Searching for migration: Estimating Japanese migration to Europe with Google Trends data

Bert Leysen (bert.leysen@vub.be)
Vrije Universiteit Brussel
Pieter-Paul Verhaeghe
Vrije Universiteit Brussel

Research Article

Keywords: Big data, migration, Japan, forecasting

Posted Date: April 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1537858/v1

Abstract

In recent research, Google Trends data has been identified as a potentially useful data source to complement or even replace otherwise traditional data for predicting migration flows. However, the research on this is in its infancy, and as of yet suffers from a distinctive Western bias both in the topics covered as in the applicability of the methods. To examine its wider utility, this paper evaluates the predictive potential of Google Trends data, which captures Google search frequencies, but applies it to the case of Japanese migration flows to Europe. By doing so, we focus on some of the specific challenging aspects of the Japanese language, such as its various writing systems, and of its migration flows, characterized by its relative stability and sometimes limit size. In addition, this research investigates to what extent Google Trends data can be used to empirically test theory in the form of the aspirations and (cap)ability approach. The results show that after careful consideration, this method has the potential to reach satisfactory predictions, but that there are many obstacles to overcome. As such, sufficient care and prior investigation are paramount when attempting this method for less straightforward cases, and additional studies need to address some of the key limitations more in detail to validate or annul some of the findings presented here.

1. Introduction

Migration studies remain hampered by several issues, key of which is the limited availability of reliable and up-to-date migration data [55]. Ahmad-Yar and Bircan [3] identify several issues, from the multiplicity of measuring flows and defining stocks by governments and organizations, to the delay with which data is published. Furthermore, there is no comprehensive information on why people migrate. These compounding issues make migration predictions difficult and have an immediate impact on both policy and research.

During the past decades, however, new forms of data have emerged, primarily centered around the use of the internet and devices connected to it. Central are so-called *big data* defined as an “information asset characterized” by “High Volume, Velocity and Variety” [23]. Important for international migration are sources where either “the primary usage is for geolocation” such as mobile device GPS signals or geotags on social media, or data with a location component as part of its “digital exhaust” [22], allowing people’s movements to be tracked.

There has been an increasing interest from migration scholars in using big data originating from social media [24], such as LinkedIn [48], Twitter [31, 57] or Facebook [46, 52, 59], but also from mobile phones [8]. *Google search data* too is fertile soil for research due to Google’s popularity as a search engine and its free-to-use analytics. Applications range from economics, tourism, medicine to health [33, 36]. Also migration research aims to improve models for predicting migration flows through Google search data with some success [9, 54].

Search data also hold promise from theoretical reasons. Scholars have started to approach migration as a combination of *aspirations* to migrate and the *ability* to do so as elements preceding any form of migration [21]. Aspirations are brought about through their interaction with the migration environment and an individual's characteristics [14]. Yet gaining comparative insight into people's migration intentions and aspirations remains challenging. Some research uses the Gallup World Poll (GWP) to measure country-level aspirations [25, 35, 39]. Yet, the GWP has problems as its few questions can be difficult to interpret (counterfactual) and only inquire about permanent migration [14]. Also, accessing these micro-data is expensive, thus barring a wider audience with limited resources.

This is where free-to-access online search data may serve as an alternative to capture migration intentions and aspirations. International migration is a major decision for individuals or households. These decisions and subsequent preparations are accompanied by a search for information to facilitate this movement [55]. These searches thus reflect to some extent the *aspirations to migrate*, taking place *before* actual mobility.

More research is necessary to determine to what extent Google search data can be a valid source of data. Previous research has mainly focused on Google searches in Western languages and migration between Western countries (with [17] as a notable exception). So, despite the critical nature of both language and writing systems for this search activity how these aspects impact data and thus research on migration using this data, has been underexplored.

Therefore, the first aim of this study is to examine how language and the writing systems used by prospective migrants impact Google Trends data and its potential for estimating migration flows. We examine this question with the case-study of Japanese immigration to Europe, more specifically to Germany, the United Kingdom and France as three major European countries of destination, in addition to Belgium and the Netherlands as two smaller ones. Japan is linguistically homogenous, but its language is complicated, having two syllabaries and one logographic system, resulting in myriad ways of looking up information online.[1] While non-Western languages are considered an additional complication in research utilizing Google Trends and are subsequently avoided [9], we purposefully include its examination as a distinctive research aim. Moreover, Japanese migration to Europe as a topic in itself is understudied, both in English and Japanese language research. All these elements combined make Japanese immigration a compelling but challenging case to examine Google search data for predicting immigration flows.

The second aim is to examine how well immigration from Japan to Europe can be estimated using Google search data based on the methods of preceding migration research. Here, we rely on the migration process framework as explained by Carling [12-13] and de Haas [20-21]. We hypothesize that migration aspirations translate into an active search for information. This search for information can partly be captured by online search activity, such as in Google. And if more people are aspiring and later planning to migrate, more people should be searching for information. So, all else equal, this increase (or decrease if aspirations temper) in searches may be reflected in Google search activity which can be interpreted by

Google Trends data. Lastly, the increase or decrease in Google search activity may consequently reflect actual (subsequent) movement.

[1]With the exception of names and romanization methods preferred by original sources, this paper uses the Traditional Hepburn romanization of the Japanese language. All translations are done by the author.

2. Migration Theories And Alternative Data Sources

Migration as a multidimensional phenomenon has long eluded coherent theorization. Traditional theories on the initiation of migration focus one-dimensionally on economic factors [37]. More recent research similarly focuses on specific drivers of migration, such as socio-economic (e.g., education), institutional (e.g., migration policies and civil rights) and socio-cultural factors (e.g., social networks, cultural ties). This multidimensionality reflects the inherent complexity of migration [7, 15, 18]. As such, theories that aim to explain why people migrate face continued criticism. Massey et al. already noted how migration studies lack a commonly accepted theoretical framework [38] and more recently Amelina and Horvarth [5] argued how linking migration studies to general social theory is a key challenge for the future of the field.

An attempt to address these critiques is the aspirations-(cap)ability approach, most notably proposed by Carling [12] and elaborated upon by de Haas [20-21]. The framework goes beyond the one-dimensional focus on migration determinants by conceptualizing migration as a combination of aspirations to migrate and the ability to do so. Patterns of aspirations develop in the interaction of the migration environment with individual characteristics [14]. Whereas Carling [12] developed the framework to deal with ‘involuntary immobility’ (i.e., aspiring but unable to migrate), de Haas [19-21] reframed ability as ‘capabilities’, based on Sen’s capabilities approach. He considers aspirations as a function of “people’s general life aspirations and perceived spatial opportunity structures” and capabilities as dependent on “positive and negative liberties” people experience [20-21]. Migration aspirations are typically seen as static factors: one either aspires to migrate or one does not. And those with the ability to do so, end up migrating.

Yet migration aspirations themselves can be influenced by capabilities, which is why a more nuanced understanding of these dynamic aspirations is paramount. Some researchers have hinted at this dynamic nature. Migali et al. [39] showed with GWP data how more people aspire (intend) to migrate than end up preparing for it in the next 12 months, illustrating that not everyone that wishes to migrate end up moving. Carling and Schewel [14] noted similarly how with increased specificity of migration-related questions in the GWP, the answers can differ greatly. While these studies recognize nuances in migration aspirations and preparations, empirical application remains limited largely due to the difficulty in capturing these distinctions.

A reason for these limited applications is the paucity of reliable and accurate data. Despite efforts by governments and organizations, traditional migration statistics have not improved notably [3]. In an attempt to find alternatives to inadequate official data sources, researchers have turned to various forms

of big data. Zagheni and Weber [58] used Yahoo! e-mail data to estimate the rates of international migration. They discovered e-mail data have the potential to complement existing data for increased accuracy in developed countries. Together with Gummadi [59], these authors later measured migration stocks with Facebook advertising data containing socio-demographic data of its users. The authors see the potential of digital data, not only for migration but for investigating all kinds of demographic elements – finding it particularly promising for countries lacking the official infrastructure to track migration in an organized way. Zagheni et al. [57] used Twitter data of a subset of about 500,000 users in OECD countries to infer migration patterns. They found that, although difficult to predict overall variability, Twitter data was useful for predicting significant turning points in migration trends. Around the same time, State et al. [48] used LinkedIn data (geolocated career histories) to examine trends in the migration of professional workers. While the authors did not focus on prediction, the data showed levels of granularity that are difficult to find in national statistics, especially since these are typically not easy to compare cross-nationally. Combinations of these new data and traditional sources have also added depth to investigations. For instance, Yildiz et al. [56] combine bilateral migrant stocks with Facebook monthly and daily active user data to construct a Bayesian hierarchical model for EU migration stocks. Other research uses the same social media data, adjusted for bias, and combines this with traditional survey data to produce so-called “nowcasts” of migrant stocks in the United States [4].

Prediction with big data entered a new chapter with the help of data generated by internet search engines. In particular Google, both due to its increasing popularity and it being the default search engine on lower-end smartphones,[1] has been used frequently in academic research. Specifically, Google Trends, a platform that maps the relative popularity of search terms across different locations has been a crucial new data source. A pioneering application was by the hand of Ginsberg et al. [28] who matched Google searches on the flu to actual levels of influenza. Since then, applications using Google Trends data to forecast events have proliferated in the fields of economics, tourism, medicine and health, to information technology [33, 36]. More recently new fields of investigation have opened up, dealing with novel topics such as forecasting unemployment insurance claims following hurricanes [1], vaccine hesitancy and anti-vaccination sentiments in the context of Covid-19 [42], and cross-national investigations in more established areas, such as disease modelling of Covid-19 for a range of European countries [49].

Applications to migration studies followed suit. In 2016, a study by Vicéns-Feliberty and Ricketts analyzed searches of Puerto Ricans on migration to the United States, and to five states popular among Puerto Rican migrants [53]. Based on Google search data, they found that different states were popular for different reasons (job-related reasons, family considerations, and political party). In 2017, Connor successfully tracked the movements of refugees by examining the internet searches in Arabic for the word ‘Greece’ within Turkey, an important migration corridor into Europe [17]. Even trends within a single day could be discovered with hourly data. Kostakos et al. [34], also focused on refugees, investigated whether search data could improve the forecasting of their arrivals in Greece.

Böhme et al. [9] showed how Google search information (in English, French, and Spanish) can successfully be used to predict bilateral migration flows as search hits seem to reflect the intention to

migrate. The predictions with these data outperformed models based solely on traditional data, such as GDP and unemployment rates. The Google Trends Index the authors constructed was also used by Golenvaux et al. [30] in the same languages to test a long short-term memory (LSTM) approach which included Google search data against a linear gravity model (a more traditional approach), and an artificial neural network (ANN) model. Both models were outperformed by the LSTM approach combining Google data, illustrating its potential. Wanner [54] opted for a simplified approach to the above studies. Instead of using a long list of possible keywords and coopting these in more elaborate models, he used one key phrase in the dominant language of the country of origin ('working in Swiss') to predict the predominant labor migration from Spain, Italy, France, and Germany to Switzerland. By linear regression and taking into account specific periods of lag (i.e., a delay between when a search action is executed and when one actually moves), he successfully predicted to some extent migration flows from Spain and Italy albeit with less convincing results for France. Avramescu and Wiśniowski [6] followed a similar approach with Google Trends Indexes in English and Romanian, constructing composite variables capturing the interest of Romanians migrating to the United Kingdom. Their indices for employment and education managed to match the trends of official migration statistics, proving the data's potential for further research. Research by Fantazzini et al. [26] on internal migration in Russia using Google data was less successful, although they did succeed in reducing forecasting errors by including the data into a larger model.[2]

These studies have advanced the examination of alternative data sources substantially. However, deeper empirical investigations of how Google search data may be integrated in furthering theory, such as the aspirations-(cap)ability framework [12-14, 21], are limited. In addition, barring a few exceptions, the research has suffered from a predominantly Western bias. Consequently, our understanding of the wider applicability covering other regions is still lacking.

[1]This is an important point of access for developing regions which increasingly rely on mobile internet infrastructures.
[2]It should be noted that Google's market share at the time of their investigation only reached approximately 45%, being outperformed by Russia's domestic provider Yandex. This naturally has implications on the external validity of the data.

3. Japanese Migration To Europe

Research into modern Japanese migration to Europe is mainly historical in nature, often dividing the narrative in a pre- and post-World War II one. Before World War II, Japanese migration flows consisted mainly of colonial migration to countries in East and Southeast Asia [11, 50], and labor migration to the Americas [2, 40]. Compared to the flows to Asia and America, migration flows to overseas communities in Europe were much smaller. Based on figures from the *Kodansha Encyclopedia of Japan*, James Stanlaw [47] estimated that the number of Japanese emigrants to Europe in the pre-World War II period (1868-1941) did not exceed 7,980. In comparison, the Korean Peninsula alone witnessed 712,583 Japanese arriving in the same period.

After the war, Japanese migration became predominantly economic in nature. Taking the center stage in this era are Japanese multinationals, typically featuring local headquarters or branch offices in Europe employing Japanese on a rotation base either as trainees or managers [44]. A clear example is Toyota's European headquarters in Belgium (founded in 1963 in Denmark) with various vehicle and engine manufacturing plants in addition to design and R&D centers across Europe [51]. Following Bonacich's theory [10], Lucie Cheng and Marian Katz [16] consider these Japanese expats as "middleman minorities" living "close by each other, establish(ing) Japanese schools for their children, giv(ing) rise to neighborhood markets that speak Japanese and stock(ing) Japanese food, and in general maintain(ing) a distinctively Japanese community" (p.60).

Little research is dedicated to these contemporary migrant groups in the context of movement to Europe. Compared to the United States, which has an established tradition of studying incoming (East-)Asian migration flows, existing European studies focus primarily on limited cases of post-colonial Asian migration patterns [43]. Migration from Japan to Europe specifically is rarely touched upon, with notable exceptions such as the pioneering work by Glebe [29] on the Japanese community in Düsseldorf, Germany. English language quantitative analyses are likewise scarce.

The research presented here aims to contribute to the three areas of study presented above. First, it aims to further the examination of the framework established by Carling [12] and de Haas [20-21] dealing with migration aspirations specifically, by making use of alternative data (Google Trends). Next, we widen the investigation of Google Trends by analyzing its applicability to other areas and languages (Japanese). Last, this research contributes to the topical lacuna of Japanese immigration to Europe.

4. Data And Methods

4.1 Data

This study makes use of several datasets. A first dataset consists of official immigration figures. The first target was to obtain the monthly data for each country to construct a detailed analysis and examine the results based on different lags (in months) between search and movement. However, these data are not always readily available. The statistical agencies of the different countries were contacted by email with the request for access, but only the representatives of Belgium replied positively to this inquiry.[1] The other countries stated that monthly numbers are not available and refer to the yearly data.[2] Other research has implicitly encountered the same limitation and has successfully used yearly data instead [9, 30]. For yearly figures, OECD data proved to be more complete than Eurostat data when consulted. For instance, entries for Germany after 2008 were missing. When necessary, the data were supplemented with numbers of the national statistical agencies.[3] The immigration flow data for the United Kingdom are based on the yearly "International Passenger Surveys". It should be noted here that these survey-numbers are not accurate immigration figures and are rounded to one hundred.

Next, Google Trends data are used (trends.google.com). This tool allows extraction of the relative search frequency of one or a set of keywords input in google in a specific geographic entity. Data is available for free from 2004 onwards. The relative frequency is indicated by a number between 0 and 100 (low to high search intensity) and is provided for each month in the time series specified. Absolute frequencies are not made public due to privacy concerns. While Google's market share in Japan is not as high as in the U.S. or Europe, it is still over 70% for the period January 2009 to December 2021 according to StatCounter (https://gs.statcounter.com/search-engine-market-share/all/japan retrieved on January 8, 2022). Also, according to the International Telecommunication Union, internet is widely accessible in Japan [32].[4]

Google Trends data are investigated for two periods: 2006-2019 and 2011-2019. Similar to previous research, 2006 is selected as a starting point because it coincides with a more widespread adoption of Google. We end with 2019 because, at the time of research, national statistics of 2020 were not yet available. We opt for this double approach because Google implemented an algorithm change in 2011. Depending on the keyword input, stark differences can be seen in the time series pre-and post-2011 data (see supplementary data). Since Google Trends is a relative index, it is not possible to use the same dataset and investigate the post-2011 numbers separately as the numbers are in relation to all the data in the set. Each new period under investigation necessitates a fresh generation and extraction of Google search frequencies.

For the sections where only country names were used (see *methods*), the data for Belgium is only considered until May 2018. Due to the popularity of the World Cup football game Japan-Belgium on July 2, 2018, any search action that only takes into account the Japanese word for 'Belgium' culminated in an excessive peak around this date, thus skewing all the relating data.

Next, we make use of an existing keyword list generated by Böhme et al. [9]. This list has been successfully used in another research too [30]. In this study, the list is modified and translated to suit the specific context of Japanese immigration. Here, the focus is solely on the Japanese language. Despite mandatory English classes in the Japanese education system, English is not routinely used by native Japanese to the extent that it could realistically be captured by online search activities. As a consequence, Google searches in English would primarily capture the search activities of foreign nationals in Japan. Since these people are typically not included in official immigration statistics counting Japanese citizens entering a country, including non-Japanese Google searches in the Google Trends data would add additional bias to the analysis. As such, the focus is on Japanese language specifically to target the searches of Japanese nationals.

In addition, the Japanese language is sufficiently complicated to warrant a standalone investigation as it has several writing systems. 1) *Kanji* originates from Chinese characters and is mainly used for *kango* or Sino-Japanese words. Most nouns and parts of adjectives are written in *kanji* (e.g., 'music' □□, or the first character of 'beautiful' □□□). 2) *Hiragana* is primarily used for grammatical suffixes of words (for instance endings to denote the past tense of adjectives or adverbs such as the aforementioned 'beautiful' □□□ □□□□□ □□□□□) and grammatical elements in sentences (e.g. □ can mark the topic of a sentence or indicate

contrast). 3) *Katakana* on the other hand is mostly used for loanwords, scientific words, and other imported terminology such as IT-related jargon. While 4) *rōmaji* is rarely used by itself, it can be used to input Japanese on digital devices. Several systems of transcribing a Japanese pronunciation to Latin script exist. We only consider the Hepburn and *Nihon Shiki* systems here. The former is used primarily by non-native speakers, and the latter is the main system used by native Japanese speakers.

The specific difficulty with applying the Japanese language to research with Google Trends is twofold: first, the different writing systems are not always mutually exclusive. For instance, the same Japanese word for 'beautiful' can be written both in *kanji* and *hiragana* (□□□□ or□□□). Both versions of this word are commonly used although for most *kanji* is preferred due to the second complication: Japanese is rife with homophones (see supplementary data for examples).[5] As such, using *kanji* would be the logical option for searching online, but being a logographic system as opposed to a simple alphabet, not all characters are equally well known. Their sheer number can make *kanji* difficult, so even well-educated Japanese typically have not memorized all of them [41].[6] In case of ambiguity or uncertainty, one may opt to use *hiragana* when searching the internet.

4.2 Methods

The method of clarifying this as it relates to our research is straightforward. Based on the keyword list by Böhme et al. [9], 20 migration-related keywords were selected. This list is supplemented with ten keywords that focus on the specific Japanese migration experience, so centering around overseas study (e.g., 'study' or 'scholarship'), expats (e.g., 'insurance,' 'work,' or 'tax'), and overseas Japanese communities (e.g., 'Japanese food' or 'Japanese Association').

Each keyword is inputted and compared in Google Trends in as many ways as possible. Concretely this means that, when possible, the same word was input in 1) *kanji*, 2) *hiragana*, 3) *katakana*, 4) *rōmaji* (Hepburn system), 5) *rōmaji* (*Nihon Shiki* system) (see figure 1 for an example). Loanwords in *katakana* do not have a *kanji*-equivalent so this option is left out for these words, resulting in two sets of words: a) *kango* or Sino-Japanese words which have a *kanji* equivalent (24 words), and b) loanwords that are predominantly *katakana* and do not have a directly corresponding *kanji* (six words). Next, the time series of the different inputs for every keyword in Google Trends are compared to come to an understanding of how these different systems impact the data that can be extracted.

To predict migration with Google search data, we start with the same keyword list by Böhme et al. [9]. Whereas research by Golenvaux et al. [30] successfully used the list unmodified to predict immigration, to use it for Japanese migration it a) needs to be adjusted to reflect the specific nature of Japanese migration and b) needs to be translated taking into account the specificity of the Japanese language. Concretely, most words dealing with topics such as 'asylum' or 'smuggling' were deleted as these are not relevant for Japanese immigration to Europe, and words such as 'insurance' or 'studying overseas' were added. Also, words such as 'migration' and 'migrating', while different in English, are differentiated in Japanese only by grammatical sentence constructions (e.g., *ijū* and *ijū suru*). The words containing the

meaning of the words do not include these grammatical differentiators. This means that these keywords are identical in Japanese.

Finally, following the findings of examining the different writing systems, the words are translated and transcribed resulting in a list of 90 words (table 1). For some words, compound search terms are also constructed, both to boost measurable search frequencies by Google where results were lacking and subsequently to promote data extraction, and to address the issue of synonyms. For instance, we combined the words 'consulate' and 'embassy', and operated the search term as follows in combination with 'Paris': □□ □□□ + □□ □□□ ('Paris consulate + Paris embassy')

For determining the predictive power of Google Trends, several approaches are examined. As a first step, a straightforward approach is used, following Wanner [54]. The keywords are inputted in Google Trends together with the Japanese word for each country. For instance 'study (in) France' would be translated into □□ □□□□. Monthly time series of Google Trends (ranging from 0 to 100) are downloaded for 2006 to 2019 and 2011 through 2019 and are aggregated for each year *t* in Japan (*ja*) as location.[1] The resulting time series are labeled as bilateral Google Trends indexes ($GTIbil_{jat}$). We estimate linear regression models via ordinary least squares method (OLS) to examine the relationship between immigration (y_t = the number of moves in year *t*), and the relative number of searches in year *t* conducted in Japan (*ja*), expressed by $GTIbil_{jat}$.

In a second step, we follow Golenveaux et al. [30] and Böhme et al. [9] and construct an interaction term consisting of additional Google Trends indexes: $GTIuni_{jat}$ x $GTIdest_{jat}$.[2] Whereas the aforementioned authors construct one Google Trends index which aggregates the frequencies of all the keywords, we maintain the frequencies per keyword to examine the possible nuances between words. Although the assumption is that all associations of the words should follow the same direction, this needs to be confirmed by considering each word individually. $GTIuni_{jat}$ is a predictor containing the Google Trends values of the keywords by themselves for Japan during year *t* (i.e., not specifying the European destination). $GTIdest_{jat}$ is the relative search intensity in Japan for the country names (e.g., 'France' but without another keyword). OLS linear regression is used for the periods 2006-2019 and 2011-2019 but with two predictors: $GTIbil_{jat}$ + $GTIuni_{jat}$ x $GTIdest_{jat}$.

Compared to moving from Germany to France for instance, migrating from Japan to Europe requires more planning both due to the distance (both Euclidean and cultural) involved and the additional paperwork compared to within-Schengen movement. To capture this preparation phase, the models are run again with a one-year time lag (y_{t-1}) for Germany, France, the Netherlands, and the United Kingdom. Because monthly data are available for Belgium, the number of lags is increased and delays of three, six, nine, and twelve months between searching and moving are examined for this country.

In a third search action, we only focus on the country and city names. Instead of examining general searches, a built-in tool by Google Trends is used that categorizes searches in specific categories. The data are extracted based on four categories: 1) all categories, 2) business and industrial, 3) jobs and

education, and 4) law and government. The resulting time series only take into account searches related to the specified categories and are thus not limited to exact words.[3] These are analyzed with OLS linear regression with predictor $GTIdest_{jat}$.

Next, the first analysis is repeated but the country names are exchanged with a key city from each country. As reflected in the literature, cities such as Paris, Düsseldorf or Brussels are known within their respective countries and Japan as featuring a relatively established Japanese community and may serve as a prime destination for Japanese immigrants. We examine if these city names can serve as proxies for country names. Some keywords practically make more sense on a regional/city level. For instance, when searching for accommodation it can be assumed that people do this at the level of a city and do not just look for a place to stay anywhere in the country. Here we focus on one predictor $GTIbil_{jat}$ and analyze the predictive strength via OLS linear regression for 2011 to 2019.

Finally, as a fifth step, the search location is changed from Japan to each of the five European countries (*cod*). This translates into searching how frequently Japanese words were searched for in European countries. In this step, only the Japanese keywords are used without the European country or city name. These are analyzed for both periods starting in 2006 and 2011 via OLS linear regression with predictor $GTIuni_{codt}$. The inspiration for this reversed approach can be found in Connor's research [17]. We assume that after people have moved, they still need to search for information that may be captured by Google (e.g., where the embassy is to arrange visa formalities, looking for a job, how tax works, and more).

Throughout the above analyses, linear regression is used for a number of reasons. One of which is that linear regression is used in comparable research [1, 6, 9, 54] and this research also aims to find out how replicable these techniques are to other cases. Another reason is that it conceptually follows the logic of migration aspirations: More people aspiring to migrate means more people searching for information. Increases/decreases in these numbers ought to be followed by increases/decreases in real mobility, potentially after some delay. Whereas other research makes use of a narrower, more targeted range of methods and data, there is no prior research which can be used as a guideline for analyzing Japanese immigration. Consequently, this research opts to explore several ways of searching for predictive strength by using a wide range of Google search terms.

[4]The United Kingdom's Office for National Statistics has monthly numbers, but these are not split between citizenship or countries of origin and are consequently not suitable for this paper.

[5]An exception are the numbers for asylum seekers and refugees which are monitored more closely.

[6]The immigration flow from Japan to Germany for 2019 was lacking in the OECD dataset at the time of consultation and was supplemented with data generated by the German Federal Statistical Office (set 12711-0007).

[7]Some selection bias is expected due to differences in digital literacy and access to technology. This second part, however, is not a concern when dealing with Japan. According to the ITU (International

Telecommunication Union), a specialized department for information and communication technologies by the UN, Japan has a 3G mobile coverage of 100%, and a 4G mobile coverage of 99% of the population (2019 and 2017 respectively), so the basic network is well established. Active subscriptions follow the same trend: 203 active mobile-broadband subscriptions per 100 inhabitants and 34 fixed subscriptions per 100 inhabitants in 2019 [32].

[8] These are words with a different meaning but the same or similar sound.

[9]Whereas the basic sets (roughly 2000 characters) are learned in school, a complete list of *kanji* existing in Japanese would range from 40,000 to more than 75,000 unique characters. Diverging proficiency is illustrated by a nationally organized kanji-exam (*kanji kentei*) aimed at Japanese of all ages and levels. 631,521 people registered for the second round in 2020, but only 10.9% could pass the most difficult first grade [41].

[10]For Belgium, the monthly time series are used. For other countries these are aggregated to yearly ones.

[11]We only follow the researchers' principle of constructing Google Trends indexes but not the analysis since they used Google Trends as part of a model rather than by itself. In this research, we are more concerned with the keywords, so our emphasis differs.

5. Results

5.1 Google Trends and the Japanese writing system

5.1.1 Kango or Sino-Japanese vocabulary

Google searches of Sino-Japanese words (24 out of 30) are predominantly performed in *kanji* (see Table 2). For 13 out of 24 keywords, the search frequency in Google for inputs in *hiragana*, *katakana*, and *rōmaji* is equal to or lower than 1 on a scale from 0 to 100 which means they are barely used relative to the *kanji* version. Seven of the 24 keywords are predominantly searched for in *kanji*, but also show some frequencies for inputs in *hiragana* albeit much lower. Each of the remaining four variations is unique in the sample of keywords: one keyword is not searched for enough so there is no result in Google Trends, another features some small fluctuations not only in *hiragana* but also *katakana*, and a third also in *rōmaji.* A final keyword, *kika*, meaning 'naturalization' (of, for instance, citizenship), results in more *hiragana* than *kanji* searches. From this initial analysis, we conclude that for *kango* or Sino-Japanese word searches in Google trends the predominant writing system for Japanese input is *kanji*.

Table 2
Number of successful Google Trends keyword extractions per writing system

1. Only *kanji*	**13**
2. Predominantly *kanji*, and small fluctuations in *hiragana*	7
3. Predominantly *kanji*, and small fluctuations in *hiragana*, *katakana*, and *romaji*	1
4. Predominantly *kanji*, and fluctuations in *hiragana* and *katakana*	1
5. Predominantly *hiragana*, and fluctuations in *kanji*	1
6. No result	1
Total	**24**

5.1.2 Katakana loanwords

The list of words for this category is more limited and includes loanwords such as ‘visa’, ‘hotel’, or ‘internship’. Here, transcription in *kanji* is not possible (a *kanji* equivalent does not exist), so only a comparison with *hiragana* and *rōmaji* can be made. Google searches of these kinds of words appear to be done overwhelmingly in *katakana*. Both inputs in *hiragana* and *rōmaji* do not show up in Google Trends.

For the next steps of the analysis, Sino-Japanese words can be input in *kanji*, and loanwords in *katakana*.

5.2 Predicting migration with Google Trends data

In this section, first the results of the search actions of keywords in Google Trends are discussed, followed by the potential for predicting migration with these data.

5.2.1 Searching for migration

The number of positive hits, that is when the input of the keyword (and country/city) in Google Trends generates a usable time series of relative search frequencies, depends on the country or city name used. Bilateral searches (both place name and keyword combined - $GTIbil_{jat}$) have the highest success rates (56%) for France and Germany. Belgium, although similar to the Netherlands in terms of population size, has a much lower success rate than its Northern neighbor in generating usable Google Trends data (10% vs 29%). Combinations with city names instead of country names, have a low success rate for Amsterdam, Brussels, and Düsseldorf (2–4%), whereas Paris and London score relatively well (33% and 36% respectively). Focusing only on the country name ($GTIdest_{jat}$) or the keywords ($GTIuni_{jat}$) always generated results. Finally, when searching for just the Japanese keywords but changing the search location from Japan to the European countries of destination ($GTIuni_{codt}$), there was a low success rate for Belgium (9%) in generating usable time series, and higher rates for the other four countries, with the United Kingdom on top (67%) (see supplementary data).

5.2.2 Predicting migration

Following the mixed quality of data that are extracted from Google Trends, we can expect considerable differences between the countries in coefficients of determination (R^2 values), indicating correlations between searches (and aspirations) and mobility. In table 3 only keywords with the highest R^2 are included to maintain the overview. First, the correlations between the keywords and country names, and Japanese immigration flow figures are examined (analysis I). Overall, the associations between searches and Japanese migration flows are best for Germany and the Netherlands. For Germany, the highest R^2 value of 0.678 is for the search terms related to 'welfare' from 2006 through 2019, 0.697 is for 'visa', and 0.621 for a compound keyword consisting of 'applicant + recruitment + employment' in the period 2011–2019. For the Netherlands in the period 2006–2019, we note the highest R^2 for the search term capturing various configurations of the term 'migration' (0.686). The results for the period 2011–2019 show an R^2 of 0.969 for 'visa' and 0.787 for 'migration' (see Fig. 2 for a visual representation of 'visa').

Figure 2 Japanese migration flow to the Netherlands (blue) versus 'visa' searches (grey)

More surprising is the lack of correlations found for keywords combined with 'France'. Despite the high success rate in extracting data from Google Trends, only the Japanese word 'migration' resulted in an R^2 higher than 0.5 (0.522). Lastly, when examining the correlations of Google searches and migration flows to the United Kingdom, only 'airline ticket' and 'studying abroad' resulted in R^2 values higher than 0.5 for the period 2011–2019 (0.598 and 0.567 respectively).

While R^2 values are informative and have been used in prior research to indicate correlations between Google Trends and migration, they do not explain the complete situation. For the logic of migration aspirations which are translated into search action and movement to make sense in this analysis, the regression coefficient should be positive since more searches lead to more movement. Table 3 shows this is not always the case. Negative coefficients are interspersed with positive ones, signifying that sometimes a *higher* search frequency correlates with a *lower* movement.

In the second approach, a predictor in the form of the interaction term $GTIuni_{jat}$ x $GTIdest_{jat}$ is added and the correlations with official immigration figures are examined. Table 3 shows the overall fit in most cases improving by adding this interaction term. For Germany, there are four words with a coefficient of determination above 0.8 in the 2006 period. And whereas for the 2011 dataset the prediction power is lower, there are significantly more words that have a high R^2 value compared to just the predictor $GTIbil_{jat}$. The predictive strength for the Netherlands is higher as well, but for several words, the added interaction term does not increase prediction power (e.g., 'visa' or 'migration'). The results for France and the United Kingdom are also mixed. The regression coefficient for the first predictor again shows opposite signs for some words, and the second predictor mainly shows coefficients of zero or close to zero. Therefore, these belie the fact that all keywords capture the same aspirations.

When repeating this analysis after introducing lag between searching and moving, there is an increase in the coefficient of determination for Germany, reaching the levels of prediction found by Golenveaux et al. and Böhme et al. [23, 9] (e.g., for 'economy'). However, the coefficients are predominantly negative. So,

translating this to searching and migrating would mean that *more* searches of these keywords result in *less* migration. For France, prediction power of the keywords decreases when the period 2006–2019 is analyzed but increases for the period 2011–2019 with the words for 'moving (between houses)' and 'contract' resulting in an R^2 value of 0.838 and 0.826 respectively. For the Netherlands, some words switch places in the ranking: 'visa' was the best predictor in the previous models but now ranks last among those words with a coefficient of determination higher than 0.5. 'Employment' now has the best predictive power (0.832). Adding lag for the United Kingdom has mixed results. Overall predictive strength remains mediocre, but the successful words are more work-related.

Table 3 Results of the OLS models for different keywords combined with different country names and with one year lag for different keywords combined with different country names (predictor GTIbiljat; predictors GTIbiljat and GTIunijat x GTIdestjat - dependent variable: number of moves from Japan to European countries)

In a third analysis, we examine if migration can be predicted by relying on Google's in-built algorithm instead of combining place names with keywords. For this, search data in specific categories as designated by Google are extracted (see supplementary data). While we managed to extract time series from Google Trends for each instance, none of the coefficients of determination are high. Only searches for 'Germany' for the period 2006–2019 in the categories jobs and education and law and government result in an R^2 value higher than 0.5 (0.582 and 0.541). However, in this case, the regression coefficients are negative (-5.741 and − 7.648), thus not matching the hypothesis that more Google searches, as a proxy for migration aspirations, result in more mobility.

Following the fourth approach, the first analysis is repeated but country names are replaced by city names Amsterdam, Brussels, Düsseldorf, London, and Paris (see supplementary data). Compared to the countries (except for Belgium), cities show a lower predictive power. Only for London can we find a correlation with an R^2 value above 0.5 (0.545 for the keyword 'studying abroad'). For the other search terms the quality of the data was too low to conduct any meaningful analysis (i.e., containing a large number of months with 0 relative search frequencies).

The last analysis substitutes Japan as the search location in Google Trends for the European countries of destination (table 4). The results regarding the strength of prediction are in line with those of the first analysis. Only several keywords result in fairly high R^2 values. Perhaps more strikingly, the regression coefficients are all positive and consequently fit the search-to-action hypothesis. The best results are for Germany and the Netherlands. In Germany, mainly the words for 'money' and 'moving' contain predictive strength (0.769 and 0.750). In the Netherlands, it is the Japanese word for 'work' or 'job' that has the best predictive power (R^2 of 0.577 and 0.808 for the periods starting in 2006 and 2011 respectively – see supplementary data for a visual representation). It should be noted that most of the words that contain predictive power differ from those in the earlier analyses. The predictive strength for the United Kingdom and France is likewise poor, with R^2 values staying below 0.6.

In none of the above analyses was there any useable result for Belgium which can be expected considering the low number of useable keywords found before.

[12]We explored the variation of these categories also with the set of key terms but found that Google would restrict the searches to a level that most results became unusable. As such, we do not cover this side-investigation here.

[13]See the supplementary material for tables split between analysis and covering additional key words.

6. Discussion And Conclusion

This study aims to examine Japanese immigration to Europe as a new context in which to empirically investigate alternative data sources such as Google Trends. A first aim was to examine to what extent language writing systems impact online searches conducted with Google, and by extension the data used for estimating immigration flows. Next, we examined whether Google Trends data can function as a tool for predicting Japanese immigration flows to European countries signaling migration aspirations.

The findings indicate that the specific writing system is of consequence in this case. For non-Latin scripts, it is advisable to conduct a preparatory study of the different writing systems, especially if the researcher is less familiar with its peculiarities. While the results for Japanese suggest that the writing system we logically expect for each word can be safely used (e.g., *kanji* for Sino-Japanese words, and *katakana* for loanwords), there are exceptions. For instance, the word □□ (*kika*) featured a higher frequency in *hiragana* than the *kanji* equivalent □□ which can be explained by the significant number of homophones for this word. In other words, it is plausible that people searching in Japanese use the *hiragana* form because the correct *kanji* are less known.

In all, research that uses Google Trends data would benefit from the additional step of analyzing the different ways of inputting keywords. This conclusion is a tentative one and may be specific to the Japanese language. Google's algorithms are proprietary, so it is difficult to come to generalizable findings. Also, this challenge can lessen over time as AI-powered translation engines become more powerful and inputs different writing systems may become combined in singular search terms.

Compared to similar exercises in other research [9, 23, 54] the success rate of the prediction analyses here is lower. A first contextual limitation can be found in the immigration flows from Japan. Not only are these flows rather stable with little variation, but they are also limited compared to the total Japanese population (126 million in 2019, World Bank). This first point makes statistical analysis challenging. The second point is relevant for using Google Trends since it relies on relative search frequencies. As such, the keywords need to be searched sufficiently compared to all other Google searches within Japan to generate usable results. Consequently, if a thousand people (roughly the yearly flow from Japan to the Netherlands or Belgium) Google information related to migrating, this could be too small a number compared to 126 million people Googling other things. This explains why several Google Trends keyword extractions were not successful.

There is also a cultural aspect to a certain subsection of Japanese migration which may limit the usability of this investigation. Japanese companies are highly regulated, compartmentalized, and process-bound [27, 45]. Consequently, most aspects of company and even private life are taken care of by departments rather than the individual. Specifically for expats, it is primarily the HR department that is in charge of the preparations for employees moving to Europe. These departments in turn consult specialized firms that deal with the paperwork for visa applications, shipping personal items and more. The specialists employed by these firms may have little need for Google: Contacts at the embassy may be stored in Outlook, so there is no need to Google 'embassy' to look up contact information. Draft and blank forms for visa applications and templates for moving companies may be stored on a local server, similarly bypassing the need for Google search. This may explain the low search frequencies and consequently the lack of predictive power of Google Trends for Japanese immigration flows. However, these limitations are not yet substantiated by any further empirical investigation so they should be approached with care.

With regards to methodological limitations, the analyses performed here are relatively straightforward. More complex modeling may reveal other useful aspects such as reductions in estimation errors which are more difficult to identify here (for instance, see [9, 30]). Future research may also opt to focus on countries with available monthly data to examine the different lags more in detail and obtain an overall more thorough picture of migration flows.

Aside from these limitations, we identify several findings. First of all, in the analyses, the useful keywords differ between countries. The word 'visa' is a strong performer, resulting in the highest predictive value for the Netherlands (0.969) but less so for other countries. This shows the need to carefully curate keywords, and not to rely solely on lists generated by previous research or online tools.

Also, this method can work for smaller migration flows. The strongest predictor was found for the Netherlands despite its much smaller immigration flow from Japan (on average yearly 1296 people compared to 6898 to Germany over the same 15-year period. While it was more challenging to find usable Google search data for smaller countries, carefully curating keywords can result in good predictive performance.

In addition, 'visa' resulted in a good prediction for the period 2011 to 2019 for the Netherlands but was not usable for the period 2006-2019. Similarly, when the interaction term was added, we found fewer results for the United Kingdom during the period starting in 2006 compared to the period starting in 2011. This shows that the time period used to analyze Google Trends data can have a distinct impact on predictive power, and consequently on research using these data in general.

Lag between search and movement also had significant effects. For France (2011-2019), the predictive power increased substantially after introducing a one-year lag. We also see that the most suitable words change after introducing this lag. In the Netherlands, words such as 'employment', 'economy' and 'expenses' seem to matter more *before* mobility, whereas without lag words such as 'visa' and 'airline ticket' suit the immigration data better. This finding may substantiate the hypothesis that specific

keywords reflect specific phases leading up to migration (changing aspirations). Words such as ‘employment’, ‘economy’ and ‘expenses’ signify a preparatory phase with individuals researching the future location (how is the economy, the employment situation, etc.) thus signaling immigration aspirations. Words such as ‘visa’ and ‘airline ticket’ searched later may signal increasingly crystalizing aspirations, leading to concrete plans and preparations (e.g., getting a visa, buying a plane ticket). With careful tuning, it appears Google Trends data hold promise as a data source contributing to the expansion of the aspirations-(cap)ability approach as proposed by Carling [12] and de Haas [20-21].

Analyzing Japanese words in the countries of destination was on average more successful, particularly considering the overall positive regression coefficient. This finding links back to the first research aim concerning language. Not only does the input matter, but also where we search. It is possible to predict migration by using a language or writing system not used by the local population to identify migrant groups and their search actions, echoing some of Connor’s findings [17]. A possible explanation for this may be that some information is (also) searched after arrival (e.g., opening hours of the embassy or job positions). Part of this could also be explained by expats searching for information for family members who often join them with some delay. Focusing on this type of analysis where keywords in the language of the migrant are examined in the arrival country is a promising avenue for future research.

Using main cities of destination as proxies for country names or just using country names and Google Trends categories to capture related search queries produced no usable results. This may be due to the relatively small immigration flows and could work better for forecasting tourism and shorter stays in Europe [36]. Also, despite having a similar Japanese immigration flow to the Netherlands, Google search frequencies and predictions for immigration to Belgium were poor in all five analyses.

Perhaps the strongest admonition is that the effect of search frequencies on mobility was often not positive. For some cases, we can imagine how increased searches may capture (or result into) anxiety about mobility. Information may confront a searcher with potential difficulties. For instance, job hunting in Japan differs considerably from that in Europe. Searching for the specific procedures online may put potential applicants off, resulting in a repression of migration aspirations. Furthermore, most of these negative effects were found in the interaction term ($GTIuni_{jat}$ x $GTIdest_{jat}$) which does not specifically capture the keywords in conjunction with the migration location as the predictor ($GTIbil_{jat}$) in the first analysis does. Considering the small migration flows, a completely unrelated trend may be captured by this predictor.

In conclusion, Google search data can be used as an alternative data source for predicting migration, but there are challenges depending on the context. The current research presents a specific case with mixed results. While we find it is possible to predict some migration flows from Japan to Europe, this prediction power is highly specific to destination, time frame, and in particular, the keywords used. Certain cultural limitations which are absent in previous research should be carefully considered as well, thus making migration research with Google Trends a worthwhile but challenging option.

Declarations

Conflict of interest statement None of the authors have any competing interests to declare.

Funding No funding was extended in exchange for this research.

Author contribution Conceptualization, B.L. and P.V.; Methodology, B.L. and P.V.; Software, B.L.; Data collection and analysis, B.L.; Supervision, P.V.; Writing original draft, B.L.; Writing review & editing, P.V. All authors have read and agreed to the published version of the manuscript.

References

1. Aaronson, D., Brave, S. A., Butters, R. A., Fogarty, M., Sacks, D. W., & Seo, B. (2022). Forecasting unemployment insurance claims in realtime with Google Trends. International Journal of Forecasting, *38*(2), 567–581. https://doi.org/10.1016/j.ijforecast.2021.04.001
2. Adachi, N. (2006). Introduction: theorizing Japanese diaspora. In *Japanese Diasporas: Unsung pasts, conflicting presents, and uncertain futures 2* (pp. 1–23). Oxon: Routledge.
3. Ahmad-Yar, A. W., & Bircan, T. (2021). Anatomy of a Misfit: International Migration Statistics. Sustainability, 13(7), 4032. https://doi.org/10.3390/su13074032
4. Alexander, M., Polimis, K., & Zagheni, E. (2022). Combining Social Media and Survey Data to Nowcast Migrant Stocks in the United States. Population Research and Policy Review, *41*(1), 1–28. https://doi.org/10.1007/s11113-020-09599-3
5. Amelina, A., & Horvath, K. (2017). Sociology of migration. In *The Cambridge Handbook of Sociology* (Vol. 1, pp. 455–464). Cambridge University Press.
6. Avramescu, A., & Wiśniowski, A. (2021). Now-casting Romanian migration into the United Kingdom by using Google Search engine data. Demographic Research, 45, 1219–1254. https://doi.org/10.4054/DemRes.2021.45.40
7. Bijak, J. (2011). *Forecasting International Migration in Europe: A Bayesian View* (Vol. 24). Springer Netherlands. https://doi.org/10.1007/978-90-481-8897-0
8. Blumenstock, J. (2012). Inferring Patterns of Internal Migration from Mobile Phone Call Records: Evidence from Rwanda. Information Technology and Development, *18*(2), 107–125.
9. Böhme, M. H., Gröger, A., & Stöhr, T. (2020). Searching for a Better Life: Predicting International Migration with Online Search Keywords.pdf. Journal of Development Economics, *142*, 1–14. https://doi.org/https://doi.org/10.1016/j.jdeveco.2019.04.002
10. Bonacich, E. (1972). A Theory of Ethnic Antagonism: The Split Labor Market. American Sociological Review, *37*(5), 547–559. https://doi.org/10.2307/2093450
11. Caprio, M., & Jia, Y. (2009). Occupations of Korea and Japan and the Origins of the Korean Diaspora in Japan. In *Diaspora Without Homeland: Being Korean in Japan* (pp. 21–38). Berkeley: University of California Press.

12. Carling, J. (2002). Migration in the age of involuntary immobility: theoretical reflections and Cape Verdean experiences. Journal of Ethnic and Migration Studies, *28*(1), 5–42. https://doi.org/10.1080/13691830120103912
13. Carling, J. (2014). The role of aspirations in migration. In *Determinants of International Migration*. Oxford: International Migration Institute.
14. Carling, J., & Schewel, K. (2018). Revisiting aspiration and ability in international migration. Journal of Ethnic and Migration Studies, *44*(6), 945–963. https://doi.org/10.1080/1369183X.2017.1384146
15. Castles, S. (2010). Understanding global migration: A social transformation perspective. Journal of Ethnic and Migration Studies, *36*(10), 1565–1586. https://doi.org/10.1080/1369183X.2010.489381
16. Cheng, L., & Katz, M. (1998). Migration and the diaspora communities. In R. Maidment & C. Mackerras (Eds.), *Culture and Society in the Asia-Pacific* (pp. 52–69). New York: Routledge.
17. Connor, P. (2017). *The Digital Footprint of Europe's Refugees*. Retrieved from http://assets.pewresearch.org/wp-content/uploads/sites/2/2017/06/08094856/Pew-Research-Center_Digital-Footprint-of-Europes-Refugees_Full-Report_06.08.2017.pdf
18. Czaika, M., & Reinprecht, C. (2020). *Drivers of migration. A synthesis of knowledge.* (Working Paper No. 163; IMI Working Papers Series). International Migration Institute.
19. de Haas, H. (2003). *Migration and development in Southern Morocco: The disparate socio-economic impacts of out-migration on the Todgha Oasis valley*. Katholieke Universiteit Nijmegen.
20. de Haas, H. (2014). *Migration Theory: Quo Vadis?* (No. 100). *IMI Working Papers Series*. Oxford.
21. de Haas, H. (2021). A theory of migration: The aspirations-capabilities framework.Comparative Migration Studies, 9(1), 8. https://doi.org/10.1186/s40878-020-00210-4
22. De Backer, O. (2014), 'Big Data and International Migration' (United Nations Global Pulse: Pulse Lab Diaries, 16 June 2014) < www.unglobalpulse.org/big-data-migration> accessed 26 March 2022.
23. De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. Library Review, 65(3), 122–135. https://doi.org/10.1108/LR-06-2015-0061
24. Dekker, R., Engbersen, G., & Faber, M. (2016). The Use of Online Media in Migration Networks. Population, Space and Place, *22*(6), 539–551. https://doi.org/10.1002/psp.1938
25. Docquier, F., Peri, G., & Ruyssen, I. (2014). The cross-country determinants of potential and actual migration. International Migration Review, *48*(s1), S37–S99. https://doi.org/10.1111/imre.12137
26. Fantazzini, D., Pushchelenko, J., Mironenkov, A., & Kurbatskii, A. (2021). Forecasting Internal Migration in Russia Using Google Trends: Evidence from Moscow and Saint Petersburg. Forecasting, *3*(4), 774–804. https://doi.org/10.3390/forecast3040048
27. Fulcher, J. (1988). The Bureaucratization of the State and the Rise of Japan. The British Journal of Sociology, *39*(2), 228–254.
28. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, *457*(February), 1012–1015. https://doi.org/10.1038/nature07634

29. Glebe, G. (1986). Segregation and intra-urban mobility of a high-status ethnic group: The case of the Japanese in Düsseldorf. Ethnic and Racial Studies, *9*(4), 461–483. https://doi.org/10.1080/01419870.1986.9993546
30. Golenvaux, N., Alvarez, P. G., Kiossou, H. S., & Schaus, P. (2020). *An LSTM approach to Forecast Migration using Google Trends*. Université catholique de Louvain. Retrieved from http://arxiv.org/abs/2005.09902
31. Hsiao, Y., Fiorio, L., Wakefield, J., & Zagheni, E. (2020). *Modeling the bias of digital data: an approach to combining digital and survey data to estimate and predict migration trends* (Vol. 49). https://doi.org/https://doi.org/10.4054/MPIDR-WP-2020-019
32. International Telecommunication Union. (2021). ITU-D ICT Statistics. Retrieved May 29, 2021, from https://www.itu.int/en/ITU-D/Statistics/Pages/links/default.aspx
33. Jun, S. P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. Technological Forecasting and Social Change, *130*(February), 69–87. https://doi.org/10.1016/j.techfore.2017.11.009
34. Kostakos, P., Pandya, A., Oussalah, M., Hosio, S., Sattari, A., Kostakos, V., … Kyriakouli,O. (2018). Correlating refugee border crossings with internet search data. In *2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science* (pp. 264–268). IEEE. https://doi.org/10.1109/IRI.2018.00048
35. Laczko, F., Tjaden, J., & Auer, D. (2017). *Measuring Global Migration Potential, 2010–2015*. Retrieved from
36. Li, X., Law, R., Xie, G., & Wang, S. (2021). Review of tourism forecasting research with internet data. *Tourism Management*, *83*(October 2020), 104245. https://doi.org/10.1016/j.tourman.2020.104245
37. Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., & Taylor, J. E. (1993). Theories of International Migration: A Review and Appraisal. Population and Development Review, *19*(3), 431–466.
38. Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., (1994). An Evaluation of International Migration Theory: The North American Case. Population and Development Review, *20*(4), 699–751.
39. Migali, S., & Scipioni, M. (2018). *A global analysis of intentions to migrate*. *JCR Technical Report*.
40. Moore, S. C. (2010). *Gender and Japanese Immigrants to Peru, 1899 through World War II*.
41. Nippon kanji nōryoku kentei 2020 nendo (2020 Japanese Kanji Proficiency Test). (2021). Retrieved July 1, 2021, from https://www.kanken.or.jp/kanken/investigation/result/2020_1.html#anc03
42. Pullan, S., & Dey, M. (2021). Vaccine hesitancy and anti-vaccination in the time of COVID-19: A Google Trends analysis. Vaccine, *39*(14), 1877–1881. https://doi.org/10.1016/j.vaccine.2021.03.019
43. Rutten, M., & Verstappen, S. (2014). Middling Migration: Contradictory Mobility Experiences of Indian Youth in London. Journal of Ethnic and Migration Studies, 40(8), 1217–1235. https://doi.org/10.1080/1369183X.2013.830884

44. Sedgwick, M. W. (2001). Positioning “globalization” at overseas subsidiaries of Japanese multinational corporations. In *Globalizing Japan: Ethnography of the Japanese presence in Asia, Europe, and America* (pp. 43–51). New York: Routledge.
45. Shimizu, Y. (2020). *The Origins of Modern Japanese Bureaucracy*. London: Bloomsbury Academic.
46. Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., & Rango, M. (2019). Quantifying international human mobility patterns using Facebook Network data. PLoS ONE, *14*(10), 1–22. https://doi.org/10.1371/journal.pone.0224134
47. Stanlaw, J. (2006). Japanese emigration and immigration: from the Meiji to the modern. In *Japanese Diasporas: Unsung pasts, conflicting presents, and uncertain futures* (pp. 35–51). Oxon: Routledge.
48. State, B., Rodriguez, M., Helbing, D., & Zagheni, E. (2014). Migration of professionals to the U.S. Evidence from linkedin data. In L. M. Aiello & D. McFarland (Eds.), *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11–13, 2014, Proceedings* (pp. 531–543). Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-13734-6_37
49. Sulyok, M., Ferenci, T., & Walker, M. (2021). Google Trends Data and COVID-19 in Europe: Correlations and model enhancement are European wide. Transboundary and Emerging Diseases, *68*(4), 2610–2615. https://doi.org/10.1111/tbed.13887
50. Tamanoi, M. A. (2006). Overseas Japanese and the challenges of repatriation in post-colonial East Asia. In *Japanese Diasporas: Unsung pasts, conflicting presents, and uncertain futures* (pp. 217–235). Oxon: Routledge.
51. Toyota Motor Europe. (n.d.). Our European Journey: From small-scale car importer to multiple manufacturing sites. Retrieved June 5, 2021, from https://www.toyota-europe.com/world-of-toyota/feel/operations
52. Vespe, M., Rango, M., Zagheni, E., Weber, I., Natale, F., Spyratos, S., & European Commission. Joint Research Centre. (2018). *Migration data using social media: a European perspective.* https://doi.org/10.2760/964282
53. Vicens-Feliberty, M. A., & Ricketts, C. F. (2016). An analysis of Puerto Rican interest to migrate to the US using Google Trends.pdf. The Journal of Developing Areas, *50*(2), 411–430. https://doi.org/10.1353/jda.2016.0090
54. Wanner, P. (2020). How well can we estimate immigration trends using Google data? Quality & Quantity. https://doi.org/10.1007/s11135-020-01047-w
55. Willekens, F., Bijak, J., Klabunde, A., & Prskawetz, A. (2017). The science of choice: an introduction. Population Studies, *71*, 1–13. https://doi.org/10.1080/00324728.2017.1376921
56. Yildiz, D., Wisniowski, A., Abel, G., Gendronneau, C., Stepanek, M., Weber, I., Zagheni, E., & Hoorens, S. (2019, April 10). *Probabilistic Methods For Combining Traditional and Social Media Bilateral Migration Data*. Population association of America. Annual meeting, Austin, Texas. http://paa2019.populationassociation.org/uploads/191592

57. Zagheni, E., Garimella, V. R. K., Weber, I., & State, B. (2014). Inferring international and internal migration patterns from twitter data. In *WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web* (pp. 439–444). https://doi.org/10.1145/2567948.2576930
58. Zagheni, E., & Weber, I. (2012). You are where you E-mail: Using E-mail data to estimate international migration rates. *Proceedings of the 4th Annual ACM Web Science Conference, WebSci'12*, *volume*(October), 348–357. https://doi.org/10.1145/2380718.2380764
59. Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook's Advertising Platform to Monitor Migrations. Population and Development Review, *43*(4), 1–19. https://doi.org/https://doi.org/10.1111/padr.12102

Tables

Tables 1 ,3 & 4 are available in the Supplementary Files section

Figures

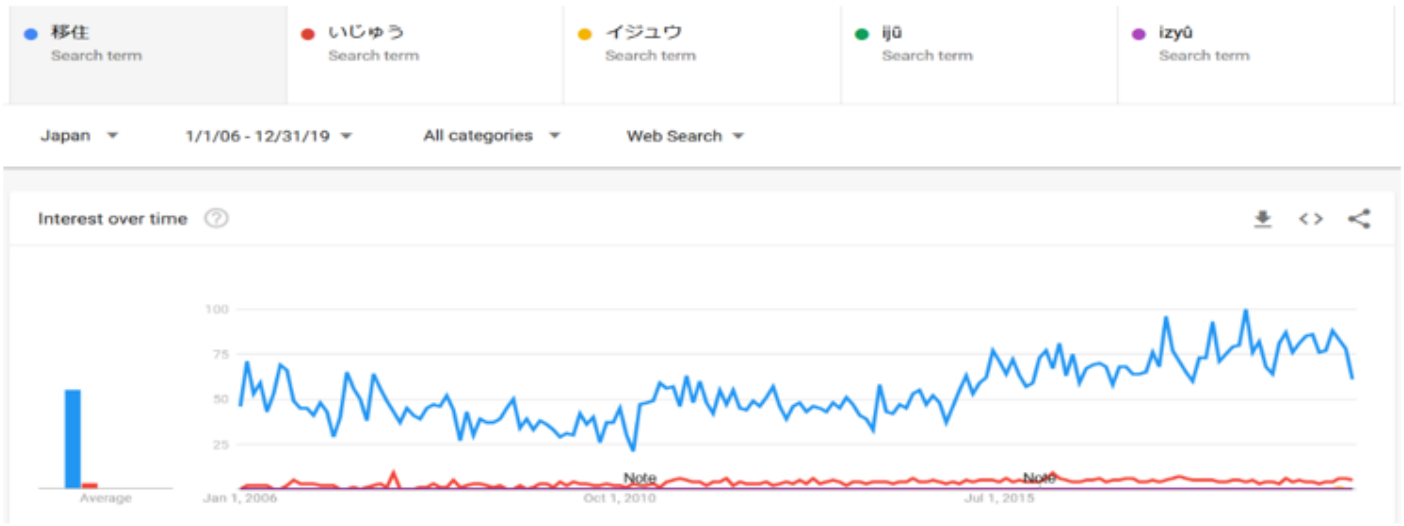

Figure 1

Google Trends image capture comparing the search frequencies of the same word in kanji, hiragana, katakana, rōmaji (Hepburn), and rōmaji (Nihon Shiki)

Source: Image captured from trends.google.com

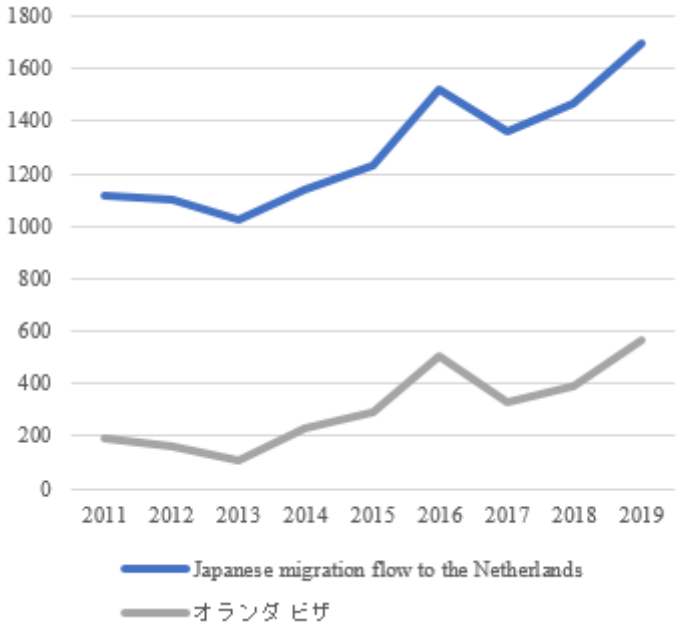

Figure 2

Japanese migration flow to the Netherlands (blue) versus 'visa' searches (grey)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Submissionwofig.pdf
- Tables.docx