Searching for a Better Life: Predicting International Migration with Online Search Keywords*

Marcus Böhme† André Gröger‡ Tobias Stöhr§

Version: April 11, 2019

Abstract

Migration data remains scarce, particularly in the context of developing countries. We demonstrate how geo-referenced online search data can be used to measure migration intentions in origin countries and to predict bilateral migration flows. Our approach provides strong additional predictive power for international migration flows when compared to reference models from the migration and trade literature. We provide evidence, based on survey data, that our measures partly reflect genuine migration intentions and that they outperform any of the established predictors of migration flows in terms of predictive power, especially in the bilateral within dimension. Our findings contribute to the literature by (1) providing a novel way for the measurement of migration intentions, (2) allowing real-time predictions of current migration flows ahead of official statistics, and (3) improving the performance of conventional models of migration flows.

JEL classification: F22, C82, J61, O15

Keywords: International Migration, Migration Intention, Google Trends

---

*We are grateful to the Editor-in-Chief Andrew Foster, to the Guest Editor Çaglar Özden, and to two anonymous referees for their insightful comments. We would like to thank Toman Barsbai, Christian Fons-Rosen, Andreas Fuchs, Stephen Hansen, Julian Hinz, Joan Llull, Juri Marcucci, Hannes Müller, Manuel Santos Silva, Claas Schneiderheinze, and Alessandro Tarozzi for useful comments and discussions. We also thank conference participants at the WIDER Development Conference on Migration and Mobility 2017, the annual conference of the German Economic Association's Research Group on Development Economics 2017, and seminar participants at Hamburg University, Goethe University Frankfurt, Pompeu Fabra University, and the Kiel Institute for the World Economy. We are grateful to Google Inc. for providing access to the Google Trends data through their API. Gröger acknowledges financial support from the Spanish Ministry of Economy and Competitiveness through grant ECO2015-67602-P and through the Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2015-0563). Stöhr acknowledges financial support from Stiftung Mercator (MEDAM, PN 14-297-0). Any remaining errors are our own.

†German Federal Ministry of Finance. The opinions expressed herein are those of the author and do not necessarily reflect those of the author's employer.

‡Corresponding author. Universitat Autònoma de Barcelona (UAB) and Barcelona Graduate School of Economics (BGSE). Contact: Dep. Economia i Història Econòmica, Edifici B, 08193 Bellaterra, Spain. E-mail: andre.groger@uab.cat.

§Kiel Institute for the World Economy (IfW) and IZA

1 Introduction

With profound effects on both origin and destination countries, migration has become one of the most important and most contested policy issues for developed and developing countries alike. There is a large body of literature dedicated to analyzing the determinants of international migration, which has identified that demographic factors, income differences, and violent conflicts are among the main push- and pull-factors. However, a lack of migration data is still plaguing academic research and effective policy making; the high costs of collecting nationally representative data on migration, inconsistent measures and definitions across data sources worldwide, as well as data publishing lags of several years still pose severe restrictions on migration research.[1] This is especially the case for developing and emerging countries in which administrative, survey-based migration flow or migration intention data are often unavailable, making many forms of analysis impossible.[2] As information technology is spreading rapidly around the world, geo-referenced online search data provides new and practically infinite opportunities for measuring and predicting human behavior through revealed demand for information (Varian 2014). The utilization of such big data sources is becoming increasingly important in applied economic research (Einav and Levin 2014), and scientific and technical advances have generated powerful tools, referred to as machine learning, that help analyze this complex data (Mullainathan and Spiess 2017).

Approaches that can help measure migration intentions and provide accurate predictions of recent flows are relevant to academics and policy makers alike. For these reasons, we propose a novel and *direct* measure of migration intentions using aggregate online search intensities, measured by what we henceforth call the Google Trends Index (GTI) for migration-related search terms.[3] Empirical evidence shows that aspiring migrants acquire relevant information about migration opportunities online, in their country of origin, prior to departure (Maitland and Xu 2015). This implies that, all else equal, demand for migration-relevant information can be used as a proxy for changes in the number of aspiring migrants. While gravity-type models can predict the between variation of international migration flows relatively well, because these depend on many static factors such as population size or distance between countries, existing approaches struggle to explain variation over time. Surges in online search intensities for specific keywords

[1] For example, in the case of the International Migration Database of the Organization for Economic Co-operation and Development (OECD), the lag is between two to three years.

[2] Apart from the coincidental existence of national surveys in some countries which include migration modules, to the best of our knowledge, there is only one multi-country survey which provides data on migration intentions for a larger set of countries of origin, the Gallup World Poll (GWP). The GWP has, however, at least two important disadvantages: First, it is not freely available and tends to be very costly. Second, it does not provide a consistent time series of migration intentions for specific origin countries.

[3] The GTI is a relatively high-frequency time series capturing the relative search intensities for any keyword performed through the Google search engine across the globe.

that relate to migration can indicate an increase in the demand for migration, reflect aspirations, and can thus help predict migration flows. Relying on Google search data, an engine which is estimated to be used by over a billion users worldwide, provides a high level of representativeness for internet users and can help offer a general tool for the prediction of migration.[4] We determine keywords related to migration based on a set of expressions which are semantically linked to this topic through their co-occurrence within the Wikipedia encyclopedia. We then extract the GTI for this list of keywords in the official language of the respective country of origin.

With our new measures, we test the predictive power of our GTI by augmenting workhorse models of international migration flows from a large range of origin countries to the destination countries of the Organisation for Economic Co-operation and Development (OECD). Using an extensive set of fixed effects and controlling for many potential push- and pull-factors typically used in the migration literature, we find that our approach yields substantial improvements in the predictive power of international migration flows models. The increase in predictive power is substantial in the unilateral setup which pools migration flows to the OECD at the level of the origin country and, therefore, focuses mainly on migrants' departure decision, independent of their destination choice. However, even when moving to a full bilateral gravity specification with an increasingly tight set of fixed effects, the predictive power of the GTI still outperforms any other benchmark model that we estimate by a large margin. To provide additional evidence of the robustness of the approach we explore the origin-country-level heterogeneity of the results and apply machine learning techniques to cope with potential concerns about in-sample overfit that might arise in selected regressions. The results confirm that our approach systematically improves the goodness-of-fit of international migration flow models. Last but not least, we also provide evidence, based on survey data, that our measures indeed reflect genuine emigration intentions.

The contribution of this article is threefold. First, we propose a novel approach to improve the measurement of migration intentions (or alternative outcomes) with consistent indicators that are freely available with a close to universal geographic coverage.[5] So far, the availability of data on migration intentions is severely restricted to selective and exclusive surveys. Easing this data constraint can help facilitate migration research, especially in the context of developing countries. Second, our approach is capable of providing short-term predictions of current migration flows ahead of official data release lags, which amount up to several years. This approach could, for example, be used for short-term

[4]Google is by far the most widely used search engine worldwide, having a market share of more than 80% on desktop devices and 97% on mobile and tablet devices. Source: https://www.netmarketshare.com/, accessed November 2017. Note that the representative character of the data is limited to internet users and many poor countries still lack internet coverage.

[5]The empirical strategy we propose here can readily be applied to the prediction of any alternative outcome variables.

policy prediction exercises in the case of humanitarian crises. Third, it can improve the predictive performance of conventional models for the determinants of migration flows.

There is a growing literature that uses big data from social networks and online search engines to predict economic outcomes across a large range of fields. In their seminal work, Choi and Varian (2012) suggest that online search data has a lot of potential to measure users' interest in a variety of economic activities in real time, and they also demonstrate how it can be used for the prediction of home and automotive sales. One of the most prominent applications so far has been published by Ginsberg et al. (2009), who show that levels of influenza activity can be predicted by the Google Flu Trend indicators with a reporting lag of only about one day. Despite a number of important criticisms (Lazer et al. 2014), the prediction literature has since grown quickly. It now includes applications to the prediction of aggregate demand (Carriére-Swallow and Labbé 2013) and private consumption (Schmidt and Vosen 2009), the number of food stamp recipients (Fantazzini 2014), stock market trading behavior and volatility (Da et al. 2011, Preis et al. 2013, Vlastakis and Markellos 2012), commodity prices (Fantazzini and Fomichev 2014), and even phenomena such as obesity (Sarigul and Rui 2014). The most frequent application to date is using Google Trends to predict unemployment, with applications in the context of France (Fondeur and Karamé 2013), Germany (Askitas and Zimmermann 2009), and the United States of America (D'Amuri and Marcucci 2017).

There is a small number of recent applications that have tried to use internet metadata to measure migration dynamics and patterns. Zagheni et al. (2014) use geo-referenced data of about half a million users of the social network "Twitter" in OECD countries, whereas Zagheni and Weber (2012) rely on the IP addresses of about 43 million users of the email service provider "Yahoo" to estimate international migration rates. The contribution of these studies is mainly methodological in the sense that they seek to provide an approach to infer trends about migration rates. However, by relying on data from highly specialized online services like Twitter and Yahoo, users' self-selection into these services severely limits the generalization of these results. Thus, those approaches are unsuitable for inferring general migration patterns. Furthermore, the data used in these studies is proprietary and, therefore, their analysis cannot be replicated or used in other contexts by external researchers.

The literature on the determinants of migration has evolved considerably in recent years[6]. We rely on some important contributions, such as the migration models of Mayda (2010) and Ortega and Peri (2013), to inform our empirical benchmark specification. Following Chort (2014), Docquier et al. (2014), and Dustmann and Okatenko (2014) we try to connect actual migration, which is limited by certain barriers such as budget constraints or migration policy, with migration intentions. Also, searching for information online might be a partial substitute for the information that is transmitted through

[6]See Docquier and Rapoport (2012) or Beine et al. (2016) for an overview of this literature

networks. Our results, thus, also relate to the previous literature on migration decisions and the role of migration networks (McKenzie and Rapoport 2010, Pedersen et al. 2008, Beine et al. 2011, Beine and Salomone 2013, Bertoli and Fernández-Huertas Moraga 2015, Bertoli and Ruyssen 2018).

The remainder of the paper is structured as follows. Section 2 describes the data used in the empirical part, with a particular emphasis on our specific GTI of migration intentions. In Section 3, we describe the panel models used to analyze the prediction of unilateral and bilateral migration. Section 4 provides estimation results. In Section 5 we compare the performance of our approach to that of the Gallup World Poll's migration intention questions. Section 6 offers our conclusion.

2 Data

2.1 Google Trends Data

Google Trends data are freely accessible at https://www.google.com/trends/ and have been generally available on a daily basis, since January 10, 2004.[7] The database provides time series of search intensities of the user's choice of keywords. In the current version of Google Trends, the GTI can be restricted by geographical area, date, a set of predefined general search categories such as "Jobs & Education" or "Travel", and by the type of search (i.e., standard web search, image, etc.).[8]

In order to match the structure of the OECD migration data that we use as the outcome variable, we download specific time series for each country of origin. The resulting GTI measure then captures the relative quantities of web searches through the Google search engine for a particular keyword in a given origin country as well as the specified time period.[9] Each subindex of the GTI provided by Google is normalized and ranges between 0 and 100. The maximum value of each country-keyword specific index is assigned to the peak of the respective time series during the selected period.[10] Thus, they do not contain useful variation of the absolute level of internet searches, rather they provide a useful signal in the within dimension.

[7]Extracting large quantities of Google Trends data through the website is, however, time consuming. Google offers access to their database through an Application Programming Interface (API) for registered users and non-commercial purposes. This approach provides an automated and efficient way of extracting the required data for our application and we rely on this API for the construction of our panel database (Google Inc. 2016). Due to the aggregate nature of the data, their use does not infringe on individual privacy rights.

[8]Depending on the country under investigation, sub-national disaggregation of the GTI is available down to the second administrative level. Note that "migration" has no predefined search category.

[9]For privacy reasons, the absolute numbers of searches are not publicly disclosed by Google. As the Google Trends database does not allow extracting yearly data directly, we extract monthly variations and aggregate them up to the annual level.

[10]Since we aggregate data at the annual level using the mean of the monthly values, most subindices of our GTI do not actually reach this maximum.

In essence, each time series reflects how the searches of a particular keyword through the Google search engine have changed over the years in a given country of origin. Geographical attribution is achieved through IP addresses, and are released only if the number of searches exceeds an – undeclared – minimum threshold. This implies that missing observations tend to occur predominantly in countries where Google search intensities are generally low, and when keywords and their combinations are rarely searched. Repeated queries from a single IP address within a short period of time are disregarded by Google to suppress potential biases arising from so-called internet bots searching the web. Finally, the index is calculated based on a sampling procedure of all IP addresses which changes over time and, thereby, introduces some measurement error into the time series. As a consequence, the indices can vary slightly according to the date of download. However, time series extracted at different times are nearly identical, with cross-correlations always above 0.99.

In order to operationalize the use of the GTI for our particular application and setting, we are faced with two non-trivial questions regarding the extraction of data: which keywords to choose and which language to extract them for? With respect to keyword selection, existing studies show a huge variety. Depending on the empirical context, the number of keywords chosen ranges between one and several thousand. For instance, D'Amuri and Marcucci (2017) simply use the term "jobs" in order to predict unemployment in the US. Carriére-Swallow and Labbé (2013) use a set of nine automobile brands in order to predict car sales. By contrast, Da et al. (2011) use a set of over 3,000 company names to predict stock prices. Technically speaking, the quantity of possible keywords and resulting data is close to infinity and only limited by the computing infrastructure and the maximum number of permitted downloads per day.

In the absence of a general pre-defined search category related to migration, we are left with the task of selecting individual keywords that we believe to be predictive of migration decisions in origin countries. Due to the multidimensionality of migration processes and motives, this task is more challenging than in other applications, where the set of potential keywords is rather narrow, such as in the case of car sales, oil prices, and unemployment registries. Given that for migration and topics of similar diversity, there is not one clear-cut search term, we rely on a broader set of keywords, whose exact composition is determined by an exogenous source.

In particular, we take advantage of semantic links between words in the Wikipedia encyclopedia related to the overarching topic of migration. We use the website "Semantic Link" (http://semantic-link.com/), which analyzes the text of English language Wikipedia and identifies pairs of keywords which are semantically related.[11] The website

[11]For that purpose the website uses a statistical measure called mutual information (MI). The higher the MI for a given pair of words, the higher the probability that they are related. The search is currently limited to words that have at least 1,000 occurrences in Wikipedia. Semantic links between words generated by this methodology change over time to the extent that Wikipedia is modified. Therefore,

displays the top 100 related words for each query and we retrieve those for the keyword "immigration". Since the majority of migration decisions are made for economic reasons such as employment, higher wages, or leaving poverty, we also retrieve a second list of semantically related terms based on the keyword "economics". For tractability reasons, we chose the subset of the top third most related keywords based on the two lists of 200 related keywords provided by the website (i.e., a total of 67). Additionally, we include the names of all OECD destinations that are in our migration data. The combination of the two sets of keywords also allows us to capture bilateral migration intentions.[12]

Finally, we are left with the empirical decision of which languages to extract the subindices of our GTI for our lists of keywords. We restrict the set of languages to the three official UN languages with Latin roots, i.e., English, French, and Spanish. For simplicity, we do not include the other official UN languages Arabic, Chinese (Mandarin), and Russian since the use of non-Latin characters imposes additional difficulties when extracting data. Based on these empirical choices and according to the "Ethnologue" database (https://www.ethnologue.com/statistics/size), we thereby capture the search behavior of an estimated 842 million speakers from 107 countries of origin in which at least one of the three selected languages is officially spoken.[13] We report the resulting lists of keywords in the three chosen languages in Tables 1 and 2.

As the spelling of keywords may differ between American and British English, we therefore include both versions in such cases. Similarly, we include both singular and plural forms of nouns where applicable because users might be searching for either. There are also differences between male and female forms, particularly in French, which we include. As expected, it turns out that most searches use male forms, especially when the plural is chosen. Furthermore, for the French and Spanish languages, we use both spellings with and without accents. Based on our selection of keywords, these different spellings only produce marginal differences with respect to the level of the corresponding GTI and their cross-correlations are very high. For our analysis, we combine the different versions of the same keyword with the Boolean operator "OR". Consequently, we capture their joint search intensity.

To investigate the predictive power of the GTI for unilateral migration decisions (i.e., all departures from a specific origin country to all OECD destinations), as well as for bilateral migration flows (i.e., towards a particular OECD destination), we extract two different types of Google trends indicators. First, we extract unilateral time series for each of our main keywords covering "'Migration & Economic" topics by country of origin and

the list retrieved today is unlikely to be identical to the one we obtained on January 16, 2015.

[12]By OECD we always refer to the 35 OECD member states that became members prior to 2018.

[13]For countries with speakers of more than one of our chosen languages, we select the language with the larger population share in the country of origin. Other languages with more than 200 million speakers that we do not cover include Hindi and Portuguese. Nevertheless, an extension into any type of language is technically feasible following our approach, provided that adequate translations and character conversions are available.

year. This translates into origin-specific time series of relative search intensities for each of the topical 67 keywords, over 12 years (2004–2015) in 101 countries of origin. Second, we extract bilateral indices combining each of our "Migration & Economics" keywords with a particular "OECD destination" by country of origin and year. The resulting data provides origin-destination-specific time series of relative search intensities for all topical keywords in 101 countries of origin and with respect to each of the 35 OECD countries of destination over 12 years.

We need to take into account a number of methodological pitfalls to which studies using Google Trends data tend to be subject to. First, it is not at all certain that people in origin countries who search for the chosen keywords online are genuinely interested in migration. They might just follow local or global search trends, which could have been ignited by news on migration or other topics in the media that spark interest regarding one or more of the chosen keywords. In other words, the change in search intensity could be driven by a diffusion of interest for an exogenous and unrelated topic, and not by genuine intentions to migrate. This argument has been put forward and illustrated by Ormerod et al. (2014) who investigated the precision of Google search activity to predict flu trends, originally proposed by Ginsberg et al. (2009). They find that social influence (i.e., the fact that people may search for a specific keyword at a specific moment simply because many others do so), may negatively affect the reliability of the GTI as a predictor for contemporaneous human behavior. This may be a problem, especially when relying on a small number of search terms. Therefore, we try to capture migration-related information demand by using a medium-sized set of keywords that are related to the topic, which can help smooth out such herding behavior in online search trends while avoiding the risk of selecting arbitrarily related keywords from hundreds of thousands of available ones.

Another potential risk of this approach, pointed out by Lazer et al. (2014), are changes in Google's search algorithms. Since Google is a commercial enterprise, it constantly adapts and changes its services in line with their business model. This could (and, if effective, should) influence the search behavior of users and, thereby, would change the data-generating process as well as the representativeness of the specific keywords chosen in this study over time. Due to this fact, we cannot rule out that search intensities increase due to adjustments made in the underlying search algorithms rather than an increased interest in migration. In other words, the index we created with the choice of our keywords in this exercise carries the implicit assumption that relative search volumes for certain search terms are statically related to external events. However, search behavior is not just exogenously determined, as it is also endogenously cultivated by the service provider. Such factors may give rise to a time-varying bias in the predictive power of our GTI variables and we account for this potential issue by including a set of yearly fixed effects in our empirical specification.

2.2 Migration and Country Data

Data on bilateral migration flows come from the OECD International Migration Database (IMD), which provides yearly immigrant inflows into the OECD countries by foreign nationality. Taking into account the geographical coverage of origin countries, the yearly frequency, and how up-to-date it is, the IMD is the most comprehensive data set available in our empirical setting. The sample includes almost all countries of origin worldwide, both from the group of developing and developed countries. Despite migration flow data being available from earlier periods, we are confined to a panel spanning 12 years because the GTI is only available from 2004.

While the OECD puts major efforts into ensuring the comparability of their IMD data across countries, inconsistencies still persist due to differences in national migrant definitions (e.g. using the place of birth or nationality) or data sources (e.g. census, residence registries, or specific surveys). This can cause analytic problems, especially in cross-country studies (Ortega and Peri 2013). However, as our preferred specification relies on changes of migration flows over time within bilateral corridors, this does not constitute a problem in our analysis.[14]

We match this panel of migration flows with macroeconomic indicators for each origin and destination country from the World Development Indicators (WDI) (World Bank 2015). These covariates help to control for the most important push- and pull-factors that have been emphasized in the migration literature (Mayda 2010, Ortega and Peri 2013). In our benchmark setup, we restrict the list of covariates to total GDP and population size to avoid losing too many origin countries for which more detailed control variables are unavailable. However, many other predictors have been used in the literature as additional control variables. In an extension, we therefore include additional origin country controls from the WDI that are inspired by the literature on migration and migration intentions such as, the unemployment rate, the share of the young population, the share of internet users (per 100 people), and mobile phone subscriptions (per 100 people). We also include the number of weather and non-weather disasters from the EM-DAT International Disaster Database (Guha-Sapir 2018). To control for political factors, we include the Polity IV Autocracy Score and the State Fragility Index (Marshall et al. 2016). These variables are meant to capture dimensions of local amenities (Dustmann and Okatenko 2014). Furthermore, since our approach relies heavily on language choice and its effective use among the native population in the countries of origin, we also use

[14] A potential concern in our setting is related to reverse causality, which is related to OECD destination countries that use residence permits to measure migration (e.g. France and Italy). This could occur if immigration amnesties in those countries lead to spikes in the migration data and if these co-occur with increases in migration-related searches in origin countries (e.g. for the keywords "legalization", "unauthorized", or "undocumented") one year earlier. In unreported regressions, we check for the robustness of our findings with respect to this issue by excluding migration flows to these two countries. The results are consistent with our main results and provide evidence that this is not an issue in our context.

data on the share of the native population that commonly speaks the official languages in origin countries (Melitz and Toubal 2014). These allow us to analyze the extent to which the respective languages spoken in the countries of origin have an influence on the performance of the GTI. Unfortunately, many of these additional control variables are missing for our sample of origin countries, leading to significantly lower sample sizes when including the set of extended controls.

2.3 Descriptive Statistics

After merging the bilateral GTI with the migration and country data outlined, we end up with a bilateral panel of migration flows from 101 origin countries to 35 OECD destination countries which covers 12 years. Accounting for missing values in our bilateral benchmark specification (introduced in Section 3), the resulting sample size is 23,947 origin-destination-year observations.

In Table 3, we report descriptive statistics for the main variables of the bilateral panel data set used in the analysis. The average bilateral migration flow is 742 individuals, with a large standard deviation. Many bilateral migration corridors have zero values. The largest flow in our sample occurred between Mexico (origin) and the USA (destination) in 2007 with almost 190,000 migrants. In the second line, we provide descriptive statistics for the bilateral GTI "destination country", the name of the latter being the keyword in this case. For the abovementioned largest flow, this variable reflects the relative search intensity in Mexico for the keyword "USA" compared to the other years. The fact that the variable has a mean of around 14 and a standard deviation of 16, indicates that these time series also have many zero observations. In other words, there are origin countries in which searches for the respective keywords are too rare to pass the undeclared threshold that Google imposes before data are reported. We report descriptive statistics for the complete list of bilateral constituents of the GTI in Appendix Table A1.[15]

As one would expect, the total GDP of the OECD destinations is almost 4 times greater than that of origin countries, which belong predominantly to the group of middle- and low-income countries. In the specifications with extended control variables, we also approximate labor market prospects at origin by using the unemployment rate and control for additional population dynamics by using the share of the young population. We also include two variables measuring state functionality: the State Fragility Index and the Polity IV Autocracy Score. We include these variables to proxy for push-factors, such as security concerns and malfunction of political systems. We also control for the penetration of mobile phone and internet technology at the origin, which are crucial prerequisites for the use of internet search engines. Last but not least, we also include the number of

[15]In fact, there is one bilateral GTI variable corresponding to the keyword "emigrant", which has missing observations for all origin countries. This variable is dropped in the bilateral analysis.

weather and non-weather-related disasters in the country of origin to proxy for additional push factors.

If our approach works, the signals extracted from the GTI should track actual migration flows relatively well. To show this graphically, we plot fitted values from a simple OLS regression that tries to explain next year's migration flow with current GDP, population size, and an origin-specific intercept in the unilateral setup as a benchmark. We then add the GTI to this model. The selected examples of origin countries differ by language, levels and changes in macroeconomic fundamentals, and show distinct behavior of the migration flow throughout the sample period. The results are depicted in Figure 1. The solid line represents log aggregate migration flows (plus one) from six origin countries to the OECD. The dashed line represents the fitted values from a regression of the migration flows on GDP, population, and origin fixed effects. This line is slow moving and shows that these predictors provide little time variation to explain year-on-year fluctuations in migration flows. In the model represented by the dotted line, we then add our unilateral GTI predictors. We observe that the fit between the dotted and solid line improves greatly, which confirms that the time variation in the GTI variables helps to track actual migration flows to a much better degree, compared to the classic migration predictors.

3 Empirical Methodology

In order to investigate whether the GTI can improve the prediction of migration decisions, we estimate a range of fixed effects panel models using two different specifications: a unilateral and a bilateral model.

In the unilateral fixed effects model, the dependent variable is the aggregate annual flow of migrants from a given country of origin to the group of all OECD countries. We first estimate a benchmark specification of this model and, subsequently, augment it with our unilateral GTI time series, capturing the internet search intensities for the selected keywords. The results from this model are informative about the predictive power of GTI for aggregate emigration decisions among the origin population, irrespective of the migrant's destination choice. The unilateral fixed effects equation we estimate is:

$$Y_{ot+1} = \beta_1 GTIuni_{ot} + \beta_2 O_{ot} + \gamma_o + \delta_t + \varepsilon_{ot} \tag{1}$$

with o indexing the country of origin and t time. The dependent variable, Y_{ot+1}, is the log of annual migration flows (plus one) from the origin country to all OECD destination countries in year $t+1$. We lag the outcome variable by one year to reflect that preparing for migration takes time and to mitigate concerns about reverse causality. $GTIuni_{ot}$ represents our unilateral GTI for a given origin country with respect to a specific keyword

in a given year.[16] O_{ot} is a vector of origin-specific control variables, γ_o stands for origin country-specific fixed effects, and δ_t are year fixed effects.[17] ε_{ot} represents a robust error term, which is clustered at the origin country level. The set of fixed effects in this specification absorbs time-invariant factors at the origin country level as well as aggregate changes over time. Therefore, the identifying variation comes from changes in the origin search intensities for selected Google keywords over time.

For the unilateral analysis to match the dimension of the GTI that vary at the country of origin level, we collapse the bilateral panel data set by OECD destination country. To put it differently, our outcome variable in this setup is the aggregate migration flows from one given origin to all OECD countries. Thus, we implicitly focus on the general migration decision of the country of origin[18] and abstract it from the sorting decision (i.e., the decision of which destination country to immigrate to).

In our benchmark case, we can thus exploit a sample of 98 countries of origin over 12 years. The corresponding total sample size is 1,068 origin-year observations. Due to some missing values, this sample size decreases to 70 origin countries and 700 observations in the extended control specification.

Second, we estimate a bilateral fixed effects model. This specification is also informative about the predictive power of the GTI with respect to migrants' destination choices. We follow Bertoli and Fernández-Huertas Moraga (2013) and Beine et al. (2016) for the specification of the gravity equation and include different sets of fixed effects which replicate the most important workhorse models of bilateral migration flows from the literature (Mayda 2010, Ortega and Peri 2013). We even go beyond the most demanding specification found in the migration literature and estimate the following gravity equation, inspired by the trade literature (Head and Mayer 2015):

$$Y_{odt+1} = \beta_1 GTIbil_{odt} + \beta_2 GTIuni_{ot} \times GTIdest_{odt} + \beta_3 O_{ot} + \beta_4 D_{dt} + \gamma_{ot} + \delta_{dt} + \tau_{od} + \varepsilon_{odt} \quad (2)$$

with d indexing the OECD destination country. The dependent variable, Y_{odt+1}, is the log annual bilateral flow of migrants (plus one) from a given origin to each OECD destination in year $t+1$. $GTIbil_{odt}$ is the vector of bilateral GTI variables. It is composed of a set of variables, each of them based on the combined query of the destination's name and one of the topical keywords ("Migration" and "Economics"). In addition, the destination name is added on its own. $GTIuni_{ot} \times GTIdest_{odt}$ is a vector of interaction terms between the unilateral GTI constituents and the "OECD destination" GTI keyword. Applying our

[16] Continuing with the previous example, $GTIuni_{ot}$ reflects the relative search intensity in Mexico for each of the topical "Migration" and "Economics" keywords.

[17] Given the use of year fixed effects, we do not include control variables at the OECD level.

[18] Lacking the respective yearly emigration panel data we have to omit all non-OECD destinations.

example, $GTIbil_{odt}$ stands for the relative search intensity in Mexico for the combined query of "work USA", in any order and possibly combined with any other terms, as in "find work in USA". $GTIdest_{odt}$ represents the relative search intensity in Mexico for "USA". We have two reasons to include two versions of bilateral GTI variables. First, as mentioned in Section 2, the combination of more complex keyword queries such as $GTIbil_{odt}$ increases the likelihood of zero observations in countries with low internet traffic due to the undeclared minimum search intensity threshold of the Google Trends database. This is at least partly alleviated by the inclusion of $GTIuni_{ot} \times GTIdest_{odt}$ which is based on single keyword queries only. Second, including an interaction term between the vector of unilateral "Migration & Economics" GTI variables and the bilateral "OECD destination" measure allows for a more flexible functional form in this setup. Both versions of the bilateral GTI capture different signals and have complementary predictive power. D_{dt} is a vector of destination-specific control variables. γ_{ot} is a vector of origin-time specific fixed effects, δ_{dt} a vector of destination-time specific fixed effects, and τ_{od} are origin-destination pair fixed effects. ε_{odt} represents the robust error term. The set of fixed effects in this specification absorb time-varying factors at the levels of both origin and destination country (e.g., economic development, population dynamics, etc.) or unilateral policy changes, as well as time-invariant factors at the bilateral level such as distance, common language, and shared borders. As a consequence, in our most rigorous specifications, the exploited variation is based exclusively on within-variation in the search intensity of bilateral keywords.

The presence of zero flows in the estimation of gravity models, as common with bilateral migration data, constitutes an empirical problem since the estimation on the reduced sample of non-zero observations or using a log-linearized model under heteroscedasticity may lead to biased parameter estimates (Santos Silva and Tenreyro 2006). Alternative estimation techniques such as the Poisson pseudo-maximum likelihood have been shown to work well in this situation (Santos Silva and Tenreyro 2011) and have found their way into the migration literature (Beine et al. 2016). The application of these techniques is of foremost importance for studies that conduct causal parameter estimation. In contrast, our focus is on the predictive power of different model specifications as represented by the coefficient of variation. Since this measure is not affected by heteroscedasticity and to facilitate comparability with the related literature, we consistently rely on scaled OLS estimation in the following analysis. Note, however, that the parameter estimates reported may be subject to bias and should therefore be interpreted with caution.

4 Panel Estimations

4.1 Unilateral model

The results from the unilateral fixed effects estimations based on equation 1 are reported in Table 4. Column (1) displays the result for our benchmark regression specification, including only the restricted set of origin control variables, i.e., log GDP and population. This basic model of migration flows results in a sample size of 1,068 observations from 98 origin countries. In this empirical setting of aggregate migration flows, the origin and year fixed effects explain most of the variation in this model as reflected by the high values of overall-R^2. Nevertheless, the crucial performance indicator for our application is represented by the within-R^2, which reflects the coefficient of variation from the mean-deviated regression, i.e., the ordinary R^2 from running OLS on the transformed data.[19] In other words, controlling for a time-invariant origin and aggregate year factors, we are interested in the predictive power of time-varying origin-specific explanatory variables. In column (1), the basic set of predictors yields a within-R^2 of 6.2%. Once we augment the basic model by the vector of unilateral GTI variables in column (2), the within-R^2 increases to 24.2%, suggesting that the GTI provides strong additional explanatory power.

In columns (3) and (4), we repeat the same exercise with an extended set of control variables. Due to missing observations in the vector of additional controls, we are left with a sample size of 700 observations including only 70 countries of origin in this specification. In the benchmark specification in column (3), the additional controls increase the within-R^2 considerably to 16.6%. However, even compared to this benchmark model of extended controls, adding the unilateral GTI in column (4) still more than doubles the within-R^2 to 35.5%. This highlights that the improvement in the predictive power of the GTI is robust to comparing their performance against an extensive set of control variables from the migration literature. In sum, the results from the unilateral model so far, show that the GTI variables substantially improve the goodness-of-fit for the estimation of international migration flows.

In the following specifications (5) through (8), we explore language- and technology-related heterogeneity across origin countries. The underlying hypothesis is that the GTI variables should become more predictive for migration decisions in countries which are linguistically more homogeneous or in which there is a high penetration of internet technology. In columns (5) and (6) we focus on the set or origin countries in which at least 50% of the population commonly use either of the three the language in which we have

[19] Other than in most out-of-the-box estimates of R^2 with higher dimensional fixed effects, we calculate the within-R^2 such that any time-varying fixed effects are not included within it. The amount of variation captured by our within-R^2 is thus always purely driven by real explanatory variables. This is implemented with the Stata package *regxfe* (Fernando Rios-Avila 2016).

extracted the GTI variables. This restriction excludes 31 linguistically heterogeneous origin countries from the sample. We are left with a sample size of 732 observations from 67 origins. In column (5), we find that the within-R^2 increases only marginally to 7.4% compared to 6.2% in column (1). Reassuringly, when augmenting this model by the vector of GTI variables, the within-R^2 increases substantially to 31.6%. This constitutes an increase in the within-R^2 by factor 4.3, which is significantly larger than the improvement in columns (1) and (2) (factor 3.9).

In the last specification in columns (7) and (8), we restrict the sample to countries in which at least 10% of the population have access to the internet.[20] The resulting sample size is 647, including 79 origins. As reported in column (7), the within-R^2 in this specification is 9.2%. When again including the vector of GTI variables, the within-R^2 increases to 42.2%, which constitutes an increase by the factor 4.7. The results from exploring origin country heterogeneity, therefore, confirms our hypothesis that the GTI is more predictive in countries which are linguistically more homogeneous and have a higher level of internet penetration.

The results from the unilateral specification presented so far have to be interpreted with caution because adding a large vector of GTI variables to the unilateral model significantly increases model complexity and decreases the ratio of observations per predictor (last line in Table 4). A low ratio of observations per predictor[21] increases the risks of in-sample overfit, i.e., of picking up a spurious correlation between the time series and the dependent variable (Babyak 2004). Different rules of thumb regarding this ratio advise that a minimum of 10-15 observations per predictor is necessary to achieve reliable estimations. Based on simulations, Babyak (2004) shows that even at intermediate values of around 13, overfit is still possible and argues for a more conservative interpretation of these rules of thumb.

In order to deal with this potential issue head-on, we apply two techniques to guard against in-sample overfit (Varian 2014, Kleinberg et al. 2015). The results of this exercise are reported in Section A of the Online Appendix. First, we conduct out-of-sample predictions using k-fold cross-validation techniques. These results, summarized in Figure A1, show that the increase in the goodness-of-fit by introducing the GTI also holds out

[20]For the data generation process, the general availability and the use of internet technology among the local population of the origin country is crucial. We observed marked differences in the number of internet users across countries, which are positively correlated with economic development at the origin. According to data from the International Telecommunication Union, the rate of internet users among the general population was only 12% for low-income economies in 2016, compared to 42% in middle- and 82% in high-income economies, respectively. Source: World Telecommunication / ICT Development Report and database, and World Bank estimates (URL: https://data.worldbank.org/indicator/IT.NET.USER.ZS, accessed: November 2017).

[21]The number of predictors is calculated as the number of variables added into the model, e.g., 2 and 2+67 in columns 1 and 2 of Table 4, respectively. Fixed effects are not counted here since the data is demeaned and, accordingly, their inclusion does not increase the effective number of predictors in the estimation.

of sample. This would not be the case if it was due merely to overfit. Second, we apply a variable selection method that penalizes larger numbers of covariates in a model. This algorithm includes a considerably larger number of regressors than what could be expected if the within variation only consisted of pure noise. Together these results provide evidence that the results from the unilateral model are not driven by overfit.

4.2 Bilateral model

The results from the bilateral fixed effects estimations based on the gravity equation 2 are reported in Table 5.[22] We estimate five different reference models with increasingly demanding sets of fixed effects (A–E), first in a benchmark setup and, second, augmenting it with the vector of bilateral GTI variables. We proceed as before and analyze the predictive power of the GTI variables in the bilateral dimension. The results from the most basic fixed effects specification model A, based on earlier work by Mayda (2010), are reported in columns (1) and (2). The benchmark model in column (1) includes separate destination, origin, and year fixed effects as well as a basic vector of destination and origin control variables (i.e., log GDP and population size). The fixed effects absorb common changes over time and time-invariant factors on each side of the migration corridor. The resulting within-R^2 in this setting is only 0.1%, which is driven exclusively by the set of basic control variables which turn out to have low predictive power for bilateral migration flows. On the other hand, the set of fixed effects explains 73.2% of the overall variation. Once we add our bilateral GTI predictors to this benchmark model in column (2), we observe a similar effect as before in the unilateral model analysis, with the within-R^2 increasing strongly to 27.2%. The overall-R^2 also increases to 80.5%.

In the second specification B, we further augment the gravity model, including fixed effects for each destination as well as each origin-year combination in the fashion of Ortega and Peri (2013). This setup explicitly accounts for multilateral resistance to migration (Bertoli and Fernández-Huertas Moraga 2013) and the fixed effects absorb all time-invariant factors at the destination country as well as changes over time at the origin country level (e.g. GDP and population). Again, as reported in column (3), the within-R^2 in the benchmark setup is close to zero (0.06%), reflecting the poor predictive power. On the other hand, the resulting overall-R^2 is similar to the one in setup B at 73.5%. When augmenting this model again with the bilateral GTI variables, the within-R^2 increases to 29.9%, while the overall-R^2 reaches 81.4%.

In specification C, we estimate the bilateral gravity model including fixed effects for each destination-year and origin-year combination. These fixed effects absorb all changes over time at the origin and destination country level, such as economic development or population dynamics, which is why the basic set of control variables drops out completely.

[22]Note that the model specified in equation 2 corresponds to the fixed effects setup E in this Table.

This implies that the resulting within-R^2 is zero. The overall-R^2 of this benchmark model is 73.9%. When augmenting this model with the bilateral GTI variables, once again, the within-R^2 reaches 31.1% and the overall-R^2 increases to 82%, signaling a significant improvement in the goodness of fit.

Specification D includes destination-origin pair and origin-year fixed effects which corresponds to the most demanding specification of the gravity equation estimated by Ortega and Peri (2013). This model absorbs all time-varying origin factors as well as time-invariant bilateral factors, such as common language, colonial ties, or land borders. The results from the benchmark specification in column (7) show that the within-R^2 again is only 0.7%, driven by the basic set of destination-specific control variables. Based on this tight set of fixed effects the overall-R^2 reaches a value of 97.1%. Moving on to column (8), we see that the within-R^2 again increases to 2.4%, while the overall-R^2 remains constant. Including bilateral fixed effects in this specification amounts to all between variation being purged out of the GTI variables. The only systematic variation left in the model is destination-year and pair-year variation. Even in this dimension, our GTI predictors contribute to an improvement in explaining the remaining variation.

The last specification E of the bilateral model we estimate includes a saturated vector of bilateral as well as destination-year and origin-year fixed effects. It is inspired by the specification of gravity models in the international trade literature and has, to the best of our knowledge, not been applied in the estimation of migration flows thus far. This model controls for time-invariant bilateral factors as well as changes over time at origin *and* destination. This accounts for all destination and origin-specific factors over time as well as for the vast majority of time-invariant bilateral factors as discussed by Beine et al. (2016). Consequently, all the signal left from the GTI comes from its changes over time within each bilateral corridor. The results of this exercise are reported in columns (9) and (10). By construction, the within-R^2 in the benchmark column (9) is zero, as all predictors are absorbed by the fixed effects, which together explain 97.4% of the overall variation. Nevertheless, when looking at the results from the augmented model in column (10) the within-R^2 increases to 1.9%. Apart from the Gallup World Poll, we are not aware of any time-varying bilateral migration predictors available for the estimation of migration flows in the literature. As described in the next section, the bilateral migration intention questions from the Gallup survey do not survive this demanding test. Given the saturated set of fixed effects and the fact that this model is far more demanding than existing gravity models used in the migration literature, this result once again confirms the predictive power of the bilateral GTI.

The GTI-based approach aims at exploiting the signal that can be extracted from people's search behavior as they look for information online. An important substitute is the information gained through diaspora networks. There is a consensus in the migration literature that networks increase migration by reducing migration costs (Pedersen et al.

2008, McKenzie and Rapoport 2010, Beine et al. 2011, Beine and Salomone 2013, Beine et al. 2015, Bertoli and Fernández-Huertas Moraga 2015). In particular, Bertoli and Ruyssen (2018) investigate the role of networks for the intention to migrate. They find that having a distance-one connection in a specific destination increases the likelihood of intending to migrate to this destination by a factor of 6 to 8, conditional on intending to migrate. Along these lines, we are interested in the predictive performance of the diaspora size at destination compared to our GTI variables. In order to investigate this, we augment the vector of control variables in equation 2 by the logarithmic transformation of the OECD migrant stock at the destination (plus one) for each bilateral corridor and replicate the specifications in Table 5.[23] The results from this robustness check can be found in Appendix Table A3. Note that the results are not directly comparable to bilateral baseline results due to different sample sizes. As expected, the coefficient on the migrant stock is positive and highly significant. The results show that the inclusion of the diaspora size leads to a jump in overall-R^2, indicating that the stock of migrants explains a large amount of variation in migration flows. At the same time, it also increases the within-R^2, especially in specifications A through C, which do not include bilateral fixed effects. Nevertheless, even when controlling for diaspora size, the GTI variables still improve the predictive power in these specifications by around 13%. This becomes more pronounced once we include bilateral fixed effects in specification D and E, focusing exclusively on the bilateral within dimension. While the diaspora size yields a within-R^2 of 3.2% in column (7), including the GTI almost doubles this value to 5.9% in column (8). Finally, in the fully saturated fixed effects specification E, the within-R^2 is close to zero, signaling a poor performance of migrant stocks when focusing on changes over time within the bilateral dimension. Notwithstanding, adding the GTI increases this low R^2 fivefold to about 2%.

As our results from this section show, the GTI offers strong additional predictive power for bilateral migration flows and outperforms any of the established predictors as well as any benchmark specification of the gravity model that we have tested from the migration and trade literature. Also, given the high ratio of observations to predictors in the bilateral specifications, which is always above 170, we can rule out that these results are driven by overfit. Although the additional predictive power from the GTI decreases with increasing saturation of fixed effects, the fact that they yield positive results (even in the most ambitious fixed effects setup) provides clear evidence in favor of its predictive power, both in the between and within dimension of bilateral migration flow models.

[23] See Llull (2016) for a description of the OECD database of bilateral migrant stocks.

5 Beyond Predictive Power?

We have presented evidence that our tailor-made GTI lead to significant improvements in the predictive power of models of international migration flows, both in the unilateral and bilateral dimension. As emphasized by Mullainathan and Spiess (2017), the prediction objective (i.e., generating a prediction of outcome y based on independent variables x) should not be confused with that of classic parameter estimations, where the focus is on the effect of x on y. In other words, the results provided so far testify to a robust correlation but are not informative about causality or the mechanism that is captured to improve prediction. This does not mean that causality is not important. In many applications, it is not enough to simply have a well-performing predictor. Gaining an understanding whether it indeed works as assumed (in our case we suggest it approximates migration aspirations) is reassuring and can reduce the risk of falling prey to any spurious correlation that would be unlikely to yield a good prediction in the future. Therefore, in what follows, we use the Gallup World Poll (GWP) survey questions on migration intentions to compare their performance to our approach.

The GWP, which started in 2006, is a private and exclusive survey conducted at varying intervals from one up to several years, with many countries now receiving attention on a yearly basis. Each sample is independent. Thus, the GWP consists of repeated cross-sections. The data are based on a stratified random sample that is considered nationally representative. Each wave consists of typically around 1,000 respondents per country (more for very large countries). The survey is implemented either as a face-to-face or telephone interview with subjects older than 15 years.[24] We rely on three migration-related questions from this survey which are designed to assess individuals' migration intentions to different degrees. These questions are:

1. *Ideally, if you had the opportunity, would you like to move permanently to another country, or would you prefer to continue living in this country?* And, if yes: *To which country would you like to move?*

2. *Are you planning to move permanently to [COUNTRY] in the next 12 months?*

3. *Have you done any preparation for this move? For example, have you applied for residency or a visa, purchased the ticket, etc.?*

These questions are framed such that they reflect an increasing migration aspiration intensity. While question one indicates the respondent's potential and abstract aspiration for migration in general, question two intends to elicit whether the individual plans to

[24] Stratification is based on population size and the geography of sampling units. Further details about the survey methodology can be accessed online at: http://www.gallup.com/178667/gallup-world-poll-work.aspx.

realize this intention in the short-term. Question three captures whether the respondent has indeed started to make concrete preparations.

The GWP data reveals that every year, the highest absolute number of people with reported migration intentions live in China, Nigeria, and India. In relative terms, small countries such as Haiti, Sierra Leone, and the Dominican Republic tend to have the highest migration intentions as a share of the adult population. The most popular destination countries have changed over time. In 2007, the most frequently mentioned preferred destinations were the USA, the UK, Saudi Arabia, Canada and Spain. By 2017 this had shifted to the USA, Canada, Germany, Australia, and France, respectively.

Aggregating the data across countries, the GWP indicates that among the nearly 5 billion people it represents worldwide, approximately 660 million people reported general migration intentions according to question one in 2010, compared to 677 million in 2015. Analyzing migration intentions, plans, and actual preparation lead to vastly different numbers. In 2010 only about 27 million out of 4.8 billion reported to have an active plan for migrating during the following 12 months and approximately half of those also reported to have started preparing their move at the time of the survey. By 2015 this had increased to 67 and 23 million people respectively. In relative terms, this amounts to almost doubling numbers of potential migrants despite a stagnation of general migration intentions. Thus, for predictive purposes, it seems advisable to use all three migration questions rather than just resorting to general intentions.

In order to compare the GWP migration intentions to our GTI, we first augment the set of basic control variables in our bilateral estimation equation 2 by the variables corresponding to the three GWP questions.[25] The results of this exercise are reported in Table 6. The sample size in this specification is 21,855 observations from 2,611 bilateral corridors, which is lower than in the main bilateral regressions, due to a later start of the GWP survey as well as non-annual coverage or exclusion of some origin countries from the GWP. Note that the results are not directly comparable to the ones from the previous section due to different sample sizes.[26] In column (1) through (6), we can observe that the GWP variables generally carry the expected positive sign and are highly significant as predictors of bilateral migration flows. The GWP variables only lose their predictive performance once pair fixed effects are added in columns (7) through (10), meanwhile, the GTI still contributes positively. In other words, in a horserace between the GTI and GWP predictors, the GTI shows strong predictive performance both in the between and within dimension, whereas the appeal of the Gallup data comes mainly from the

[25]We do not implement the same exercise in the unilateral setting because, first, the GWP variables have a bilateral dimension. Second, missing data in the GWP leads to a further drop in the sample size compared to Table 5. This results in the ratio of observations per predictor being far below the critical threshold of 10, thereby, magnifying the problem of in-sample overfit.

[26]Results from estimations on the correct comparison sample are reported as robustness check in Table A3.

between dimension and for descriptive statistics (e.g. comparing migration intentions across countries).

Given the lack of predictive power of the GWP data in the bilateral dimension and the positive results found for the GTI, it shows that the latter hold promise for further alleviating the lack of migration data. For example, fine-tuning our migration GTI to specific bilateral corridors through the inclusion of bilateral keyword choices and contextual language could also allow the implementation of specific policy predictions. The Gallup variables are unsuitable for such kind of prediction tasks because their collection requires substantial field work, which implies a considerable time lag before they become available.

The question remains: can the GTI also be used directly to measure migration intentions? In order to shed some light on this question, we regress each of the GWP variables on the set of bilateral GTI variables. Results are reported in Table A4 in the Online Appendix. As we can observe, many GTI coefficients show a statistically significant correlation with the GWP, far more of them than we would expect if this was up to pure chance.[27] These coefficients carry both positive and negative signs. We interpret this as being a consequence of cross-correlations among the set of GTI variables in the presence of non-linearities. Reassuringly, in most cases, the sign and significance level is relatively constant when comparing across specifications (1) through (3). As expected, the magnitude of these coefficients also decreases from left to the right, due to fewer respondents reporting to "prepare" or "plan" for migration, compared to general intentions as captured by "demand". As reflected in the overall-R^2 values, the GTI explains almost 17% of the GWP variation in specification (1) and 7% and 6% in specification (2) and (3), respectively.[28] This signals that there seems to be a partial overlap in the predictive power between the two sets of variables, but also that the variation captured by them is not congruent.

In summary, these tests provide the first evidence that our GTI indeed captures meaningful variation for the prediction of migration flows and also outperform the predictive performance of established survey-based measures. Whether and to what extent the GTI can also be applied as a direct measure of migration intentions by itself cannot be answered conclusively here. This would require a larger national sample of bilateral migration intentions, for example, from a specialized survey on migration or census, which is also representative of migrant households. The answer to this question is thus left for future research.

[27] There is one bilateral GTI keyword ("emigrant", as in "emigrant to Canada") which did not yield any positive observations for our sample of countries. Consequently, the variable is excluded from the analysis.

[28] Note that the values of overall-R^2 appear generally lower in this specification, compared to the main analysis, because the regressions do not include fixed effects or control variables.

6 Conclusion

In this paper, we have presented evidence that using information on internet search intensities for specific keywords in migrants' countries of origin can help estimate international migration flow models. This holds true both in a setup that focuses on aggregate emigration decisions from a specific origin country and in the bilateral dimension when accounting for destination choices. In line with our expectations, these results become stronger when restricting our sample to origin countries where the internet is more widely used and where the languages for which we test our approach are more widely spoken. Using survey data we provide evidence that our GTI partly reflects genuine migration intentions.

Our findings contribute to different parts of the migration literature related to measuring and predicting international as well as internal migration. First, our methodology might be able to help improve data availability on migration intentions by offering freely available, high-frequency indicators that even have sub-national geographic coverage. By selecting GTI constituents based on keywords with semantic links to other topics, our methodology could even serve as a general guideline of how to make use of the GTI in other prediction contexts. This could be particularly helpful when the availability of additional control variables is poor, for example, in the context of sub-national or real-time data requirements. Second, our approach could be used to generate short-term predictions of current migration flows ahead of official data releases, which in practice can have lags amounting up to several years. This could be used for policy applications in the case of humanitarian crises in order to deliver real-time monitoring of migration intentions ahead of their realization so as to be able to design responsive policies. This is comparable to recent advances in the political economy literature demonstrating that newspaper text can be used to predict armed conflict ahead of time (Mueller and Rauh 2018). Similar to macroeconomic forecasting models such a nowcasting application of our approach would start with a model that uses lagged dependent variables and the GTI to predict changes in the outcome of interest in close to real time (D'Amuri and Marcucci 2017).

Can a GTI-based approach be feasible for the prediction of international migration flows in the long-run? The experience of *Google Flu Trends* for the United States has shown that there are several obstacles (Ginsberg et al. 2009, Lazer et al. 2014). The predictive power of the keywords we employ in this study to capture migration intentions may change over time. Evolving associations between individual keywords and the outcome variable are likely to affect the composition of the "optimal" prediction model in the future, the creation of which is beyond the scope of this paper. Surging interest in a particular keyword may cause its worth for a prediction to plummet. Therefore, we advocate an approach based on a broader set of keywords in order to smooth out potential

biases that could occur for specific keywords over time. Furthermore, our approach can be further tailored to the empirical context, especially when concerned with short-run predictions of migration flows in a subset of countries, for example, by optimizing the selection of keywords. Here, a combination with text analysis tools, e.g. based on media reports, could also be helpful to capture other sources of semantic links. An interesting empirical test for future work could be to investigate the impact of an exogenous shock on migration-specific GTI and on migration flows in a sub-national setting. This would allow calibrating the coefficients and to measure the association between the shock on the one hand and migration intentions according to the GTI and real-life migration realizations on the other.

Bibliography

Askitas, N. and Zimmermann, K. F.: 2009, Google Econometrics and Unemployment Forecasting, *Applied Economics Quarterly* **55**(2), 107–120.

Babyak, M. A.: 2004, What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models, *Psychosomatic Medicine* **66**(3), 411–421.

Beine, M., Bertoli, S. and Fernández-Huertas Moraga, J.: 2016, A Practitioners' Guide to Gravity Models of International Migration, *The World Economy* **39**(4), 496–512.

Beine, M., Boucher, A., Burgoon, B., Crock, M., Gest, J., Hiscox, M., Mcgovern, P., Rapoport, H., Schaper, J. and Thielemann, E.: 2015, Comparing Immigration Policies: An Overview from the IMPALA Database.

Beine, M., Docquier, F. and Özden, Ç.: 2011, Diasporas, *Journal of Development Economics* **95**(1), 30–41.

Beine, M. and Salomone, S.: 2013, Network Effects in International Migration: Education versus Gender, *Scandinavian Journal of Economics* **115**(2), 354–380.

Bertoli, S. and Fernández-Huertas Moraga, J.: 2013, Multilateral resistance to migration, *Journal of Development Economics* **102**, 79–100.

Bertoli, S. and Fernández-Huertas Moraga, J.: 2015, The size of the cliff at the border, *Regional Science and Urban Economics* **51**, 1–6.

Bertoli, S. and Ruyssen, I.: 2018, Networks and migrants' intended destination, *Journal of Economic Geography* **18**(4), 705–728.

Carriére-Swallow, Y. and Labbé, F.: 2013, Nowcasting with Google trends in an emerging market, *Journal of Forecasting* **32**(4), 289–298.

Choi, H. and Varian, H.: 2012, Predicting the Present with Google Trends, *Economic Record* **88**(SUPPL.1), 2–9.

Chort, I.: 2014, Mexican migrants to the US: What do unrealized migration intentions tell us about gender inequalities?, *World Development* **59**, 535–552.

Da, Z., Engelberg, J. and Gao, P.: 2011, In Search of Attention, *Journal of Finance* **66**(5), 1461–1499.

D'Amuri, F. and Marcucci, J.: 2017, The predictive power of Google searches in forecasting US unemployment, *International Journal of Forecasting* **33**(4), 801–816.

Docquier, F., Peri, G. and Ruyssen, I.: 2014, The cross-country determinants of potential and actual migration, *International Migration Review* **48**(S1), S37–S99.

Docquier, F. and Rapoport, H.: 2012, Globalization, Brain Drain, and Development, *Journal of Economic Literature* **50**(3), 681–730.
URL: *http://pubs.aeaweb.org/doi/10.1257/jel.50.3.681*

Dustmann, C. and Okatenko, A.: 2014, Out-migration, wealth constraints, and the quality of local amenities, *Journal of Development Economics* **110**, 52–63.

Einav, L. and Levin, J.: 2014, Economics in the age of big data, *Science* **346**(6210), 1243089.

Fantazzini, D.: 2014, Nowcasting and forecasting the monthly food stamps data in the US using online search data, *PLoS ONE* **9**(11).

Fantazzini, D. and Fomichev, N.: 2014, Forecasting the real price of oil using online search data, *International Journal of Computational Economics and Econometrics* **4**(1/2), 4–31.

Fernando Rios-Avila: 2016, REGXFE: Stata module to fit a linear high-order fixed-effects model.

Fondeur, Y. and Karamé, F.: 2013, Can Google data help predict French youth unemployment?, *Economic Modelling* **30**(1), 117–125.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. and Brilliant, L.: 2009, Detecting influenza epidemics using search engine query data, *Nature* **457**(7232), 1012–1014.

Google Inc.: 2016, Google Trends Application Programming Interface.

Guha-Sapir, D.: 2018, EM-DAT: The Emergency Events Database.
URL: *www.emdat.be/*

Head, K. and Mayer, T.: 2015, Gravity Equations: Workhorse, Toolkit, and Cookbook, *Handbook of International Economics*, Vol. 4, pp. 131–195.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z.: 2015, Prediction Policy Problems, *American Economic Review: Papers & Proceedings* **105**(5), 491–495.

Lazer, D., Kennedy, R., King, G., Vespignani, A., Butler, D., Olson, D. R., McAfee, A., Brynjolfsson, E., Goel, S., Tumasjan, A., Bollen, J., Ciulla, F., Metaxas, P. T., Lazer, D., Vespignani, A., King, G., Boyd, D., Crawford, K., Ginsberg, J., Cook, S., Copeland, P., Viboud, C., Thompson, W. W., Hall, I. M., Ong, J. B. S., Ortiz,

J. R., Mustafaraj, E., Metaxas, P., Ratkiewicz, J., King, G., Voosen, P., Lazarus, R., Chunara, R., Balcan, D., Chao, D. L., Shaman, J., Karspeck, A., Shaman, J., Nsoesie, E. O., Hannak, A. and Berinsky, A. J.: 2014, Big data. The parable of Google Flu: traps in big data analysis., *Science (New York, N.Y.)* **343**(6176), 1203–5.

Llull, J.: 2016, Understanding international migration: evidence from a new dataset of bilateral stocks (1960–2000), *SERIEs* **7**(2), 221–255.

Maitland, C. and Xu, Y.: 2015, A Social Informatics Analysis of Refugee Mobile Phone Use : A Case Study of Za'atari Syrian Refugee Camp, *TPRC*.

Mallows, C. L.: 1973, Some Comments on Cp, *Technometrics* **15**(4), 661.

Marshall, M. G., Gurr, T. R. and Jaggers, K.: 2016, Polity IV project: Political Regime Characteristics and Transitions, 1800-2016 and State Fragility Index and Matrix.
URL: *www.systemicpeace.org*

Mayda, A. M.: 2010, International migration: A panel data analysis of the determinants of bilateral flows, *Journal of Population Economics* **23**(4), 1249–1274.

McKenzie, D. and Rapoport, H.: 2010, Self-selection patterns in Mexico-U.S. migration: The role of migration networks, *Review of Economics and Statistics* **92**(4), 811–821.

Melitz, J. and Toubal, F.: 2014, Native language, spoken language, translation and trade, *Journal of International Economics* **93**(2), 351–363.

Mueller, H. and Rauh, C.: 2018, Reading Between the Lines: Prediction of Political Violence Using Newspaper Text, *American Political Science Review* **112**(2), 358–375.

Mullainathan, S. and Spiess, J.: 2017, Machine Learning: An Applied Econometric Approach, *Journal of Economic Perspectives* **31**(2), 87–106.
URL: *http://pubs.aeaweb.org/doi/10.1257/jep.31.2.87*

Ormerod, P., Nyman, R. and Bentley, R. A.: 2014, Nowcasting economic and social data: when and why search engine data fails, an illustration using Google Flu Trends, *arXiv preprint arXiv:1408.0699* .

Ortega, F. and Peri, G.: 2013, The Effect of Income and Immigration Policies on International Migration, *Migration Studies* **1**(1), 1–35.

Pedersen, P. J., Pytlikova, M. and Smith, N.: 2008, Selection and network effects-Migration flows into OECD countries 1990-2000, *European Economic Review* **52**(7), 1160–1186.

Preis, T., Moat, H. S. and Stanley, H. E.: 2013, Quantifying trading behavior in financial markets using Google Trends., *Scientific reports* **3**, 1684.

Santos Silva, J. M. and Tenreyro, S.: 2006, The log of gravity, *Review of Economics and Statistics* **88**(4), 641–658.

Santos Silva, J. M. and Tenreyro, S.: 2011, Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator, *Economics Letters* **112**(2), 220–222.

Sarigul, S. and Rui, H.: 2014, Nowcasting Obesity in the U.S. Using Google Search Volume Data, number 166113, Agricultural and Applied Economics Association.

Schmidt, T. and Vosen, S.: 2009, Forecasting Private Consumption, *Economic Papers* **155**, 23.

Tibshirani, R.: 1996, Regression Selection and Shrinkage via the Lasso, *Journal of the Royal Statistical Society B* **58**(1), 267–288.

Varian, H. R.: 2014, Big Data: New Tricks for Econometrics, *Journal of Economic Perspectives* **28**(2), 3–28.

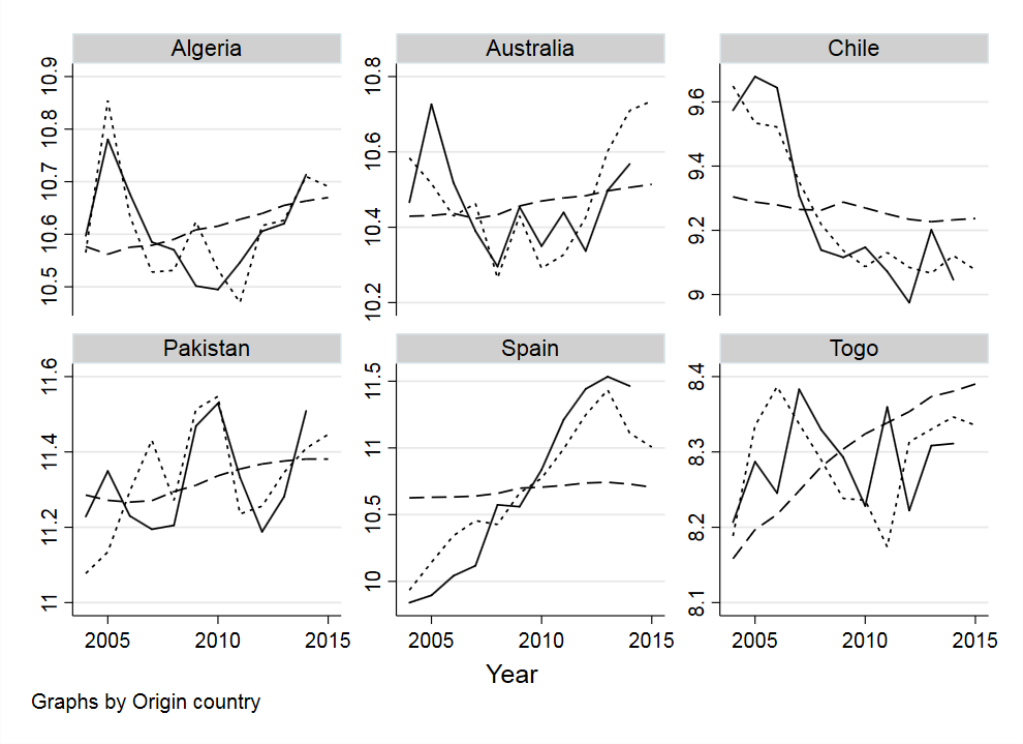
Vlastakis, N. and Markellos, R. N.: 2012, Information demand and stock market volatility, *Journal of Banking and Finance* **36**(6), 1808–1821.

World Bank: 2015, World Development Indicators.
URL: *http://data.worldbank.org/data-catalog/world-development-indicators*

Zagheni, E., Garimella, V. R. K., Weber, I. and State, B.: 2014, Inferring international and internal migration patterns from Twitter data, *Proceedings of the companion publication of the 23rd international conference on World wide web companion. International World Wide Web Conferences Steering Committee.* .

Zagheni, E. and 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.* .

Figures and Tables



Notes: The figure shows log migration flows (plus one) from six origin countries to the OECD (solid line) and fitted values of two simple regressions that use log GDP, log population size, origin-specific intercepts and fixed effects (dashed line) plus the GTI (dotted line). The regressions are estimated on the full sample including all countries, the model used to fit the data is thus identical across panels. Differences between dotted and dashed lines are thus based on changes in GTI search intensities. As the dashed line shows, GDP and population size change too slowly to explain large short term fluctuations in migration flows.

Figure 1: Descriptive illustration of GTI in predicting migration flows

Table 1: List of Main Keywords

English	French	Spanish
applicant	candidat	solicitante
arrival	arrivee	llegada
asylum	asile	asilo
benefit	allocation sociale	beneficio
border control	controle frontiere	control frontera
business	entreprise	negocio
citizenship	citoyennete	ciudadania
compensation	compensation	compensacion
consulate	consulat	consulado
contract	contrat	contrato
customs	douane	aduana
deportation	expulsion	deportacion
diaspora	diaspora	diaspora
discriminate	discriminer	discriminar
earning	revenu	ganancia
economy	economie	economia
embassy	ambassade	embajada
emigrant	emigre	emigrante
emigrate	emigrer	emigrar
emigration	emigration	emigracion
employer	employer	empleador
employment	emploi	empleo
foreigner	etranger	extranjero
GDP	PIB	PIB
hiring	embauche	contratacion
illegal	illegal	ilegal
immigrant	immigre	inmigrante
immigrate	immigrer	inmigrar
immigration	immigration	inmigracion
income	revenu	ingreso
inflation	inflation	inflacion
internship	stage	pasantia
job	emploi	trabajo
labor	travail	mano de obra
layoff	licenciement	despido
legalization	regularisation	legalizacion
migrant	migrant	migrante
migrate	migrer	migrar
migration	migration	migracion
minimum	minimum	minimo
nationality	nationalite	nacionalidad
naturalization	naturalisation	naturalizacion
passport	passeport	pasaporte
payroll	paie	nomina
pension	retraite	pension
quota	quota	cuota
recession	recession	recesion
recruitment	recrutement	reclutamiento
refugee	refugie	refugiado
remuneration	remuneration	remuneracion
required documents	documents requis	documentos requisito
salary	salaire	sueldo
Schengen	Schengen	Schengen
smuggler	trafiquant	traficante
smuggling	trafic	contrabando
tax	tax	impuesto
tourist	touriste	turista
unauthorized	non autorisee	no autorizado
undocumented	sans papiers	indocumentado
unemployment	chomage	desempleo
union	syndicat	sindicato
unskilled	non qualifies	no capacitado
vacancy	poste vacante	vacante
visa	visa	visa
waiver	exemption	exencion
wage	salaire	salario
welfare	aide sociale	asistencia social

Note: For GTI data retrieval, both singular and plural as well as male and female forms of these keywords are used where applicable. In the English language, both British and American English spelling is used. All French and Spanish keywords were included with and without accents.

Table 2: List of OECD Destinations

English	**French**	**Spanish**
Australia	Australie	Australia
Austria	Autriche	Austria
Belgium	Belgique	Belgica
Canada	Canada	Canada
Chile	Chili	Chile
Czech Republic	Republique Tcheque	Republica Checa
Denmark	Danemark	Dinamarca
Estonia	Estonie	Estonia
Finland	Finlande	Finlandia
France	France	Francia
Germany	Allemagne	Alemania
Greece	Grece	Grecia
Hungary	Hongrie	Hungria
Iceland	Islande	Islandia
Ireland	Irlande	Irlanda
Israel	Israel	Israel
Italy	Italie	Italia
Japan	Japon	Japon
Latvia	Lettonie	Letonia
Luxembourg	Luxembourg	Luxemburgo
Mexico	Mexique	Mexico
Netherlands	Pays-Bas	Paises Bajos
New Zealand	Nouvelle-Zelande	Nueva Zelanda
Norway	Norvege	Noruega
Poland	Pologne	Polonia
Portugal	Portugal	Portugal
Slovak Republic	Republique Slovaque	Republica Eslovaca
Slovenia	Slovenie	Eslovenia
South Korea	Coree du Sud	Corea del Sur
Spain	Espagne	Espana
Sweden	Suede	Suecia
Switzerland	Suisse	Suiza
Turkey	Turquie	Turquia
United Kingdom	Royaume-Uni	Reino Unido
United States	Etats-Unis	Estados Unidos

Note: For GTI data retrieval, both singular and plural as well as male and female forms of these keywords are used where applicable. In the English language, both British and American English spelling is used. All French and Spanish keywords were included with and without accents. Additionally, English acronyms in the case of the United Kingdom (UK) and the United States of America (USA) were included.

Table 3: Descriptive statistics for the main variables of the bilateral panel data set

	Count	Mean	SD	Min	Max
Bilateral migration flow	23,947	741.91	4,675.16	0.00	189,989
Bilateral GTI "destination"	23,947	14.38	15.76	0.00	94
Total GDP (destination)	23,947	1,139.81	2,149.69	12.79	14,797
Total population (destination)	23,947	34.99	52.47	0.29	319
Total GDP	23,947	297.69	1,490.84	0.02	14,797
Total population	23,947	34.86	134.42	0.01	1295
Unemployment rate	19,101	7.82	5.32	0.60	38
Share of young population	21,075	32.79	9.70	11.72	50
State Fragility Index	20,001	9.56	5.95	0.00	24
Polity IV Autocracy Score	20,001	-1.18	13.89	-88.00	9
Mobile cellular subscriptions (per 100 people)	22,549	71.88	42.89	0.66	200
Internet users (per 100 people)	21,822	26.00	25.39	0.19	95
Number of weather-related disasters	23,947	1.71	3.21	0.00	27
Number of non-weather-related disasters	23,947	0.52	0.95	0.00	9

Sources: OECD International Migration database 2004–2015, World Development Indicators, Polity IV, State Fragility Index, and EM-DAT International Disasters Database. *Notes*: Bilateral migration flows according to the OECD IMD. All other variables refer to the origin country, unless otherwise indicated. Total GDP in billion USD (constant 2005 USD). Total population in millions..

Table 4: Unilateral model including Google Trends Indices

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Benchmark	GTI	Benchmark	GTI	Benchmark	GTI	Benchmark	GTI
Sample	All		Extended controls		Spoken Language > 50%		Internet Access > 10%	
Log GDP (origin)	-0.641***	-0.486**	-0.276	-0.340	-0.996***	-0.801***	-1.038***	-0.821**
	(0.231)	(0.191)	(0.245)	(0.233)	(0.283)	(0.211)	(0.390)	(0.314)
Log Population (origin)	2.161***	1.681***	0.690	0.729	1.730**	1.368*	2.154**	1.817**
	(0.597)	(0.624)	(0.880)	(0.847)	(0.788)	(0.719)	(0.859)	(0.759)
Unemployment rate			0.0225	0.00671				
			(0.0161)	(0.00959)				
Share of young population			0.0306	-0.00707				
			(0.0293)	(0.0307)				
State Fragility Index			0.000270	0.00534				
			(0.0109)	(0.0122)				
Polity IV Autocracy Score			-0.000771	-0.000536				
			(0.000950)	(0.000873)				
Mobile cellular subscriptions			-0.00191	-0.00117				
			(0.00160)	(0.00149)				
Internet users			-0.00903***	-0.00767***				
			(0.00265)	(0.00288)				
No. weather-related disasters			-0.00137	-0.00251				
			(0.00580)	(0.00654)				
No. non-weather-related disasters			-0.0189*	-0.0109				
			(0.0111)	(0.00915)				
GTI (unilateral)	–	√	–	√	–	√	–	√
Joint significance GTI (p-value)	–	0.000	–	0.000	–	0.000	–	0.000
Fixed effects								
Origin	√	√	√	√	√	√	√	√
Year	√	√	√	√	√	√	√	√
Observations	1,068	1,068	700	700	732	732	647	647
Number of origins	98	98	70	70	67	67	79	79
R^2 (within)	0.062	0.242	0.166	0.355	0.074	0.316	0.092	0.422
R^2 (overall)	0.988	0.991	0.987	0.990	0.990	0.992	0.992	0.995
Observations per predictor	534	15.5	107	9.1	366	10.6	324	9.4

Sources: Authors' calculations based on OECD IMD 2004–2015, World Development Indicators, Google Trends, Polity IV, State Fragility Index, and EM-DAT International Disasters Database. *Notes*: Each column displays the result of a separate regression based on equation 1. Dependent variable is the logarithm of the annual migration flow (plus one) from a given origin country to all OECD destinations. Robust standard errors, clustered at the origin country level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Bilateral model including Google Trends Indices

	(1) Benchmark	(2) GTI	(3) Benchmark	(4) GTI	(5) Benchmark	(6) GTI	(7) Benchmark	(8) GTI	(9) Benchmark	(10) GTI
Specification	(A)		(B)		(C)		(D)		(E)	
Log GDP (destination)	0.523***	0.154	0.494**	0.147			0.832***	0.791***		
	(0.203)	(0.222)	(0.207)	(0.230)			(0.185)	(0.182)		
Log Population (destination)	-1.845***	-0.0920	-1.802***	-0.0427			-1.911***	-1.899***		
	(0.364)	(0.433)	(0.357)	(0.424)			(0.325)	(0.324)		
Log GDP (origin)	-0.450***	-0.0814								
	(0.0989)	(0.117)								
Log Population (origin)	0.456*	1.299***								
	(0.276)	(0.320)								
GTI (bilateral)	–	√	–	√	–	√	–	√	–	√
Joint significance GTI (p-value)	–	0.000	–	0.000	–	0.000	–	0.000	–	0.000
Fixed effects										
Destination	√	√	√	√	–	–	–	–	–	–
Origin	√	√	–	–	–	–	–	–	–	–
Year	√	√	–	–	–	–	–	–	–	–
Destination-year	–	–	–	–	√	√	–	–	√	√
Origin-year	–	–	√	√	√	√	√	√	√	√
Destination-origin	–	–	–	–	–	–	√	√	√	√
Observations	23,947	23,947	23,947	23,947	23,947	23,947	23,947	23,947	23,947	23,947
Number of pairs	2,627	2,627	2,627	2,627	2,627	2,627	2,627	2,627	2,627	2,627
R^2 (within)	0.001	0.272	0.001	0.299	0.000	0.311	0.007	0.0244	0.000	0.014
R^2 (overall)	0.732	0.805	0.735	0.814	0.739	0.820	0.971	0.971	0.974	0.974
Observations per predictor	5,987	174	11,974	176	-	179	11,974	176	-	179

Sources: Authors' calculations based on OECD International Migration Database 2004–2015, World Development Indicators, and Google Trends. *Notes*: Each column displays the result of a separate regression based on equation 2. Dependent variable is the logarithm of the annual flow of migrants (plus one) from a given origin country to a specific OECD destination. Robust standard errors, clustered at the origin country level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 6: Bilateral model including Google Trends Indices and Gallup World Poll migration intentions

	(1) Benchmark	(2) GTI	(3) Benchmark	(4) GTI	(5) Benchmark	(6) GTI	(7) Benchmark	(8) GTI	(9) Benchmark	(10) GTI
Specification	(A)		(B)		(C)		(D)		(E)	
Log GDP (destination)	1.230***	0.675***	1.625***	0.810***			0.767***	0.750***		
	(0.242)	(0.239)	(0.258)	(0.246)			(0.199)	(0.197)		
Log Population (destination)	-3.384***	-0.764*	-3.816***	-0.742*			-1.821***	-1.868***		
	(0.433)	(0.425)	(0.456)	(0.417)			(0.328)	(0.331)		
Log GDP (origin)	-0.915***	-0.503***								
	(0.125)	(0.121)								
Log Population (origin)	0.280	0.735**								
	(0.328)	(0.323)								
Log GWP1 (intention)	0.0704***	0.0655***	0.0911***	0.0786***	0.119***	0.0838***	-0.00159	2.10e-05	-0.00262	-0.000900
	(0.00455)	(0.00397)	(0.00652)	(0.00516)	(0.00722)	(0.00566)	(0.00286)	(0.00232)	(0.00272)	(0.00233)
Log GWP2 (plan)	0.0347***	0.0375***	0.0355***	0.0386***	0.0484***	0.0409***	-0.00544	-0.00608	-0.00555	-0.00591
	(0.00902)	(0.00801)	(0.0104)	(0.00869)	(0.0104)	(0.00892)	(0.00469)	(0.00453)	(0.00457)	(0.00455)
Log GWP3 (preparation)	0.0712***	0.0484***	0.0705***	0.0426***	0.0808***	0.0441***	-0.00128	0.00222	-0.000429	0.00277
	(0.0126)	(0.0105)	(0.0137)	(0.0109)	(0.0135)	(0.0110)	(0.00561)	(0.00574)	(0.00536)	(0.00555)
GTI (bilateral)	–	√	–	√	–	√	–	√	–	√
Joint significance GTI (p-value)	–	0.000	–	0.000	–	0.000	–	0.000	–	0.000
Fixed effects										
Destination	√	√	√	√	–	–	–	–	–	–
Origin	√	√	–	–	–	–	–	–	–	–
Year	√	√	–	–	–	–	–	–	–	–
Destination-year	–	–	–	–	√	√	–	–	√	√
Origin-year	–	–	√	√	√	√	√	√	√	√
Destination-origin	–	–	–	–	–	–	√	√	√	√
Observations	21,855	21,855	21,855	21,855	21,855	21,855	21,855	21,855	21,855	21,855
Number of pairs	2,611	2,611	2,611	2,611	2,611	2,611	2,611	2,611	2,611	2,611
R^2 (within)	0.075	0.326	0.080	0.350	0.098	0.355	0.008	0.027	0.001	0.016
R^2 (overall)	0.753	0.820	0.760	0.830	0.769	0.835	0.972	0.973	0.975	0.976
Observations per predictor	3,122	153	4,371	155	7,285	157	4,371	155	7,285	157

Sources: Authors' calculations based on OECD International Migration Database 2004–2015, World Development Indicators, Gallup World Poll, and Google Trends. *Notes*: Each column displays the result of a separate regression based on equation 2. Dependent variable is the logarithm of the annual flow of migrants (plus one) from a given origin country to a specific OECD destination. Robust standard errors, clustered at the origin country level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Online Appendix

to accompany

Searching for a Better Life: Predicting Migration with Online Search Keywords

by

Marcus Böhme, André Gröger, and Tobias Stöhr

A Variable Selection and Out-of-sample Estimations

Any attempt to link an arbitrary keyword to an outcome variable without providing strong evidence of a causal link may rightly be criticized for suffering from an underlying and undeclared variable selection problem. That would result, among other issues, in standard errors that are too small. Particularly for the case of the unilateral analysis, the problem can be summarized as "large X, small N, small T", with the number of countries or origin N with yearly migration data and a short panel dimension T being the main data restrictions, while the number of potential predictors X can be considerably larger than the number of observations $N \cdot T$. In such a setting, overfit can occur for purely mechanical reasons when a large number of potential predictors X with a low signal-to-noise ratio are used to fit a model. As discussed in the Data section, we use a set of keywords, which is determined by an exogenous algorithm which selects keywords and reduces the number of predictors considerably before starting estimations. In what follows, we first use a variable selection procedure to show that an algorithm which internally prices complexity also suggests an added value of adding data on search volumes. Following this, we conduct the most important test: We show that the improvements in the goodness-of-fit our approach achieves in the within dimension are not due to in-sample overfit, but also holds out-of-sample.

A.1 Variable selection

A way of receiving an external assessment of the importance of our right-hand side variables are *variable selection models*. In these procedures, the underlying algorithms are designed to optimize models while incorporating a penalty term serving as the "price" of additional complexity. This can help in choosing parsimonious specifications. Many such approaches, however, can yield unstable results when many of the variables to choose

from are highly correlated. When the main risk of additional predictors is to include statistical noise, these approaches can be very helpful.

The least absolute selection and shrinkage operator (LASSO), proposed by Tibshirani (1996), is a popular technique of variable selection. It is an OLS-based method with a penalty on the regression coefficients that systematically shrinks small coefficients towards zero in order to reduce the high variance commonly introduced when predicting outcomes with a linear regression model. Therefore, LASSO combines the idea of shrinkage with variable selection using an absolute, linear penalty.[29]

Just as OLS and other standard techniques, LASSO relies on correlations and thus does not typically yield a model of causal relationships when used with observational data. Multicollinearity of independent variables is likely to result in actually relevant relationships being biased towards zero. This method does not "build" optimal models, for example by testing non-linearities and interactions as curve fitting approaches. It is far blunter and only provides an indication of whether extra variance can be explained by adding specific variables.

We follow the literature by using Mallows' Cp as the main information criterion.[30] It optimizes the mean squared prediction error and thus trades off the number of extra predictors and the residual sum of squares. To reflect the panel approach used in the main parts of the analysis, we demean all variables before running the model. LASSO suggests that a model retaining 51 out of 67 predictors is the combination that yields the lowest mean squared prediction error. In addition, log population, log GDP, and the constant are kept. This underlines the view that the GTI can systematically predict migration flows even if extra predictors are penalized. However, the resulting reduction in the number of predictors of 16 (i.e. 67-51) is still insufficient to bring the ratio of observations per predictor above the critical threshold for all specifications in the main unilateral regressions (see Table 4. For example, it would only increase the ratio in column (8) to slightly above 12). This implies that variable selection models such as the one used here are insufficient in our setup to rule of overfit completely. In the next subsection, we therefore study out-of-sample performance.

A.2 Out-of-sample exercise

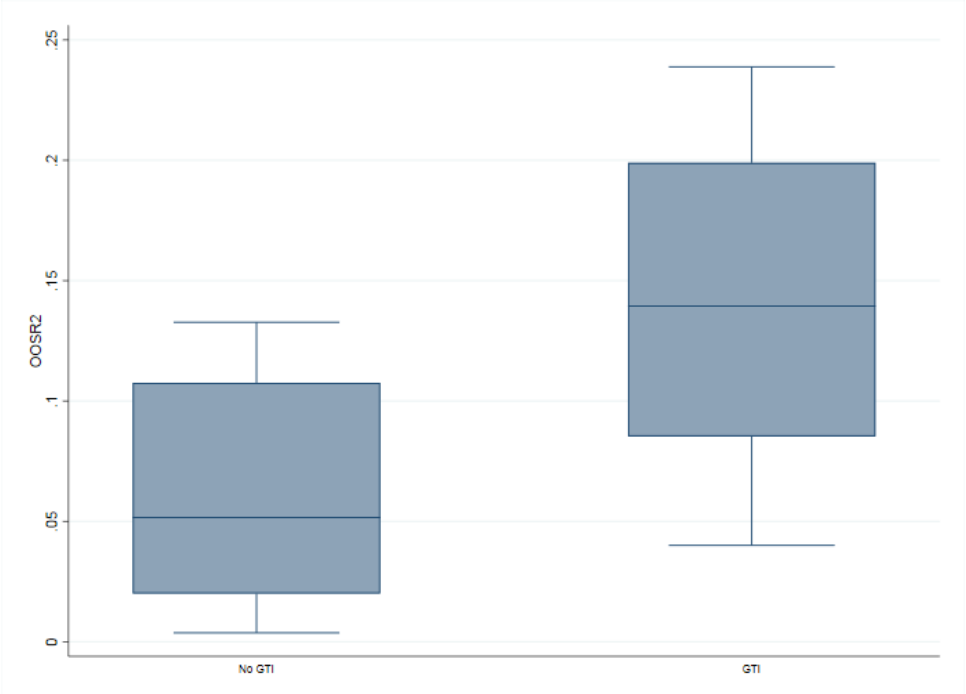
The impact of mechanical in-sample overfit can be reduced by using out-of-sample measures of fit such as the out-of-sample R^2 (OOS-R2). Overfitting the model by including variables with a low signal-to-noise-ratio would typically not improve OOS-R2, compared to a baseline model without GTI, even if it yielded higher in-sample R^2.

[29]When allowing an intercept, the LASSO is defined as $\hat{\beta}^{lasso} = argmin|y - \beta_0 - X\beta|_2^2 + \lambda|\beta|^1$, where λ is the tuning parameter which controls the parsimony of the model.

[30]Mallow's Cp is a technique for model selection in regression proposed by Mallows (1973). The Cp-statistic is defined as a criterion to compare fit across models with different numbers of parameters.

In order to provide evidence of the out-of-sample performance of our models, we apply a standard technique from the machine learning literature: k-fold cross-validation. This procedure is closely related to the idea of bootstrapping. Choosing an arbitrary number $k = 10$, we split up our data into 10 random folds. We then train the regression model on 90% of the data and calculate the in-sample and out-of-sample R^2, the latter on the remaining 10 percent of the data. This is done for each of the ten folds, yielding ten estimates of out-of-sample performance.

We use the same benchmark model from the previous section, a demeaned OLS representation of the panel model from Table 4, consisting of the basic control variables of origin countries (GDP and population size), fixed effects for origin countries and years as well as the GTI. Figure A1 depicts the out-of-sample R^2 results from this exercise for the benchmark model ("no GTI") and the augmented model ("GTI"). This is a rigid test as the model needs to perform well in the time dimension in order to improve upon the baseline specification. The results show that the GTI model explains, on average, three times as much of the variance than the benchmark model. The results thus suggest that even in the unilateral case, which is prone to overfit due to the low ratio of observations per predictor, the improvements in R^2 also hold out-of-sample and are, therefore, not driven by overfit.



Notes: The figure reports out-of-sample estimates from 10-fold cross-validation. Each boxplot thus reflects ten out-of-sample R^2s. The no-GTI model contains controls for log GDP and log population, and origin as well as year fixed effects. In addition, unilateral GTI variables are added in the GTI-model.

Figure A1: Out-of-sample within-R^2 from 10-fold cross-validation of the unilateral model

Table A1: Descriptive statistics for bilateral GTI variables

	Count	Mean	SD	Min	Max
Keyword: applicant	23,947	0.02	0.61	0.00	34
Keyword: arrival	23,947	3.08	10.20	0.00	67
Keyword: asylum	23,947	0.18	1.79	0.00	45
Keyword: benefit	23,947	3.16	12.30	0.00	77
Keyword: border control	23,947	0.59	3.73	0.00	50
Keyword: business	23,947	1.42	6.24	0.00	83
Keyword: citizenship	23,947	0.58	3.55	0.00	82
Keyword: compensation	23,947	0.13	1.90	0.00	65
Keyword: consulate	23,947	1.64	6.91	0.00	85
Keyword: compensation	23,947	0.13	1.90	0.00	65
Keyword: contract	23,947	4.55	14.32	0.00	83
Keyword: customs	23,947	0.52	3.76	0.00	83
Keyword: deportation	23,947	1.07	4.92	0.00	44
Keyword: diaspora	23,947	0.02	0.61	0.00	34
Keyword: discriminate	23,947	0.00	0.25	0.00	18
Keyword: earning	23,947	1.06	5.14	0.00	61
Keyword: economy	23,947	1.35	5.53	0.00	69
Keyword: embassy	23,947	3.81	9.95	0.00	83
Keyword: emigrant	23,947	0.00	0.00	0.00	0
Keyword: emigrate	23,947	0.21	1.98	0.00	57
Keyword: emigration	23,947	0.10	1.53	0.00	63
Keyword: employer	23,947	2.66	10.14	0.00	71
Keyword: employment	23,947	0.70	4.40	0.00	86
Keyword: foreigner	23,947	3.78	10.88	0.00	68
Keyword: GDP	23,947	0.95	4.44	0.00	69
Keyword: hiring	23,947	0.24	2.53	0.00	76
Keyword: illegal	23,947	0.37	2.97	0.00	70
Keyword: immigrant	23,947	2.27	8.45	0.00	66
Keyword: immigrate	23,947	0.25	2.32	0.00	56
Keyword: immigration	23,947	1.45	6.07	0.00	89
Keyword: income	23,947	0.77	4.27	0.00	66
Keyword: inflation	23,947	0.30	2.64	0.00	64
Keyword: internship	23,947	2.42	9.05	0.00	64
Keyword: job	23,947	7.37	17.69	0.00	89
Keyword: labor	23,947	0.36	2.95	0.00	59
Keyword: layoff	23,947	1.79	7.23	0.00	64
Keyword: legalization	23,947	0.02	0.64	0.00	41
Keyword: migrant	23,947	1.62	6.07	0.00	52
Keyword: migrate	23,947	0.17	1.87	0.00	52
Keyword: migration	23,947	2.87	9.65	0.00	66
Keyword: minimum	23,947	0.71	3.67	0.00	69
Keyword: nationality	23,947	0.48	2.87	0.00	60
Keyword: naturalization	23,947	0.83	4.16	0.00	53
Keyword: passport	23,947	1.49	5.91	0.00	68
Keyword: payroll	23,947	0.10	1.89	0.00	72
Keyword: pension	23,947	4.38	14.14	0.00	86
Keyword: quota	23,947	1.28	5.87	0.00	54
Keyword: recession	23,947	0.19	1.93	0.00	60
Keyword: recruitment	23,947	0.39	3.14	0.00	76
Keyword: refugee	23,947	1.10	4.24	0.00	42
Keyword: remuneration	23,947	0.02	0.65	0.00	50
Keyword: requirement	23,947	0.06	1.15	0.00	52
Keyword: salary	23,947	2.64	10.04	0.00	76
Keyword: Schengen	23,947	0.58	3.40	0.00	53
Keyword: smuggler	23,947	0.77	3.60	0.00	34
Keyword: smuggling	23,947	0.05	1.01	0.00	45
Keyword: tax	23,947	4.89	14.54	0.00	79
Keyword: tourist	23,947	3.11	9.76	0.00	71
Keyword: unauthorised	23,947	0.01	0.43	0.00	41
Keyword: undocumented	23,947	0.02	0.48	0.00	29
Keyword: unemployment	23,947	0.45	3.50	0.00	81
Keyword: union	23,947	3.82	13.00	0.00	86
Keyword: unskilled	23,947	0.00	0.17	0.00	20
Keyword: vacancy	23,947	0.06	1.10	0.00	52
Keyword: visa	23,947	4.20	10.40	0.00	81
Keyword: waiver	23,947	1.72	7.29	0.00	54
Keyword: wage	23,947	4.47	13.17	0.00	70
Keyword: welfare	23,947	0.14	1.81	0.00	68

Sources: Google Trends. *Notes*: Each of these bilateral GTI reflects the joint search intensity for each of the main keywords as listed in Table 1 in combination with the OECD destination country as listed in Table 2. Consequently, the bilateral GTI capture queries such as "migrate USA". Maxima are below 100 because we take the mean of weekly search volumes to calculate the yearly value.

Table A2: Robustness I: Bilateral model including Google Trends Indices and diaspora size

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Benchmark	GTI	Benchmark	GTI	Benchmark	GTI	Benchmark	GTI	Benchmark	GTI
Specification	(A)		(B)		(C)		(D)		(E)	
Log GDP (destination)	0.425	0.303	0.510*	0.380			0.781***	0.648**		
	(0.319)	(0.322)	(0.307)	(0.314)			(0.295)	(0.290)		
Log Population (destination)	-2.514***	-2.446***	-2.340***	-2.367***			-4.267***	-4.194***		
	(0.558)	(0.568)	(0.557)	(0.569)			(0.448)	(0.445)		
Log GDP (origin)	-0.473***	-0.307***								
	(0.108)	(0.110)								
Log Population (origin)	-0.466	-0.456								
	(0.310)	(0.324)								
Log migrant stock	0.705***	0.658***	0.708***	0.656***	0.726***	0.672***	0.0897***	0.0753***	0.0768***	0.0703***
	(0.0152)	(0.0163)	(0.0154)	(0.0167)	(0.0148)	(0.0167)	(0.0147)	(0.0149)	(0.0183)	(0.0182)
GTI (bilateral)	–	√	–	√	–	√	–	√	–	√
Joint significance GTI (p-value)	–	0.000	–	0.000	–	0.000	–	0.000	–	0.000
Fixed effects										
Destination	√	√	√	√	–	–	–	–	–	–
Origin	√	√	–	–	–	–	–	–	–	–
Year	√	√	–	–	–	–	–	–	–	–
Destination-year	–	–	–	–	√	√	–	–	√	√
Origin-year	–	–	√	√	√	√	√	√	√	√
Destination-origin	–	–	–	–	–	–	√	√	√	√
Observations	15,004	15,004	15,004	15,004	15,004	15,004	15,004	15,004	15,004	15,004
Number of pairs	1,943	1,943	1,943	1,943	1,943	1,943	1,943	1,943	1,943	1,943
R^2 (within)	0.544	0.612	0.551	0.623	0.563	0.635	0.0320	0.0587	0.0041	0.0203
R^2 (overall)	0.912	0.925	0.915	0.929	0.920	0.933	0.979	0.979	0.981	0.981
Observations per predictor	3,000	110	5,001	112	15,004	114	5,001	112	15,004	114

Sources: Authors' calculations based on OECD International Migration Database 2004-2015, World Development Indicators, Gallup World Poll, and Google Trends. *Notes*: Each column displays the result of a separate regression based on equation 2. Dependent variable is the logarithm of the annual flow of migrants (plus one) from a given origin country to a specific OECD destination. Robust standard errors, clustered at the origin country level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1. In this model and subsample, the bilateral keywords "applicant", "benefit", and "discriminate" (each of them combined with the destination country name as in "applicant Germany" are redundant. The number of GTI in the model is thus lower than in Table 5.

Table A3: Robustness II: Bilateral model including Google Trends Indices on Gallup World Poll sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Benchmark	GTI	Benchmark	GTI	Benchmark	GTI	Benchmark	GTI	Benchmark	GTI
Specification	(A)		(B)		(C)		(D)		(E)	
Log GDP (destination)	0.438**	0.0784	0.491**	0.114			0.809***	0.760***		
	(0.219)	(0.241)	(0.224)	(0.250)			(0.200)	(0.196)		
Log Population (destination)	-1.875***	-0.0817	-1.817***	-0.00667			-1.896***	-1.869***		
	(0.360)	(0.438)	(0.355)	(0.432)			(0.335)	(0.332)		
Log GDP (origin)	-0.562***	-0.156								
	(0.104)	(0.120)								
Log Population (origin)	0.634**	1.437***								
	(0.284)	(0.326)								
GTI (bilateral)	–	√	–	√	–	√	–	√	–	√
Joint significance GTI (p-value)	–	0.000	–	0.000	–	0.000	–	0.000	–	0.000
Fixed effects										
Destination	√	√	√	√	–	–	–	–	–	–
Origin	√	√	–	–	–	–	–	–	–	–
Year	√	√	–	–	–	–	–	–	–	–
Destination-year	–	–	–	–	√	√	–	–	√	√
Origin-year	–	–	√	√	√	√	√	√	√	√
Destination-origin	–	–	–	–	–	–	√	√	√	√
Observations	21,855	21,855	21,855	21,855	21,855	21,855	21,855	21,855	21,855	21,855
Number of pairs	2,611	2,611	2,611	2,611	2,611	2,611	2,611	2,611	2,611	2,611
R^2 (within)	0.001	0.276	0.001	0.302	0.000	0.313	0.007	0.026	0.000	0.016
R^2 (overall)	0.733	0.807	0.739	0.817	0.744	0.824	0.972	0.973	0.975	0.976
Observations per predictor	5464	158	10928	161	-	163	10928	161	-	163

Sources: Authors' calculations based on OECD International Migration Database 2004-2015, World Development Indicators, and Google Trends. *Notes*: Each column displays the result of a separate regression based on equation 2. Dependent variable is the logarithm of the annual flow of migrants (plus one) from a given origin country to a specific OECD destination. Robust standard errors, clustered at the origin country level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A4: Bilateral regression predicting Gallup World Poll variables from Google Trends Indices, continued on next page

	(1) GWP1 (Demand)	(2) GWP2 (Plan)	(3) GWP3 (Prepare)
Keyword: destination country	0.0202***	0.00372***	0.00636***
	(0.00236)	(0.00143)	(0.00116)
Keyword: applicant	-0.334**	0.142	0.0704
	(0.147)	(0.0886)	(0.0719)
Keyword: arrival	0.0320**	0.0180*	0.0173**
	(0.0157)	(0.00947)	(0.00769)
Keyword: asylum	0.0163	0.0188	0.0352***
	(0.0263)	(0.0159)	(0.0129)
Keyword: border control	-0.149***	0.0213	0.0268*
	(0.0280)	(0.0169)	(0.0137)
Keyword: citizenship	-0.0280	-0.0320***	-0.0156*
	(0.0178)	(0.0107)	(0.00872)
Keyword: consulate	-0.0208***	-0.0161***	-0.00581
	(0.00760)	(0.00459)	(0.00373)
Keyword: customs	-0.00372	-0.000159	0.00301
	(0.0165)	(0.00996)	(0.00809)
Keyword: deportation	-0.00144	-0.0333***	-0.0110
	(0.0165)	(0.00999)	(0.00811)
Keyword: diaspora	0.237***	0.122**	0.118***
	(0.0816)	(0.0493)	(0.0400)
Keyword: embassy	0.0144***	-0.0122***	-0.00557**
	(0.00532)	(0.00322)	(0.00261)
Keyword: emigrant = o,	-	-	-
Keyword: emigrate	-0.0307*	-0.00927	-0.000487
	(0.0185)	(0.0112)	(0.00908)
Keyword: emigration	-0.0612**	-0.0247	-0.0290**
	(0.0297)	(0.0180)	(0.0146)
Keyword: foreigner	-0.0541***	-0.000185	-0.000477
	(0.00945)	(0.00571)	(0.00463)
Keyword: illegal	-0.0596***	-0.0302**	-0.0300***
	(0.0210)	(0.0127)	(0.0103)
Keyword: immigrant	-0.0961***	-0.0363***	-0.0286***
	(0.0131)	(0.00794)	(0.00644)
Keyword: immigrate	0.0108	-0.0123	0.00780
	(0.0216)	(0.0131)	(0.0106)
Keyword: immigration	-0.0882***	-0.0157***	-0.0199***
	(0.00848)	(0.00512)	(0.00416)
Keyword: legalization	0.0384	0.0550*	0.0691***
	(0.0472)	(0.0285)	(0.0232)
Keyword: migrant	0.0850***	0.0220**	0.0262***
	(0.0175)	(0.0106)	(0.00858)
Keyword: migrate	0.0756***	0.00881	-0.0124
	(0.0269)	(0.0163)	(0.0132)
Keyword: migration	-0.0412***	-0.0403***	-0.0132*
	(0.0140)	(0.00844)	(0.00685)
Keyword: nationality	0.104***	0.0445***	0.0413***
	(0.0180)	(0.0109)	(0.00884)
Keyword: naturalization	0.0843***	-0.0128	-0.00558
	(0.0195)	(0.0118)	(0.00955)
Keyword: passport	0.128***	0.0502***	0.0342***
	(0.0111)	(0.00672)	(0.00545)
Keyword: quota	0.0146	0.105***	0.0598***
	(0.0193)	(0.0117)	(0.00947)
Keyword: refugee	-0.0542***	0.0295**	0.00672
	(0.0210)	(0.0127)	(0.0103)
Keyword: requirement	0.163***	-0.0497	-0.0224
	(0.0589)	(0.0356)	(0.0289)
Keyword: Schengen	0.0220*	-0.00176	-0.00848
	(0.0120)	(0.00725)	(0.00588)
Keyword: smuggler	-0.0631***	-0.0344***	-0.0537***
	(0.0212)	(0.0128)	(0.0104)
Keyword: smuggling	0.0222	0.0197	-0.0294
	(0.0452)	(0.0273)	(0.0222)
Keyword: tourist	-0.0452***	-0.00749	-0.0116**
	(0.0103)	(0.00620)	(0.00503)
Keyword: unauthorised	0.304*	0.185*	0.113
	(0.173)	(0.105)	(0.0848)
...	...	...	...

...	...	...	...
Keyword: undocumented	-0.380***	-0.0450	-0.00754
	(0.100)	(0.0606)	(0.0492)
Keyword: unskilled	-0.0814	0.167	0.0585
	(0.183)	(0.110)	(0.0895)
Keyword: visa	0.117***	0.0661***	0.0487***
	(0.00585)	(0.00353)	(0.00287)
Keyword: waiver	0.0516***	-0.00620	0.00323
	(0.0158)	(0.00952)	(0.00773)
Keyword: benefit	-0.00582	-0.0179**	-0.0209***
	(0.0128)	(0.00772)	(0.00627)
Keyword: business	-0.101***	-0.0360***	-0.0273***
	(0.0116)	(0.00701)	(0.00569)
Keyword: compensation	-0.0948***	-0.0377*	-0.0375**
	(0.0355)	(0.0214)	(0.0174)
Keyword: contract	-0.0597***	-0.0119	-0.00718
	(0.0136)	(0.00819)	(0.00665)
Keyword: discriminate	-1.146***	-0.524**	-0.433**
	(0.361)	(0.218)	(0.177)
Keyword: earning	-0.0417*	-0.0518***	-0.0332***
	(0.0233)	(0.0141)	(0.0114)
Keyword: economy	0.0178	-0.0252***	-0.0122**
	(0.0111)	(0.00668)	(0.00542)
Keyword: employer	0.0595***	0.00622	0.00629
	(0.0152)	(0.00917)	(0.00745)
Keyword: employment	-0.0519***	0.00285	-0.00996*
	(0.0120)	(0.00724)	(0.00588)
Keyword: GDP	0.00888	0.00221	-0.00949
	(0.0146)	(0.00880)	(0.00714)
Keyword: hiring	0.0620***	0.0511***	0.0560***
	(0.0211)	(0.0127)	(0.0103)
Keyword: income	-0.0724***	-0.0374***	-0.0308***
	(0.0154)	(0.00929)	(0.00754)
Keyword: inflation	0.0495**	-0.0107	-0.00760
	(0.0213)	(0.0129)	(0.0105)
Keyword: internship	0.00692	0.0225***	0.00765
	(0.0135)	(0.00817)	(0.00663)
Keyword: job	0.00386	-0.0224***	-0.0184***
	(0.00606)	(0.00366)	(0.00297)
Keyword: labor	-0.0515***	-0.0310***	-0.0199**
	(0.0179)	(0.0108)	(0.00876)
Keyword: layoff	0.124***	-0.00583	-0.00930
	(0.0141)	(0.00850)	(0.00690)
Keyword: minimum	0.107***	0.0837***	0.0672***
	(0.0135)	(0.00813)	(0.00659)
Keyword: payroll	-0.00800	0.0135	0.0294*
	(0.0313)	(0.0189)	(0.0154)
Keyword: pension	0.00790	0.00543	0.00484
	(0.0111)	(0.00668)	(0.00542)
Keyword: recession	0.119***	-0.0385***	-0.0437***
	(0.0196)	(0.0118)	(0.00961)
Keyword: recruitment	0.0409**	0.0372***	0.0351***
	(0.0161)	(0.00972)	(0.00788)
Keyword: remuneration	-0.0713	-0.108**	-0.0760*
	(0.0857)	(0.0518)	(0.0420)
Keyword: salary	0.102***	0.0356***	0.0225***
	(0.00978)	(0.00591)	(0.00479)
Keyword: tax	-0.0318***	0.000411	0.00280
	(0.0107)	(0.00648)	(0.00526)
Keyword: unemployment	0.0162	0.0140	0.00799
	(0.0161)	(0.00970)	(0.00788)
Keyword: union	-0.00743	-0.0106	-0.00415
	(0.0110)	(0.00663)	(0.00538)
Keyword: vacancy	-0.0582	-0.0372*	-0.0298*
	(0.0360)	(0.0217)	(0.0176)
Keyword: wage	0.0482***	0.0271***	0.0163***
	(0.0112)	(0.00678)	(0.00550)
Keyword: welfare	0.0912***	0.00451	-0.0146
	(0.0300)	(0.0181)	(0.0147)
Constant	1.799***	0.618***	0.362***
	(0.0391)	(0.0236)	(0.0192)
Observations	21,855	21,855	21,855
Overall-R^2	0.166	0.069	0.062

Sources: Authors' calculations based on Gallup World Poll and Google Trends. *Notes*: Each column displays the result of a separate regression of the GWP variable on the set of bilateral GTI, including only a constant. Term "emigrant" is automatically dropped from this specification. Robust standard errors, clustered at the origin country level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.