

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

Marcus Böhme<sup>†</sup>      André Gröger<sup>‡</sup>      Tobias Stöhr<sup>§</sup>

Version: June 8, 2018

## Abstract

Migration data remains scarce, largely inconsistent across countries, and often outdated, particularly in the context of developing countries. Rapidly growing internet usage around the world provides geo-referenced online search data that can be exploited to measure migration intentions in origin countries in order to predict subsequent outflows. Based on fixed effects panel models of migration as well as machine learning and prediction techniques, we show that our approach yields substantial predictive power for international migration flows, while reducing prediction errors considerably. We provide evidence based on survey data that our measures indeed reflect genuine emigration intentions. Our findings contribute to different literature by providing 1) a novel way for the measurement of migration intentions, 2) an approach to generate close to real-time predictions of current migration flows ahead of official statistics, and 3) an improvement in the performance of conventional migration models that involve prediction tasks, such as in the first stage of a linear instrumental variable regression.

*JEL classification:* F22, C53, C80

*Keywords:* Emigration, Migration Intention, Machine Learning, Big Data

---

\*We would like to thank Toman Barsbai, Christian Fons-Rosen, Stephen Hansen, 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 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. 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). Any remaining errors are our own.

<sup>†</sup>Organisation for Economic Co-operation and Development (OECD)

<sup>‡</sup>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. Fax: +34-93581-2012, phone: +34-93581-4324, e-mail: andre.groger@uab.cat.

<sup>§</sup>Kiel Institute for the World Economy (IfW) and IZA

# 1 Introduction

With profound effects on both origin and destination countries, the topic of migration has become one of the most important and most contested policy issues for developed and developing countries alike. There is a large literature dedicated to analyzing the determinants of international migration, which has identified demographic factors, income differences, and violent conflicts to be among the main push- and pull-factors. However, a lack of migration data is still plaguing the discipline: 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. This is especially the case for developing and emerging countries in which administrative or survey-based indicators are often unavailable, making many forms of analysis impossible.<sup>1</sup>

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 information demand (Varian 2014). The use 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 analyzing this complex data (Mullainathan and Spiess 2017).

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, which was first released in 2009, Choi and Varian (2012) suggest that online search data has a large potential to measure users' interest in a variety of economic activities in real time, and demonstrate how it can be used for the prediction of home and automotive sales as well as tourism. 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 initial criticisms (Lazer et al. 2014), the literature has since grown quickly, including applications to the prediction of aggregate demand (Carrière-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

---

<sup>1</sup>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 survey which provides consistent data for a larger set of countries of origin: the Gallup World Poll (GWP). The GWP data has, however, two big disadvantages: First, it is not freely available and tends to be very costly. Second, it does not provide consistent time series of migration intentions for origin countries.

the context of France (Fondeur and Karamé 2013), Germany (Askitas and Zimmermann 2009), and the USA (D’Amuri and Marcucci 2017).

There is a small number of recent applications that have tried to use internet meta data 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 and Zagheni and Weber (2012) relies on 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 from highly selective samples obtained from online sources. Their user bases are heavily self-selected. These rather specialized online services thus cannot be used to infer general migration patterns.<sup>2</sup> Furthermore, the data used in these studies is proprietary and, therefore, their analysis cannot be replicated or used in other contexts by external researchers.

Approaches that can help measuring migration intentions and providing 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 the Google Trends Index (GTI) for migration-related search terms.<sup>3</sup> 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 demand for information can be used as a proxy for changes in the number of aspiring migrants. Consequently, surges in online search intensities for specific keywords related to the topic of migration can indicate an increase in the demand for migration, reflecting aspirations, and can thus help predicting 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 and, therefore, can help offering a general tool for the prediction of migration. We define keywords related to the topic of migration based on a set of expressions which is semantically linked to the topic of “migration” and “economics” through their co-occurrence within the Wikipedia encyclopedia. We then extract the GTI indicators for each individual keyword in the official language of the respective country of origin.

We test the predictive power of our GTI migration indicators first by augmenting a standard fixed effects panel model of international migration decisions from a large range of origin countries to the OECD destination countries with our tailor-made measures. Controlling for a large number of potential push- and pull-factors from the migration

---

<sup>2</sup>The Twitter sample is constituted predominantly by young male users and the user profile of Yahoo seems to be selected on factors such as age, sex, and level of internet penetration in the country.

<sup>3</sup>The GTI data consists of high-frequency time series capturing the relative search intensities for any keyword performed through the Google search engine across the globe. The GTI is by far the most representative data source for online searches worldwide with Google having a market share of more than 80% on desktop devices. This figure increases to 97% once considering mobile and tablet devices. Source: <https://www.netmarketshare.com/>, accessed November 2017.

literature, we find that our approach yields substantial improvements in the predictive power of international migration flows. In the most conservative specification, the inclusion of our measures yields a 100% increase in the explained variability of migration flows as measured by the within- $R^2$ . We also explore the heterogeneity of these results with respect to origin country characteristics. Reassuringly, we find that this performance improves further when restricting the sample to relatively homogeneous origin countries with respect to their official language, to middle- and high-income origins, and those with high internet penetration.<sup>4</sup> Using machine learning techniques, we also test the robustness of these results to in-sample overfit by applying dimension reduction algorithms and out-of-sample predictions. The results confirm that our approach systematically yields substantial improvements in the goodness of fit for international migration models. Last but not least, we also provide evidence based on survey data that our measures indeed reflect genuine emigration intentions.

The contribution of our paper is threefold. First, we propose a universal approach to improve the measurement of migration intentions with consistent and representative indicators that are freely available at close to universal geographic coverage. So far, the availability of data on migration intentions is severely restricted to selective and exclusive surveys. Easing this data constraint can help facilitating 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.<sup>5</sup> This approach could, for example, be used for short-term policy prediction exercises in the case of humanitarian crises. Third, it can improve the performance of conventional models of the determinants of migration flows<sup>6</sup> in application that involve prediction tasks, such as in the first stage of a linear instrumental variable regression, when estimating heterogeneous treatment effects, or flexibly controlling for observed confounders.

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 measures of migration intentions. In Section 3, we describe the panel estimation model used in the analysis of the determinants of migration and, subsequently, introduce machine learning techniques, which help dealing with the econometric challenges from the former approach. Section 4 provides the results from the panel estimations and Section 5 those from the machine learning techniques. We discuss the value of our findings for empirical applications and policy recommendations in Section 6, and Section 7 concludes.

---

<sup>4</sup>The rationale for these trade-offs being that our measures can be expected to perform better in countries in which the official language is more representative of the total population, in countries with relatively low financial migration barriers, and those with high internet penetration.

<sup>5</sup>For example, in the case of the International Migration Database of the Organisation for Economic Co-operation and Development (OECD), the lag is between two to three years.

<sup>6</sup>See, for example, Beine et al. (2016) and Docquier and Rapoport (2012) for an overview of this literature, and Mayda (2010) and Ortega and Peri (2013) for specific applications.

## 2 Data

### 2.1 Google Trends Data

Google Trends data are freely accessible at <https://www.google.com/trends/> and generally available on a daily basis, starting on January 10, 2004.<sup>7</sup> The database provides time series of the search intensities of the user's choice of keywords, which we call the Google Trends Index (GTI). In the current version of Google Trends, the GTI can be restricted by geographical area, date, a set of general search categories such as "Jobs & Education" or "Travel", and by the type of search, i.e. standard web search, image, etc. We use the first two restrictions based on web searches to create a country-specific, yearly time series of online search intensity. We proceed as follows.

The GTI captures the relative quantities of web searches through the Google search engine for a particular keyword in a given geographical area ( $r$ ) and during a specific day ( $d$ ) in a specified time period. For privacy reasons, the absolute numbers of searches are not publicly released by Google. The share  $S_{d,r}$  of searches for a specific keyword in geographical area  $r$  and during day  $d$  is given by the total number of web searches containing that keyword ( $V_{d,r}$ ), divided by the total number of web searches in that area and during a specific day ( $T_{d,r}$ ), i.e.  $S_{d,r} = \frac{V_{d,r}}{T_{d,r}}$ . Since migration flows are typically recorded in yearly intervals between countries, we adapt our GTI measure accordingly to reflect yearly variations as well, based on the simple average of the daily shares per year ( $a$ ) in the country of origin ( $o$ ):  $S_{a,o} = \frac{1}{d} \sum_{d=1}^d \sum_{r=1}^r S_{d,r}$ . In addition, the indicator provided is normalized and effectively ranges between 0 and 100, with the top value being assigned to the time period during which it reaches the maximum level of search intensity over the selected timespan. Consequently, the GTI measure for a specific keyword in year  $a$  and country of origin  $o$  used in this paper is calculated by:  $GTI_{a,o} = \frac{100}{\max_a(S_{a,o})} S_{a,o}$ .

In essence, our measure of internet search intensity reflects the probability of a random user inquiring a particular keyword through the Google search engine in a given country of origin and in a given year. Geographical attribution is achieved through IP addresses and are released only if the number of searches exceeds a certain - undeclared - minimum threshold. Repeated queries from a single IP address within a short period of time are disregarded by Google, for example 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 measurement error into the time series. As a consequence, the indices can vary according

---

<sup>7</sup>Extracting large quantities of Google Trends data through the website is, however, time consuming. Google offers access to their the 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.

to the day of download. However, time series extracted during different periods 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 decisions regarding the extraction of data: which keyword to choose and in which language to extract them for? With respect to keyword selection, existing studies show a huge variety, depending on each context, which can range between one to several thousand keywords for which time series of the GTI are extracted. For instance, D'Amuri and Marcucci (2017) simply use the term "jobs" in order to predict unemployment in the US. Carrière-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 computing infrastructure.

In the absence of a general pre-defined search category related to migration, we are left with the task of selecting individual keywords, which 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, the identification of a specific search term is ambiguous, we rely on a broader set of keywords, the exact composition of which 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.<sup>8</sup> The website displays the top 100 related words for each query and we retrieve those for the keyword "immigration". Since the majority of migration decisions tend to follow economic motives, we also retrieve a second list of semantically related words based on the keyword "economic". Based on the two lists of 200 semantically related words in total, for tractability reasons, we chose the subset of the top third most related keywords from each list (i.e. a total of 68). As for the English language there may be varying spellings for the same keyword in the American and British form, we include both versions if applicable. Similarly, users might be searching for both singular as well as plural forms of a keyword, we include both forms for nouns. Different versions of the same keyword can be combined with the Boolean operator "OR", which allows us to retrieve the joint search intensity

---

<sup>8</sup>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. Note that semantic links between words generated by this methodology change over time to the extent that Wikipedia is modified. Therefore, the list retrieved today is not identical to the one we obtained on January 16th, 2015.

from Google Trends.

Finally, we are left with the empirical decision in which languages to extract GTI data for our list 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 an additional difficulty when extracting data. Based on this restriction 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. 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 are available. The final list of keywords in the three chosen languages is included in the Appendix Section B.1. Based on the operational procedure described above, we proceed to download GTI time series data for 68 keywords, in 107 countries of origin, and over 10 years each, which amounts to a total of 72,760 keyword-country-year observations. For countries with speakers of at least two of English, French and Spanish, we select the time series in the language with the larger respective number of speakers.

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 searching for information online, based on the list of keywords chosen, in a given country of origin and at a given moment in time, are genuinely interested in emigration. They may as well just follow a local or global search trend, which could eventually have been ignited by news on migration or other topics on the media that spark interest in that direction. 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 in a specific moment simply because many others are, 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 smoothing 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 adopts and changes its services in line with their business model. This could (and if effec-



tive should) affect the search behavior of users and, thereby, change the data-generating process as well as the representativeness of the specific keywords chosen in this study over time. Due to this issue, we cannot rule out that search intensities increase due to adjustments made in the underlying search algorithms rather than increased interest in migration. In other words, the index we create by the choice of our keywords in this exercise is carrying 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. This 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 year dummies in our empirical specification.

## 2.2 Migration and Country Data

We merge data from a panel of bilateral migration flows with macroeconomic indicators and other information on the respective origin countries for which we intend to capture migration intentions through the GTI. Migration data comes from the OECD International Migration database, which provides yearly immigrant inflows into the OECD countries by foreign nationalities. Since this database is fed by population, residence, and employment registers from the OECD member countries, it covers only legal immigration, i.e. workers, asylum seekers, and other types of legal immigrants. The sample includes almost all countries of origin worldwide, both from the group of developing and developed countries. One issue in the use of such flow data is the presence of zeros, which are particularly prevalent in the case of small countries of origin with low population. Despite migration flow data being available for earlier periods, we focus on the period starting in 2004, the year the GTI data starts, until 2015, which is the last year of OECD migration flow data available.

We match this panel of migration flows with macroeconomic indicators of the origin country from the World Development Indicators (WDI) (World Bank 2016). In the benchmark setup, we use only GDP and population control variables in order to not restrict our sample of origin countries. By including these covariates we intend to control for the most important push- and pull-factors that have been emphasized in the migration literature (Mayda 2010). Many other predictors have been used in the literature as additional control variables. In an extension, we include additional origin country controls 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) from the WDI. We also include the number of weather and non-weather disasters from the EM-DAT database (Guha-Sapir 2016). To control for political factors, we include the Polity IV Autocracy Score and the State Fragility Index (Marshall et al. 2016). 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 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 restrict the analysis to a subset of countries of origin, which is particularly homogeneous in terms of the use of the official language in which we extracted the GTI time series for. However, most of these indicators are partly unavailable, especially for smaller countries.

Given that the GTI data we rely on vary at the country of origin level, we collapse the matched panel data set at the level of the OECD destination countries. In other words, we consider all migration flows from a given origin to all OECD countries simultaneously. Thus, we implicitly focus on the general migration decision of the country of origin and abstract from the sorting decision, i.e. the decision which destination country to immigrate to. This provides the advantage that we can discard the problem of multilateral resistance related to gravity models of international migration (Bertoli and Fernández-Huertas Moraga 2013). Furthermore, it also helps alleviating issues related to the presence of zero observations in the flow of migrants (Beine et al. 2016). Proceeding along these lines and accounting for missing values in the GTI data, in our benchmark sample we are left with the aggregated migration decisions towards the OECD countries from a sample of 98 countries of origin over 12 years (2004–2015). Due to the inclusion of a one year lag in our preferred specification (equation 1 below), the corresponding total sample size is 1,068 country of origin-year observations. Due to missing values in our control variables for certain countries, in the extended specification, we are left with 70 origin countries and 680 observations.

### 3 Methodology

In order to investigate the predictive power of our GTI measures for migration-related keywords in origin countries for the estimation of migration decisions, we proceed as follows: First, as a benchmark specification, we estimate a standard fixed effects model of migration flows from approximately 100 origin countries to the OECD. This model includes destination-year fixed effects as well as origin fixed effects, thus eliminating all explanatory factors at the destination country level as well as time invariant ones. Subsequently, we augment this benchmark specification with our GTI time series of origin country-specific variables, capturing the internet search intensities for the selected keywords. The estimated regression equation is:

$$Y_{ot+1} = \alpha + \beta T_{ot} + \gamma O_{ot} + \eta D_t + \delta_o + \tau_t + \varepsilon_{ot}, \quad (1)$$

with  $o$  indexing the country of origin and  $t$  is time. The dependent variable,  $Y_{ot+1}$ , is the

logarithmic transformation of migration flows from the origin country to the OECD in a given year. All right hand side variables are lagged by one period in order to account for concerns about reverse causation.  $T_{ot}$  represents our GTI measures for a given origin country with respect to a specific keyword in a given year.  $O_{ot}$  is a vector of origin-specific control variables,  $D_t$  a vector of destination-specific controls, and  $\delta_o$  stands for origin country-specific fixed effects.  $\tau_t$  are time dummies and  $\varepsilon_{ot}$  represents a robust error term, which is clustered at the origin country level. Given the use destination-year fixed effects, we do not include control variables at the destination level. They would only increase model complexity and are statistically insignificant.

Adding the GTI variables for a large number of single keywords to this model increases the risks of in-sample overfit, i.e. of picking up a spurious correlation between the time series and the outcome variable. Adding several time series that contain only statistical noise would be likely to yield some statistically significant predictors, reducing the predictive power of our model out-of-sample. In order to deal with this potential problem, we apply two techniques to guard against in-sample overfit (Varian 2014, Kleinberg et al. 2015). First, we estimate out-of-sample predictions using k-fold cross-validation techniques. Second, we apply shrinkage methods to show that, when penalizing larger numbers of covariates in a model, the applied algorithms tend to include a considerably larger number of regressors than what could be expected if the within-variation only consisted of noise.

## 4 Panel estimation

The results from the fixed effects estimations based on equation 1 are reported in Table 1. Column (1) displays the coefficients for our benchmark regression specification, without any GTI predictors. Based on this basic model of migration flows, the resulting within- $R^2$  is relatively low (7.7%). However, once we augment this model by our migration-related GTI variables in column (2), the  $R^2$  increases to 20.8%, suggesting that the additional covariates possess substantial predictive power. Column (3), in turn, reports the results when including GTI predictors related to economic keywords. As we can observe, the  $R^2$  also increases substantially to 16.7%, but with a smaller magnitude compared to the previous specification. Finally, when augmenting the model by all GTI variables including both migration- and economic-specific keywords, the fit of the model increases even further to 25.8%. Taken together, these results suggest that the predictive power of our benchmark model as measured by the within- $R^2$  can be improved strongly (ranging between 115 to 235%) when including the internet search intensities in origin countries for migration and economic search terms.

In Table 2, we repeat the same exercise for the group of origin countries which are relatively homogeneous in terms of their spoken languages. Since our GTI measures

depend on a certain term in a specific language, it is important for the estimation that the official language is also commonly used when performing online searches. In other words, we expect the predictive power of our GTI variables to increase with the share of the native population in the country of origin that commonly uses the official language. Therefore, in panel A, we restrict the sample to countries in which at least 20% of the native population uses the official language commonly. This results in the exclusion of 16 countries of origin compared to the benchmark specification, such that the remaining number of countries included in the sample is 82 in this specification. In column (1), we find that the coefficient of determination in the basic setup only increases slightly to 9% due to sample composition effects. Comparing columns (2) to (4) to the same columns in Table 1, however, we observe a general increase in  $R^2$  in line with our expectations. In other words, this suggests that the GTI variables are more predictive for migration flows in countries where the three languages we use are more widely spoken. Comparing the results in columns (2) to (4) with column (1), we find that the relative increase in the predictive power resembles the one from Table 1, ranging between 120 to 220%, with the combined keywords for migration and economic terms yielding the highest predictive power.

In panel B, we restrict our sample even further, focusing only on the origin countries in which the majority of the population commonly speaks the official language. Doing so excludes 31 countries of origin from our benchmark sample, which do not fulfill this criterion. Similarly as in panel A, we observe that the resulting levels of  $R^2$  increase once again for all specifications including the GTIs, while they remain constant in column (1), the basic setup. Comparing the coefficients of determination across columns (2) to (4), consequently, we find that they increase more strongly, here between 140 to 250%, with the combined model including both migration and economics search terms performing the best.

For the data generation process we rely on the general availability and the use of the internet technology among the local population of the origin country is crucial. We observe marked differences in the number of internet users across countries, which are positively correlated with the 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.<sup>9</sup> Since internet search intensity turns out to be zero or is measured noisily in countries with low internet usage, we expect the predictive power of the GTI's to be stronger in countries with higher internet penetration. In order to test this hypothesis we perform an additional exercise in which we drop the subsample of low-income origin countries or restrict it to those countries with

---

<sup>9</sup>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).

at least 10% of the general population having access to internet (Table 3, panels A and B respectively). Focusing on panel A, we find similar results as in Table 2 (panel B) in the sense that there is a general increase of the absolute level of within- $R^2$  as well as a stronger relative increase with respect to the basic setup in column (1). Panel B shows that this trend becomes even more pronounced when focusing on the countries with at least 10% of the general population having internet access. Absolute levels of  $R^2$  relative to any previous table and relative increases across the different are once again stronger, with the latter ranging between 160 to 265% here.

In Table 4 we also test the robustness of our results with respect to an extended set of country of origin controls. Compared to the benchmark specification in Table 1 with only origin GDP and population size as control variables, here we add the unemployment rate, the share of young population, the State Fragility Index, the Polity IV Autocracy Score, the share of mobile phone subscriptions and internet users over the general population, as well as the incidence of weather and non-weather related disasters over time. In this extended control setup, we find that the level of within- $R^2$  increases especially for the basic specification in column (1), absorbing some of the relative performance increase of the GTI variables. However, this comes at a high price: due to the lack of control variables for around 28 countries, our sample size shrinks to only 70 origins. Nevertheless, the relative performance in the predictive power of the GTI remains strong, ranging between 50 to 115%, with the combined keywords for migration and economics still yielding the highest predictive power.

In sum, the results show that in the panel regression framework, the GTI variables provide substantial increases in the goodness of fit for international migration flows estimations. Reassuringly, we find evidence that this effect becomes stronger when restricting our sample to origins for which we expect the GTIs to be more predictive based on economic development, the penetration of information technology, and the common use of the official language. This increase is also robust to including a host of additional control variables, despite with a lower magnitude as the controls absorb some of the signal which is captured in the GTIs. However, one important thing to note in this context is that the availability of many additional control variables is often poor, especially among classic countries of origin in the developing world. This implies that, beside the additional predictive power of internet search activity, data availability is another crucial argument in favor of our approach. This applies particularly to prediction settings, i.e. when the objective of the application is to obtain an estimation of the outcome variable  $y$  as compared to parameter estimations where the focus is on the effect of  $x$  on  $y$ .

## 5 Machine Learning and Prediction Methods

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. Essentially, the problem we are trying to solve 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 helps determining the choice of keywords as well to reduce the number of predictors considerably before starting estimations. In what follows, we first use a variable selection procedure to show that an algorithm that internally prices added complexity also suggests 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 model achieves in the within dimension are not due to in-sample overfit, but also holds out-of-sample.

### 5.1 Variables 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 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.

Shrinkage methods such as the least absolute selection and shrinkage operator (LASSO) and the least-angle regression (LARS) algorithm<sup>10</sup> systematically shrink small coefficients towards zero in order to reduce the high variance commonly introduced when predicting outcomes with a linear regression model.<sup>11</sup> Thereby, LASSO combines the idea of

---

<sup>10</sup>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, which tends to produce simpler models. LARS, proposed by Efron et al. (2004), is a method that can be viewed as a vector-based version of the LASSO procedure to accelerate computations.

<sup>11</sup>Ridge regression cannot perform variable selection because it never shrinks coefficients to non-zero values by using a squared penalty function. This makes it not ideal if we expect coefficients to be exactly zero and will therefore not be considered here. For our purpose, our choice is thus more conservative. Furthermore, we do not use naïve stepwise model selection (such as the "step" package in R) because it is known to yield unstable models across datasets and folds. Instead we use penalized regression, which

shrinkage with variable selection using an absolute, linear penalty.<sup>12</sup>

Just as OLS and other standard techniques, LASSO and LARS rely on correlations and thus do typically not 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. The methods we use in this section do not "build" models, for example by testing non-linearities and interactions as curve fitting approaches. They are blunter and only provide 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.<sup>13</sup> 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, we calculate first differences of all variables before running the model. Both the LASSO and the LARS models suggests a model with 15 migration keywords out of 37 as the combination that yields the lowest mean squared prediction error. In addition log population, log GDP and year fixed effects are kept. Adding economic keywords does not result in a model with higher R<sup>2</sup> and lower Cp at the same time. The results from these variable selection approaches, thus, support the view that migration-related GTI predictors are systematically related to migration flows. However, variable selection models such as the ones used here potentially overfit the model in-sample. In the next subsection, we therefore study out-of-sample performance.

## 5.2 Out-of-sample exercise

The potential impact of overfit can be reduced by using out-of-sample measures of fit, for example, the out-of-sample R<sup>2</sup> (OOS-R2) and the out-of-sample root mean squared error (OOS-RMSE). Imprecise out-of-sample predictions lead to a particularly high penalty when using the OOS-RMSE due to the error terms being squared. In contrast to in-sample estimations, unrelated predictors are less likely to yield any improvement in predictive power out-of-sample, because a spurious relationship would only continue to hold in this setting by mere chance. Overfitting variables with a low signal-to-noise-ratio, by contrast, would be likely to lead to systematically higher OOS-RMSE's and typically no improvement in OOS-R2, compared to a baseline model without GTI predictors, even if having a higher in-sample R<sup>2</sup>.

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 yields far more stable results.

---

<sup>12</sup>When allowing an intercept, the LASSO is defined as  $\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} |y - \beta_0 - X\beta|_2^2 + \lambda|\beta|_1$ , where  $\lambda$  is the tuning parameter which controls the parsimony of the model.

<sup>13</sup>Mallow's Cp is a technique for model selection in regression proposed by Mallows (1973). The Cp-statistic is defined as a criteria to assess fits when models with different numbers of parameters are being compared.



procedure is closely related to the idea of bootstrapping that is well known in economics. 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, consisting of the basic control variables of origin countries (GDP and population size), dummies for origin countries and years, as well as the different sets of GTI's. Figure 1 depicts the out-of-sample  $R^2$  results from this exercise for different models and Figure 2 plots the root mean squared error. Note that 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 label "basic" indicates that the model includes only the benchmark variables without any GTI. Migration and economic keywords as well as a combination of both are added successively in the same way as in the previous estimations.

All models including GTI controls perform significantly better than in the basic model as indicated by the increased out-of-sample  $R^2$  in Figure 1. The relative gains across the different models are similar in magnitude compared to the in-sample estimations in Table 1, with the combined migration and economic keywords performing the best. Furthermore, as depicted in Figure 2, the models using GTI variables have a lower root mean squared error than the basic model. Combining both goodness of fit measures, Figure 3 shows that there is no trade-off between higher predictive power and lower prediction error involved. The models with GTI included perform better, on average, than the basic one, explaining more of the variation in the migration outcome measure and, at the same time, producing fewer errors. Again, the best performance comes from the model including both sets of GTI. Hence, the predictive power of online search intensities for next year's migration flow remains strong, even in the artificial out-of-sample experiment. This suggests that the GTI provide genuine predictive power for migration outcomes in the within dimension.

In Figure 4 we add the extended set of control variables and rerun the k-fold cross-validation on the smaller sample. The results show that the model with extended controls generally produces fewer prediction errors compared to those with basic controls depicted in Figure 2. This is partly due to the additional control variables and partly to sample selection as several countries of origin drop out, for which the extended controls are unavailable. Simultaneously, the inclusion of the GTI variables again leads to a significant increase in the out-of-sample  $R^2$ , while the additional reduction in the RMSE is less pronounced compared to the extended benchmark model without GTI. Again, the best performance is obtained from the combined set of migration and economic GTI. This highlights another advantage of using GTI: data availability is not an issue and potentially data are also available at far higher frequency and a finer spatial resolution, margins at



which standard macro variables are often unavailable, especially in developing countries.

## 6 Beyond Predictive Power?

We have presented evidence that our tailor-made GTI measures lead to significant increases in the predictive power of models of current international migration flows, both in-sample and out-of-sample. In this setting, machine learning techniques are helpful to deal with the high dimensionality of the GTI data and provide a solid benchmark to quantify the predictive power of these additional regressors. As emphasized by Mullaithan and Spiess (2017), the prediction objective (i.e. generating a prediction of outcome  $y$  based on independent variables  $x$ ) should not be confounded with the one 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 agnostic about causality. A remaining open question is therefore the one about the underlying causal mechanism between changes in the GTI and real-life migratory movements. In other words, what are the GTI measures effectively capturing: demand for or supply of migration?

Relating to recent criticism in the context of the Google Flu Index, Lazer et al. (2014) and Ormerod et al. (2014) have shown that such models are susceptible to over-prediction due to herding behavior. With respect to migration decisions, this translates into a situation in which many people start searching for migration-related topics despite having any personal migration intentions a priori (e.g. due to media reports about the Syrian refugee crisis). Such a situation might lead to an erosion of the predictive power of our approach. However, such a phenomenon might also occur in an environment of high migration prevalence, i.e. can be the result of reverse causality (e.g. people searching for migration topics because many of their fellows have left the country). If that situation led to an increase in migratory movements, it would usually be described as a migration network effect or *chain migration* in the literature. On the other hand, it might also purely be driven by curiosity without any realization of migration. The same can happen in a low migration environment, due to an unrelated third event that might trigger a general interest in the topic. In essence, from the causal perspective of parameter estimation, it is an empirical challenge to distinguish these cases in our context and to separate demand from supply as well other third factors that might determine the search behavior for migration-related keywords.

In order to shed some light on these questions, we use a global dataset on migration intentions. This analysis relies on individual-level data from the Gallup World Poll (GWP), which has been implemented starting in 2006. Each survey is conducted in varying intervals of one up to several years, depending on the country. Note that each sample is independent in the sense that it constitutes a repeated cross-section instead of a panel. The data consists of a stratified random sample of typically around 1,000 respondent per

country and is deemed nationally representative.<sup>14</sup> We rely on three migration related questions in the Gallup World Poll which are designed to assess individuals' migration intentions to different degrees. In particular, 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.?*

Note that the framing of these questions is such that they reflect an increasing migration aspiration intensity.<sup>15</sup> While question one indicates the respondent's potential and abstract demand for migration in general, number two indicates whether individuals plan to realize their this intention in the short-term, and number three whether they have started to prepare already. Aggregating the data across countries, the descriptive statistics indicate that approximately 675 million people worldwide had general migration intentions according to question one in 2008, compared to 703 million in 2014. In terms of absolute migration demand China, Nigeria, and India lead the ranking in each year. In relative terms of the share of adult population at origin, it is most often small countries such as Haiti, Sierra Leone, and the Dominican Republic that have the highest migration intentions among the general population. The most popular destination countries tend to be the United States, Great Britain, and Saudi Arabia. In 2010 only about 4% of the sample stated to actively plan 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. Hence, out of 675 Million individuals who indicate a general intention to migrate in 2008, 2% or 14 million individuals were reportedly in a stage of preparation at that time. In 2014, this share increased to about 3.5% of the sample or 25 million individuals worldwide.

In order to compare the Gallup survey data of migration intentions to our GTI measures, we augment our regression specification 1 to include each of the variables corresponding to the three questions one by one. Given time gaps in the survey data for

---

<sup>14</sup>Stratification is based on population size and the geography of sampling units. The survey is implemented either as face-to-face or telephone interview with subjects older than 15 years. Further details about the survey methodology can be accessed online at: <http://www.gallup.com/178667/gallup-world-poll-work.aspx>.

<sup>15</sup>Note that there are a number of important caveats that have to be borne in mind when using this data. First, question number one explicitly asks about permanent migration. However a large number of people might misunderstand the question thinking they could not come back. Hence, it is possible that the actual demand for migration is even bigger than what we observe in this survey. Second, a substantial number of people are already migrants (either internally or internationally) and, therefore, part of the data might represent return migration in fact.

certain countries, we follow a recommendation from Gallup and compute rolling averages based on the three questions over time and match them on our main data set. Note that the results are not directly comparable to the ones from the panel specification for two reasons: first, due to the time gaps, the sample size is reduced massively such that we have to rely only on only 330 observations in this exercise. Second, the Gallup data is a repeated cross-section and its within dimension is not very accurate. For these reasons, the findings from this exercise should rather be interpreted as suggestive evidence.

In cross-sectional regressions without our GTI measures, we find that the GWP variables are generally positively and significantly correlated with migration flows from our sample countries. Reassuringly, this correlation is increasing with the intensity of migration intentions as captured by above questions. The point estimates indicate that a 1 percentage point increase in the GWP variables is associated with a 0.18 to 0.26 point increase in migration flows from the origin to the OECD countries. When including our GTI measures simultaneously, the magnitudes of the Gallup coefficients decreases considerably to 0.09 to 0.11 points, but remain statistically significant. This indicates that there is a positive correlation between the GTI and the GWP variables, but also that they are not collinear. In other words, one possible interpretation is that part of the GTI appears to reflect “real” demand for migration as measured by the Gallup data. When estimating the same regression in a panel specification with fixed effects, however, the coefficients for the Gallup variables become insignificant and close to zero. This seems to be mainly due to the low accuracy of the within-variation, which prevents us from directly comparing our GTI prediction results to the Gallup specification in this section.

In summary, these preliminary tests provide some evidence that our GTI measures are indeed capturing a *demand* for migration or, in other words, genuine migration intentions among the origin population. On the other hand, this exercise also demonstrates that, despite the increasing importance of international migration, there is still a general lack of data on migration intentions across countries. The GWP as the only existing survey with near universal coverage worldwide (147 countries) provides a good overview across countries, but is not very useful when comparing country trends over time. Furthermore, the dataset is proprietary. Given the general absence of reliable and comparable data, our GTI approach offers a promising way for improvement along these lines.

## 7 Conclusion

We have presented evidence that GTI-based indicators for migration-related online search terms provide substantial predictive power for international migration decisions, both in- and out-of-sample. In line with our expectations, these results become stronger when restricting our sample to more developed origins, and those with higher penetration of information technology and common use of the official language. We also provided

suggestive evidence based on observational survey data that our GTI measures partly reflect genuine migration intentions.

Can a GTI-based approach be feasible for the prediction of international migration flows in the long-run? The experience of the Google Flu Trends for the United States has shown that there are several obstacles, even if predictive power can be established convincingly. The predictive power of the composition of keywords that we employ in this study to capture migration intentions is changing over time. Changing associations between individual keywords and the outcome variable are likely to affect the composition of the “optimal” prediction model in the future. Surging interest in a particular keyword may cause its worth for prediction to plummet. Therefore, we advocate to apply 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, especially when concerned with short-run predictions of migration flows in a particular country context, it should be worthwhile to refine both the semantic links of migration-related words in that particular language context as well as for the particular time period to increase or update the predictive power of the GTI indicators. Here, a combination with text analysis tools, e.g. based on media reports, could 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 measures and on migration flows in a sub-national setting, which would allow us to calibrate 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.

Our findings contribute to different areas in the migration literature related to measurement and prediction of international and domestic migration. First, we propose this methodology as a universal approach to improve existing data on migration intentions with consistent, representative, and high-frequency indicators that are freely available at close to universal geographic coverage. So far, the availability of data on migration intentions is severely restricted to selective and exclusive surveys, often providing inconsistent data at very low-frequency. By constructing GTI measures 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 to be applied for prediction purposes in other contexts. Second, our approach can be used to generate short-term predictions of current migration flows ahead of official data release lags, which can amount up to several years.<sup>16</sup> This could, for example, 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 to organize humanitarian responses. This is comparable to recent applications in the political economy literature demonstrating that newspaper text can be used to predict

---

<sup>16</sup>For example, in the case of the International Migration Database of the Organisation for Economic Co-operation and Development (OECD), the lag is between two to three years.

armed conflict ahead of time (Mueller and Rauh forthcoming). Third, our approach also contributes to conventional models of parameter estimation in migration studies that involve prediction tasks. For example, it can help improve performance in the first stage of a linear instrumental variable regression, when estimating heterogeneous treatment effects, or flexibly controlling for observed confounders.

## Bibliography

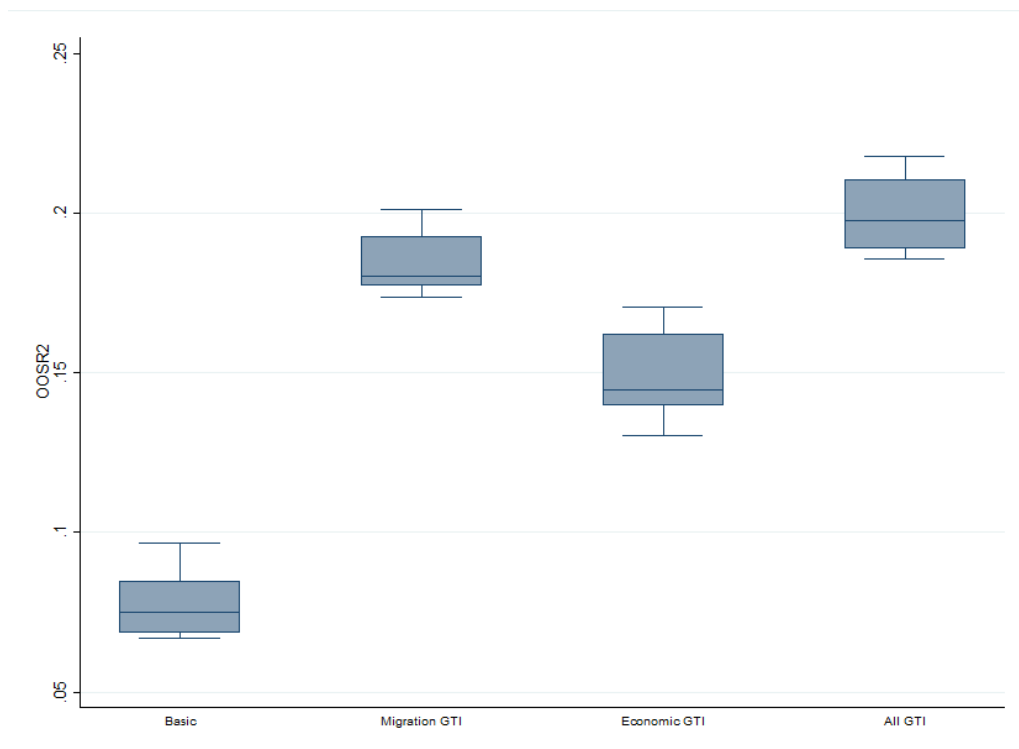
- Askatas, N. and Zimmermann, K. F.: 2009, Google Econometrics and Unemployment Forecasting, *Applied Economics Quarterly* **55**(2), 107–120.
- Beine, M., Bertoli, S. and Fernández-Huertas Moraga, J.: 2016, A Practitioners' Guide to Gravity Models of International Migration, *The World Economy* **39**(4), 496–512.
- Bertoli, S. and Fernández-Huertas Moraga, J.: 2013, Multilateral resistance to migration, *Journal of Development Economics* **102**, 79–100.
- Carrière-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.
- 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. and Rapoport, H.: 2012, Globalization, Brain Drain, and Development, *Journal of Economic Literature* **50**(3), 681–730.
- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., Ishwaran, H., Knight, K., Loubes, J. M., Massart, P., Madigan, D., Ridgeway, G., Rosset, S., Zhu, J. I., Stine, R. A., Turlach, B. A., Weisberg, S., Hastie, T., Johnstone, I. and Tibshirani, R.: 2004, Least angle regression, *Annals of Statistics* **32**(2), 407–499.
- 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.
- 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.: 2016, EM-DAT: International Disaster Database.  
**URL:** <http://www.emdat.be/>
- 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.
- 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 Jagers, K.: 2016, Polity IV project: Political Regime Characteristics and Transitions, 1800-2016 and State Fragility Index and Matrix.  
**URL:** <http://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.
- 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.: forthcoming, Reading Between the Lines: Prediction of Political Violence Using Newspaper Text, *American Political Science Review* .
- Mullainathan, S. and Spiess, J.: 2017, Machine Learning: An Applied Econometric Approach, *Journal of Economic Perspectives* **31**(2), 87–106.
- 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* .



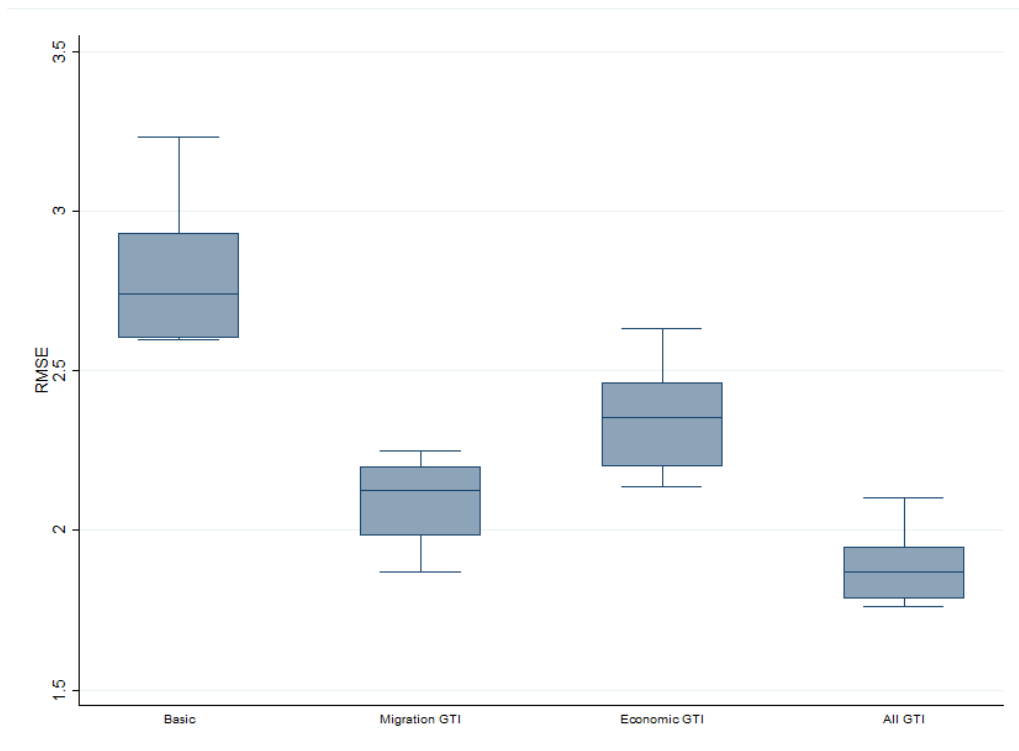
- Ortega, F. and Peri, G.: 2013, The Effect of Income and Immigration Policies on International Migration, *Migration Studies* **1**(1), 1–35.
- Preis, T., Moat, H. S. and Stanley, H. E.: 2013, Quantifying trading behavior in financial markets using Google Trends., *Scientific reports* **3**, 1684.
- 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: 2016, 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.* .

## A Figures and Tables



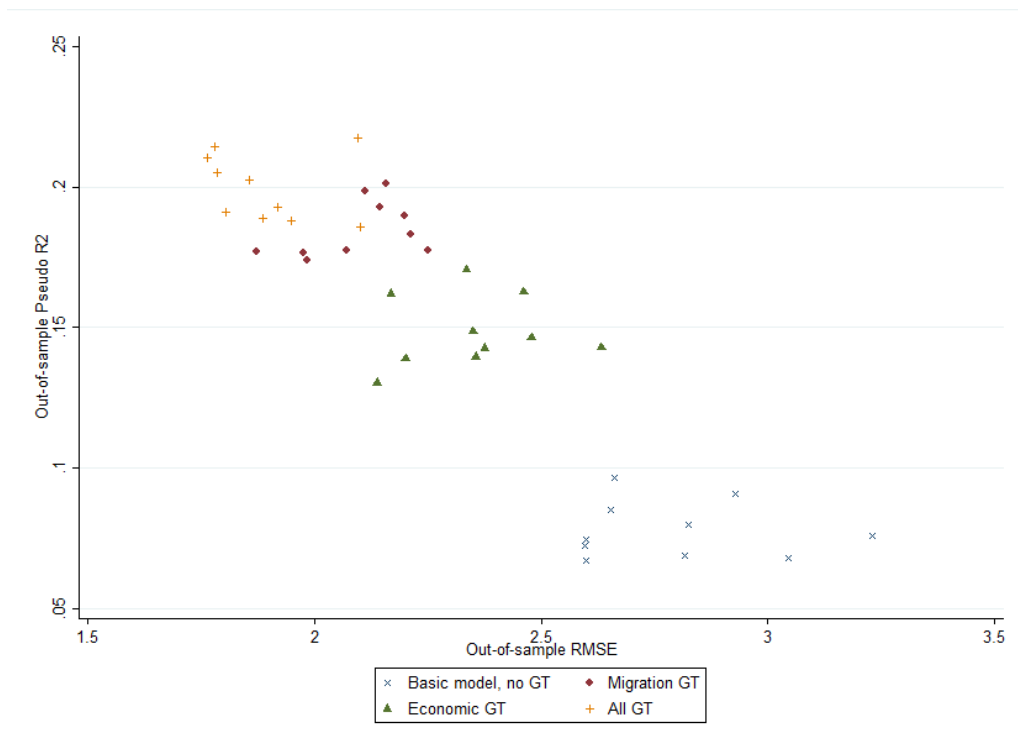
Notes: The figure reports out-of-sample estimates from 10-fold cross-validation. Each boxplot thus covers ten out-of-sample  $R^2$ s. The basic model contains controls for log GDP and log population, and origin as well as year dummies. In addition, migration keywords, economic keywords or both are added in the respective models.

Figure 1: Out-of-sample within- $R^2$  based on 10-fold cross validation



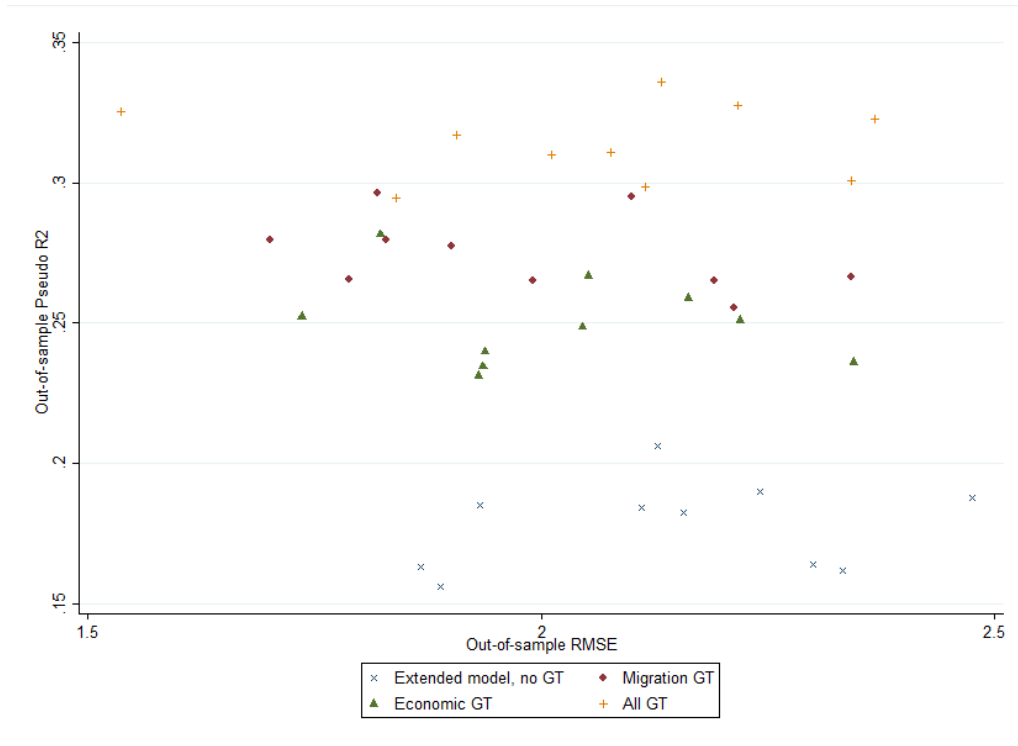
Notes: The figure reports out-of-sample estimates from 10-fold cross-validation. Each boxplot thus covers ten out-of-sample RMSEs. The basic model contains controls for log GDP and log population, origin fixed effects and destination-year fixed effects. In addition, migration keywords, economic keywords or both are added in the respective models.

Figure 2: Out-of-sample RMSE based on 10-fold cross validation



Notes: The figure reports out-of-sample estimates from 10-fold cross-validation. There are thus 10 estimates of  $OOS-R^2$  and  $OOS-RMSE$  for each model. The basic model contains controls for log GDP and log population and origin as well as year dummies. In addition, migration keywords, economic keywords or both are added in the respective models.

Figure 3: Little evidence of a trade-off between explained variance and noisy predictions (both out-of-sample)



Notes: The figure reports out-of-sample estimates from 10-fold cross-validation. There are thus 10 estimates of  $OOS-R^2$  and  $OOS-RMSE$  for each model. The extended contains controls for log GDP, log population, unemployment rate, share of young population, state fragility, the Polity IV autocracy score, population percentage of cell phone and internet users, the number of weather as well as non-weather related catastrophes, origin and year dummies. In addition, migration keywords, economic keywords or both are added in the respective models.

Figure 4: Large gains in out-of-sample within- $R^2$  compared to extended model

Table 1: Benchmark fixed effects model including GTI

Google Trends	(1) None	(2) Migration	(3) Economic	(4) Mig+Econ
Log GDP (origin)	-0.641*** (0.231)	-0.399** (0.182)	-0.445** (0.198)	-0.344* (0.176)
Log Population (origin)	2.161*** (0.597)	1.626*** (0.561)	1.793*** (0.672)	1.432** (0.612)
GTI Migration keywords (37)		✓		✓
GTI Economic keywords (37)			✓	✓
Origin FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Observations	1,068	1,068	1,068	1,068
Joint significance GTI keywords (p-value)	–	0.000	0.0002	0.000
within- $R^2$	0.077	0.2080	0.167	0.258
Number of Origins	98	98	98	98

*Sources:* Authors' calculations based on OECD International Migration database 2004–2015, World Development Indicators, and Google Trends Indices. *Notes:* Each column displays the result of a separate regression based on equation 1. Dependent variable is the logarithm of the annual aggregated flow of migrants from a given origin country to OECD. Given the within transformation of the estimator, the dependent variable captures the change in migration flows between the origin country and the OECD between period  $t$  and  $t - 1$ , while the independent variables capture the change with a lag of one year, i.e. between period  $t - 1$  and  $t - 2$ . Heteroskedasticity-robust standard errors, clustered at the origin country level, in parentheses. \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .

Table 2: Fixed effects model including Google Trends by spoken language

<b>Panel A: Spoken Language Share &gt; 20% at Origin</b>				
Google Trends	(1) None	(2) Migration	(3) Economic	(4) Mig+Econ
Log GDP (origin)	-0.860*** (0.256)	-0.584*** (0.196)	-0.634*** (0.219)	-0.471** (0.200)
Log Population (origin)	2.208*** (0.707)	1.830*** (0.659)	1.922** (0.776)	1.642** (0.719)
GTI Migration keywords (37)		✓		✓
GTI Economic keywords (37)			✓	✓
Origin FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Observations	894	894	894	894
Joint significance GTI keywords (p-value)	–	0.000	0.000	0.000
within- $R^2$	0.090	0.235	0.199	0.292
Number of Origins	82	82	82	82
<b>Panel B: Spoken Language Share &gt; 50% at Origin</b>				
Google Trends	(1) None	(2) Migration	(3) Economic	(4) Mig+Econ
Log GDP (origin)	-0.996*** (0.283)	-0.675*** (0.203)	-0.738*** (0.235)	-0.569*** (0.214)
Log Population (origin)	1.730** (0.788)	1.237* (0.708)	1.659* (0.853)	1.224 (0.788)
GTI Migration keywords (37)		✓		✓
GTI Economic keywords (37)			✓	✓
Origin FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Observations	732	732	732	732
Joint significance GTI keywords (p-value)	–	0.000	0.000	0.000
within- $R^2$	0.089	0.255	0.213	0.312
Number of Origins	67	67	67	67

*Sources:* Authors' calculations based on OECD International Migration database 2004–2015, World Development Indicators, Google Trends Indices, and Melitz Toubal language data. *Notes:* We restrict the samples to countries in which the share of the population which is commonly speaking the official language for which the Google Trends data has been extracted (English, French, or Spanish) is larger than the 20% and 50% threshold in panel A and B, respectively. Each column displays the result of a separate regression based on equation 1. Dependent variable is the logarithm of the annual aggregated flow of migrants from a given origin country to OECD. Given the within transformation of the estimator, the dependent variable captures the change in migration flows between the origin country and the OECD between period  $t$  and  $t - 1$ , while the independent variables capture the change with a lag of one year, i.e. between period  $t - 1$  and  $t - 2$ . Heteroskedasticity-robust standard errors, clustered at the origin country level, in parentheses. \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .



Table 3: Fixed effects model including Google Trends by population and income levels

<b>Panel A: Middle &amp; High Income Origins</b>				
Google Trends	(1) None	(2) Migration	(3) Economic	(4) Mig+Econ
Log GDP (origin)	-0.819*** (0.272)	-0.563*** (0.201)	-0.579** (0.224)	-0.485** (0.199)
Log Population (origin)	2.023*** (0.734)	1.121 (0.740)	1.286 (0.940)	0.916 (0.801)
GTI Migration keywords (37)		✓		✓
GTI Economic keywords (37)			✓	✓
Origin FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Observations	875	875	875	875
Joint significance GTI keywords (p-value)	–	0.000	0.000	0.000
within- $R^2$	0.092	0.244	0.207	0.312
Number of Origins	80	80	80	80
<b>Panel B: At least 10% of population with internet access</b>				
Google Trends	(1) None	(2) Migration	(3) Economic	(4) Mig+Econ
Log GDP (origin)	-1.038*** (0.390)	-0.638** (0.286)	-0.666** (0.303)	-0.438* (0.255)
Log Population (origin)	2.154** (0.859)	1.541** (0.673)	1.630* (0.850)	1.231* (0.627)
GTI Migration keywords (37)		✓		✓
GTI Economic keywords (37)			✓	✓
Origin FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Observations	647	647	647	647
Joint significance GTI keywords (p-value)	–	0.000	0.000	0.000
within- $R^2$	0.129	0.372	0.342	0.466
Number of Origins	79	79	79	79

*Sources:* OECD International Migration database 2004–2015, World Development Indicators, and Google Trends Indices.  
*Notes:* We restrict the samples to countries categorized as middle and high income economies according to the World Bank threshold of 1,025 USD per capita GDP. Each column displays the result of a separate regression based on equation 1. Dependent variable is the logarithm of the annual aggregated flow of migrants from a given origin country to OECD. Given the within transformation of the estimator, the dependent variable captures the change in migration flows between the origin country and the OECD between period  $t$  and  $t - 1$ , while the independent variables capture the change with a lag of one year, i.e. between period  $t - 1$  and  $t - 2$ . Heteroskedasticity-robust standard errors, clustered at the origin country level, in parentheses. \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .

Table 4: Fixed effects model including Google Trends with additional controls

Google Trends	(1) None	(2) Migration	(3) Economic	(4) Mig+Econ
Log GDP	-0.211 (0.285)	-0.238 (0.260)	-0.185 (0.268)	-0.256 (0.258)
Log Population	0.681 (0.911)	0.573 (0.807)	0.818 (0.941)	0.550 (0.838)
Unemployment rate	0.0228 (0.0167)	0.00547 (0.00933)	0.0115 (0.0101)	-0.00258 (0.00703)
Share of young population	0.0297 (0.0310)	0.00639 (0.0292)	0.00758 (0.0278)	0.00376 (0.0298)
State Fragility Index	0.00244 (0.0129)	0.00364 (0.0124)	0.00512 (0.0133)	0.00793 (0.0125)
Polity IV Autocracy Score	-0.0165 (0.0174)	0.0118 (0.0200)	-0.00403 (0.0183)	0.0215 (0.0188)
Mobile cellular subscriptions (per 100 people)	-0.00205 (0.00168)	-0.00110 (0.00163)	-0.00126 (0.00167)	-0.000547 (0.00152)
Internet users (per 100 people)	-0.00903*** (0.00280)	-0.00749*** (0.00260)	-0.00872*** (0.00251)	-0.00737*** (0.00262)
Weather-related disasters	-0.00169 (0.00585)	-0.00388 (0.00578)	0.00109 (0.00631)	-0.00327 (0.00566)
Non-Weather-related disasters	-0.0191* (0.0114)	-0.0235** (0.00932)	-0.0147 (0.00988)	-0.0194** (0.00954)
GTI Migration keywords (37)		✓		✓
GTI Economic keywords (37)			✓	✓
Origin FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Observations	680	680	680	680
Joint significance GTI keywords (p-value)	–	0.000	0.003	0.000
within- $R^2$	0.181	0.316	0.276	0.388
Number of Origins	70	70	70	70

*Sources:* OECD International Migration database 2004–2015, World Development Indicators, Polity IV, State Fragility Index, EMDAT disasters and Google Trends Indices. *Notes:* All explanatory variables refer to the origin country. Each column displays the result of a separate regression based on equation 1 with additional controls. The unemployment rate and the share of young population cause the largest reduction in sample size. Disaster variables count the number of events of each type as covered in the EMDAT dataset. The dependent variable is the logarithm of the annual aggregated flow of migrants from a given origin country to OECD. Given the within transformation of the estimator, the dependent variable captures the change in migration flows between the origin country and the OECD between period  $t$  and  $t - 1$ , while the independent variables capture the change with a lag of one year, i.e. between period  $t - 1$  and  $t - 2$ . Heteroskedasticity-robust standard errors, clustered at the origin country level, in parentheses. \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .

## B Appendix

### B.1 List of Keywords

*Keywords: Migration*

English	French	Spanish
applicant	candidat	solicitante
arrival	arrivee	llegada
asylum	asile	asilo
border control	controle frontiere	control frontera
citizenship	citoyennete	ciudadania
consulate	consulat	consulado
customs	douane	aduana
deportation	expulsion	deportacion
diaspora	diaspora	diaspora
embassy	ambassade	embajada
emigrant	emigre	emigrante
emigrate	emigrer	emigrar
emigration	emigration	emigracion
foreigner	etranger	extranjero
illegal	illegal	ilegal
immigrant	immigre	inmigrante
immigrate	immigrer	inmigrar
immigration	immigration	inmigracion
legalization	legalisation	legalizacion
migrant	migrant	migrante
migrate	migrer	migrar
migration	migration	migracion
nationality	nationalite	nacionalidad
naturalization	naturalisation	naturalizacion
passport	passeport	pasaporte
quota	quota	cuota
refugee	refugie	refugiado
required documents	documents requis	documentos requisito
Schengen	Schengen	Schengen
smuggler	contrebandier	traficante
smuggling	contrebande	contrabando
tourist	touriste	turista
unauthorized	non autorisee	no autorizado
undocumented	sans papiers	indocumentado
unskilled	non qualifie	no capacitado
visa	visa	visa
waiver	exemption	exencion

*Keywords: Economics*

English	French	Spanish
benefit	allocation sociale	beneficio
business	entreprise	negocio
compensation	compensation	compensacion
contract	contrat	contrato
discriminate	discriminer	discriminar
earning	revenu	ganancia
economic	economique	economico
economy	economie	economia
employer	employer	empleador
employment	emploi	empleo
GDP	PIB	PIB
hiring	embauche	contratacion
income	revenu	ingreso
inflation	inflation	inflacion
internship	stage	pasantia
job	emploi	trabajo
labor	travail	mano de obra
layoff	licenciement	despido
minimum	minimum	minimo
payroll	paie	nomina
pension	retraite	pension
recession	recession	recesion
recruitment	recrutement	reclutamiento
remuneration	remuneration	remuneracion
salary	salaire	suelo
tax	tax	impuesto
unemployment	chomage	desempleo
union+unions	syndicat	sindicato
vacancy	poste vacante	vacante
wage	salaire	salario
welfare	aide sociale	asistencia social

Notes: All keywords used with and without accents as well as British English and American English spelling.