Searching for a Better Life: Now-casting International Migration with Online Search Keywords*

Marcus Böhme André Gröger Tobias Stöhr

Work in progress: Comments welcome.
Version: May 29, 2017

Abstract

Despite recent improvements in data collection and measurement, data on migration intentions and out-flows remains scarce, largely inconsistent across countries, and often outdated. 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 well-ahead of official data publication (now-casting). We contribute to the literature by projecting flows using Google Trend Index data (GTI) on migration-specific search terms. 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. We provide evidence based on survey data that our measures indeed reflect genuine emigration intentions. They can hence not only be used in research but may also inform border control or humanitarian aid management and thus matter for policy-makers in both developing and developed countries.

JEL classification: F22, C53, C80
Keywords: Big Data, Emigration, Machine Learning, Migration Intention, Now-casting, Panel Data

*Böhme: OECD, Gröger: Goethe University Frankfurt, Stöhr: Kiel Institute for the World Economy and IZA. Corresponding author: André Gröger, Faculty of Economics and Business Administration, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60629 Frankfurt, e-mail: agroeger@wiwi.uni-frankfurt.de. We would like to thank Toman Barsbai, Christian Fons-Rosen, Juri Marcucci, Manuel Santos Silva, Claas Schneiderheinze and Alessandro Tarozzi for useful comments and discussions. We also thank seminar participants at Goethe University, Pompeu Fabra University, and the Kiel Institute for the World Economy. We are grateful to the Google Inc. for providing access to the Google Trends data. Any remaining errors are our own.

1 Introduction

In an increasingly globalized world, the topic of international migration is gaining momentum as it unfolds profound effects, both in origin and destination countries. 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 drivers of this phenomenon. However, 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 make it difficult to maintain an up-to-date picture of global migration intentions and flows. This is especially the case in developing and emerging countries in which administrative or survey-based indicators are often unavailable.[1] Approaches that can provide accurate estimates of current migration flows could thus be very helpful for policy makers and academics alike, interested in recent migration dynamics. As internet usage is increasing rapidly around the world, geo-referenced online search data provides new opportunities for predicting current human behavior ahead of official data publication (*now-casting*). Despite the existence of applications to other fields, so far, there is still a lack of evidence regarding the potential of this approach to the topic of migration.

For these reasons, we propose a novel and *direct* measure of worldwide migration intentions using aggregate online search intensities, measured by the Google Trends Index (GTI) for migration-related search terms. Recent empirical evidence has shown that people who are intending to migrate acquire relevant information about migration opportunities online in their country of origin, prior to departure (Maitland and Xu 2015). This revealed demand for information can be used as a proxy of changes in the number of aspiring migrants. Therefore, surges in online search intensities for specific keywords can indicate an increase in interest for migration-relevant information and thus reflect migration intentions directly. The GTI data consists of time series covering the relative search intensities for specific keywords through the Google search engine, which can be disaggregated down to the sub-national level on a daily basis.

Given the general lack of contemporaneous migration data, we believe that our GTI measure is useful for several reasons. First, internet searches are predominantly performed using search engines, and Google is by far the most commonly used one worldwide with a market share of 80%.[2]. Therefore, compared to other engines and online search modes,

[1] Apart from the coincidental existence of national surveys in some countries which include questions about migration intentions, 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 very costly. Second, it does not provide consistent time series of migration intentions at the country of origin level.

[2] This figure reflects the market share of Google on desktop device searches worldwide. The figure is 95% considering mobile and tablet devices. Source: https://www.netmarketshare.com/, accessed February 2017. Market shares differ by country and are well above 95% in many countries even for desktop computers. The major exception with an impact on the global market share is China, where

the GTI measure is likely to be the most representative of the internet search behavior among origin countries' general population, on average. Second, online search behavior is capturing revealed demand for information, which can help identifying migration intentions with high frequency, in a consistent and direct way, and in near to real-time.

We construct a range of country of origin-specific GTI measures based on a set of keywords which is semantically linked to the topic of "immigration" and "economics" through their co-occurrence within the Wikipedia encyclopedia. We test the predictive power of the resulting 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. We find that the augmented model outperforms standard models of migration decisions by large margins in terms of statistical fit (within-R^2). We also find evidence that the model's performance is even stronger when restricting our sample to those countries of origin in which the official language for which we extract our GTI measures, is commonly used among the general population. In other words, the more homogeneous the official language use within a country, the higher the predictive power of our measure using this language. The same applies when restricting the sample on the subset of middle- and high-income economies in which the availability and usage of internet technology is, on average, substantially higher than for low-income countries.

In order to test the robustness of our results to in-sample overfit, i.e. spurious correlations between our GTI measures and migration decisions, we apply a range of machine learning and prediction techniques such as dimension reduction, out-of-sample predictions, and variable selection methods as suggested by Varian (2014) and Kleinberg et al. (2015). When reducing the number of GTI variables from 68 to 14 by linking the keywords using the Boolean operator "OR", the performance in within-R^2 is reduced somewhat, but still outperforms the benchmark specification by around 20%. Applying a principal component analysis and reducing the dimension of our predictors to only 5 yields similar results. In a second step, we also perform a range of out-of-sample predictions using the k-fold cross-validation routine. The results from this exercise confirm that our findings also hold out-of-sample, in the sense that the augmented model performs uniformly better than the benchmark one in terms of predictive power. Last but not least, we also check the performance of our approach when using variable selection models such as the least absolute selection and shrinkage operator and the least angle regression algorithm, which penalize model complexity. The findings, again, support the view that our GTI indicators are systematically related to migration flows and yield a superior prediction performance compared to the benchmark specification.

There is a growing now-casting literature that uses big data from social networks and online search engines to predict economic outcomes across a large range of fields.

Baidu dominates the search engine market.

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 now-casting 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 substantial criticisms against their approach (Lazer et al. 2014), the now-casting literature has since grown quickly, including 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 now-casting unemployment, with applications in the context of France (Fondeur and Karamé 2013), Germany (Askitas and Zimmermann 2009), and the USA (D'Amuri and Marcucci 2012).

There is a small number of recent applications that have tried to use internet meta data to measure migration dynamics and patters. Zagheni et al. (2014) use geo-located data of about half a million users of the social network "Twitter" in OECD countries. A second application by Zagheni and Weber (2012) uses geo-referenced 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 biased samples obtained from online sources. Their main shortcoming is that self-selection into these rather specialized online services implies that results are not representative of the whole population and cannot be used to infer general migration patterns.[3] Furthermore, the data used in these studies is proprietary and, therefore, their analysis cannot be replicated or used in other contexts by external researchers. Relying on the GTI data, which is estimated to be used by over a billion users worldwide, provides a much higher level of representativeness and, therefore, can help offering a general tool for the prediction of migration.

The contribution of our paper is threefold. First, we propose a universal approach to improve existing data on migration intentions with consistent and representative indicators that are freely available. Our approach is capable of providing short-run predictions of current migration intentions which has, so far, only been captured imperfectly by selective survey data. This approach could, for example, be used for short-term now-casting analyses in the case of humanitarian crises. Second, it improves upon conventional mod-

[3]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.

els of the determinants of migration, which are frequently used to assess the elasticity of international migration intentions to changes in origin country conditions (Dustmann and Okatenko 2014). Third, it has the practical merit of helping overcome the scarcity of consistent and up-to-date data on migration intentions.

The remainder of the paper is structured as follows. Section 2 describes the data used to predict international migration flows between origin and destination countries, with a particular emphasis on our specific GTI measure of migration intentions. In Section 3, we describe the panel model used in the analysis of the determinants of migration and, subsequently, introduce machine learning techniques, which help dealing with some econometric challenges from the former approach. Section 4 provides the results from the panel estimation framework, and Section 5 reports the findings from the machine learning and prediction techniques. We discuss the value of our findings for empirical applications beyond the now-casting applications in Section 6, and Section 7 concludes.

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.[4] 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). For privacy reasons, the absolute numbers of searches are not publicly released by Google. In particular, 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

[4]Extracting large quantities of Google Trends data through the website is, however, time consuming. Google offers access to their Trends 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.

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 reference 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 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 chose 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 (2012) 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 through pure computational performance.

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 most existing now-casting 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 the English Wikipedia and identifies pairs of keywords which are semantically related.[5] 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 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.

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

[5]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 chance 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.

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 pitfall of this now-casting 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 effective 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 origin and destination countries of each migration corridor. 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 for which the GTI data is available.

We match this panel of migration flows with macroeconomic indicators of the origin and destination country from the World Development Indicators (World Bank 2015).[6] 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). Furthermore, since our approach relies heavily on language choice and its effective use among the native population in the countries of origin, we also include data on the share of the native population that commonly speaks the official languages in origin countries (Melitz and Toubal 2014). We use this data in our estimations in order 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.

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, we are left with the aggregated migration decisions towards the OECD from a sample of 95 countries of origin over ten years (2004–2013). Due to the inclusion of a one year lag in our preferred specification (equation 1 below), the corresponding total sample size is 855 country of origin-year observations.

3 Methodogy

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. Subsequently,

[6]The main setup in this paper uses GDP and population. Many other predictors have been used in the literature, for example, unemployment rates, the Human Development Index, or other proxies for the development of the origin and destination country. However, most of these indicators are often unavailable, especially for smaller countries, due to a lack of data. In order to not restrict our sample of origin countries further, we rely on our main setup based on the most frequently used predictors in the literature, unless otherwise indicated.

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:

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

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 δ_o stands for origin country-specific fixed effects. τ_t are time dummies and ε_{ot} represents a robust error term, which is clustered at the origin country level.

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 a number of prediction and machine learning techniques that have been proposed to guard against in-sample overfit (Varian 2014, Kleinberg et al. 2015).

In the literature concerned with the estimation of causal effects, some confounding factors can often be eliminated by estimating models with origin country and year fixed effects such as in equation 1. The predictive power of our GTI variables is robust to such fixed effects, as we will show in the next section. However, by definition, year fixed effects cannot be included in a now-casting or forecasting model in a similar fashion. Also, country fixed effects do not convey the same meaning in a forecasting model than in a panel model, where δ_o includes all periods t. In the second part of the empirical analysis, where we proceed to estimating out-of-sample now-casts, we therefore do not include year and country fixed effects. Instead, we apply a model which is close to the above panel model, but provides more flexibility in capturing between-variation at the country of origin level. At the same time, we ensure that no information, which is not yet available at time t, is used in the regression. Consequently, in the latter part of the paper we estimate the following now-casting equation:

$$Y_{ot+1} = \alpha + \delta_1 Y_{ot} + \delta_2 \Delta Y_{ot} + \beta T_{ot} + \gamma O_{ot} + \eta D_t + \varepsilon_{ot}, \tag{2}$$

with Y_{ot} being log migration flows from a country of origin to the OECD in a given year as above, $\Delta Y_{ot} = Y_{ot} - Y_{ot-1}$ is the change in the outcome variable before period t. The other parts of this specification are identical to the former equation. The term δ_1 allows

us to capture the role of the previous year's flow as a good starting point for predicting next year's flow. In addition, δ_2 can capture phenomena such as network effects due to the log transformation of the outcome variable.

Using equation 2, we apply three different approaches to provide evidence on the robustness of our results. All three are mainly motivated by the risk of overfitting data. First, we use dimension reduction. Here we try to keep as much of the relevant variation while reducing the number of potential regressors such as to decrease the risk of overfit. Second, we estimate out-of-sample predictions using k-fold cross-validation techniques. Third, 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) in panel A 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 with 5.7%. However, once we augment this model by our migration-related GTI variables in column (2), the R^2 almost doubles to 10.4 %, 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 increases quite substantially as well to 9.5%. 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 13.8%. Taken together, these results suggest that the predictive power of our benchmark model as measured by the within-R^2 can be improved by between 70 to 140% when including the internet search intensities in origin countries for migration and economic search terms.

In panel B, we proceed analogously as in panel A, however, starting off from a more restrictive benchmark model including the lagged dependent variable as a covariate additionally. In this augmented benchmark specification, the R^2 without any GTI predictors in column (1) increases to 10.2%, indicating that the magnitude of migration flows from the previous year is a fairly good predictor of current flows. When repeating the exercise from panel A, i.e. augmenting the model by the same set of GTI variables we still observe a substantial improvement of the model's predictive power. Despite the magnitude of the increase in R^2 being lower compared to panel A, it is still substantial, ranging between 30 to 65%.

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 applied 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 15 countries, such that the remaining number of countries included in the sample is reduced to 80 in this specification. Comparing column (2) to (4) with the benchmark specification in column (1), we find that the predictive power increases by between 80 to 180%, with the combined keywords for migration and economic terms yielding the highest predictive power. Compared to panel A in Table 1, the resulting R^2's are generally higher in this specification, consistent with our expectations regarding the language use at origin.

In panel B, we restrict our sample even further, by focusing on the origin countries in which the majority of the population commonly speaks the official language. By doing so, we drop almost 30 countries from our sample, which do not fulfill this criterion. Comparing the coefficients of determination across the different specifications, we find that they increase more strongly once again, compared to panel A in Table 1. This increase ranges between 110 to 210%, with the combined model including both migration and economics search terms performing the best.

Another empirical issue of our setup is the general availability and the use of the internet technology among the local population of the origin country. We observe strong 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 per 100 people was only 9.5 for low-income economies in 2015, compared to 39.8 in middle- and 81.0 in high-income economies, respectively.[7] 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 high internet penetration. In order to test this hypothesis, in Table 3, we perform an additional exercise in which we drop the subsample of low-income countries (29 in total). Comparing the benchmark specification in column (1) with the other specifications including the GTI predictors, we find the increase in the within-R^2 to be substantially stronger compared to the one in Table 1: depending on the specification the coefficient of determination increases between 120 to 280%, with the highest increase being achieved from the combination of migration and economic keywords in column (4) again.

[7]Source: World Telecommunication / ICT Development Report and database, and World Bank estimates (URL: http://data.worldbank.org/indicator/IT.NET.USER.P2).

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 correctly be criticized to suffer from an underlying and undeclared variable selection problem. That would result, among other issues, in standard errors that being 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 are 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 above, we use a set of keywords, which is determined by an exogenous source to restrict the number of predictors considerably before starting estimations, as a first step of addressing this issue. In what follows, we apply additional dimension reduction techniques in order to provide robust evidence that the increase in the R^2 is caused by the predictive power of the GTI measures and not simple due to mere in-sample overfit.

5.1 Dimension reduction

One obvious way to highlight the extent to which online search intensities have predictive power, while not falling prey to in-sample overfit, is by reducing dimensions mechanically before fitting the model. Short time variations in single keywords' search intensities can potentially be noisy. Reducing that noise can be achieved mechanically by linking the search terms with the Boolean "OR"-operator and then downloading chained GTI indicators.[8] These chained GTI's are less likely to provide spurious fluctuations over time. Linking keywords in such a way, however, means that the more commonly searched terms, which are not necessarily the better predictors, carry more weight within each keyword chain. If infrequently searched keywords or those that carry variation, which is relatively uncorrelated with the other keywords were the best predictors, this would result in the GTI variables appearing less relevant compared to an ideal composition of keywords. Using chained keywords is thus likely to provide a more conservative picture of the predictive power of our approach. We create chains of five keywords each by simply using their alphabetical order in English. This thus also reduces the number of regressors in T_{ot} in equation 2 to a fifth.

Another technique for dimension reduction before the estimation stage is the *principal component analysis (PCA)*. It also allows capturing more relevant variation in a limited number of explanatory variables. Principal components can be used to reduce dimensions

[8] However, Google only allows to combine up to five terms, which puts an upper limit on the extent to which this technique can be applied.

by rotating the underlying coordinate system to capture more of the variation provided by single keywords with only a few newly constructed right hand side variables. Individual principal components can then also be used as outcome variables itself to study the determinants of the relevant variation.

In Table 4, we use the two techniques of dimension reduction described above. In column (1), the GTI variables are excluded in order to derive a benchmark measure of R^2. The resulting R^2 of 0.53 indicates that approximately half of the variation is explained by the basic setup consisting of the previous mean flow, the change of the flow leading to the last period, as well as the GDP and population. Adding all single GTI variables as independent variables increases the R^2 to over 0.7. Using the mechanically dimension-reduced keyword chains instead, one can achieve a substantial improvement in the explained variance and a large number of the GTI's are statistically significant at conventional p-value cutoffs. In column (4), we use the first five principal components, which, as discussed above, cover about two thirds of the variance corresponding to the total number of single keywords. Including these principal components jointly, yields a similar increase in the in-sample R^2 compared to the keyword chains, even though their number is lower and they are thus less likely to cause a spurious R^2 increase through overfitting. The PCA model's explanatory power does not come from the first component, which accounts for almost half the variation in the underlying keywords' search intensities but does not improve prediction of migration flows. Rather, the component most strongly predictive of migration is the fifth PC, which yields an increase in in-sample R^2 as high as the combined increase from the arbitrarily formed 14 chains of keywords that are included in column (3). On its own, the fifth PC improves the R^2 by over 8 percentage points compared to the benchmark model. Combining this principal component with the keyword chains, only leads to a marginal increase in R^2, indicating that the PC covers most of that model's prediction-improving signal.

A drawback of using individual principal components is that they are relatively abstract. A way of understanding what kind of variation they might be picking up is observing the factor loadings of each component. The first principal component of migration keywords, in our example, carries 49 percent of the variation in search volumes. The first five components together amount to 68 percent of overall variation, i.e. between and within dimension. The first principal component of our keywords' search volumes might thus not only pick up interest in migration but also other sources of level effects such as the general development of search intensities in a country. Indeed all migration and economic keywords are positively related to it, which highlights that this component is most likely picking up overall search activity rather than interest in migration related terms. By contrast, the most important keywords underlying the variation captured in PC 5 are "undocumented", "applicant", "naturalization", "layoffs", "required documents", while "refugee" is the most negatively associated keyword.

Using both the manual dimension reduction technique and the PCA thus suggests that our GTI measures are not purely predictive of migration outcomes for mechanical reasons related to the large number of regressors included in the panel estimations above.

5.2 Out-of-sample exercise

The potential impact of overfit can also be reduced by using out-of-sample measures of fit, for example, the out-of-sample R^2 (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^2.

In order to provide evidence of the out-of-sample performance of our models, we use a standard technique from the machine learning literature, k-fold cross-validation. This 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 previous mean of the dependent variable, its change leading up to the previous period, the GDP and population sizes for destination and origin countries, as well as the different sets of GTI's. Figure 1 provides boxplots of the ten R^2's 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 origin-country controls O_{ot} while "empty" indicates that this term is excluded.

All models that we provide perform at least weakly better than the basic model or the empty model, which consists only of the previous periods' mean flow and the previous period's trend. In general, the basic models are slightly more consistent in their performance out-of-sample. The model using single keywords has the highest average OOS-R2 and the lowest OOS-RMSE, while the PC5 ones provide medium performance in line with our expectations. Figure 3 shows that the models with GTI's included perform better, on average, than the ones without, by explaining more of the variance in our migration outcome measure and, at the same time, producing fewer outliers, which

are heavily penalized in this setup by the OOS-RMSE. Hence, the predictive power of online search intensities for next year's migration flow remains strong, even in the artificial out-of-sample experiment, suggesting that the GTI provide genuine predictive power for migration outcomes.

5.3 Variables selection

Another 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[9] systematically shrink small coefficients towards zero in order to reduce the high variance commonly introduced when predicting outcomes with a linear regression model.[10] Thereby, LASSO combines the idea of shrinkage with variable selection using an absolute, linear penalty.[11]

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.

We follow the literature by using Mallows' Cp as the main criterion,[12] which optimizes the mean squared prediction error and thus trades off the number of extra predictors and the residual sum of squares. Using the same empirical setup applied in the previous section, the LASSO suggests a model with 36 regressors, 34 of which are single GTI's, as the

[9]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.

[10]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 yields far more stable results.

[11]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.

[12]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.

combination that yields the lowest mean squared prediction error. The LARS procedure, on the other hand, proposes 45 regressors, out of which 42 are single GTI's (chosen from a total of 68). Using the chained instead of single keywords to reduce multicollinearity before fitting, both algorithms arrive at a model with 19 regressors, including 13 of the 14 keyword chains. The results from these variable selection approaches, thus, support the view that migration-related GTI predictors are systematically related to migration flows.

6 Beyond Predictive Power?

We have presented evidence that our tailor-made GTI measure can help to increase the predictive power of economic models of current international migration flows (*now-casting*). However, an important open question regarding this approach is the one about the underlying causal mechanism between changes in the GTI and real-life migratory movements. In other words, what is our measure effectively capturing: demand for or supply of migration? Relating to recent criticism in the context of the Google Flu Index, several authors have shown that such models are susceptible to over-prediction due to herding behavior (Lazer et al. 2014, Ormerod et al. 2014). In terms of 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). On the one hand, the same might happen 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 finally led to a migratory movement, it would usually be described as a migration network effect or *chain migration* in the literature. However, 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, 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.

Nevertheless, 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 conducted 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 1,000 respondent per country and is deemed nationally representative.[13] We rely on three migration related questions

[13]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

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 intention:[14] 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 demand. 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 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 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

details about the survey methodology can be accessed online at: http://www.gallup.com/178667/gallup-world-poll-work.aspx.

[14]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.

to the time gaps, the sample size is reduced massively such that we have to rely only on only 330 observations (out of 855 in the main specification). Second, the Gallup data is a repeated cross-section and its within dimension is not very accurate. Therefore, the findings from this exercise should rather be interpreted as suggestive evidence.

In cross-sectional regressions without our GTI's, we find that the GWP variables are generally positively and significantly correlated with migration flows from our sample countries and that this correlation is increasing with the intensity of migration intentions, 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 log points, but remain significant. This indicates that there is a positive correlation between the GTI's and the GWP variables, but 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 data 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 provides substantial predictive power for estimating international migration flows. We also provided preliminary evidence based on observational survey data that our GTI measures indeed reflect genuine migration intentions. Based on these results, we propose our methodology as a universal approach to improve existing data on migration intentions worldwide with consistent and representative indicators that are freely available. By constructing GTI measures based on keywords with semantic links to other topics, our methodology could potentially serve as a general guideline of how to make use of the GTI

to be applied for prediction purposes in other contexts.

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 constitution 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.

Bibliography

Askitas, 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.

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.

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.: 2012, The predictive power of Google searches in forecasting unemployment, *Bank of Italy Temi di Discussione (Economic working papers)* .

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

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.

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.

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.

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.

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* .

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: 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.* .

A Figures and Tables

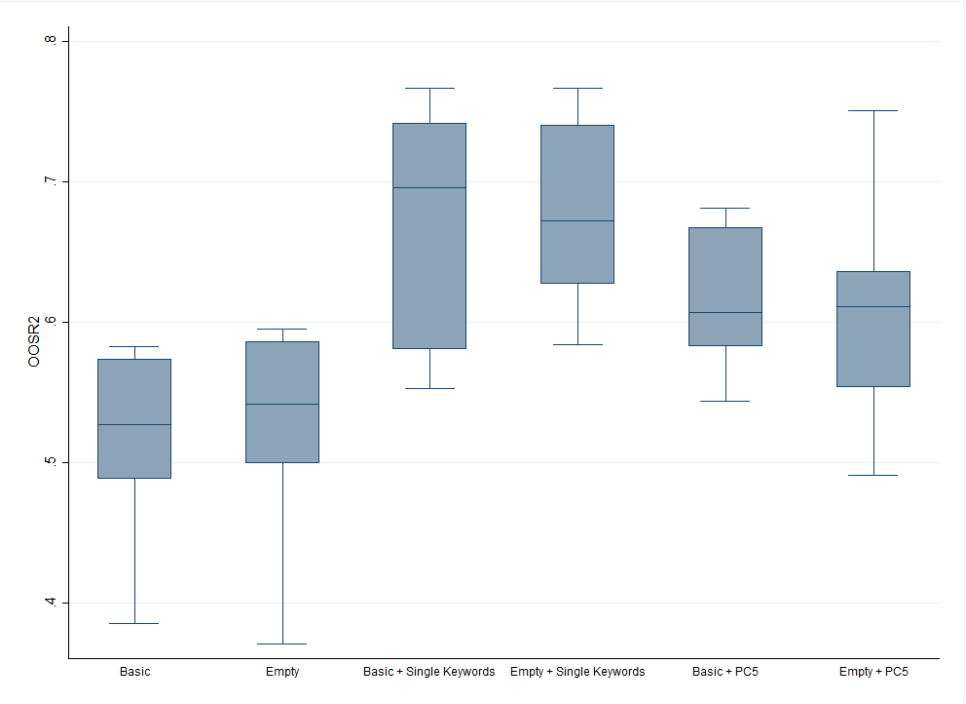

Figure 1: Out-of-sample Pseudo R2 based on 10-fold cross validation

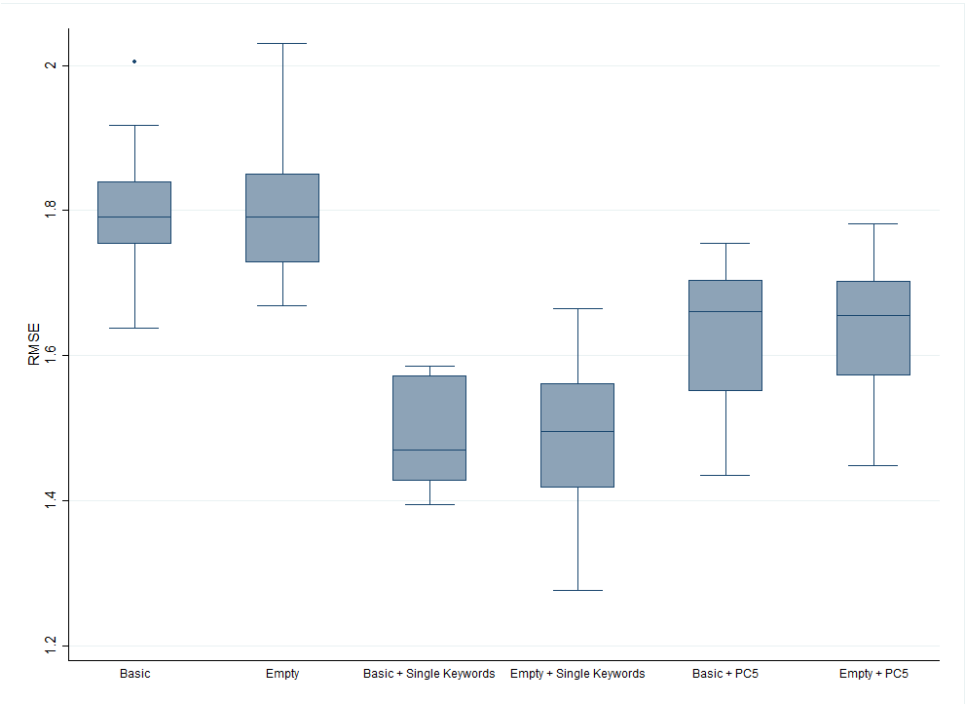

Figure 2: Out-of-sample RMSE based on 10-fold cross validation

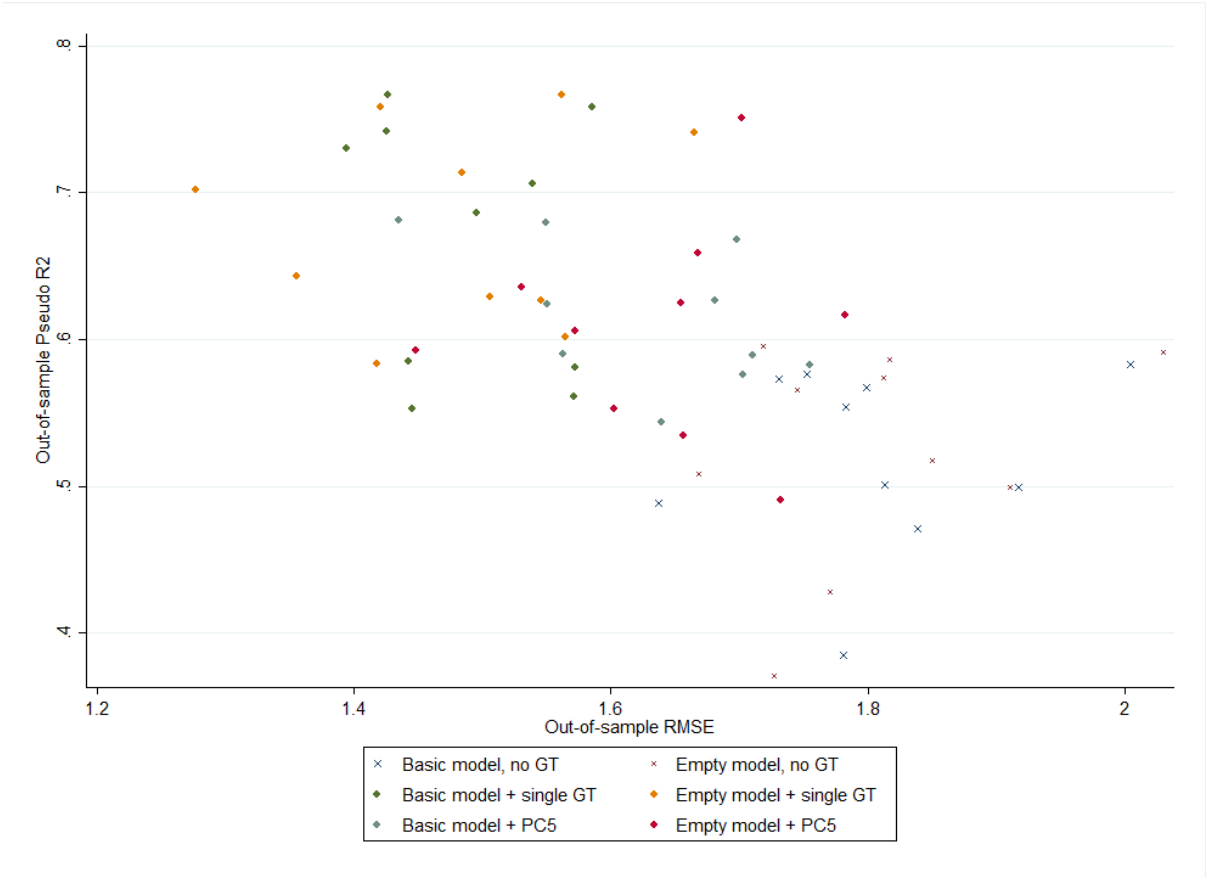

Figure 3: Little evidence of a trade-off between explained variance and noisy predictions (both out-of-sample)

Table 1: Fixed effects model including GTI

Panel A: Full Sample				
	(1)	(2)	(3)	(4)
GTI	None	Migration	Economic	Mig+Econ
Log GDP (origin)	-0.322	-0.254	-0.415	-0.279
	(0.333)	(0.354)	(0.359)	(0.380)
Log GDP (OECD)	-25.02	-19.93	-25.73	-21.14
	(19.27)	(19.59)	(20.75)	(21.55)
Log Population (origin)	-0.0339	0.0942	0.0675	-0.0739
	(1.142)	(1.066)	(1.131)	(1.148)
Log Population (OECD)	55.66	42.99	55.43	44.93
	(36.36)	(36.97)	(39.30)	(40.85)
Migration keywords		√		√
Economic keywords			√	√
Origin FE	√	√	√	√
Year FE	√	√	√	√
Observations	855	855	855	855
Number of Google Keywords	None	34	34	68
P-val (joint significance of Google keywords)		0.000	0.166	0.000
within-R^2	0.057	0.104	0.095	0.138
Number of Origins	95	95	95	95

Panel B: With Lagged Dependent Variable				
	(1)	(2)	(3)	(4)
GTI	None	Migration	Economic	Mig+Econ
Lagged Dependent Variable	0.222***	0.201***	0.198***	0.186***
	(0.0520)	(0.0561)	(0.0558)	(0.0599)
Log GDP (origin)	-0.183	-0.149	-0.288	-0.170
	(0.276)	(0.306)	(0.309)	(0.337)
Log GDP (OECD)	-29.74	-23.58	-29.57	-24.33
	(20.47)	(20.64)	(21.51)	(22.10)
Log Population (origin)	-0.310	-0.120	-0.0704	-0.225
	(1.011)	(0.956)	(1.022)	(1.042)
Log Population (OECD)	62.70	48.39	60.70	49.31
	(38.61)	(38.93)	(40.64)	(41.75)
Migration keywords		√		√
Economic keywords			√	√
Origin FE	√	√	√	√
Year FE	√	√	√	√
Observations	855	855	855	855
Number of Google Keywords	None	34	34	68
P-val (joint significance of Google keywords)		0.000	0.482	0.000
within-R^2	0.102	0.139	0.129	0.166
Number of Origins	95	95	95	95

Source: Authors' calculations based on OECD International Migration database 2004–2013, World Development Indicators, and Google Trends Indices. *Note*: 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

Panel A: Spoken Language Share > 20% at Origin				
	(1)	(2)	(3)	(4)
Google Trends	None	Migration	Economic	Mig+Econ
Log GDP (origin)	-0.671*	-0.641	-0.761*	-0.628
	(0.379)	(0.422)	(0.416)	(0.457)
Log GDP (OECD)	-24.28	-19.22	-25.67	-21.87
	(20.43)	(20.70)	(21.62)	(22.51)
Log Population (origin)	-0.152	0.0806	0.0432	-0.156
	(1.487)	(1.455)	(1.459)	(1.554)
Log Population (OECD)	55.41	42.36	57.55	48.47
	(38.30)	(38.92)	(40.79)	(42.46)
Migration keywords		√		√
Economic keywords			√	√
Origin FE	√	√	√	√
Year FE	√	√	√	√
Observations	720	720	720	720
Number of Google Keywords	None	34	34	68
P-val (joint significance of Google keywords)		0.000	0.134	0.000
within-R^2	0.055	0.115	0.098	0.152
Number of Origins	80	80	80	80

Panel B: Spoken Language Share > 50% at Origin				
	(1)	(2)	(3)	(4)
Google Trends	None	Migration	Economic	Mig+Econ
Log GDP (origin)	-0.606	-0.620	-0.817*	-0.750
	(0.437)	(0.490)	(0.477)	(0.507)
Log GDP (OECD)	-20.42	-7.368	-22.44	-7.233
	(22.63)	(23.08)	(24.16)	(25.62)
Log Population (origin)	-1.287	-1.156	-0.923	-1.197
	(1.722)	(1.746)	(1.676)	(1.769)
Log Population (OECD)	49.47	21.16	52.78	21.81
	(42.28)	(43.46)	(45.80)	(48.83)
Migration keywords		√		√
Economic keywords			√	√
Origin FE	√	√	√	√
Year FE	√	√	√	√
Observations	594	594	594	594
Number of Google Keywords	None	34	34	68
P-val (joint significance of Google keywords)		0.000	0.004	0.000
within-R^2	0.063	0.135	0.133	0.204
Number of Origins	66	66	66	66

Source: Authors' calculations based on OECD International Migration database 2004–2013, World Development Indicators, Google Trends Indices, and Melitz Toubal language data. *Note*: 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

	Middle & High Income Origins			
	(1)	(2)	(3)	(4)
Google Trends	None	Migration	Economic	Mig+Econ
Log GDP (origin)	-0.524	-0.491	-0.596	-0.467
	(0.436)	(0.446)	(0.434)	(0.456)
Log GDP (OECD)	-21.73	-13.65	-21.66	-9.875
	(23.59)	(23.41)	(22.85)	(22.06)
Log Population (origin)	-0.228	-0.0156	-0.197	-0.335
	(1.724)	(1.619)	(1.592)	(1.487)
Log Population (OECD)	49.70	30.50	49.67	24.52
	(44.35)	(43.90)	(43.08)	(41.70)
Migration keywords		√		√
Economic keywords			√	√
Origin FE	√	√	√	√
Year FE	√	√	√	√
Observations	567	567	567	567
Number of Google Keywords	None	34	34	68
P-val (joint significance of Google keywords)		0.000	0.033	0.000
within-R^2	0.055	0.150	0.120	0.213
Number of Origins	66	66	66	66

Source: OECD International Migration database 2004–2013, World Development Indicators, and Google Trends Indices. *Note*: 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: Explanation of migration levels using dimension reduction techniques

	(1)	(2)	(3)	(4)	(5)	(6)
Setup	No GTI	Single	Chained	PCA	PCA	PCA
Basic setup	yes	yes	yes	yes	yes	yes
Single keywords (68)		yes				
Chained GTI keywords (14)			yes			
Principal components 1-5				yes		
Only principal component 1					yes	
Only principal component 5						yes
Observations	855	855	855	855	855	855
R-squared	0.529	0.719	0.624	0.630	0.536	0.618

Source: OECD International Migration database 2004–2013, World Development Indicators, and Google Trends Indices. *Note*: Each column displays the results of a separate regression based on equation 2. 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. O_{ot} and D_t include log GDP and log population at origin and destination, respectively.

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 qualifies	no capacitado
visa	visa	visa
waiver	exemption	exencion

Keywords: Economic

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	sueldo
tax	tax	impuesto
unemployment	chomage	desempleo
union+unions	syndicat	sindicato
vacancy	poste vacante	vacante
wage	salaire	salario
welfare	aide sociale	asistencia social