated that the CRV model was generally superior in terms of family welfare even in cases where the families had a certain order of preferences. In the study conducted by Nair et al. (2019), simulated experiments were examined to estimate labor force export in Taiwan, Japan, and Korea, it was seen that in comparison to the real data in Taiwan, Korea, and Japan, Back-Propagation Neural Network findings yielded the closest results.

Immigration affects not only immigrants but also those left behind. In the analyses performed by Martey and Armah (2021), it was determined that money transfers made by immigrants to the household left behind caused a decrease in weekly working hours in the labor force and reduced household poverty by a moderate degree.

2.1.4 *Climate Immigration*

One of the causes of immigration is the climate crisis which has gained rapidity in recent years. Best et al. (2022) ran analyses to estimate the immigration status in Bangladesh, which is considered one of the countries most vulnerable to climate change. Their results showed that the total number of individuals in a household had a negative effect on immigration, and the number of unemployed individuals in a household increased immigration by the head of the household. In a study conducted around Lake Urmia in Iran, it was revealed that individuals had to migrate due to reasons such as unemployment and financial difficulties, the emergence of various diseases, the depletion of drinking water, and air pollution (Khangahi and Kiani, 2021). In the study conducted by Kiossou et al. (2020), artificial neural networks were used to understand how immigration was affected by driving forces, and the effect of draught in destination countries on immigration flow was examined.

2.1.5 *Biopsychosocial Health and Socioeconomic Status*

Immigration affects the psychological, biological, cultural, and economic aspects of the lives of refugees and immigrants to different extents. In a previous study, whether there was an increase in risky behaviors among individuals who were displaced and settled in Germany was examined, and it was found that displaced individuals had experienced at least one trauma, and they were exposed to various forms of violence, with physical violence in the first place (Augsburger & Elbert, 2017). The mental health of children, as a vulnerable group affected by immigration, was examined using the LASSO estimation method based on their drawings. The results showed that children's drawings could be used as a diagnostic tool in crisis environments (Baird et al., 2022).

In the immigration process, in addition to experiencing traumatic events, new crises such as the COVID-19 pandemic have affected the mental health of immigrants, asylum seekers, and refugees. Using artificial neural networks, it was found that the psychological resilience of these groups was an important determinant in their overcoming of mental problems (Choi et al., 2020).

In addition to mental health, in studies conducted on physiological health, the prevalence of infectious diseases was investigated. The most frequently diagnosed disease involved intestinal parasites (Fernández-Martínez et al., 2021).

It is generally believed that the status problem experienced by immigrants, asylum seekers, and refugees affects their health service use and health expenditures. Wilson et al. (2020) reported that these groups had lower health expenditures compared to individuals born in the US where they immigrated, it was estimated that approximately half of unregistered immigrants worked without insurance, and it was determined that this rate was significantly higher compared to the rates of registered immigrants and individuals born in the US (Wilson et al., 2020). Using regression models, Simionescu (2021) determined that immigrants with education degrees higher than high school had lower chances of being employed compared to those with education degrees of high school or below, and those who came from households with higher income preferred to wait until they found a better-paying job.

2.1.6 Legal Status Problem

Ruhnke et al. (2012) showed that using machine learning, especially the Random Forest Algorithm, to determine the legal status of refugees provided more accurate results in identifying undocumented refugees and minimized prejudice. Additionally, it was found that the undocumented population included in the study were younger on average and had lower socioeconomic status compared to those with documents.

The problem of legal status emerges in different guises as well. In the study conducted by Raman et al. (2022), immigration court processes and cases in the US were examined, and it was revealed that refugees who were not represented by lawyers had lower levels of access to immigration rights and were kept under custody longer compared to those represented by lawyers. Katsikolui et al. (2022) analyzed refugee cases in court decisions in Denmark and found that the Random Forest Classification was used over information such as the cultural background of the applicants, the year when they entered the country, and/or the year when the cases were appealed, and it was used to estimate whether the Immigration Appeals Board would reverse the judgment or not.

2.1.7 Discriminatory Attitudes

Among the leading problems that immigrants, asylum seekers, and refugees experience are discrimination, labeling, and ostracization. In the study conducted by Alsan et al. (2019), comments made on news websites were examined to determine attitudes towards these groups, and it was found through the Perception Analysis Model that labeling and derogatory language were used in most comments.

Daily discrimination experienced by immigrants, asylum seekers, and refugees has countless negative effects on their mental health in numerous contexts (Choi et al., 2020). These adversities differ according to the region where this population lives. Immigrants in rural areas experience discrimination at a higher level compared to local residents, and they are mostly deprived of opportunities for integration with other groups (Zhou, 2021).

Political arena is an important determinant in attitudes towards immigrants, asylum seekers, and refugees. Especially election processes should be evaluated in this sense. For example, whether the attitudes of the media towards refugees and immigrants changed before, during, and after the

election campaign of Donald Trump who took office in the US elections of 2016 was investigated, and it was found that following the elections, the volume of the news on immigration almost doubled compared to other news stories, and there was a striking increase in the use of the theme of crime in these news stories (Krupenkin and Rothschild, 2019).

2.1.8 Organizations Providing Services

In a study conducted to define immigrant-oriented non-profit organizations, it was found that basic NLP models were weak in terms of differentiation and categorization between immigrant and non-immigrant non-profit organizations, and the machine learning categorization strategy performed better and accurately estimated 90% and 67% of immigrant-oriented non-profit organizations (Ren & Bloemraad, 2022). Additionally, in a study that examined health service use and health expenditures in the context of unregistered immigrants, registered immigrants, and those born in the US, lower mean annual health expenditures per person and lower rates of receiving in-patient services were found among unregistered immigrants.

2.2 Datasets

The datasets used in the articles were subcategorized as “survey”, “demographic data”, “social media data”, “topographical data”, “website data”, “organization and institution data”, and “visual data.” In the included studies, various datasets were used to analyze issues related to immigrants, asylum seekers, and refugees. In these datasets, in addition to the use of surveys (Augsburger & Elbert, 2017; Choi et al., 2020; Lee & Song, 2022) and basic demographic data of populations (Aydemir, 2022; Fernández-Martínez et al., 2021) as in conventional research, big data such as social media data were utilized. In this context, GPS data of smart phones (Gao, Nan & Song, 2022), Facebook (Garha & Domingo, 2019), Twitter (Havas et al., 2021), YouTube (Juric, 2022b), Google Trend data (Juric, 2022a), telecommunication service provider company data (Kilic et al., 2019; Szocka et al., 2021), and Google Earth data (Rahaman et al., 2022) were used. Moreover, datasets such as sets of geological data on regions where immigrants, asylum seekers, and refugees were located (Ahmed, Firoze & Rahman, 2020; Emami et al., 2020), weather data (Molina et al., 2022), satellite images used to track and estimate the mobility of these groups (Ahmed et al., 2019; Quinn et al., 2018), and news websites and news contents (Krupenhin & Rothschild, 2019) were also used. Datasets were also obtained from national institutions such as ministries and international organizations such as the World Bank (Giang et al., 2022; Martey & Armah, 2021), projects (Azizi, Ngwaba & Ekhator-Mobayode, 2021; Carammia, Iacus & Wilkin, 2022), national research (Best et al., 2022), and national statistics institutions (Kilic et al., 2019; Ruhnke et al., 2022). Furthermore, data obtained from organizations working on immigration (UNHCR, IOM) (Havas et al., 2021; Micevska, 2021) were employed. In addition to textual data, visual data were also analyzed. In this context, children’s drawings (Baird et al., 2022) and satellite images (Rahaman et al., 2022) were used.

2.3 Analyses

Among the models that were used in the articles that were examined, it was seen that descriptive, predictive, diagnostic, and prescriptive analyses were used to process, refine, and analyze the data. Descriptive analyzes were made to reveal the current situation or past experiences about immigration and immigrants. Aslan et al. (2019) examined comments in the newspapers and revealed the opinions of people about refugees. Jurić, (2022c) carried out the research to obtain sociodemographic information about Ukrainian. Various analyzes of migration processes have been used in predictive studies. Ahmet et al. (2020) machine learning analysis was used to predict the

landslide risk of the camp infrastructure. Weber, (2020) carried out the research with the aim of developing a model that predicts both domestic and international migration. Azizi et al., (2021) carried out research to predict how many people will immigrate and how long they will stay in the United States. Diagnostic analyzes were made to reveal the causes of migration. Kiossou et al., (2020) established a model to gain deeper insights into how migration is affected by its drivers. It has been tried to deal with whether the coverage of the media regarding immigration has changed before the campaign, during the campaign and after Donald Trump was inaugurated in Krupenkin & Rothschild (2019)'s research. Various analyzes of migration processes have been used in prescriptive analytics. Gao et al., (2022) examined Tencent big dataset to provide new causal evidence about the impact of improved transportation on urban tourism. Huang & Shao, (2022) focused on the application of big data statistics to solve the construction of the Arab migration and entrepreneurship data system.

Besides these in the studies, machine learning analyses were frequently used. In this context, among supervised machine learning algorithms, regression algorithms (Linear Regression), classification algorithms (Support Vector Machine, Support Vector Classifier, K-nearest neighbors models, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest), and among unsupervised machine learning algorithms, clustering algorithms (K-means algorithm), Artificial Neural Network analyses, Prediction algorithms (XGBoost algorithms) were used.

2.4 Software

The subcategorization related to the software used in the articles were as follows: "geographic information system software" (ARCGIS, ArcMap, ENVI, QGIS, Google Maps, Google Places API), "social media applications" (Facebook, Instagram, Twitter, YouTube), "big data analysis software" (JavaScript, R, SQL, NoSQL, MySQL Google Colab, ResNet101, Stanford CoreNLP, TensorFlow), "data visualization" (Microsoft Power BI Pro, Timemap), "websites" (GOHAKKA), "conventional data analysis tools" (PSPP, SPSS, STATA), "screening tools" (Google Trends, Spotlight), "visual object identifying software" (ImageNet), and "network analysis software" (Pajek, Gephi).

3. DISCUSSION

This systematic review presents significant evidence regarding the studies on immigrants, asylum seekers, and refugees by using big data analysis tools. The included studies offer indicators on how information can be obtained more easily about the population in question, a group difficult to access due to their numbers, by utilizing these tools. In the final analysis, it is seen that various opportunities provided by big data sets are attractive to researchers due to limitations such as language differences, ambiguity due to lack of documentation and lack of clear information in studies conducted with these groups. Although international migration data provided by governments is limited in the relevant studies, it is seen that data sets such as data accumulated by public institutions over the years, population indicators, etc. -used in traditional analyses- are used, as well as different big data sets such as social media data and geological data.

The findings of the study revealed significant findings in terms of both strengthening the immigration and refugee policies of states and developing the literature on immigration in a scientific sense thanks to the big data sets that allow data acquisition and analysis on large masses of people quickly and at less cost compared to traditional social science research. The data obtained in studies through big data analysis tools point to a series of positive effects ranging from the access of immigrants, asylum seekers, and refugees to health services to eliminating uncertainties and facilitating their integration. Choi et al. (2020) presented the psychological resilience and coping

mechanisms of immigrants and emphasized the importance of developing policies that aim to strengthen the mental health of immigrants. Simionescu (2021) used microeconomic and macroeconomic big data and revealed the possibilities of immigrants to be employed with high wages. Ahmed et al. (2020) shared the geographical structure of immigrant camp sites and drew attention to the risks in these areas and improvements that should be made. Ruhnke et al. (2022) shed light on policies that deal with the importance of a new legal status for the access of undocumented immigrant groups to health services.

On the other hand, it should be noted that as per its nature, big data use involves various risks as well. In particular, ethical problems can be encountered in terms of the violation of the confidentiality of the study groups, labeling, and increasing the extent of surveillance. Same problems are also experienced in studies based on big data that aim to contribute to immigration and refugee management practices. In some cases, beyond ethical problems, collecting, processing, and including data on immigrants in algorithm production can lead to more devastating outcomes for them, including the endangerment of their lives. In fact, data on immigrants can be obtained by many public institutions from various sources such as censuses, smart phones, social media networks, and border passes. When access to big data becomes easy and open to the public, there arises a risk that these data could be obtained illegally or by uncontrolled companies, organizations engaged in human trafficking and illegal immigration, and even by states at war with one another. When this predicament is considered, it becomes critically important that big data analyses and production be performed with more sensitivity and based on the principle of "do no harm" (ICRS & Privacy International, 2018). The findings of this study revealed that some studies included in the sample showed vulnerabilities among these groups in terms of these sensitivities. According to Raman et al. (2022), if a minority group is statistically disproportionately marked as an outlier, and statistical minorities are combined with social minorities, this could lead to discriminatory findings. In a study conducted in the US to design an immigration policy that would prevent crimes, in which the risk of retrial of non-US citizens who were sentenced due to various crimes was estimated, the risk of juxtaposing immigrants with the phenomenon of crime emerged (Bertsimas & Fazel-Zarandi, 2021). As in the data used in the study performed by Szocska et al. (2021), the use of datasets detailed enough to track the changes at the level of individual settlements in smaller geographical regions shows that surveillance tools have become more powerful.

Another risk of the big data phenomenon is that studies conducted with big data analyses have significant methodological handicaps. These can be seen as problems in representation, accuracy, excessive homogenization, and easy generalization. The findings of this study showed that some of the studies included in analyses had some of these handicaps. In the study conducted by Garha and Domingo (2019), while Facebook data were used for diaspora studies in the context of India, the researchers ignored the possibility that these data could have been affected by the problem of representation of women and elderly who had limited internet access. The data used in the study carried out by Juric (2022a) depended on internet access in a certain group and did not cover all age groups to the same extent. Another problem in their study was that users could have multiple unconnected Facebook and Instagram accounts, which would lead to data discrepancies. In another study by Juric (2022c), it was pointed out that Youtube search index analyses do not show the exact number of searches in a given country, and that the exact number of potential migration flows cannot be calculated using this tool, only the increase in the trend can be recognized with certainty. In the study conducted by Szocska et al. (2021), call detail records were used, and these were compared against Google's Community Reports. However, as Google detects only Android users and among them, only those whose location history is turned on, some users were excluded. In the study performed by Wilson et al. (2020), as the LAFANS data that were used had been collected more than a decade ago, the possibility that the characteristics of immigrants could have changed since

then was not considered. Furthermore, the representation ratio of undocumented immigrants was quite low.

Based on all this information, while interpreting the findings of the present review, three factors should be considered. The first factor is that big data-based studies on migrants, asylum seekers and refugees contribute to a better quality of life for these groups. The second factor is that it was revealed that these studies may create harmful consequences for these groups. Finally, it was shown that these studies have various methodological handicaps.

Limitations

This study had certain limitations. Following the matching of key terms, the references of the studies that were selected could have been examined, and other relevant articles could have been reviewed through hand-search. Secondly, the research pool could have been expanded by including more databases in the search strategy. Thirdly, the number of key terms subjected to search could have been increased, and more matches/combinations could have been utilized. For example, terms such as artificial intelligence and simulation that could be relevant to such studies could have been used as key terms. Moreover, only scientific research articles written in English were included in the pool of studies to be examined, but articles written in other languages could have been included within the scope of the study. Finally, articles published before 1 January 2023 were included in the review. Other articles published in 2023 could also have been examined.

4. CONCLUSION

The main emphasis of socio-historical transformation pointed out by sociological theories such as "information society", "information age" and "post-ideologies", which describe the post-industrial society, is about the digitalization of our age. This digitalization is of interest to the discipline of sociology, because it not only affects factors such as relationship networks, consumption habits, and production styles in societies but also transforms the understanding of social problems. The phenomenon of migration, which is the most prominent social problem of our day, is also affected by digitalization processes. In addition to negative factors such as surveillance, social control, and labeling brought by digitalization at all stages of the migration process, positive factors such as facilitating access to services, rapid access to data, and easy identification of masses can be kept in mind.

This study demonstrated that scientific studies conducted on immigrants, asylum seekers, and refugees using big data analysis tools could yield beneficial results for these groups, but they could also involve various risks. Therefore, while revealing data analyses that would facilitate the immigration of these groups and their integration into the host starting from immigration mobility, studies to be conducted in this regard should adhere to ethical considerations so that the rights of these groups stemming from national and international law are not violated or infringed upon. Additionally, there is a need for descriptive and predictive studies to be conducted with sets of big data and big data analysis tools on issues such as access to welfare services, family reunions, fight against discrimination, cultural adaptation, and social inclusion for immigrants, asylum seekers, and refugees, especially vulnerable groups. Consequently, the results of this study showed that the models used in big data analyses in the studies examined here were still in the testing phase. In this context, new models should be established, and by strengthening existing models, models that have high accuracy in these populations need to be shared. Using models whose accuracy has been ensured and that are commonly used is an important factor in terms of methodological consistency and considering that two of the most important problems regarding big data are easy generalizability and representativeness, there is a need for models with high representativeness.

REFERENCES

Ahmed, N., Diptu, N.A., Shadhin, M.S.K., Jaki, M.A.F., Hasan, M.F., Islam, M.N. and Rahman, R.M. (2019). Artificial neural network and machine learning based methods for population estimation of Rohingya refugees: Comparing data-driven and satellite image-driven approaches. Vietnam Journal of Computer Science, 6(04), 439-455. DOI: 10.1142/S2196888819500246.

Ahmed, N., Firoze, A. and Rahman, R.M. (2020). Machine learning for predicting landslide risk of Rohingya refugee camp infrastructure. Journal of Information and Telecommunication, 4(2), 175-198. DOI: 10.1080/24751839.2019.1704114.

Aslan, P. and Ertem Eray, T. (2019). How to analyze big data: a study on understanding what the Turkish think about Syrian refugee crisis. Journal of Selçuk Communication, 12(2), 763-780. DOI: 10.18094/josc.596301.

Atar, E. (2021). Systematic analysis of the advantages and disadvantages of using big data in the context of international migration and refugees. Alternative Policy, 13(1), 146-174.

Augsburger, M. and Elbert, T. (2017). When do traumatic experiences alter risk-taking behavior? A machine learning analysis of reports from refugees. PLoS ONE, 12(5), e0177617. DOI: 10.1371/journal.pone.0177617.

Aydemir, B., Aydın, H., Çetinkaya, A. and Polat, D.Ş. (2022). Predicting the Income Groups and Number of Immigrants by Using Machine Learning (ML). International Journal of Multidisciplinary Studies and Innovative Technologies, 6(2), 162-168. DOI: 10.36287/ijmsit.6.2.162.

Azizi, S., Ngwaba, C.A. and Ekhator-Mobayode, U.E. (2021). Can Machine Learning Predict Quantity and Duration of Migration to the USA? The Journal of Prediction Markets, 15(1), 97-107. DOI: 10.5750/jpm.v15i1.1859.

Baird, S., Panlilio, R., Seager, J., Smith, S. and Wydick, B. (2022). Identifying psychological trauma among Syrian refugee children for early intervention: Analyzing digitized drawings using machine learning. Journal of Development Economics, 156 (1), 102822. DOI: 10.1016/j.jdeveco.2022.102822.

Baym, N. (2010). Personal Connections in a Digital Age. Cambridge: Polity Press.

Bell, D. (1999). The coming of post-industrial society. New York: Basic Books.

Bertsimas, D. and Fazel-Zarandi, M.M. (2021). Prescriptive machine learning for public policy: The case of immigration enforcement. Computer Sciences. Under Review.

Best, K., Gilligan, J., Baroud, H., Carrico, A., Donato, K. and Mallick, B. (2022). Applying machine learning to social datasets: a study of migration in southwestern Bangladesh using random forests. Regional Environmental Change 22(2), 52. DOI: 10.1007/s10113-022-01915-1.

Beyer, M.A. and Laney, D. (2012). The importance of 'big data': A definition. Gartner Report. Available at: https://www.gartner.com/doc/2057415/importance -big-data-definition.

Carammia, M., Iacus, S.M. and Wilkin, T. (2022). Forecasting asylum-related migration flows with machine learning and data at scale. Scientific Reports 12(1), 1-25. DOI: 1457. 10.1038/s41598-022-05241-8.

Castells, M. (1996). The rise of the network society. Cambridge. Blackwell.

Chang, C.C. (2018). Hakka genealogical migration analysis enhancement using big data on library services. Library Hi Tech, 36(3), 426-442. DOI: 10.1108/LHT-08-2017-0172.

Chen, Y., Li, K., Zhou, Q. and Zhang, Y. (2022). Can Population Mobility Make Cities More Resilient? Evidence from the Analysis of Baidu Migration Big Data in China. International Journal of Environmental Research and Public Health, 20(1), 36. DOI: 10.3390/ijerph20010036.

Choi, S., Hong, J.Y., Kim, Y.J. and Park, H. (2020). Predicting psychological distress amid the COVID-19 pandemic by machine learning: discrimination and coping mechanisms of Korean immigrants in the US. International Journal of Environmental Research and Public Health, 17(17), 6057. DOI: 10.3390/ijerph17176057.

Cox, M. and Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. Report NAS-97-010, MS T27A-2. Moffett Field, CA: NASA Ames Research Center.

Davenport, T.H. (2014). How strategists use “big data” to support internal business decisions, discovery and production. Strategy & Leadership, 42(4), 45-50.

Diebold, F.X. (2021). What’s the big idea? Big data and its origins. Significance, 18(1), 36-37. DOI: 10.1111/1740-9713.01490

Emami, S.N., Yousefi S., Pourghasemi, H.R., Tavangar, S. and Santosh, M. (2020). A comparative study on machine learning modeling for mass movement susceptibility mapping (a case study of Iran). Bulletin of Engineering Geology and the Environment, 79, 5291-5308. DOI: 10.1007/s10064-020-01915-7.

Fernández-Martínez, J.L., Boga, J.A., de Andrés-Galiana, E., Casado, L., Fernández, J., Menéndez, C., ... Rodríguez-Guardado, A. (2021). A Machine Learning Model for Evaluating Imported Disease Screening Strategies in Immigrant Populations. The American Journal of Tropical Medicine and Hygiene, 105(5), 1413-1419. DOI: 10.4269/ajtmh.20-1443.

Gahi, Y., Guennoun, M. and Mouftah, H.T. (2016). Big data analytics: security and privacy Challenges. 2016 IEEE Symposium on Computers and Communication (ISCC), 952-957. Messina. Italy: IEEE.

Gao, Y., Nan, Y. and Song, S. (2022). High-speed rail and city tourism: Evidence from Tencent migration big data on two Chinese golden weeks. Growth and Change, 53(3), 1012-1036. DOI: 10.1111/grow.12473.

Garha, N.S. and Domingo, A. (2019). Indian diaspora population and space: national register, UN Global Migration Database and Big Data. Diaspora Studies, 12(2), 134-159. DOI: 10.1080/09739572.2019.1635390.

Giang, N.H., Nguyen, T.T., Tay, C.C., Phuong, L.A. and Dang, T.T. (2022). Towards predictive Vietnamese human resource migration by machine learning: A case study in northeast Asian countries. Axioms, 11(4), 151. DOI: 10.3390/axioms11040151.

Havas, C., Wendlinger, L., Stier, J., Julka, S., Krieger, V., Ferner, C., ... & Resch, B. (2021). Spatio-Temporal Machine Learning Analysis of Social Media Data and Refugee Movement Statistics. ISPRS International Journal of Geo-Information, 10(8), 498. DOI: 10.3390/ijgi10080498.

Huang, Y. and Shao, M. (2022). Challenges and Countermeasures of Arab Immigrants and International Trade in the Era of Big Data. Mathematical Problems in Engineering, 1(1), 1-11. DOI: 10.1155/2022/1025453.

International Committee of the Red Cross (ICRC) & Privacy International (2018). The humanitarian metadata problem: "Doing no harm" in the digital era. Available at: https://privacyinternational.org/report/2509/humanitarian-metadata-problem-doing-no-harm-digital-era

Juric, T. (2022b). Predicting refugee flows from Ukraine with an approach to Big (Crisis) Data: a new opportunity for refugee and humanitarian studies. Athens Journal of Technology and Engineering, 9(3), 159-184.

Juric, T. (2022c). Ukrainian refugee integration and flows analysis with an approach of Big Data: Social media insights. MedrXiv, Under review.

Jurić, T. (2022a). Big (Crisis) Data in Refugee and Migration Studies–Case Study of Ukrainian Refugees. Comparative Southeast European Studies, 70(3), 540-553. DOI: 10.1515/soeu-2022-0048.

Katsikouli, P., Byrne, W.H., Gammeltoft-Hansen, T., Høgenhaug, A.H., Møller, N.H., Nielsen, T.R., ... & Slaats, T. (2022). Machine Learning and Asylum Adjudications: From Analysis of Variations to Outcome Predictions. IEEE Access, 10(1), 130955-130967.

Khangahi, F.D. and Kiani, F. (2021). Social Mobilization and Migration Predictions by Machine Learning Methods: A study case on Lake Urmia. International Journal of Innovative Technology and Exploring Engineering, 10(6), 123-127. DOI: 10.35940/ijitee. F8833.0410621.

Kılıç, Ö.O., Akyol, M.A., Işık, O., Kılıç, B.G., Aydınoğlu, A.U., Sürer, E., ... & Temizel, T.T. (2019). Data analytics without borders: multi-layered insights for Syrian refugee crisis. In Data for refugees challenge workshop.

Kiossou, H.S., Schenk, Y., Docquier, F., Houndji, V.R., Nijssen, S. and Schaus, P. (2020). Using an interpretable Machine Learning approach to study the drivers of International Migration. arXiv preprint, Available at: https://aiforgood2020.github.io/papers/AI4SG_paper_46.pdf

Korkmaz, E. (2020). Using big data in migration and refugee studies. Science & Enlightenment, 4(3), 241-248.

Krupenkin, M. and Rothschild, D. (2019). Using Machine Learning to Measure Changes in Cable News Coverage of Immigration (2014-2019). Computation+ Journalism. Available at: https://bpb-us-w2.wpmucdn.com/sites.northeastern.edu/dist/0/367/files/2020/02/CJ_2020_paper_41.pdf

Lai, J. and Pan, J. (2020). China's City Network Structural Characteristics Based on Population Flow during Spring Festival Travel Rush: Empirical Analysis of "Tencent Migration" Big Data. Journal of Urban Planning and Development, 146(2), 04020018.

Lee, H. and Song, E. (2022). Peer discrimination toward rural migrant students and academic performance in urban China: A machine learning approach. Cities, 131 (1), 104027. DOI: 10.1016/j.cities.2022.104027.

Lu, Y. (2022). Detecting Imperfect Substitution between Comparably Skilled Immigrants and Natives: A Machine Learning Approach. International Migration Review. Under Review.

Martey, E. and Armah, R. (2021) Welfare effect of international migration on the left-behind in Ghana: Evidence from machine learning. Migration Studies, 9(3), 872-895. DOI: 10.1093/migration/mnaa025.

Masuda, Y. (1990). Managing in the information society. Cambridge: Blackwell.

Micevska, M. (2021). Revisiting forced migration: A machine learning perspective. European Journal of Political Economy, 70 (1), 102044. DOI: 10.1016/j.ejpoleco.2021.102044.

Molina, M.D., Chau, N., Rodewald, A.D. and Garip, F. (2022). How to model the weather-migration link: a machine-learning approach to variable selection in the Mexico-US context. Journal of Ethnic and Migration Studies, 49(4), 1-27. DOI: 10.1080/1369183X.2022.2100549.

Nair, R., Madsen, B.S., Lassen, H., Baduk, S., Nagarajan, S., Mogensen, L.H., ... & Urbak, S. (2019). A machine learning approach to scenario analysis and forecasting of mixed migration. IBM Journal of Research and Development, 64 (1/2), 7-10. DOI: 10.1147/JRD.2019.2948824.

Olberg, N. and Seuken, S. (2022). Enabling trade-offs in machine learning-based matching for refugee resettlement. arXiv, 1 (11), 1-19. DOI: 10.48550/arXiv.2203.16176.

Quinn, J.A., Nyhan, M.M., Navarro, C., Coluccia, D., Bromley, L. and Luengo-Oroz, M. (2018). Humanitarian applications of machine learning with remote-sensing data: Review and case study in refugee settlement mapping. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376 (2128), 20170363 DOI: 10.1098/rsta.2017.0363.

Rahaman, M., Morshed, M.M., and Bhadra, S. (2022). An integrated machine learning and remote sensing approach for monitoring forest degradation due to Rohingya refugee influx in Bangladesh. Remote Sensing Applications: Society and Environment, 25(1), 100696. DOI: 10.1016/j.rsase.2022.100696.

Raman, V., Vera, C. and Manna, C.J. (2022). Bias, Consistency, and Partisanship in US Asylum Cases: A Machine Learning Analysis of Extraneous Factors in Immigration Court Decisions. In Equity and Access in Algorithms, Mechanisms, and Optimization, 4(1), 1-14. DOI: 10.1145/3551624.3555288.

Ran, X., Xu, Y., Liu, Y. and Jiang, J. (2022). Examining online social behavior changes after migration: An empirical study based on OSN big data. Computers in Human Behavior 129 (1), 107158. DOI: 10.1016/j.chb.2021.107158.

Ren, C. and Bloemraad, I. (2022). New Methods and the Study of Vulnerable Groups: Using Machine Learning to Identify Immigrant-Oriented Nonprofit Organizations, Socius, 8(1), 1-14. DOI: 10.1177/23780231221076992.

Ruhnke, S.A., Reynolds, M.M., Wilson, F.A. and Stimpson, J.P. (2022). A healthy migrant effect? Estimating health outcomes of the undocumented immigrant population in the United States using machine learning. Social Science & Medicine, 307(1), 115177. DOI: 10.1016/j.socscimed.2022.115177.

Ruhnke, S.A., Wilson, F.A. and Stimpson, J.P. (2022). Using machine learning to impute legal status of immigrants in the National Health Interview Survey. MethodsX, 8(9), 101848. DOI: 10.1016/j.mex.2022.101848.

Simionescu, M. (2021). The status of immigrants on Italian labour market in the context of economic decline: Evidence from survey, macroeconomic and big data. Economics, Management and Sustainability, 6(1), 34-48. DOI: 10.14254/jems.2021.6-1.3.

Štular, B., Lozić, E., Belak, M., Rihter, J., Koch, I., Modrijan, Z., ... Gutjahr, C. (2022). Migration of Alpine Slavs and machine learning: Space-time pattern mining of an archaeological data set. PLoS ONE, 17(9), 0274687. DOI: 10.1371/journal.pone.0274687.

Szocska, M., Pollner, P., Schiszler, I., Joo, T., Palicz, T., McKee, M., ... Gaal, P. (2021). Countrywide population movement monitoring using mobile devices generated (big) data during the COVID-19 crisis. Scientific Reports, 11 (1), 5943. DOI: 10.1038/s41598-021-81873-6

Weber, H. (2020). How well can the migration component of regional population change be predicted? A machine learning approach applied to German municipalities. Comparative Population Studies-Zeitschrift für Bevölkerungswissenschaft, 45 (1), 143-178. DOI: 10.12765/CPoS-2020-08en.

Wilson, F.A., Zallman, L., Pagán, J.A., Ortega, A.N., Wang, Y., Tatar, M. and Stimpson, J.P. (2020). Comparison of use of health care services and spending for unauthorized immigrants vs authorized immigrants or US citizens using a machine learning model. JAMA network open, 3 (12), e2020.29230. DOI: 10.1001/jamanetworkopen.2020.29230.

Xiao Z, Bi M, Zhong Y, Feng X and Ma H (2021) Study on the evolution of the source-flow-sink pattern of China's Chunyun population migration network: Evidence from Tencent big data. Urban Science, 5(3): 66. DOI: 10.3390/urbansci5030066.

Yılmaz, B. and Ozcan, E. (2021). Big data problematic in social sciences of fourth generation human rights, Turkish Studies, 17(3), 473-393. DOI: 10.7827/TurkishStudies.58072.

Zhou, X., Chen, Z., Yeh, A.G. and Yue, Y. (2021). Workplace segregation of rural migrants in urban China: A case study of Shenzhen using cellphone big data. Environment and Planning B: Urban Analytics and City Science, 48 (1), 25-42. DOI: 10.1177/2399808319846903.