Qi and Bircan *EPJ Data Science* (2023) 12:41
https://doi.org/10.1140/epjds/s13688-023-00419-0

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Can Google Trends predict asylum-seekers' destination choices?

Haodong Qi[1*] and Tuba Bircan[2]

*Correspondence:
haodong.qi@mau.se
[1]Malmö University, Nordenskiöldsgatan 1, 205 06 Malmö, Sweden
Full list of author information is available at the end of the article

Abstract

Google Trends (GT) collate the volumes of search keywords over time and by geographical location. Such data could, in theory, provide insights into people's *ex ante* intentions to migrate, and hence be useful for predictive analysis of future migration. Empirically, however, the predictive power of GT is sensitive, it may vary depending on geographical context, the search keywords selected for analysis, as well as Google's market share and its users' characteristics and search behavior, among others. Unlike most previous studies attempting to demonstrate the benefit of using GT for forecasting migration flows, this article addresses a critical but less discussed issue: when GT *cannot* enhance the performances of migration models. Using EUROSTAT statistics on first-time asylum applications and a set of push-pull indicators gathered from various data sources, we train three classes of gravity models that are commonly used in the migration literature, and examine how the inclusion of GT may affect models' abilities to predict refugees' destination choices. The results suggest that the effects of including GT are highly contingent on the complexity of different models. Specifically, GT can only improve the performance of relatively simple models, but not of those augmented by flow Fixed-Effects or by Auto-Regressive effects. These findings call for a more comprehensive analysis of the strengths and limitations of using GT, as well as other digital trace data, in the context of modeling and forecasting migration. It is our hope that this nuanced perspective can spur further innovations in the field, and ultimately bring us closer to a comprehensive modeling framework of human migration.

Keywords: Asylum-seeker; International migration; Destination choices; Internet search data

1 Introduction

The potential for Google Trends (GT) data to predict asylum-seekers' destination choices is a burgeoning area of research, drawing upon the nexus among big data analytics, migration studies, and the digital sociology of search behaviors. Asylum-seekers, like many other demographic groups, use online tools to inform their migration decisions, including information on potential destinations. GT, which collates search volumes of specific terms over time and by geographic location, could, in theory, provide insights into where asylum-seekers plan to move, revealed by what they are searching online. Empirically, the inclusion of GT-based indicators for migration intentions tend to enhance the performances of mi-

Springer

gration models [1–5]. However, what has been less explored in this literature is when GT *cannot* improve the accuracy of migration forecasts.

A distinctive feature of GT, compared to other digital traces, is that it may capture *ex ante* rather than *ex post* migration outcome. For example, geo-referenced emails, social media posts, LinkedIn profiles, among others, can help map actual migration flows by observing the changes in users' locations (see e.g., [6–13]). GT, on the other hand, may capture migration intentions rather than actions through their revealed demand for information [14]. This *ex ante* feature could be valuable for predictive analysis, e.g., a growing demand for migration-related information may predict an increase in migration volume. However, empirically, the predictive power of GT is not unequivocal; while some demonstrate that GT can outperform any of the established predictors of migration flows [1], others argued that it is not always the case [3, 4]. For example, GT's ability to predict can differ depending on geographical context, the search keywords selected for analysis, as well as Google's market share and its users' characteristics and search behavior, among others.

The predictive power of GT may also be sensitive to the complexity of migration models. For example, in [1], it is evident that GT can only increase the R-square values for relatively simple models, but not for those with a large set of fixed-effects (i.e., including origin-year and origin-destination constants). To gain more insights into such sensitivity, this article conducts a systematic examination of how GT's predictive power may vary depending on the capacity of migration models. Our analysis follows two steps. First, we estimate three classes of models that are commonly used in the migration literature: i) pooled regression (PL), see e.g., [15, 16]; ii) auto-regressive regression (AR), see e.g., [4]; and iii) flow fixed-effects regression (FE), see e.g., [17, 18]. We then add GT to these models and compare how their performances are changing, respectively.

It is noteworthy that these three model classes are inherently different in terms of complexity. PL is most parsimonious; it only includes a set of push and pull factors. AR is slightly more complex; it allows the migration outcome variable to be linearly dependent on its own previous values, which can potentially capture some unobserved time-varying processes of migration flows. FE is much more complex than PL and AR, as it includes a constant for each dyad flow to capture some time-invariant factors, such as colonial ties, language proximity, geographical distance, among others. Given these differences, we argue that the extent to which GT may predict migration flows might not be universal. Specifically, if unobserved factors (such as destination preferences or migration intentions) captured by GT are time-invariant, they will be fully absorbed into the dyad-specific constants in the FE model. Moreover, if unobserved factors that GT captures are time-varying but proportional to the migration outcome in the last period, they will be adjusted for by the AR component. In such cases, GT might only add predictive power to the relatively simple PL model.

Using GT and a set of push-pull indicators gathered from various data sources, we train three model classes described above to predict refugees' choices of EU destinations (using EUROSTAT statistics on first-time asylum applications). The results suggest that the inclusion of GT can only enhance the performance of relatively simple model (PL), but not the more complex ones (AR and FE).

2 Gravity models: micro foundation and empirical specifications

Migration models have undergone a substantial transformation in the last decade, with a pronounced shift towards micro-founded gravity models. Such models build upon a the-

oretical foundation that combines the gravity model of international trade with concepts from the random utility framework. This blending of principles allows researchers to interpret macro-level migration flows through the lens of micro-level decision-making, as well as to infer individual migration behavior while using aggregated dyad flow data (see e.g., [15, 17, 18]).

2.1 Micro foundation

Micro-founded gravity models treats migration flows as collective outcomes of individual decisions made under uncertainty. These models estimate the probability of migrating to a specific destination as a function of the characteristics of the origin and destination locations, the distance as well as costs associated with moving between different locations, and the characteristics of the individual migrant or household [18]. Furthermore, they account for the multilateral resistance to migration, which refers to the way in which the attractiveness of a potential destination is influenced not only by its characteristics and those of the origin country, but also by the characteristics of other alternative destinations [19].

Micro-founded gravity models provide a more nuanced understanding of migration decisions and can offer more accurate forecasts of migration flows. They allow researchers to explore how changes in variables such as income levels, employment rates, or immigration policies in one country can affect migration flows to and from other countries. However, as with any model, they are simplifications of reality and should be used with an understanding of their assumptions and limitations.

In general, there are two versions of the micro-founded gravity model. A restrictive version assumes that the unobserved component of the random utility function is independent and irrelevant across alternative destinations, i.e., the IIA assumption [20]. Under this assumption, an individual's preference of migrating to a given destination d and that of remaining in the origin o can be respectively represented by,

$$\begin{aligned} U_{o,d,t} &= V_{o,d,t} + W_{o,d,t}, \\ U_{o,o,t} &= V_{o,o,t} + W_{o,o,t}, \end{aligned} \tag{1}$$

where, V's are the deterministic utility and W's are the unobserved utility.

Since W's are assumed to be random draws from the type I extreme value distribution with the same mean and variance $\pi^2/6$ [21], the probabilities of migrating to d and remaining in o are,

$$\begin{aligned} \Pr(M_{o,d,t}) &= \frac{e^{U_{o,d,t}}}{e^{U_{o,o,t}} + \sum_{d \in D} e^{U_{o,d,t}}}, \\ \Pr(M_{o,o,t}) &= \frac{e^{U_{o,o,t}}}{e^{U_{o,o,t}} + \sum_{d \in D} e^{U_{o,d,t}}}. \end{aligned} \tag{2}$$

To link the probabilities above to macro flow data, we take the ratio of $\Pr(M_{o,d,t})$ to $\Pr(M_{o,o,t})$ which gives the migration rate, i.e., the number of people who decided to migrate from o to d relative to those who remained in o,

$$E[m_{o,d,t}] = \frac{\Pr(M_{o,d,t})}{\Pr(M_{o,o,t})} = \frac{e^{U_{o,d,t}}}{e^{U_{o,o,t}}}. \tag{3}$$

By taking logarithm, Eq. (3) can be linearized as,

$$\ln\big(E[m_{o,d,t}]\big) = U_{o,d,t} - U_{o,o,t} = V_{o,d,t} - V_{o,o,t} + W_{o,d,t} - W_{o,o,t}, \tag{4}$$

where, $W_{o,d,t} - W_{o,o,t}$ is expected to have a zero mean, as the difference between two random variables from the type I extreme value distribution with the same mean has itself a mean of zero [21].

A less restrictive version of the micro-founded gravity model was introduced in [17] which relaxes the IIA assumption and assumes that the unobserved utility are correlated across alternative destinations. Under this assumption, Eq. (1) becomes,

$$\begin{aligned} U_{o,d,t} &= V_{o,d,t}^{1/\tau} + W_{o,d,t}, \\ U_{o,o,t} &= V_{o,o,t} + W_{o,o,t}, \end{aligned} \tag{5}$$

where, τ is a dissimilarity parameter with range $(0, 1]$.

The dissimilarity parameter τ is inversely related to the correlation in the unobserved utility $W_{o,d,t}$ across alternative destinations [18]; the higher τ is, the less correlation in $W_{o,d,t}$. When the destinations in the choice set are completely different, the dissimilarity parameter τ will equal to one, and hence Eq. (5) will be the same as Eq. (1).

Substitute Eq. (5) into Eq. (2), the probabilities of migrating to d and remaining in o becomes,

$$\begin{aligned} \Pr(M_{o,d,t}) &= \frac{e^{V_{o,d,t}^{1/\tau} + W_{o,d,t}}}{e^{V_{o,o,t}+W_{o,o,t}} + e^{V_{o,d,t}^{1/\tau}+W_{o,d,t}} + \sum_{l\in D, l\neq d} e^{V_{o,l,t}+W_{o,l,t}}}, \\ \Pr(M_{o,o,t}) &= \frac{e^{V_{o,o,t}+W_{o,o,t}}}{e^{V_{o,o,t}+W_{o,o,t}} + \sum_{l\in D} e^{V_{o,l,t}+W_{o,l,t}}}. \end{aligned} \tag{6}$$

Due to the presence of τ, the denominators in Eq. (6) are no longer the same, and therefore cannot be cancelled out when taking the ratio of the two probabilities, the migration rate becomes,

$$E[m_{o,d,t}] = \frac{e^{V_{o,d,t}^{1/\tau}+W_{o,d,t}}}{e^{V_{o,o,t}+W_{o,o,t}}} \frac{e^{V_{o,o,t}+W_{o,o,t}} + \sum_{l\in D} e^{V_{o,l,t}+W_{o,l,t}}}{e^{V_{o,o,t}+W_{o,o,t}} + e^{V_{o,d,t}^{1/\tau}+W_{o,d,t}} + \sum_{l\in D, l\neq d} e^{V_{o,l,t}+W_{o,l,t}}}. \tag{7}$$

Using log-transformation, Eq. (7) can be linearized as,

$$\ln\big(E[m_{o,d,t}]\big) = 1/\tau\, V_{o,d,t} - V_{o,o,t} + W_{o,d,t} - W_{o,o,t} + \Omega_{o,d,t}, \tag{8}$$

$$\Omega_{o,d,t} = \ln\left(\frac{e^{V_{o,o,t}+W_{o,o,t}} + \sum_{l\in D} e^{V_{o,l,t}+W_{o,l,t}}}{e^{V_{o,o,t}+W_{o,o,t}} + e^{V_{o,d,t}^{1/\tau}+W_{o,d,t}} + \sum_{l\in D, l\neq d} e^{V_{o,l,t}+W_{o,l,t}}}\right). \tag{9}$$

$\Omega_{o,d,t}$ is known as the multilateral resistance to migration [17], which implies that the unobserved utility of moving from o to d can be influenced by the attractiveness of alternative l in the choice set. Ignoring $\Omega_{o,d,t}$ may generate biases in the estimation of the coefficients in the deterministic utility V's [18]. More importantly, this could, in turn, bias the prediction of the expected migration rate. One way to account for the influence of l's

attractiveness is to obtain information about people's actual preference for d, which can then be modeled as a deterministic utility. Specifically, we can decompose Eq. (9) as,

$$\Omega_{o,d,t} = 1/\tau\, V^*_{o,d,t} - V^*_{o,o,t} + \eta_{o,d,t}, \tag{10}$$

where, $1/\tau\, V^*_{o,d,t} - V^*_{o,o,t}$ represents a deterministic component of the multilateral resistance to migration. $\eta_{o,d,t}$ is an unobserved component. If $1/\tau\, V^*_{o,d,t} - V^*_{o,o,t}$ can be specified sufficiently, the remaining unobserved $\eta_{o,d,t}$ is essentially white noise [21].

In this article, we seek to examine whether Google Trends (GT) may help account for people's actual preferences for migration destinations, hence adjust for the multilateral resistance $\Omega_{o,d,t}$. The intuition behind GT is that people's preference for a given country may be revealed by their demand for this country's information, and this demand could be potentially captured by the intensity of Google searches.

2.2 Empirical specifications

Based on the discussion above, we specify two baseline models in this article. The first model uses GT as linear predictors for migration rates,

$$\begin{aligned}
&\ln(m_{o,d,t}) = 1/\tau\, V_{o,d,t} - V_{o,o,t} + 1/\tau\, V^*_{o,d,t} - V^*_{o,o,t} + \epsilon_{o,d,t},\\
&\epsilon_{o,d,t} = W_{o,d,t} - W_{o,o,t} + \eta_{o,d,t} + \zeta_{o,d,t},\\
&V_{o,d,t} = \beta \ln Y_{d,t-1},\\
&V_{o,o,t} = \beta \ln Y_{o,t-1} - \ln\theta \ln D_{o,t-1},\\
&V^*_{o,d,t} = \gamma \ln G_{o,d,t-1} + \lambda \ln G_{tr,d,t-1},\\
&V^*_{o,o,t} = \gamma \ln G_{o,o,t-1} + \lambda \ln G_{tr,tr,t-1},
\end{aligned} \tag{11}$$

where, τ is a dissimilarity parameter with range $(0, 1]$. $\epsilon_{o,d,t}$ is assumed to be normally distributed with zero mean and $\sigma_{\epsilon_{o,d,t}}$. $Y_{d,t-1}$ and $Y_{o,t-1}$ are economic variables in destination and origin, respectively. $D_{o,t-1}$ is a measure of conflicts or geopolitical tensions in origin. $G_{o,d,t-1}$ and $G_{o,o,t-1}$ are the volume of Google searches in origin o about destination d and origin o, respectively. $G_{tr,d,t-1}$ and $G_{tr,tr,t-1}$ are the volume of Google searches in transit country tr about destination d and transit country tr, respectively.

The second baseline model seeks to investigate the nonlinearity in the relationship between migration rates and Google searches. To do so, we expand $V^*_{o,d,t}$ and $V^*_{o,o,t}$ by p order polynomials fitted by the Generalized Additive Model (GAM),

$$\begin{aligned}
V^*_{o,d,t} &= \sum \gamma_p (\ln G_{o,d,t-1})^p + \sum \lambda_p (\ln G_{tr,d,t-1})^p,\\
V^*_{o,o,t} &= \sum \gamma_p (\ln G_{o,o,t-1})^p + \sum \lambda_p (\ln G_{tr,tr,t-1})^p.
\end{aligned} \tag{12}$$

2.3 Augmented models: flow fixed-effects (FE) and auto-regressive (AR) effects

As stressed at the outset, the ability of GT to predict human migration is not unequivocal; even if within a single study, the predictive power of GT may be inconsistent. For example, in [1], the results indicate that GT can only increase the R-square values for the models that are relatively simple, but not for those with a large set of fixed-effects. To make our study comparable to [1], we augment our baseline models by including the

origin-destination fixed-effects, and assess whether the explanatory power of GT weakens when model complexity increases.

Another important note in [1] is that the authors merely showed how GT may improve the goodness-of-fit to the training data without assessing its predictive power in the testing data. In this article, we will use cross-validation to assess whether the ability of GT to predict migration flows becomes weaker when our models are augmented with flow-specific dummies. As recent research showed that the flow fixed-effects gravity models can neither explain, nor predict migration patterns [22, 23], we expect that both the explanatory and predictive power of GT in the augmented models will be weaker than that in the baseline models.

For the purpose of short-term forecasting of migration flows, a good alternative to the flow fixed-effects could be to include an auto-regressive component [4]. To build on this literature, we augment our baseline models Eq. (11) and Eq. (12) by including $\rho \ln(m_{o,d,t-1})$, i.e., allowing the migration rate to be linearly dependent on its own previous values. The auto-regressive (AR) model is slightly more complex than our baseline models, however, it can potentially capture some unobserved time-varying processes of migration flows. It may also capture sudden shifts in migration trends, which is critical as migration flows, particularly asylum-related, are highly volatile and uncertain [24, 25].

3 Data

To estimate our baseline models Eq. (11) and Eq. (12) as well as their augmented versions, we make use of various data sources. To measure the intensity of asylum-related migration, we compute the Asylum Seeking Rate (ASR) using EUROSTAT statistics on first-time asylum applications lodged to different EU member states normalized by the population sizes in origin countries. The economic push ($Y_{o,t-1}$) and pull ($Y_{d,t-1}$) factors are measured by per capita GDP from the World Development Index. To measure push factors associated with geopolitical tensions $D_{o,t-1}$, we make use of the Uppsala Conflict Data Program (UCDP).

To measure people's preferences for migration destinations, we rely on Google Trends (GT). The rationale for using GT is that what people search for may reflect their intentions, preferences, and other human behavior through revealed demand for information [14]. Such data has been adopted for a variety of applications, e.g. analyzing patterns of influenza [26], consumer demand [27, 28], stock and commodity prices [29–31], unemployment [32, 33], and suicide occurance [34], among others.

As stressed earlier, the usability of GT can vary depending on geographical context, the search keywords selected for analysis, as well as Google's market share, users' characteristics and their search behavior, among others. To reduce some heterogeneity, we limit our analysis to five sending countries: AFG, ARM, GEO, IRQ, and SYR, which, to some extent, share certain characteristics, e.g., geographical location and/or language. More importantly, they are also the major sending countries in terms of asylum-related migration to Europe during the last decade [23].

GT can be used in numerous ways as predictors for migration flows. Typically, researchers construct country-specific time-series based on a set of predefined keywords on GT. For example, migration-related terms (e.g. "passports", "visa", "asylum", etc.) searched on Google can be proxies for interest to seek asylum or to migrate. When the intensities of these searches increase, it can be an indication of rising intentions to out-migrate from a given origin.

The intentions measured by Google Trends could be sensitive to the choice of keywords and language [1]. Moreover, the possible number of keyword-language combinations can approach infinity. Since this article emphasizes on asylum-seekers' destination preferences, we thus make use of the search trends related to country topics; Google's proprietary algorithm generates country-specific topic by grouping search keywords that share the same concept in any language. For example, the trends on the topic of "Sweden" includes all the search terms related to Sweden in all languages, e.g. Stockholm, Sverige (Sweden in Swedish), IKEA, among others.

It is important to note that GT normalizes the search frequencies of each keyword to an index between zero and one. While this may reveal how the interest in a given country evolve overtime, it masks how the level of interest may differ across destinations. For example, for an average Syrian, the level of interest in Germany might be higher than in other destinations. To recover the level differences, we normalize each time-series of a country topic and then adjust its level as,

$$S_{k,t} = \frac{S^{*}_{k,t}}{\sum_{t=1}^{T} S^{*}_{k,t}/T} \times \frac{\sum_{t=1}^{T} S^{**}_{k,t}/T}{\sum_{t=1}^{T} S^{**}_{o,o,t}/T}, \quad S_{k,t} \in [S_{o,o,t}, S_{o,d,t}, S_{tr,tr,t}, S_{tr,d,t}], \tag{13}$$

where, $S^{*}_{k,t}$ is a single download of Google Trend's index value for country k. $S^{**}_{k,t}$ is the index values for country k that are downloaded together with the index values for origin $S^{**}_{o,o,t}$. The first term normalizes the index by its mean, and the second term adjusts the index level in relation to the reference (i.e., the index value for origin o). o and d are origin and destination country, respectively. tr denotes transit country, which, in this article, refers to Turkey (as it has been widely documented that many refugees from SYR and AFG chose Turkey as a transit country before moving to Europe, hence the searches about EU countries in Turkey might also capture refugees' destination preferences).

While a rise of search intensity of a given country may indicate growing interest in migration to that country, it could also be driven by certain events unrelated to migration. For example, a sudden increase in the searches of Germany could be due to a winning game played by the German football team in a major tournament. To exclude the noises as such, we decompose $S_{k,t}$ into two components,

$$S_{k,t} = F\big(S^{\text{Migration}}_{j,t}, S^{\text{Travel}}_{j,t}, S^{\text{Refugee}}_{j,t}, S^{\text{Visa}}_{j,t},\big) + \Omega_{S_{k,t}}, \quad j \in [o, tr], \tag{14}$$

where, F is a regression model which includes four-way interactions among the GT indices for Migration, Travel, Refugee, and Visa. This model seeks to capture part of the variation in $S_{k,t}$ that is related to migration, whereas $\Omega_{S_{k,t}}$ is the remaining variation that is assumed to be unrelated to migration.

Figure 1 depicts the difference between the observed searches ($S_{k,t}$) and the searches predicted by the function $F(S^{\text{Migration}}_{j,t}, S^{\text{Travel}}_{j,t}, S^{\text{Refugee}}_{j,t}, S^{\text{Visa}}_{j,t},)$. In general, the predicted searches resemble the observed ones very well. This suggests that when people in AFG, ARM, GEO, IRQ, SYR as well as in the transit country (Turkey) Google different European countries, they are mostly looking for information related to migration, travel, refugee and visa.

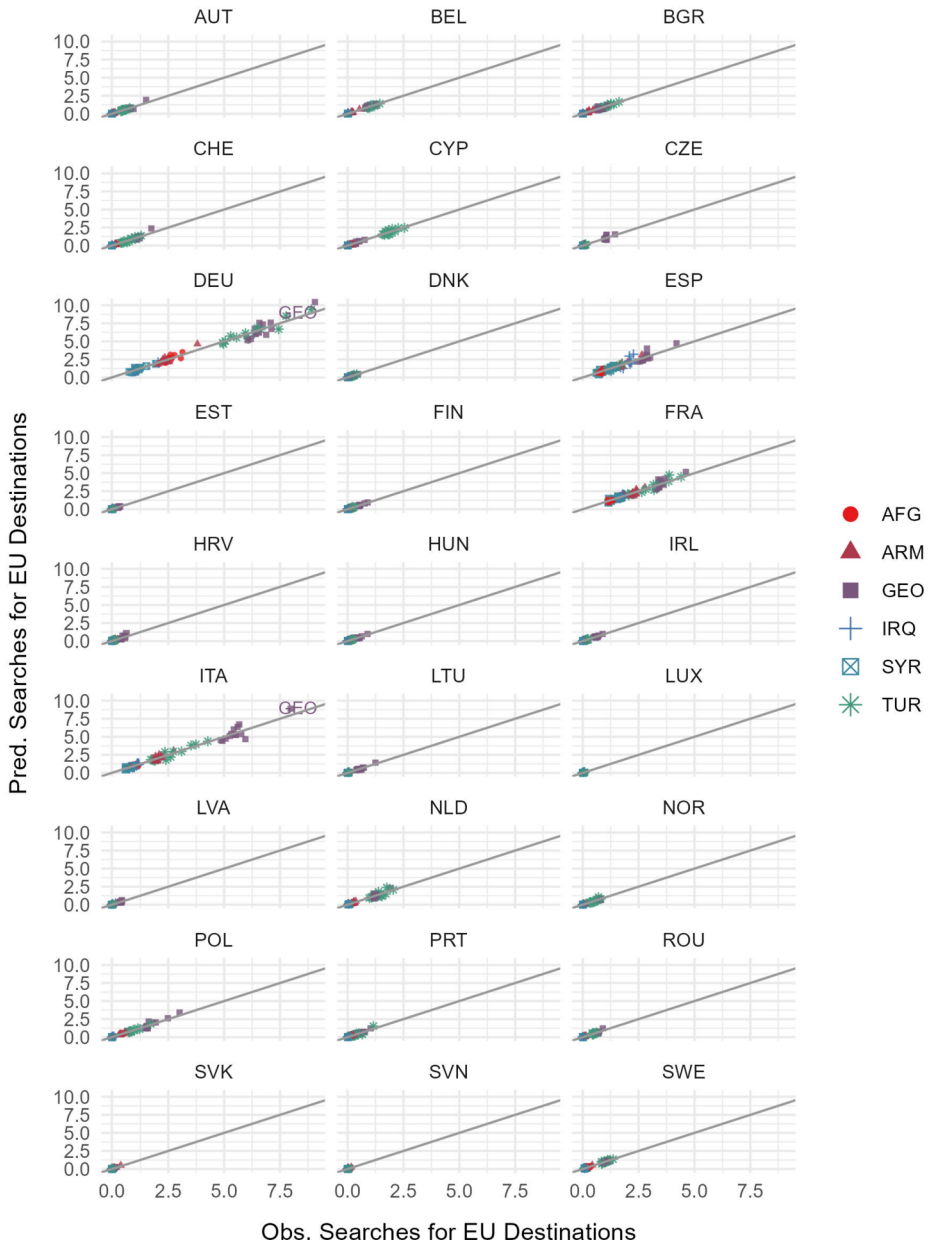

Figure 1 *Observed and Predicted Google Searches.* Each point corresponds to the observed and predicted Google Searches for EU destinations in a given country and year. Points' colors differentiate the geo-locations of Google searches. The dark solid lines represent 1:1 relationships

Using the predicted values from Eq. (14), we define people's preferences for migration destinations in Eq. (11) and Eq. (12) as,

$$\begin{aligned} &G_{k,t} = \widehat{S_{k,t}}, \\ &G_{k,t} \in [G_{o,o,t}, G_{o,d,t}, G_{tr,tr,t}, G_{tr,d,t}], \\ &\widehat{S_{k,t}} \in [\widehat{S_{o,o,t}}, \widehat{S_{o,d,t}}, \widehat{S_{tr,tr,t}}, \widehat{S_{tr,d,t}}]. \end{aligned} \tag{15}$$

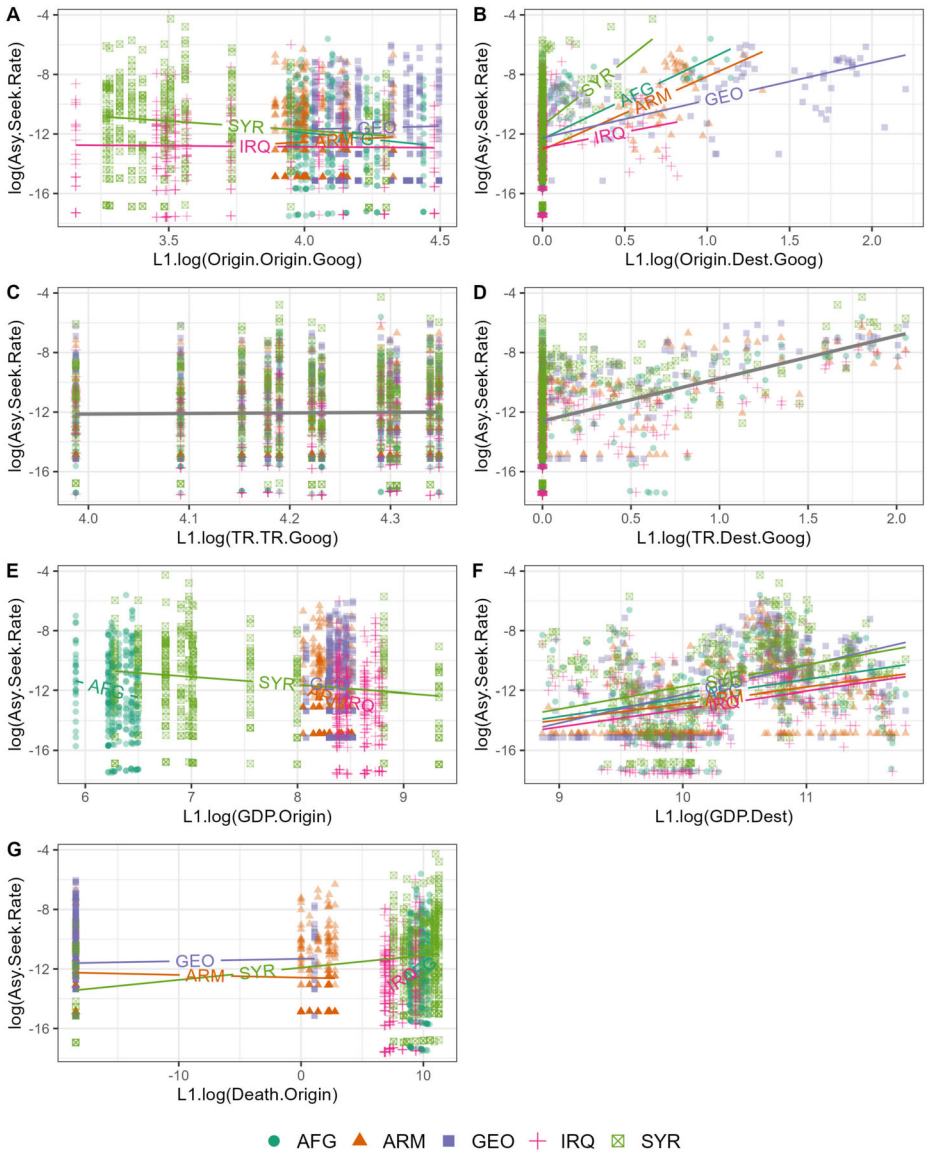

Figure 2 *Correlations between Asylum Seeking Rate and Predictors.* A: Log asylum-seeking rate and one year lag of log predicted Google searches about origin country in the origin country. B: Log asylum-seeking rate and one year lag of log predicted Google searches about destination country in the origin country. C: Log asylum-seeking rate and one year lag of log predicted Google searches about transit country in the transit country. D: Log asylum-seeking rate and one year lag of log predicted Google searches about destination country in the transit country. E: Log asylum-seeking rate and one year lag of log per capita GDP in the origin country. F: Log asylum-seeking rate and one year lag of log per capita GDP in the destination country. G: Log asylum-seeking rate and one year lag of log conflict-induced mortality in the origin country

Figure 2 illustrates how the Asylum Seeking Rates (ASR) are associated with GT searches about EU country topics, as well as with different push and pull factors. It is evident that ASR tend to increase with Google searches about EU destinations. This positive relation holds not only for the searches from origin country (Panel B), but also from transit country—Turkey (Panel D).

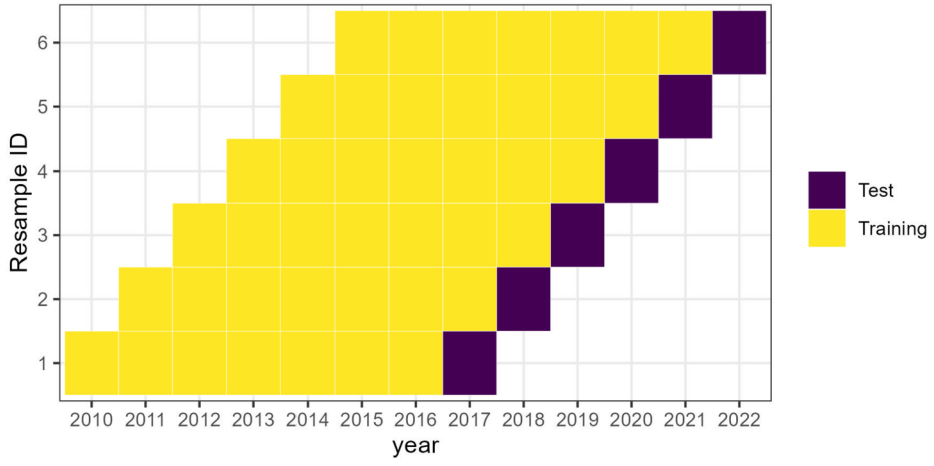

Figure 3 *Data Splits for Cross-validation*

4 Cross-validation

As stated at the outset, the overarching aim of this article is to systematically analyze how the predictive power of GT may vary depending on the capacity of migration models. To achieve this, we apply a cross-validation procedure to the baseline models Eq. (11) and Eq. (12), as well as their augmented versions (AR and FE). We then compare the effects of including GT on model performances (measured by R-square).

Figure 3 depicts how we split the data into training and testing sets. This cross-validation strategy is known as the "rolling forecasting origin" approach [35]. We choose this approach for two reasons. First, it can help us assess whether the improvements in models' performances by augmenting the baseline models are consistent overtime (over different training periods). Second, the "rolling forecasting origin" approach is also considered to be more appropriate for predicting future values of the outcome of interest [35]. It is important to stress that the choice of this cross-validation method is also to adapt to the flow Fixed-Effects (FE) model in which the values of the flow-specific constants are assumed to be fixed. Should these fixed parameter values be random draws from certain probability distributions, other cross-validation procedures might be more appropriate, such as K-fold, Leave-One-Out, or Re-sampling with Replacement.

5 Results

Figure 4 illustrates how the model performances have changed before and after including Google Trends (GT), i.e., from models with No GT ("No.Goog") to models with Linear GT ("Goog.Lin") and with Nonlinear GT ("Goog.NL"). It is evident that the effects of the inclusion of GT are not consistent across different model classes. In the Pooled regression (i.e., our baseline models Eq. (11) and Eq. (12)), there are noticeable improvements in models' performances; the R-Square values tend to increase in both training and testing sets. However, in the models augmented by an auto-regressive (AR) component, and by the flow Fixed-Effects (FE), the improvements become small in the training set, and negligible in the testing set. Another important note in Fig. 4 is that the nonlinear model ("Goog.NL") tend to increase explanatory power, but reduce the predictive power. This diverging pattern suggests that Eq. (11) is a preferred model, compared to Eq. (12), as it maintains a better balance between variance and bias.

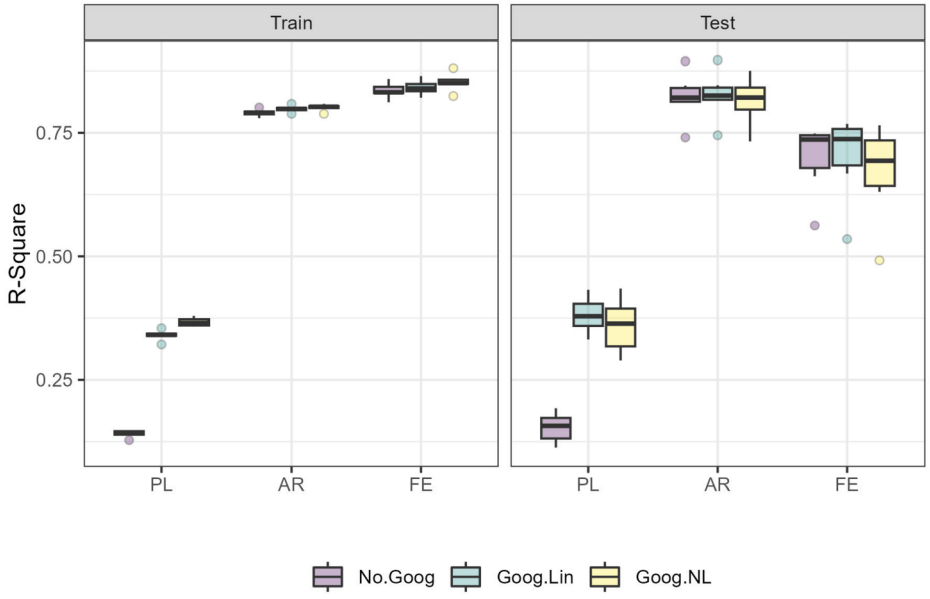

Figure 4 *Models' Performances Before and After the Inclusion of Google Trends* Each box depicts the distribution of the R-square values from six cross-validation samples (as shown in Fig. 3). Model classes are denoted by PL (Pooled Regression); AR (Auto-regressive Regression); and FE (Fixed-effects Regression). Colors are differentiated by No.Goog (models without indicators from Google Trends); Goog.Lin (models include indicators from Google Trends as linear predictors); Goog.NL (models include indicators from Google Trends as predictors in nonlinear polynomials)

To test whether the differences seen from Fig. 4 are statistically significant, we conduct a simple regression analysis in which the R-Square value is a function of a set of dummies indicating whether a model includes linear GT ("Goog.Lin") or nonlinear GT ("Goog.NL"). The results (shown in Fig. 5) clearly indicate that the effects of including GT is only significant for the Pooled (PL) regression, but not for those augmented by AR and FE. The test also confirms that in the PL model, the linear functional form of GT is more balanced in terms of variance-bias trad-off, whereas the nonlinear one is prone to over-fitting.

The model comparisons presented above shed some important light on when GT *can* and *cannot* improve the accuracy of migration forecasts. Specifically, they demonstrate that GT can significantly increase the prediction accuracy when the gravity model is kept relatively simple (e.g., when applying Pooled regression). However, when gravity model's complexity increases (e.g., augmented by AR or FE), GT can neither add explanatory power, nor enhance predictive performances. These patterns imply that the information captured by GT might be, partially, overlapping with what captured by the AR component and the flow-specific constants. As a result, GT becomes non-predictive in these more complex models.

Nevertheless, it is important to stress that, while the AR and FE models outperform the simple PL in terms of prediction accuracy, they are less valuable from an explanatory perspective. Specifically, many interesting determinants of migration are masked in these models. For example, language, cultural, and geographical proximity, and other time-invariant factors are vanished into the flow Fixed-Effects. Moreover, many time-varying drivers, such as migrant networks, migration policies and costs, are blended into the auto-regressive component. Should such factors be the primary interest, PL model is preferable.

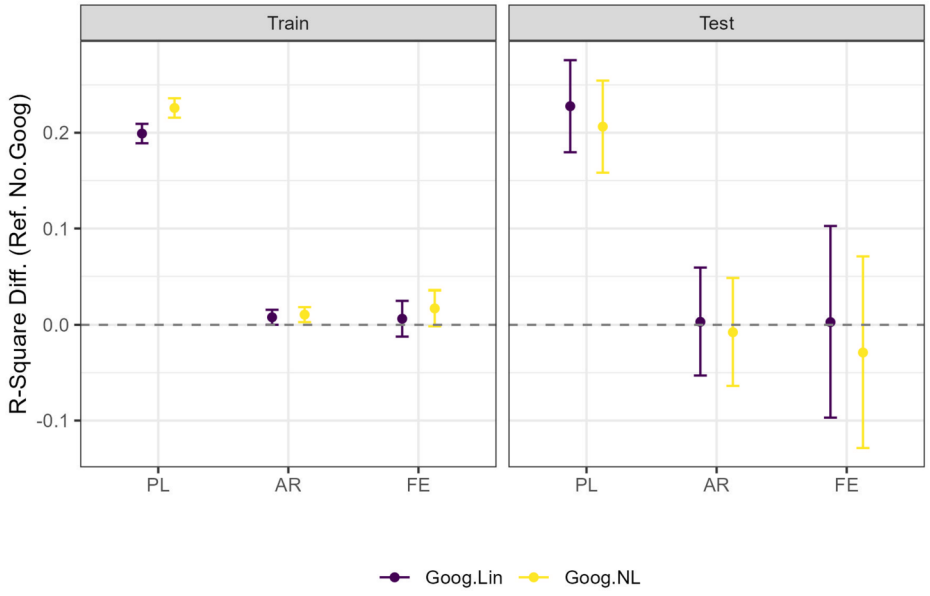

Figure 5 *Effects of Including Google Trends on Models' Performances* R-square differences between models with and without indicators from Google Trends. Model classes are denoted by PL (Pooled Regression); AR (Auto-regressive Regression); and FE (Fixed-effects Regression). Colors are differentiated by Goog.Lin (models include indicators from Google Trends as linear predictors); and Goog.NL (models include indicators from Google Trends as predictors in nonlinear polynomials)

In this regard, the trends in Google searches about EU countries are highly valuable, as it can, to a large extent, capture asylum-seekers' destination preferences, and hence adjust for the biases that may arise from the multilateral resistance in Eq. (9).

6 Conclusion

Google Trends (GT) has been increasingly adopted in migration forecasting, as it may contain information about *ex ante* intentions to migrate which are valuable for predictive analysis. While some demonstrated that GT can outperform any of the established predictors of migration flows, others argued that it is not always the case. In particular, even if within a single study, the predictive power of GT may vary depending on models' assumptions. Sensitivity as such highlights the need for recognizing the limits of GT's ability to predict future migration. Unlike most previous studies attempting to demonstrate the usability of GT for forecasting migration flows, this article emphasizes the importance of model complexity and the contextual influences that can affect the efficacy of GT as a predictive tool. Specifically, we address a critical but less discussed issue: when GT *cannot* improve the accuracy of migration forecasts.

Using EUROSTAT statistics on first-time asylum applications and a set of push-pull indicators gathered from various data sources, we train three classes of gravity models to predict refugees' choices of EU destinations. We then examine how the inclusion of GT affect the performances of these three model classes. The results suggest that the effects of including GT is highly heterogeneous and contingent on the complexity of these models. Specifically, in a simple Pooled (PL) regression model, the inclusion of GT significantly increased the R-Square value in both training and testing sets. However, the corresponding tests show no significant improvements in R-Square when the models are augmented by

an auto-regressive (AR) component or by flow Fixed-Effects (FE). These results challenge the view that GT can outperform any of the established predictors of migration flows [1].

In a broader context, our findings in this study provide a counterbalance to the prevailing optimism about the potential of big data for advancing migration research, and hence call for a more comprehensive analysis of the strengths and limitations of big data in predicting migration flows. The empirical framework used here not only contributes to the theoretical understanding of migrants' destination choices, it also offers a foundation for future research to explore a broader range of digital trace data and to more carefully evaluate their sensitivity to different methodological approaches. It is our hope that this nuanced perspective can spur further innovations in the field, and bring us closer to a comprehensive modeling framework of human migration patterns.

We would like to conclude this paper by stressing a key caveat: the findings presented here are specific to the asylum-seeking population from selected countries (AFG, ARM, GEO, IRQ, and SYR). Consequently, it is essential to extend and replicate this analytical approach to different geographic and demographic contexts, as the usability and predictive power of GT may be contextually contingent. For instance, in regions with high Google usage, GT data may offer valuable predictive insights even for complex models. Conversely, in regions with limited Google services, alternatives such as China's Baidu Index, Microsoft's Bing keyword research, or Russia-based Yandex might prove to be more effective.

Acknowledgements
The article has benefited from valuable comments of the journal editor, and anonymous referees. Financial support by European Unions Horizon 2020 Research and Innovation Program under grant agreement No. 870661 and No. 101004535, and the Swedish Research Council Vetenskapsradet under grant agreement 2022-06012-3 are gratefully noted.

Funding
Open access funding provided by Malmö University. Haodong Qi has received support from the Swedish Research Council Vetenskapsradet (grant agreement 2022-06012-3) and from the European Union's Horizon 2020 Research and Innovation Programme (grant agreement 101004535). Tuba Bircan has received support from the European Union's Horizon 2020 Research and Innovation Programme (grant agreement 870661). The funders did not play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abbreviations
GT, Google Trends; PL, Pooled; AR, Auto-regressive; FE, Fixed-effects; EU, European Union; IIA, Independent and Irrelevant Alternative; GAM, Generalized Additive Model; ASR, Asylum Seeking Rate; GDP, Gross Domestic Product; UCDP, Uppsala Conflict Data Program; AFG, Afghanistan; ARM, Armenia; GEO, Georgia; IRQ, Iraq; SYR, Syria.

Availability of data and materials
The datasets generated and/or analysed during the current study are available in the https://github.com/andykin888/GT-predict-refugee-flows.git.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author contributions
Conceptualization: HQ and TB; Data collection: HQ and TB; Methodology: HQ; Formal analysis and investigation: HQ; Writing: HQ and TB; Funding acquisition: HQ and TB. Both authors read and approved the final manuscript.

Author details
[1]Malmö University, Nordenskiöldsgatan 1, 205 06 Malmö, Sweden. [2]Vrije Universiteit, Brussels, Belgium.

Received: 29 November 2022 Accepted: 19 September 2023 Published online: 02 October 2023

References

1. Böhme MH, Gröger A, Stöhr T (2020) Searching for a better life: predicting international migration with online search keywords. J Dev Econ 142:102347
2. Carammia M, Iacus SM, Wilkin T (2022) Forecasting asylum-related migration flows with machine learning and data at scale. Sci Rep 12(1):1–16
3. Wanner P (2020) How well can we estimate immigration trends using Google data? Qual Quant 55(4):1181–1202. https://doi.org/10.1007/s11135-020-01047-w
4. Avramescu A, Wiśniowski A (2021) Now-casting Romanian migration into the United Kingdom by using Google search engine data. Demogr Res 45:1219–1254
5. Bosco C, Grubanov-Boskovic S, Iacus S, Minora U, Sermi F, Spyratos S (2022) Data Innovation in Demography, Migration and Human Mobility, EUR 30907 EN. Publications Office of the European Union, Luxembourg. https://doi.org/10.2760/027157
6. Adema J, Guha M, Adema JAH (2022) Following the online trail of Ukrainian refugees through Google Trends. In: CESifo forum, vol 23, pp 62–66. ifo Institut-Leibniz-Institut für Wirtschaftsforschung an der …, München
7. Spence PR, Lachlan KA, Rainear AM (2016) Social media and crisis research: data collection and directions. Comput Hum Behav 54:667–672
8. Nakamura N, Suzuki A (2021) Covid-19 and the intentions to migrate from developing countries: evidence from online search activities in southeast Asia. J Asian Econ 76:101348
9. Zagheni E, Weber I (2012) You are where you e-mail: using e-mail data to estimate international migration rates. In: Proceedings of the 4th annual ACM web science conference, pp 348–351
10. State B, Weber I, Zagheni E (2013) Studying inter-national mobility through IP geolocation. In: Proceedings of the sixth ACM international conference on web search and data mining, pp 265–274
11. Zagheni E, Weber I et al (2015) Demographic research with non-representative internet data. Int J Manpow 36(1):13–25
12. Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C (2014) Geo-located Twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 41(3):260–271
13. State B, Rodriguez M, Helbing D, Zagheni E (2014) Migration of professionals to the US. In: Social informatics: 6th international conference, SocInfo 2014, Barcelona, Spain, November 11–13, 2014. Proceedings, Lecture notes in computer science. Springer, Cham, pp 531–543
14. Varian HR (2014) Big data: new tricks for econometrics. J Econ Perspect 28(2):3–28
15. Grogger J, Hanson GH (2011) Income maximization and the selection and sorting of international migrants. J Dev Econ 95(1):42–57. https://doi.org/10.1016/j.jdeveco.2010.06.003
16. Pedersen PJ, Pytlikova M, Smith N (2008) Selection and network effects—migration flows into OECD countries 1990–2000. Eur Econ Rev 52(7):1160–1186. https://doi.org/10.1016/j.euroecorev.2007.12.002
17. Bertoli S, Moraga JF-H (2013) Multilateral resistance to migration. J Dev Econ 102:79–100. https://doi.org/10.1016/j.jdeveco.2012.12.001
18. Beine M, Bertoli S, Moraga JF-H (2015) A practitioners' guide to gravity models of international migration. World Econ 39(4):496–512. https://doi.org/10.1111/twec.12265
19. Anderson JE, Van Wincoop E (2003) Gravity with gravitas: a solution to the border puzzle. Am Econ Rev 93(1):170–192
20. Mcfadden D (1974) Conditional logit analysis of qualitative choice behavior. In: Zarembka P (ed) Frontiers in econometrics, vol 105. Academic Press, New York, p 42, Chap. 4
21. Train KE (2009) Discrete choice methods with simulation. Cambridge University Press, Cambridge
22. Beyer R, Schewe J, Lotze-Campen H (2022) Gravity models do not explain, and cannot predict, international migration dynamics. Humanit Soc Sci Commun 9(1):56. https://doi.org/10.1057/s41599-022-01067-x
23. Qi H, Bircan T (2023) Modelling and predicting forced migration. PLoS ONE 18(4):0284416
24. Bijak J, Disney G, Findlay AM, Forster JJ, Smith PW, Wiśniowski A (2019) Assessing time series models for forecasting international migration: lessons from the United Kingdom. J Forecast 38(5):470–487
25. Napierala J, Hilton J, Forster JJ, Carammia M, Bijak J (2022) Toward an early warning system for monitoring asylum-related migration flows in Europe. Int Migr Rev 56(1):33–62
26. Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski M, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014
27. Carrière-Swallow Y, Labbé F (2013) Nowcasting with Google Trends in an emerging market. J Forecast 32(4):289–298. https://doi.org/10.1002/for.1252
28. Vosen S, Schmidt T (2011) Forecasting private consumption: survey-based indicators vs. Google Trends. J Forecast 30(6):565–578
29. Da Z, Engelberg J, Gao P (2011) In search of attention. J Finance 66(5):1461–1499
30. Fantazzini D, Fomichev N (2014) Forecasting the real price of oil using online search data. Int J Comput Econ Econom 4(1–2):4–31
31. Vlastakis N, Markellos RN (2012) Information demand and stock market volatility. J Bank Finance 36(6):1808–1821
32. Askitas N, Zimmermann KF et al (2009) Google econometrics and unemployment forecasting. Appl Econ Q 55(2):107–120
33. D'Amuri F, Marcucci J (2017) The predictive power of Google searches in forecasting US unemployment. Int J Forecast 33(4):801–816
34. Miranda-González A, Aref S, Theile T, Zagheni E (2020) Scholarly migration within Mexico: analyzing internal migration among researchers using scopus longitudinal bibliometric data. EPJ Data Sci 9(1):34
35. Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practice. OTexts

Publisher's Note