DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 45, ARTICLE 40, PAGES 1219–1254
PUBLISHED 2 DECEMBER 2021
http://www.demographic-research.org/Volumes/Vol45/40/
DOI: 10.4054/DemRes.2021.45.40

Research Article

Now-casting Romanian migration into the United Kingdom by using Google Search engine data

Andreea Avramescu

Arkadiusz Wiśniowski

Contents

Demographic Research: Volume 45, Article 40
Research Article

Now-casting Romanian migration into the United Kingdom by using Google Search engine data

Andreea Avramescu[1]

Arkadiusz Wiśniowski[2]

Abstract

BACKGROUND
Short-term forecasts of international migration are often based on data that are incomplete, biased, and reported with delays. There is also a scarcity of migration forecasts based on combined traditional and new forms of data.

OBJECTIVE
This research assessed an inclusive approach of supplementing official migration statistics, typically reported with a delay, with the so-called big data from Google searches to produce short-term forecasts ("now-casts") of immigration flows from Romania to the United Kingdom.

METHODS
Google Trends data were used to create composite variables depicting the general interest of Romanians in migrating into the United Kingdom. These variables were then assessed as predictors and compared with benchmark results by using univariate time series models.

RESULTS
The proposed Google Trends indices related to employment and education, which exhaust all possible keywords and eliminate language bias, match trends observed in the migration statistics. They are also capable of moderate reductions in prediction errors.

CONCLUSIONS
Google Trends data have some potential to indicate up-to-date current trends of interest in mobility, which may serve as useful predictors of sudden changes in migration. However, these data do not always improve the accuracy of forecasts. The usability of Google

[1] Alliance Manchester Business School, University of Manchester, United Kingdom.
[2] Department of Social Statistics, School of Social Sciences, University of Manchester, United Kingdom. Email: a.wisniowski@manchester.ac.uk.

Trends is also limited to short-term migration forecasting and requires understanding of contexts surrounding origin and destination countries.

CONTRIBUTION
This work provides an example on combining Google Trends and official migration data to produce short-term forecasts, illustrated with flows from Romania to the UK. It also discusses caveats and suggests future work for using these data in migration forecasting.

1. Introduction

In the current century, the most significant changes seen in migration flows within Europe are credited to the 2004 and 2007 enlargements of the European Union. One of the first countries that welcomed immigrants from the new European Union member states was the United Kingdom. Between 2004 and 2017, the number of foreign nationals living in the United Kingdom almost doubled, reaching 9.4 million people, or approximately 14% of the current population (Vargas-Silva and Rienzo 2020).

One country of origin with a notable increase in the number of migrants residing in the United Kingdom is Romania. The United Kingdom's 2011 Census counted 83,168 people born in Romania who declared themselves UK residents. This is almost 12 times more than the amount counted in the 2001 Census and around five times fewer than 392,000 captured in the Annual Population Survey (APS) in 2018.[3]

The large emigration flows resulted in a decrease in labour supply and thus had a significant impact on the Romanian economy and pension system. The Romanian government was then forced to initiate policies that sought to make working-age Romanians, including emigrants, pay an additional "special" pension to first-grade dependants (e.g., parents or legal guardians) residing in Romania (digi24.ro 2019).[4] The proposed legislation was similar to policies implemented in several Asian countries (Popa 2019).

The rapid changes that occurred alongside the movement of individuals also led to significant concerns amongst the UK public about the potential negative impacts of immigration. Moreover, the "immigration threat" was a key factor leading to the British public voting in favour of leaving the European Union in the June 2016 referendum (also known as the Brexit referendum) (Dennison and Geddes 2018). In contrast to the enlargement of the European Union, the United Kingdom's departure in January 2020 (actual Brexit)

[3] The numbers are recorded by the 2011 Census, 2001 Census, and 2018 APS, respectively, by using the variable "country of birth" with the value "Romania". The 2011 Census reference covers the whole United Kingdom, while the 2001 reference is for England and Wales due to different collection frames.

[4] The original proposed legislation (in Romanian) is available at `https://senat.ro/legis/PDF/2019/19b338FG.pdf`

is expected to act as a pushing factor discouraging immigration from EU countries and leading to emigration of EU nationals back to the European Union (Sredanovic 2020).

In these circumstances, the prediction of current and future migration levels and changes in trends becomes of greater importance to policymakers and the research community. Accurate and timely measurements are needed for policy- and decision-making on multiple matters, such as allocation of public funds, estimation of labour supply and demand, or the designation of programmes for infrastructure and well-being (Sjaastad 1962; Galgoczi, Leschke, and Watt 2011), and evaluation of such policies by the local and national authorities.

The measurement and prediction of migration is exacerbated by the fact that migration is a complex event influenced by external and highly fluctuating factors (Massey 2003; Bijak and Czaika 2020). Given the current trend in technological expansion and access to new technologies (McAuliffe 2018), it would be futile to believe that the mobility phenomenon will cease within this century (Gerland et al. 2014). Further, migration has no unified definition (Raymer et al. 2013), and the data sources currently used to measure it are in most cases inconsistent and inaccurate (Parkins 2010; Bijak et al. 2019). There are also significant delays between data collection and dissemination (Abel and Sander 2014; Alexander, Polimis, and Zagheni 2020). This leads to a situation where no existent data source is able to accurately and timely measure migration (Willekens 1994, 2019).

In this article, we address this gap by assessing the approach of combining traditional surveys and big data to advance migration forecasting for short time frames (sometimes referred to as now-casting) and by predicting sudden changes in observed levels of migration flows. In particular, we propose a method of deriving novel data from the search engine Google. We then evaluate whether the inclusion of this information in a time series model that contains officially reported migration data can help reduce forecasting errors. Hence, this work differs from the majority of approaches, which rely solely on traditional datasets and have longer-term horizons (Bijak et al. 2019), and approaches that test past correlations between search engine data and observed migration (Böhme, Gröger, and Stöhr 2020). We illustrate the approach through assessing recent (2012-–2019) migration flows from Romania into the United Kingdom. One particular challenge about this flow is detecting a sudden decrease that occurred in 2018.

The rest of the paper is structured as follows. Section 2 evaluates current practices and their limitations in measuring migration between Romania and the United Kingdom; we also provide an overview of the most common big data sources used in migration research, with a focus on data from Google Trends, and also look at approaches for integrating data from various sources. The approach for creating the dataset underlying our work is presented in Section 3, while the now-casting model is explained in Section 4. Section 5 contains a brief description of the results for the forecasting exercise, and finally, Section 6 discusses the results and concludes with recommendations for further research.

2. Background

2.1 Imperfect measurement of migration from Romania to the United Kingdom

In Romania, the collection of data on international migration is problematic (Herm 2010). The main source is a population register used to measure both stocks and flows of migrants. Prior to 2009, the underlying definition used to qualify migrants was based on an intention to change residency permanently. This temporal criterion was replaced by a 12-month minimum duration of stay criterion in 2009, as required by Regulation (EC) No 862/2007 of the European Parliament and of the Council (James 2014). However, at least prior to Brexit, Romanian citizens were not legally required to register their departure to any formal authority. This is very likely to be a significant determinant of underreporting emigration (Herm 2010; Raymer et al. 2013).

The measurement of migration was further exacerbated by the fact that residence permits for Romanian nationals were not required to relocate to the United Kingdom (or elsewhere within the European Union) after Romanian accession to the European Union on 1 January 2007. The United Kingdom imposed seven years of transitional restrictions on Romanian nationals, and work permits were required for "authorised" employment until 31 December 2013 (Gower and Hawkins 2013). Free movement for employment between the European Union and the United Kingdom ceased after Brexit.

In the United Kingdom, there is no population register; the primary sources of data on migrant stocks are censuses and the Labour Force Survey with the sample boost from the APS (Raymer, Rees, and Blake 2015). Nevertheless, using these datasets to capture levels of migration is not without its problems. For example, the APS excludes persons living in communal establishments, such as student halls of residence. As the United Kingdom has been a popular education hub for European students, this exclusion criterion likely leads to an underreporting of immigration. Further, UK census data are the most comprehensive in terms of geographic granularity, providing information on economic, demographic, and social characteristics for the entire UK population (Thorvaldsen 2019: Chapter 8). However, the decennial sparsity of this source is an impediment for studying more rapid changes in both migrant stocks and migration flow trends over time. It cannot account for any immigrants entering or leaving the country in the intercensal periods (Hughes et al. 2016), making the data virtually redundant for short-term predictions.

Until March 2020, the United Kingdom's migration flow data ("Long-Term International Migration", or LTIM) were based on the International Passenger Survey (IPS), which estimated the number of people entering and leaving the United Kingdom for a period of at least 12 months based on interviews carried out at points of entry to the country. To obtain the final LTIM figures, data provided by the IPS were given further corrections using data from the Home Office and the Department of Work and Pensions (James

2020). The IPS-based LTIM estimates were also limited in not being able to provide a breakdown by country of origin (country of either birth, last residence, or citizenship).

These limitations of the IPS have been acknowledged by the United Kingdom's Office for National Statistics (ONS). The COVID-19 pandemic necessitated a suspension of the IPS and subsequent transformation of migration statistics, which began to rely on administrative data. Since November 2020, the LTIM data have been derived from a combination of figures from the Registration and Population Interaction Database (RAPID) of the Department for Work and Pensions (DWP) and visas and border data from the Home Office (Blake 2020; James 2021). However, at the time of preparing this manuscript, these data were labelled as experimental. Therefore our analysis is limited to series ending in 2019, as more information is needed to ensure that data on international migration to the United Kingdom before and after the transformation are comparable.

2.2 New forms of data

Considering the importance of accurately understanding migration trends, the development and implementation of estimation methods that can overcome the limitations of traditional data sources are of high interest. One such approach concentrates on using new forms of information, such as big data. The use of big data becomes potentially attractive, particularly when innovations in fields such as data science gather pace and the sheer amount of data available expands. For example, Sagiroglu and Sinanc (2013) noted that until 2003 the world had created only five exabytes of data, whereas by 2013 this amount was being collected every second day.

In the context of migration and mobility measurement, even though data are typically constrained by temporal, spatial, and legal dimensions, there is a wide variety of available resources. These range from mobile phone records to Internet-based approaches towards various communication channels that are either geo-tagged or have relevant user self-declared information (for a review see, e.g., Hughes et al. 2016; Rango and Vespe 2017; Righi 2019; Gendronneau et al. 2019). In the context of Romanian migration, Zagheni and Weber (2012) were the first to show that data based on emails bare many similarities to official migration statistics. More recently, social media platforms, such as Facebook and Twitter, became widely used in demographic studies. The popularity of these platforms as sources of data is largely determined by the coverage and accessibility that can be provided, as well as their potential to yield accurate measures (Zagheni, Weber, and Gummadi 2017; Fatehkia, Kashyap, and Weber 2018; Palotti et al. 2020; Alexander, Polimis, and Zagheni 2020; Fiorio et al. 2021).

Social media data provide timely documentation of movements (Hughes et al. 2016; Cesare et al. 2018). Nevertheless, their primary limitation is a lack of representativeness; many individuals will not use the platforms or do not have access to them (Blank and Lutz

2017). Other aspects of concern involve ethical and legal obstacles, such as user privacy (Boyd and Crawford 2012; Taylor 2016) and algorithmic changes. The platform owners control the data – that is, any changes in the way the information is collected and what is disclosed to the public remain uncertain and prone to sudden revisions. Moreover, the varying popularity between platforms makes the data difficult to compare over time and across countries (Hughes et al. 2016).

As a result, search engine data, such as Google Trends, have been seen as a more stable substitute. Fodness and Murray (1997, 1998) emphasised that persons wishing to engage in a travel-related activity would collect relevant social and economic information to allay any fears and doubts they may possess. This theory was confirmed by Maitland and Xu (2015), who, on a case study of educated Syrian refugees, concluded that migrants used online technologies to acquire information about the destination country before relocating. Similar findings were also obtained by Tirosh and Schejter (2017) on Israeli refugees. In the European context, there is no such study, but promising results have been obtained in the context of international tourism by Jansen, Ciamacca, and Spink (2008) and Önder and Gunter (2016).

2.3 Google Trends data

The data derived from Google Trends have been applied in various areas, from stock market trading (Preis, Moat, and Stanley 2013; Vlastakis and Markellos 2012; Kristoufek 2013) to price fluctuations (Afkhami, Cormack, and Ghoddusi 2017; Yu et al. 2019) to health and fertility (Ginsberg et al. 2009; Chan et al. 2011; Sarigul and Rui 2014; Wilde, Chen, and Lohmann 2020). The spatial-temporal dimensions of these data have mostly been explored in studies related to job searches and unemployment in Europe (Askitas and Zimmermann 2009; Fondeur and Karamé 2013; Borup and Schütte 2020) and the United States (Ettredge, Gerdes, and Karuga 2005; D'Amuri and Marcucci 2017), as well as through extensive research in tourism studies (Li et al. 2017; Dergiades, Mavragani, and Pan 2018; Siliverstovs and Wochner 2018). The pioneering paper belongs to Choi and Varian (2012), who showed that the data on searches were correlated with official tourism statistics.

In migration studies, United Nations (2014) used the Google Correlate tool (Mohebbi et al. 2011) and managed to create a dataset based on the Google Trends platform to measure labour mobility in Australia. This work was further extended by Wladyka (2017) and Wanner (2020). Wladyka (2017) tested the engine's search power in predicting international migration through a case study of labour mobility from Latin America to Spain. The author created an index by connecting work-related queries to general keywords such as "España" and "embajada de España".[5] Since Spain requires all vis-

[5] Translated as "Spain" and the "Spanish embassy", respectively.

itors from Latin American countries to have a visa, the words are likely to have added popularity due to interest in tourism and short-term visits. Moreover, the study does not account for the biases that arise due to linguistic differences between countries (Escobar 2012), with Spanish being spoken by only 64% of the population in the origin countries. Similarly, Wanner (2020) assessed predictive power of searches related to migration into Switzerland from France, Italy, Germany, and Spain. However, the study did not explicitly assess forecast errors and relied on a single search keyword related only to potential migrants' interest in employment in Switzerland.

The caveats related to linguistic differences and single search keywords were eliminated by Böhme, Gröger, and Stöhr (2020), who used the same data in a more universal model. They used terms belonging to the lexical families of keywords "immigration" and "economics". These keywords were run through the Semantic Link[6] website and one-third of the results were retrieved and used in the model. By constructing their sample by using the official languages of different countries, they offered valuable insights into the understanding of linguistic barriers in decision-making about migration. However, the actual migration forecasting was out of the scope of their article.

2.4 Migration forecasting and data integration

Recent developments in migration forecasting focused on integrating data from various sources (Willekens et al. 2016). One of the approaches to data integration is the use of Bayesian inference. For example, Bijak and Wiśniowski (2010) and Bijak (2011) forecast immigration to seven European countries using time series models that combine past data on immigration with qualitative expert knowledge obtained using a two-round Delphi survey. However, they concluded that even though this type of approach is beneficial for particular forecasts, it is not a method that can be generalised. Their inference is further reinforced by Bijak et al. (2019), who stated that expert opinion is valuable when there is a limited number of observations in the data or when a structural break is expected. Other notable examples of data integration were forecasts of net migration for all UN countries by Azose and Raftery (2015).

Studies that integrate data from traditional and social media sources to forecast migrant stocks include Gendronneau et al. (2019) for the European Union and Alexander, Polimis, and Zagheni (2019, 2020) for the United States. They supplemented the officially reported data with information derived from Facebook Advertising Platform, to produce short-term forecasts (or now-casts, or forecasts of the present) of migrant stocks. Their methods relied on Bayesian hierarchical models with a time series component and also utilised informative prior distributions to correct for biases in the Facebook data. A similar approach to now-casting, yet relying solely on the Facebook-derived data, was

[6] `https://semantic-link.com/`

used by Palotti et al. (2020) to estimate displacement of Venezuelans that resulted from recent crises taking place in Venezuela.

Econometric models are also useful tools in forecasting, although the forecasts are conditional on the predictors used in the models (Bijak et al. 2019; Cappelen, Skjerpen, and Tønnessen 2015). In their application with Google Trends data, Böhme, Gröger, and Stöhr (2020) utilised an approach for panel data with fixed effects. However, they stopped at demonstrating the explanatory power of their Google Trends Index (GTI) and recommend extending their model to include an autoregressive component for the purpose of forecasting. In this article, we fill in this gap by explicitly assessing the ability of Google Trends data to predict migration, illustrated by flows from Romania into the United Kingdom.

In summary, Google Trends data have been shown to be a valuable source in other fields. However, more research is required to advance their practical applications in migration studies, especially in short-term forecasting of migration and detecting sudden changes in trends (now-casting). This is also relevant in the lack of timely data provided by official and traditional sources. We also follow the recommendations from literature that involve integrating new forms of data and traditional sources (see, e.g., the criticism of using Google Flu Trends by Lazer et al. 2014). We do so by using autoregressive models with an additional predictor (Section 4).

3. Data

Google controls over 92% of the global search engine market (StatCounter 2019b) and close to 98% for Romania (StatCounter 2019a). Worldwide, over 40,000 searches on Google are registered every second. All these queries are recorded in the company's database by the keywords used and are openly available (presented as "intensity") on the Google Trends (GT) platform. The feature was created in 2004, and after a period of questionable validity in 2008, it has seen a continuous growth in usage. The current version of the platform was created in 2012 by merging GT with Google Insights. As explained by Google, GT measures the number of searches as a time series on a predetermined temporal scale and for a given geographical area.

To avoid the algorithm's dependency on high-density places, the relative popularity of a keyword is computed by dividing the number of searches for it by the total number of searches in a given area and period. The queries are aggregated and the final outcome is scaled between 0 and 100. This process ensures the elimination of duplicate searches by adding a recording latency of a single IP address and minimises the effect of repeated request submissions in a short time frame by Internet robots.

We posit that GT can be a useful source of information in aiding the measurement of international migration. The relatively infrequent algorithm updates compared to social media data (e.g. Palotti et al. 2020) make it possible to collect and analyse historical data for longer periods compared to other social media sources. Moreover, it can be adjusted to measure a considerable number of topics at various spatial and temporal resolutions. The only notable restriction relates to the geographical area of interest passing the minimum limit of searches set by the engine.

For our application, the GTI was created using monthly data between January 2012 and December 2019 and aggregated into annual observations. The aggregation effectively smoothed out the relatively high Google Trends values observed only in June 2016 that can be most likely attributed to the EU membership referendum and its outcome. To minimise the known biases of GT, we used a lexical approach. As the trends are sensitive to different semantics and word constructions, we used the WordNet corpus reader provided by the Natural Language Toolkit (NLTK).[7] A collection of synsets of four defined hypernyms[8] was used as a starting point: “Pound” (29 keywords; presented later as “currency”), “employment” (seven keywords), “education” (six keywords), and “housing” (six keywords). The four categories were chosen based on official statistics provided by the UK Data Service on the main reasons given by Romanian immigrants when moving to the United Kingdom (cf. Sides and Citrin 2007). Next, to eliminate the language bias and characteristics of the search engine data (Dergiades, Mavragani, and Pan 2018), we used keywords in both English and Romanian. Finally, to check that the relationship between GT data and migration counts is not spurious, we also considered a control cluster containing ten common keywords.

The keywords resulting from the lexical approach (see Table A-2 in the appendix) were then used to collect information in a programmable way using the `gtrendsR` package (Massicotte and Eddelbuettel 2020). Keywords can be either extracted individually or obtained through an endogenous relationship, dependent on time and relative popularity to each other. The first of these approaches was preferred for two reasons. Firstly, we assumed that a person searching on all keywords when gathering information on migration is unrealistic and that a combination of multiple search keywords could lead to errors in the GTI. Secondly, the differences in search interest shown by GT between some of the queries were substantially larger and resulted in no variability (zero interest) for the queries with less overall interest. Table A-2 also presents the correlations between the individual, annually aggregated GT synsets and the IPS data.

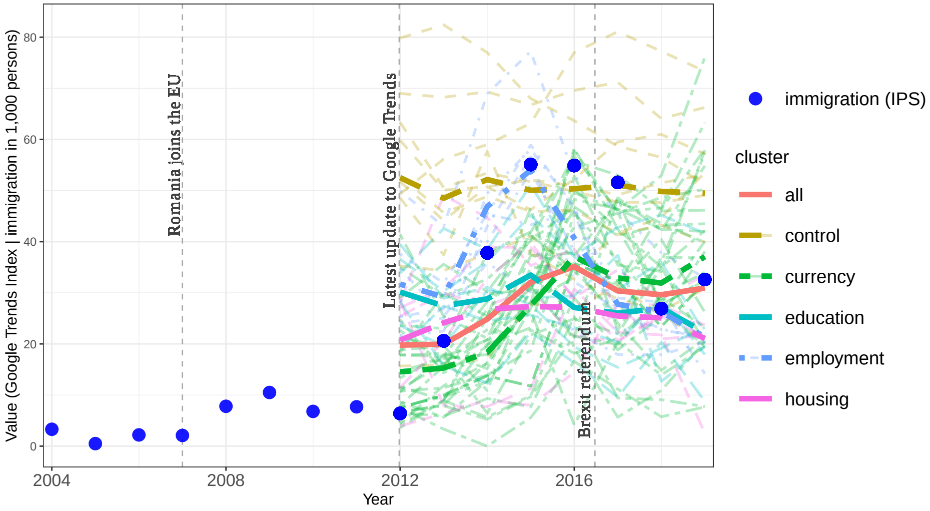
As shown in Figure 1, the selected individual synsets, aggregated into annual series, display varying trends over the 2012–2019 period, with the “control” cluster displaying the least variability over time. In Figure 1, we also present immigration flows data from the IPS. These data refer to the number of individuals (rounded to the nearest hundred)

[7] Available at `https://www.nltk.org/howto/wordnet.html`. Last accessed 1 May 2021.

[8] Synsets are unordered sets of synonyms, i.e., words that denote the same concept and are interchangeable in many contexts. Hypernyms are words whose meaning includes a group of other words, e.g. “animal” is a hypernym for “cat”.

born in Romania and entering the United Kingdom with an intention to stay for at least 12 months.

Figure 1: Immigration of Romanians to the United Kingdom (2004–2019) and Google Trends Data (2012–2019)



Source: Own calculations using International Passenger Survey (IPS) data by country of birth obtained from the Office for National Statistics; Google Trends. "Brexit" denotes the United Kingdom European Union membership referendum in June 2016.

4. Methods

In this work, we relied on relatively simple autoregressive time series models with an additional predictor (that is, a sub-class of ARIMAX models),[9] implemented in R (R Core Team 2020) via the Stan probabilistic programming language (Stan Development Team 2020). All models used two chains, with a 500 burn-in sample and 1,000 iterations per chain to produce posterior characteristics for the model parameters and forecasts. Convergence was assessed using $\hat{R}$ statistic. Computer code and data for the analysis are available at https://doi.org/10.5281/zenodo.5508213.

We tested six specifications of the Google Trends Index: four based on the clusters related to searches on employment, education, currency, and housing (see Section 3), one

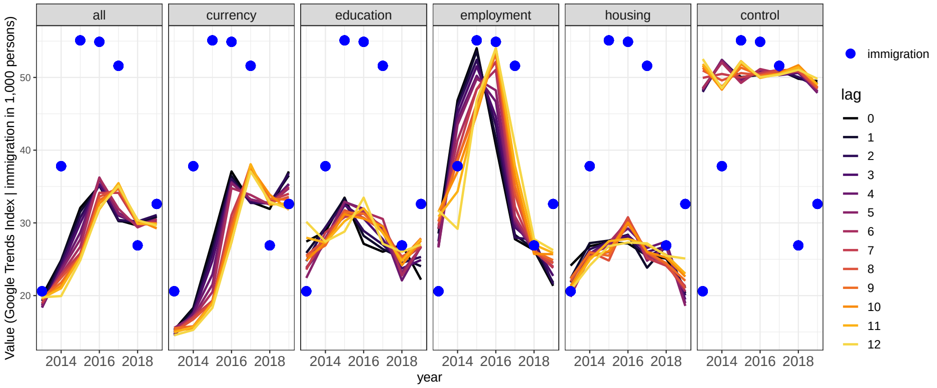
[9] These models are also known as autoregressive distributed lag (ADL) models.

related to the cluster of control keywords, and one calculated as the average GTI based on all synsets except for the control keywords.

Further, the GTI data were collected on a monthly basis to permit investigating the effect of lagging in the GTI at a finer temporal scale. We did so by calculating the annual average by using 12-month windows, over which we averaged the GTI observations for each of the synsets and then calculated annual averages for clusters and the overall GTI. This approach reflected a delay between revealing intentions to migrate by searching for relevant information and a potential decision to migrate. In principle, if monthly data on international migration were available (as, e.g., in South Korea), comparisons between monthly GTI and migration series would be possible. However, the exact lag between searches and potential migration decisions would still be unknown and a moving window lag approach could be implemented. Another limitation of working with a monthly frequency would be the relatively large variability inherent in the monthly GT search data.

In Figure 2, we present all the lagged values for the four specifications of the GTI variable, denoted further as $x_t^{(k)}$. Superscript (k), $k = 0, \ldots, 12$, denotes the window for calculating the annual average GTI. At lag $k = 12$ for a given year t, the pattern from the preceding year $t-1$ is repeated. Table A-3 further presents the correlations between the GTI clusters and the IPS data. The trend of the GTI based on the keyword “education” had the highest correlation with the IPS data of 0.98 at the lag of seven months.

Figure 2: Immigration of Romanians into the United Kingdom using Google Trends Data (2013–2019)



Note: Lag 0 in year 2019 denotes the GTI calculated as the average of indices between January and December 2019, lag 1 is the average GTI over December 2018 and November 2019, and so on. Lag 12 for 2019 is the average GTI over January to December 2018.
Source: Own calculations using International Passenger Survey (IPS) data on immigration flows by country of birth obtained from the Office for National Statistics; Google Trends.

A general specification of the forecasting model is

$$\log y_t \sim \mathrm{Normal}\left(\phi_0 + \phi_1 \log y_{t-1} + \phi_2 \log x_t^{(k)}, \sigma_t^2\right), \tag{1}$$

where y_t and $x_t^{(k)}$ are immigration flows and GTI, respectively, in year t, ϕ_0 denotes the intercept, ϕ_1 is an autoregressive parameter, ϕ_2 is a distributed lag parameter, and σ_t^2 is variance.

We assumed two variants of the model, a stationary autoregressive model ($-1 \leq \phi_1 < 1$) and a random walk model ($\phi_1 = 1$) (cf. Bijak and Wiśniowski 2010). Prior distributions are $\phi_0 \sim \mathrm{Normal}(0, 4^2)$, $\phi_1 \sim \mathrm{Normal}(0, 0.4^2)1(0 \leq \phi_1 < 1)$, and $\phi_2 \sim \mathrm{Normal}(0, 1^2)$. Due to the short time series at hand, we included a data-driven informative prior distribution for the variance parameter σ_t^2. We constructed it as follows. Since the model in (1) was specified for logarithms of migration, the prior for σ_t could be expressed in terms of relative uncertainty. In Table A-1, the IPS point estimates of immigration flows are presented together with their survey-based standard errors, expressed as a percentage of the point estimate (relative standard errors). We observed a clear pattern of underdispersion – that is, the larger the point estimate, the lower the relative uncertainty was. In particular, as shown in Figure A-1 and Table A-4, the relationship between the logarithm of the IPS point estimate and logit-transformed relative standard errors (proportions) was linear. Thus we constructed the prior for σ_t based on this relationship as

$$logit\left(\frac{\sigma_t}{y_t}\right) \sim \mathrm{Normal}(\alpha_1 \log y_t, 0.286), \tag{2}$$

$$\alpha_1 \sim \mathrm{Normal}(-0.49, 0.026). \tag{3}$$

In essence, this allowed reproducing the relative standard errors as predicted by linear regression Model (B) in Table A-4 and applying them to the forecasted values so that the relative posterior uncertainty resembled the one observed in the IPS (that is, the underdispersion).

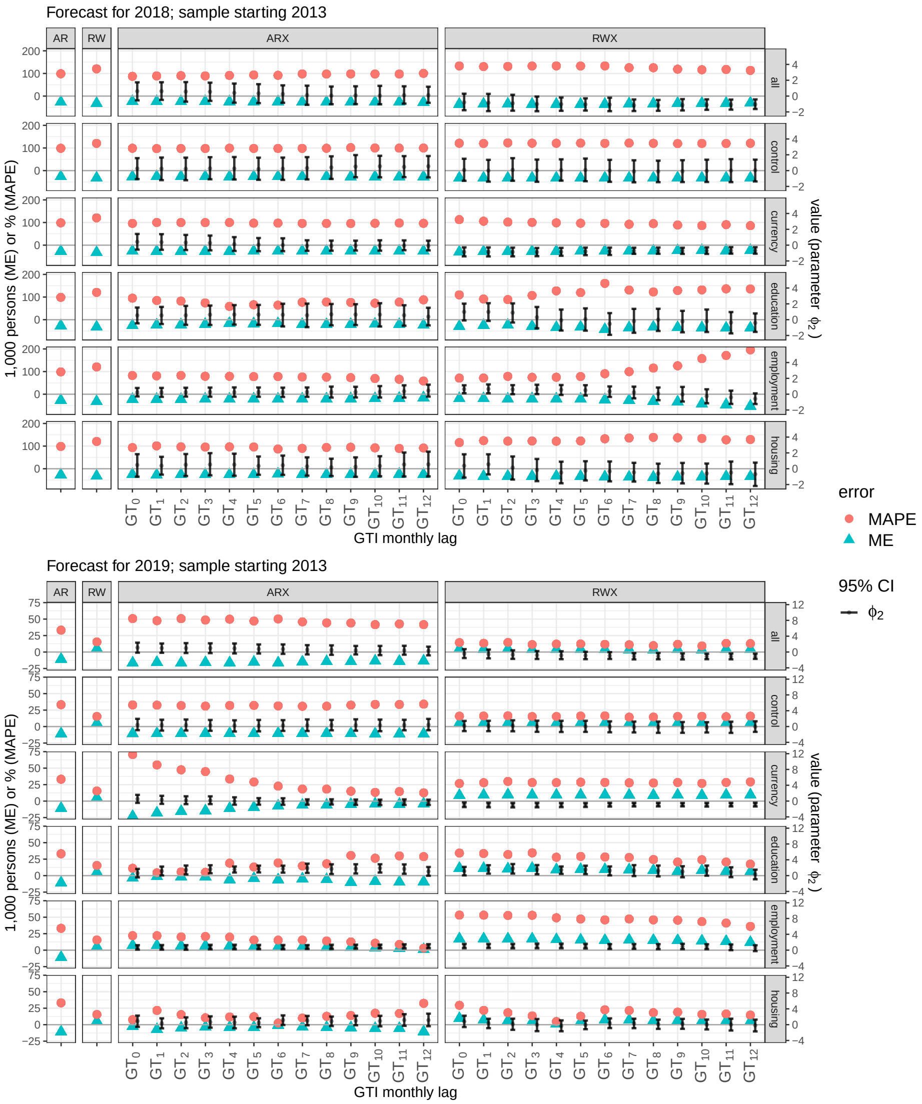
To assess the predictive performance of the GTI in short-term forecasting of immigration, we used the autoregressive and random walk models, denoted by AR and RW, respectively, without the GTI predictor ($\phi_2 = 0$) as the benchmark models. The benchmarks were compared with the models (denoted by ARX and RWX) containing a single predictor from all the $4 \times 13 = 52$ combinations of the cluster keyword-based and overall GTIs that were calculated for each monthly window lag. Performance was assessed by using two common metrics, mean error (ME) and mean absolute percentage error (MAPE). The ME captures average systematic *bias* – that is, under- or over-prediction of flow, and is quantified in the number of persons; MAPE is a measure of forecast *accuracy* – that is the average error, expressed in relative terms. We compared our forecasts with the officially reported IPS estimates of international migration from Romania to the United Kingdom. We focused on the short-term forecasts (now-casts) of immigration, by using the 2013–2017 training sample to forecast immigration in 2018 and the 2013–2018 training sample to forecast immigration in 2019. As a robustness check, we also tested if forecast accuracy improved by using first differences of logged GTIs as a predictor and by adding one more year to the data (2012).

5. Results

The results of the forecasting exercise for the sample starting in 2013 are presented in Figure 3; the two robustness checks are in Figure A-2 in the appendix. The errors (ME and MAPE) are presented for the reference models AR and RW, and for the ARX and RWX models that included a predictor from all combinations of the GTI for all lags and clusters as described in Section 4. We also present the mean and 95% credible intervals (CI) for the parameter ϕ_2 of the additional GTI predictor in a given model.

For the 2018 forecasts based on 2013–2017 data (Figure 3), we observe that the benchmark model AR performed better than the RW. However, both yielded relatively large errors (MAPE was 99% and 120% for AR and RW, respectively, while ME showed under-prediction of migration flows by 26,500 and 32,400, respectively). The inclusion of the total average GTI predictor (denoted as “all” keywords in the figure) only slightly improved the forecast errors in the ARX model for lags $k \leq 8$. However, the ARX models with the hypernym cluster “employment” visibly reduced the MAPE to around 60-75% for lags $k \geq 8$ (the average MAPE for all lags was 75%; ME was $-20,300$). Similar reductions but only for lags $k \leq 5$ were obtained by using the RWX models with the same cluster GTI. The average MAPE yielded by the “education” cluster GTI was 76% for the ARX and 122% for the RWX model. We also observe that the posterior 95% CI for the parameter ϕ_2 included zero for all models. The forecasts for 2018 and the benchmark (IPS data) for selected models are presented in Figure A-3, while reductions in MAPE averaged over all GTI lags are shown in Figure A-6 in the appendix.

Figure 3: Forecast errors and 95% credible intervals (CI) for 2018 and 2019, based on samples 2013–2017 and 2013–2018, respectively, for benchmark models (AR and RW) and corresponding models with Google Trends Index (GTI) predictors (ARX and RWX)



Note: Subscripts at GT denote a monthly lag in calculating the GTI. For example, for 2019, GT_0 refers to GTI calculated as an average over January–December 2019; GT_2 is an average over November 2018–October 2019.

For the 2019 forecasts, the RW benchmark model yielded a relatively small 15% MAPE and 5,000 ME, while for the AR these measures were 33% and $-10,800$, respectively. Reductions in errors when using GTI variables were only modest, comparing to this benchmark. Most notably, the ARX models with the largest lags of the "employment" GTI produced the smallest MAPEs (around 3–10%), along with small lags of the "education" and "housing" GTIs (MAPEs of 2–12%). The RWX models containing the total GTI ("all") showed reduction in the MAPE to 9–12%. However, the ϕ_2 coefficients were different from zero only for the "education" and "employment" GTI clusters. The forecasts for 2019 based on these models are shown in Figure A-4 in the appendix.

The uncertainty of the forecasts was assessed by comparing 95% posterior predictive intervals (PIs) of the forecasts for 2018 and 2019 with the respective 95% confidence intervals of the IPS data in Figure A-5. We observed that the resulting PIs for 2018 were wider than the confidence intervals for the IPS data. However, the predicted medians were also relatively larger. This is not surprising given the sudden drop in the reported level of migration flows from Romania to the United Kingdom in 2018. For 2019 forecasts (based on 2013–2018 data), we observed that the uncertainty was slightly larger than the one reported by the IPS for the ARX models but similar to the IPS for the RWX models.

The robustness analyses using additional observation in 2012 from the IPS (top panel in Figure A-2) yielded similar results to the ones described above. Also, lags $k > 5$ of the "currency" GTI predictor performed well when compared with the "education" and "housing" GTIs in 2019 forecasts. When using the differenced series of the GTI as a predictor, the results were more diverse. For 2018, most of the GTI coefficient (ϕ_2) CI crossed zero, and smaller gains in accuracy than for the previous two models were observed. For 2019, the RWX model with the "education" GTI reduced the forecast errors for selected monthly lags but no consistent reductions were observed; the ARX model with the "currency" GTI also noted minor improvements in forecast errors.

6. Discussion

6.1 Summary

This article builds on scarce literature relating to migration forecasting where both traditional and new forms of data are used. In particular, the ability of Google Trends data was assessed to examine whether improvements could be made to very short-term forecasts of migration flows from Romania to the United Kingdom. An obvious advantage of using search engine data in now- and short-term forecasting is their almost instantaneous availability, as opposed to more traditional official migration datasets that are typically produced with a delay of at least several months.

The accession of Romania to the European Union in 2007, alongside political and social instability in the country, significantly influenced individuals in Romania to move. Towards the end of 2011 and the beginning of 2012, Romania saw major protests against the political elite, who were taking drastic job and wage cuts (BBC 2012). The ongoing economic situation in the country, together with the withdrawal of work permits required for Romanian nationals in the United Kingdom, which came in two years later, created an environment where the propensity to migrate was likely to increase. This expansion ceased with the result of the United Kingdom's EU membership referendum in 2016. The uncertainty it brought about and the subsequent treatment of the EU nationals after the United Kingdom left could have been an important reason for people to become reluctant to choose the United Kingdom as their migration destination. This premise is also supported by the Department for Work and Pensions (2020), which confirmed a decrease in allocations of National Insurance Numbers for EU workers.

Our results indeed showed that the composite indices of search keywords constructed by using a lexical approach and related to seeking employment and education can improve short-term now-casts (one year ahead). These indices can complement official migration statistics and thus help detect potential sudden changes in trends of migration flows. Also, keywords related to employment and education are in line with the main reasons for migration as discussed in the literature (e.g., Jordan 2019). The other composite indices, related to housing and the exchange rates of Romanian and British currencies, showed mixed patterns in the ability to improve forecasts. These two indices are also difficult to interpret and may be related to short-term or seasonal migration. The overall index based on all search keywords showed small reductions in forecast errors when predicting both 2018 and 2019 levels of migration, though the errors varied depending on the specification of the models.

This work suggests that the significant associations between the time series based on search engine data and migration flows (Böhme, Gröger, and Stöhr 2020) may not be sufficient to consistently improve forecast accuracy, especially in cases where sudden changes in levels of migration flows take place. This is in line with findings by Bijak et al. (2019), who point out that various types of flows (for example, foreigners and returning nationals) may be subject to more sudden shifts in trends and driven by different push and pull factors. They also recommend that particular types of flows may be analysed more meaningfully in terms of their potential predictability, as well as risks related to the large uncertainty of the forecasts, rather than in terms of exact future levels.

6.2 Limitations and recommendations for future work

The major limitation of the proposed approach is that there is no single way of constructing a composite index based on Google Trends data. For instance, the index based on searches related to employment and education reduced forecast errors in 2018 and 2019.

However, mixed results were obtained for the indices based on other keywords, as well as the all-keywords index. This reflects that as the situation in the sending and receiving countries changes, intentions about migration and actual decisions can follow suit. Additional predictors that capture this changing situation, such as changes in GDP or unemployment, could be introduced in the forecasting models. However, their usability would depend on the availability and timeliness of the data on these covariates. Alternatively, approaches that forecast multiple variables simultaneously, such as vector autoregressive models, could be employed. These models, however, typically require longer time series than those available in this study.

The choice of a lag between the intentions expressed in searches and actual migration is also a source of uncertainty. We demonstrated that results depend on the way annual aggregated GTI predictors are created from monthly data and moving-window averages. Here, Bayesian model and/or forecast averaging may be used to account for this uncertainty.

The overall Google Trends Index constructed from all keywords showed relatively small variability in comparison with the observed migration flows. This demonstrates that, in general, the construction of such an index requires theoretical foundations, detailed knowledge of the characteristics of flows and the specific contexts in which they occur, and an understanding of temporal dynamics (cf. Klugman 2009; Bijak et al. 2019). The relative importance of the keywords may change over time (Böhme, Gröger, and Stöhr 2020).

Further, in 2016, 30% of Romanians were reported to never have used the Internet (second largest percentage in the European Union), though significant increases in the share of households with Internet access over the last decade have been observed (Eurostat 2018). This, in general, may lead to a bias that would need to be taken into account, especially in multi-country analyses and comparisons. In extensions to other countries, one would also need to consider the heterogeneity of populations, as migrants may be using their native languages in search engines. More work is also needed in differentiating intentions related to short-term and seasonal mobility and long-term migration when analysing search engine data.

Finally, key areas for future research are “hot topics” in Google searches and statistical methods that can be used to eliminate such biases. Google Trends can be influenced to a great extent by the instability of interest within the general population in topics such as Brexit or the recent COVID-19 pandemic. In such situations, information on changes in the intentions under study, for example, about relocating to other countries, may be useful.

6.3 Conclusions

The uncertain economic, political, and social environment affects migration decisions and resulting observed migration trends. Thus it is important for migration estimates and forecasts to be timely and accurate. However, the traditional data sources rarely provide timely information on rapid changes in levels of migration flows.

The data derived from Google Trends can be useful in detecting changes in national-level measurements of migration, such as migration from Romania to the United Kingdom. The Google Trends Index, based on an ensemble of search keywords, has a key advantage of being available earlier than the officially reported migration statistics. However, an important limitation is that the usage of the search engine data depends on the context surrounding sending and receiving countries, and characteristics of the migration flows being forecasted.

7. Acknowledgements

Arkadiusz Wiśniowski acknowledges funding received from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870649, the project Future Migration Scenarios for Europe (FUME). Both authors would like to thank Gregory Campbell as well as two anonymous reviewers and the associate editor for their helpful comments and suggestions.

References

Abel, G.J. and Sander, N. (2014). Quantifying Global International Migration Flows. *Science* 343(6178): 1520–1522. doi:10.1126/science.1248676.

Afkhami, M., Cormack, L., and Ghoddusi, H. (2017). Google search keywords that best predict energy price volatility. *Energy Economics* 67: 17–27. doi:10.1016/j.eneco.2017.07.014.

Alexander, M., Polimis, K., and Zagheni, E. (2020). Combining social media and survey data to Nowcast migrant stocks in the United States. *Population Research and Policy Review* doi:10.1007/s11113-020-09599-3.

Alexander, M., Polimis, K., and Zagheni, E. (2019). The impact of Hurricane Maria on out-migration from Puerto Rico: Evidence from Facebook data. *Population and Development Review* 45(3): 617–630. doi:10.1016/j.eneco.2017.07.014.

Askitas, N. and Zimmermann, K.F. (2009). Google econometrics and unemployment forecasting. *SSRN Electronic Journal*. doi:10.2139/ssrn.1480251.

Azose, J.J. and Raftery, A.E. (2015). Bayesian probabilistic projection of international migration. *Demography* 52(5): 1627–1650. doi:10.1007/s13524-015-0415-0.

BBC (2012). Romania protests: PM Emil Boc calls for dialogue. British Broadcasting Corporation. `https://www.bbc.co.uk/news/world-europe-16570860`. Accessed 12/12/2020.

Bijak, J. (2011). Forecasting international migration in Europe: A Bayesian view. *The Springer Series on Demographic Methods and Population Analysis* doi:10.1007/978-90-481-8897-0.

Bijak, J. and Czaika, M. (2020). Assessing uncertain migration futures – A typology of the unknown. QuantMig Project Deliverable D1.1. `http://quantmig.geodata.soton.ac.uk/res/files/QuantMig%20D1.1%20Uncertain%20Migration%20Futures%20V1.1%2030Jun2020.pdf`. Accessed 14/09/2021.

Bijak, J., Disney, G., Findlay, A.M., Forster, J.J., Smith, P.W., and Wiśniowski, A. (2019). Assessing time series models for forecasting international migration: Lessons from the United Kingdom. *Journal of Forecasting* 38(5): 470–487. doi:10.1002/for.2576.

Bijak, J. and Wiśniowski, A. (2010). Bayesian forecasting of immigration to selected European countries by using expert knowledge. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173(4): 775–796. doi:10.1111/j.1467-985x.2009.00635.x.

Blake, A. (2020). Population and migration statistics system transformation - overview. Office for National Statistics. `https://bit.ly/38SP3u9`. Accessed 09/09/2021.

Blank, G. and Lutz, C. (2017). Representativeness of social media in Great Britain: Investigating Facebook, Linkedin, Twitter, Pinterest, Google, and Instagram. *American Behavioral Scientist* 61(7): 741–756. doi:10.1177/0002764217717559.

Böhme, M.H., Gröger, A., and Stöhr, T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics* 142: 102347. doi:10.1016/j.jdeveco.2019.04.002.

Borup, D. and Schütte, E.C.M. (2020). In search of a job: Forecasting employment growth using Google Trends. *Journal of Business and Economic Statistics*. doi:10.1080/07350015.2020.1791133.

Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication and society* 15(5): 662–679. doi:10.1080/1369118X.2012.678878.

Cappelen, Å., Skjerpen, T., and Tønnessen, M. (2015). Forecasting immigration in official population projections using an econometric model. *International Migration Review* 49(4): 945–980. doi:10.1111/imre.12092.

Cesare, N., Lee, H., McCormick, T., Spiro, E., and Zagheni, E. (2018). Promises and pitfalls of using digital traces for demographic research. *Demography* 55(5): 1979–1999. doi:10.1007/s13524-018-0715-2.

Chan, E.H., Sahai, V., Conrad, C., and Brownstein, J.S. (2011). Using web search query data to monitor Dengue epidemics: A new model for neglected tropical disease surveillance. *PLoS Neglected Tropical Diseases* 5(5). doi:10.1371/journal.pntd.0001206.

Choi, H. and Varian, H. (2012). Predicting the present with Google Trends. *Economic Record* 88: 2–9. doi:10.1111/j.1475-4932.2012.00809.x.

Dennison, J. and Geddes, A. (2018). Brexit and the perils of 'Europeanised' migration. *Journal of European Public Policy* 25(8): 1137–1153. doi:10.1080/13501763.2018.1467953.

Department for Work and Pensions (2020). National insurance number allocations to adult overseas nationals entering the uk to june 2020. `http://tiny.cc/ml5juz`. Accessed 26/11/2020.

Dergiades, T., Mavragani, E., and Pan, B. (2018). Google Trends and tourists arrivals: Emerging biases and proposed corrections. *Tourism Management* 66: 108–120. doi:10.1016/j.tourman.2017.10.014.

digi24.ro (2019). Românii sunt uimiți de proiectul care i-ar obliga pe copii să plătească pensii părinților. “Chiar lege să dea?”. https://tinyurl.com/digiro4719. Accessed 21/11/2020.

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. doi:10.1016/j.ijforecast.2017.03.004.

Escobar, A.M. (2012). Bilingualism in Latin America. In: Tej, K.B. and William, C.R. (eds.). *The handbook of bilingualism and multilingualism*. Chichester: Blackwell Publishing: 725–744. 2 ed. doi:10.1002/9781118332382.

Ettredge, M., Gerdes, J., and Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM* 48(11): 87–92. doi:10.1145/1096000.1096010.

Eurostat (2018). Archive: Internet access and use statistics - households and individuals. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Archive:Internet_access_and_use_statistics_-_households_and_individuals&oldid=379591. Accessed 01/12/2020.

Fatehkia, M., Kashyap, R., and Weber, I. (2018). Using Facebook ad data to track the global digital gender gap. *World Development* 107: 189–209. doi:10.1016/j.worlddev.2018.03.007.

Fiorio, L., Zagheni, E., Abel, G., Hill, J., Pestre, G., Letouzé, E., and Cai, J. (2021). Analyzing the effect of time in migration measurement using georeferenced digital trace data. *Demography* 58(1): 51–74. doi:10.1215/00703370-8917630.

Fodness, D. and Murray, B. (1997). Tourist information search. *Annals of Tourism Research* 24(3): 503–523. doi:10.1016/s0160-7383(97)00009-1.

Fodness, D. and Murray, B. (1998). A typology of tourist information search strategies. *Journal of Travel Research* 37(2): 108–119. doi:10.1177/004728759803700202.

Fondeur, Y. and Karamé, F. (2013). Can Google data help predict French youth unemployment? *Economic Modelling* 30: 117–125. doi:10.1016/j.econmod.2012.07.017.

Galgoczi, B., Leschke, J., and Watt, A. (2011). Intra-EU labour migration: Flows, effects and policy responses. Tech. rep., European Trade Union Institute. doi:10.2139/ssrn.2264049.

Gendronneau, C., Wiśniowski, A., Yildiz, D., Zagheni, E., Florio, L., Hsiao, Y., Stepanek, M., Weber, I., Abel, G., and Hoorens, S. (2019). Measuring labour mobility and migration using Big Data: Exploring the potential of social-media data for measuring EU mobility flows and stocks of EU movers. Tech. rep., European Commission. https://www.rand.org/pubs/external_publications/EP68038.html. Accessed 10/09/2021.

Gerland, P., Raftery, A.E., Ševčíková, H., Li, N., Gu, D., Spoorenberg, T., Alkema, L., Fosdick, B.K., Chunn, J., Lalic, N., and et al. (2014). World population stabilization unlikely this century. *Science* 346(6206): 234–237. doi:10.1126/science.1257469.

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. doi:10.1038/nature07634.

Gower, M. and Hawkins, O. (2013). Ending of transitional restrictions for Bulgarian and Romanian workers. *House of Commons Library* 29. `https://ecas.issuelab.org/resources/18606/18606.pdf`. Accessed 10/09/2021.

Herm, A. (2010). Country Report Romania. Tech. rep., PROMINSTAT: Promoting Comparative Quantitative Research in the Field of Migration and Integration in Europe. `http://research.icmpd.org/fileadmin/Research-Website/Project_material/PROMINSTAT_File_Exchange/PROMINSTAT_Romania.pdf`. Accessed 14/09/2021.

Hughes, C., Zagheni, E., Abel, G.J., Wiśniowski, A., Sorichetta, A., Weber, I., and Tatem, A. (2016). Inferring migrations: traditional methods and new approaches based on mobile phone, mocial media, and other big data: Feasibility study on inferring (labour) mobility and migration in the European Union from big data and social media data. Tech. rep., European Commission. `https://op.europa.eu/en/publication-detail/-/publication/1f66f928-f307-4c1f-9bec-fde0d2008c69`. Accessed 10/09/2021.

James, M. (2014). International migration of Romania. Office for National Statistics. `https://bit.ly/2SX0QmP`. Accessed 26/09/2020.

James, M. (2020). Long-term international migration estimates methodology. Office for National Statistics. `https://bit.ly/36t69xb`. Accessed 26/09/2020.

James, M. (2021). International migration: Developing our approach for producing admin-based migration estimates, April 2021 release. Office for National Statistics. `http://bitly.ws/dHpR`. Accessed 26/11/2020.

Jansen, B.J., Ciamacca, C.C., and Spink, A. (2008). An analysis of travel information searching on the web. *Information Technology & Tourism* 10(2): 101–118. doi:10.3727/109830508784913121.

Jordan, B. (2019). Mobility and migration. In: *Authoritarianism and how to counter It*. Palgrave Macmillan, Cham, Switzerland: 51–62. doi:10.1007/978-3-030-17211-4.

Klugman, J. (2009). Human development report 2009. overcoming barriers: Human mobility and development. Tech. rep., United Nations Development Programme. https://ssrn.com/abstract=2294688. Accessed 14/09/2021.

Kristoufek, L. (2013). BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports* 3(1). doi:10.1038/srep03415.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science* 343(6176): 1203–1205. doi:10.1126/science.1248506.

Li, X., Pan, B., Law, R., and Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism Management* 59: 57–66. doi:10.1016/j.tourman.2016.07.005.

Maitland, C. and Xu, Y. (2015). A social informatics analysis of refugee mobile phone use: A case study of Zaaatari Syrian refugee camp. *SSRN Electronic Journal* doi:10.2139/ssrn.2588300.

Massey, D.S. (2003). Patterns and processes of international migration in the 21st century. In: *Conference on African Migration in Comparative Perspective, Johannesburg, South Africa*. vol. 4(7), 1–41. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.925&rep=rep1&type=pdf. Accessed 09/09/2021.

Massicotte, P. and Eddelbuettel, D. (2020). *gtrendsR: Perform and Display Google Trends Queries*. https://CRAN.R-project.org/package=gtrendsR. R package version 1.4.7.

McAuliffe, M. (2018). The link between migration and technology is not what you think. https://www.weforum.org/agenda/2018/12/social-media-is-casting-a-dark-shadow-over-migration/. Accessed 28/01/2019.

Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H., and Kumar, S. (2011). Google Correlate Whitepaper. *Google* https://static.googleusercontent.com/media/research.google.com/ro//pubs/archive/41695.pdf. Accessed 09/09/2021.

Önder, I. and Gunter, U. (2016). Forecasting tourism demand with Google Trends for a major European city destination. *Tourism Analysis* 21(2-3): 203–220. doi:10.3727/108354216X14559233984773.

Palotti, J., Adler, N., Morales-Guzman, A., Villaveces, J., Sekara, V., Garcia Herranz, M., Al-Asad, M., and Weber, I. (2020). Monitoring of the Venezuelan exodus through Facebook's advertising platform. *Plos one* 15(2): e0229175. doi:https://doi.org/10.1371/journal.pone.0229175.

Parkins, N.C. (2010). Push and pull factors of migration. *American Review of Political Economy* 8(2): 6–24. https://www.proquest.com/docview/912208903. Accessed 14/09/2021.

Popa, D. (2019). Legea recunoştinţei între generaţii: În ce țări mai există și cum funcționează. https://bit.ly/36jlKQ0. Accessed 21/11/2020.

Preis, T., Moat, H.S., and Stanley, H.E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports* 3(1). doi:10.1038/srep01684.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Accessed 09/09/2021.

Rango, M. and Vespe, M. (2017). Big Data and alternative data sources on migration: From case-studies to policy support. Summary report, European Commission - Joint Research Centre (JRC). https://knowledge4policy.ec.europa.eu/sites/default/files/BD4M-workshop-2017-summary-report.pdf. Accessed 14/09/2021.

Raymer, J., Rees, P., and Blake, A. (2015). Frameworks for guiding the development and improvement of population statistics in the United Kingdom. *Journal of Official Statistics* 31(4): 699–722. doi:10.1515/JOS-2015-0041.

Raymer, J., Wiśniowski, A., Forster, J.J., Smith, P.W., and Bijak, J. (2013). Integrated modeling of European migration. *Journal of the American Statistical Association* 108(503): 801–819. doi:10.1080/01621459.2013.789435.

Righi, A. (2019). Assessing migration through social media: A review. *Mathematical Population Studies* 26(2): 80–91. doi:10.1080/08898480.2019.1565271.

Sagiroglu, S. and Sinanc, D. (2013). Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)* doi:10.1109/cts.2013.6567202.

Sarigul, S. and Rui, H. (2014). Nowcasting obesity in the US using Google search volume data. Tech. Rep. 327-2016-12719, Western Coordinating Committee on Agribusiness. doi:10.22004/ag.econ.166113. http://ageconsearch.umn.edu/record/166113.

Sides, J. and Citrin, J. (2007). European Opinion about immigration: The role of identities, interests and information. *British Journal of Political Science* 37(3): 477–504. doi:10.1017/s0007123407000257.

Siliverstovs, B. and Wochner, D.S. (2018). Google Trends and reality: Do the proportions match?: Appraising the informational value of online search behavior: Evidence from Swiss tourism regions. *Journal of Economic Behavior & Organization* 145: 1–23. doi:10.1016/j.jebo.2017.10.011.

Sjaastad, L.A. (1962). The costs and returns of human migration. *Journal of Political Economy* 70(5, Part 2): 80–93. doi:10.1086/258726.

Sredanovic, D. (2020). Brexit as a trigger and an obstacle to onwards and return migration. *International Migration* Early View. doi:10.1111/imig.12712.

Stan Development Team (2020). Stan modeling language users guide and reference manual, 2.25. `https://mc-stan.org`. Accessed 09/09/2021.

StatCounter (2019a). Search Engine Market Share Romania - July 2019. `https://gs.statcounter.com/search-engine-market-share/all/romania`. Accessed 10/12/2019.

StatCounter (2019b). Search Engine Market Share Worldwide - July 2019. `https://gs.statcounter.com/search-engine-market-share`. Accessed 10/12/2019.

Taylor, L. (2016). The ethics of big data as a public good: Which public? Whose good? *SSRN Electronic Journal* doi:10.2139/ssrn.2820580.

Thorvaldsen, G. (2019). *Censuses and census takers: A global history*. Oxon: Routledge.

Tirosh, N. and Schejter, A. (2017). “Information is like your daily bread”: The role of media and telecommunications in the life of refugees in Israel. *Hagira—Israel Journal of Migration* 7: 1–25. `http://tiny.cc/0k5juz`. Accessed 10/01/2021.

United Nations (2014). Estimating migration flows using online search data - UN Global Pulse. Global Pulse Project Series No. 4, 2014. `https://www.unglobalpulse.org/wp-content/uploads/2014/04/UNGP_ProjectSeries_Search_Migration_2014_0.pdf`. Accessed 09/09/2021.

Vargas-Silva, C. and Rienzo, C. (2020). Migrants in the UK: an overview. *Migration Observatory briefing, COMPAS, University of Oxford* `http://tiny.cc/8k5juz`. Accessed 09/12/2020.

Vlastakis, N. and Markellos, R.N. (2012). Information demand and stock market volatility. *SSRN Electronic Journal* doi:10.2139/ssrn.1558434.

Wanner, P. (2020). How well can we estimate immigration trends using Google data? *Quality & Quantity* 55: 1181–1202. doi:10.1007/s11135-020-01047-w.

Wilde, J., Chen, W., and Lohmann, S. (2020). COVID-19 and the future of US fertility: What can we learn from Google? *IZA Discussion Paper Series* No. 13776.

doi:10.31235/osf.io/2bgqs.

Willekens, F. (1994). Monitoring international migration flows in Europe. *European Journal of Population/Revue européenne de Démographie* 10(1): 1–42.

Willekens, F. (2019). Evidence-based monitoring of international migration flows in Europe. *Journal of Official Statistics* 35(1): 231–277. doi:10.2478/jos-2019-0011.

Willekens, F., Massey, D., Raymer, J., and Beauchemin, C. (2016). International migration under the microscope. *Science* 352(6288): 897–899. doi:10.1126/science.aaf6545.

Wladyka, D. (2017). Queries to Google search as predictors of migration flows from Latin America to Spain. *Journal of Population and Social Studies* 25(4): 312–327. doi:10.25133/jpssv25n4.002.

Yu, L., Zhao, Y., Tang, L., and Yang, Z. (2019). Online big data-driven oil consumption forecasting with Google Trends. *International Journal of Forecasting* 35(1): 213–223. doi:10.1016/j.ijforecast.2017.11.005.

Zagheni, E. and Weber, I. (2012). You are where you e-mail. *Proceedings of the 3rd Annual ACM Web Science Conference on - WebSci 12* doi:10.1145/2380718.2380764.

Zagheni, E., Weber, I., and Gummadi, K. (2017). Leveraging Facebook advertising platform to monitor stocks of migrants. *Population and Development Review* 43(4): 721–734. doi:10.1111/padr.12102.

Appendix

Table A-1: International Passenger Survey estimates of immigration of Romanian-born immigrants to the United Kingdom

year	estimate	±95% CI	SE	SE/estimate %
2004	3.3	2.6	1.3	40.2
2005	0.5	0.6	0.3	61.2
2006	2.2	2.2	1.1	51.0
2007	2.1	2.4	1.2	58.3
2008	7.8	4.8	2.4	31.4
2009	10.5	4.5	2.3	21.9
2010	6.8	2.9	1.5	21.8
2011	7.7	3.1	1.6	20.5
2012	6.4	3.0	1.5	23.9
2013	20.6	9.5	4.8	23.5
2014	37.8	9.3	4.7	12.6
2015	55.1	13.8	7.0	12.8
2016	54.9	12.2	6.2	11.3
2017	51.6	13.4	6.8	13.2
2018	26.9	8.8	4.5	16.7
2019	32.6	10.5	5.4	16.4

Note: “Estimate” denotes a point estimate in thousands of persons, CI stands for confidence interval, and SE is a resulting standard error.

Table A-2: Google search keywords used to construct Google Trends Index with clusters and Pearson's correlation coefficient ρ

Keyword	2012	2013	2014	2015	2016	2017	2018	2019	ρ
GTI	19.8	19.9	24.8	32.1	35.1	30.4	29.6	30.9	0.74
CURRENCY CLUSTER									
evolutie curs lira sterlina	7.5	3.3	0.0	5.6	16.6	11.2	5.8	7.7	0.59
curs valutar lira	10.8	11.4	13.9	18.7	45.5	36.2	27.7	32.4	0.51
curs valutar lira sterlina	17.8	18.2	15.2	30.3	57.8	45.8	27.4	32.8	0.69
currency UK	3.6	8.9	25.3	46.6	19.1	5.7	10.9	20.5	0.46
curs lira sterlina	17.8	17.5	20.8	27.9	58.3	42.5	28.2	40.1	0.62
cat e lira	11.6	9.1	19.9	23.2	52.8	35.4	34.2	57.9	0.31
bani UK	13.3	17.9	6.4	21.8	24.9	22.0	35.0	20.2	-0.01
bani anglia	24.8	14.3	19.5	37.2	37.8	31.6	40.4	26.2	0.55
British pound	6.6	5.8	9.3	7.7	25.3	9.4	9.1	7.8	0.53
1 lira	23.7	23.0	29.7	30.8	49.6	46.2	52.9	76.1	0.00
1 pound	35.0	29.3	32.7	40.2	46.8	41.2	43.2	43.5	0.51
one pound	18.8	19.6	10.9	31.2	21.2	48.6	27.4	19.4	0.45
o lira	18.3	29.8	25.4	41.8	56.8	45.9	46.6	63.3	0.28
lira sterlina	14.7	17.1	17.9	26.9	50.3	42.4	35.0	42.2	0.48
money UK	29.1	35.6	30.2	41.3	34.3	25.7	36.0	57.3	-0.23
lei to pound	4.2	8.9	5.7	9.9	29.0	18.8	25.4	38.1	0.01
lei to GBP	4.5	11.8	16.2	28.2	33.0	48.6	42.2	54.8	0.20
GBP to ron	11.2	16.1	22.8	30.9	42.8	45.6	48.7	50.2	0.25
GBP to lei	7.7	9.7	12.6	26.8	48.2	46.4	42.4	42.0	0.44
salary UK	21.2	19.5	37.8	41.9	37.2	47.9	43.6	39.8	0.57
salariu Anglia	6.6	8.1	16.3	34.9	35.7	24.1	14.8	21.0	0.93
salariu UK	8.0	15.6	26.8	26.6	4.2	20.4	13.9	19.5	0.03
quid	41.4	25.4	21.1	38.2	34.6	31.3	46.0	46.2	-0.01
ron to pound	7.5	10.0	13.8	11.8	35.8	32.3	23.4	32.0	0.36
ron to GBP	10.6	17.7	28.2	40.5	55.8	53.0	46.3	58.1	0.50
pound to lei	6.8	7.9	5.2	9.7	21.5	19.9	29.6	34.0	-0.10
pret lira	22.7	16.1	20.8	31.3	46.4	22.7	36.9	36.8	0.36
pound to ron	7.5	10.0	13.8	11.8	35.8	32.3	23.4	32.0	0.36
pound sterling	8.7	5.7	12.8	22.8	18.2	21.1	29.0	23.2	0.27
HOUSING CLUSTER									
house England	28.4	7.9	7.8	11.2	30.8	22.3	21.8	2.5	0.49
casa Anglia	28.2	49.0	44.8	35.3	38.5	27.1	21.0	21.6	-0.05
residence UK	3.9	6.4	9.0	8.0	14.4	27.8	27.6	23.7	-0.02
casa UK	25.4	29.9	27.5	30.6	21.6	20.4	26.1	23.5	-0.37
house UK	27.1	30.9	39.0	38.8	29.1	34.5	28.8	33.5	0.35
rent UK	11.4	20.6	32.8	39.8	28.8	20.7	25.3	21.6	0.52
EMPLOYMENT CLUSTER									
munca UK	28.5	29.6	40.0	53.8	38.2	27.4	26.2	19.2	0.59
munca Anglia	25.3	26.8	49.4	58.1	46.7	29.0	28.2	20.6	0.65
work UK	37.6	31.9	45.5	43.3	45.8	21.8	27.8	14.2	0.39
locuri de munca UK	21.7	22.5	43.0	56.6	32.2	26.7	22.9	21.2	0.60
locuri de munca Anglia	24.2	25.0	45.4	58.9	46.7	25.6	22.8	21.1	0.68
jobs UK	49.9	43.0	68.9	77.2	52.8	46.0	37.8	32.9	0.57
jobs England	34.8	25.4	35.8	30.2	22.5	17.9	17.9	20.3	0.10
EDUCATION CLUSTER									
study UK	23.6	19.7	30.4	29.0	29.4	37.3	12.7	19.5	0.79
universities UK	36.6	30.3	36.0	45.2	37.1	30.7	36.0	22.1	0.50
school UK	33.8	42.6	38.5	51.1	37.3	42.8	45.2	41.7	0.09
student UK	19.2	27.4	24.5	25.7	23.2	20.3	28.8	17.9	-0.35
school England	28.2	15.7	12.3	22.2	16.3	9.6	20.3	8.8	0.08
education UK	39.5	28.9	31.1	27.7	19.4	15.4	19.2	23.2	-0.31
CONTROL CLUSTER									
tree	63.3	50.6	53.4	51.8	49.8	47.8	47.7	49.1	0.09
sport	49.3	56.2	59.0	63.2	77.0	81.1	77.2	73.3	0.36
Spain	51.8	44.2	50.8	46.2	52.2	59.5	61.0	57.2	0.01
rocket	15.7	15.9	17.2	27.8	26.9	31.3	25.7	20.7	0.74
mother	79.8	82.4	77.0	68.0	63.7	58.7	56.5	51.2	-0.21
horse	48.6	37.4	57.7	37.7	32.7	28.8	26.5	30.8	-0.02
Greece	35.2	34.3	37.7	41.1	34.6	41.4	43.2	39.9	0.08
money	59.8	50.1	50.8	52.1	50.9	49.2	51.5	53.1	-0.09
grass	52.8	45.7	48.8	45.9	46.2	41.7	44.8	53.5	-0.28
flower	69.0	68.2	69.3	66.8	69.6	71.2	64.1	66.2	0.48

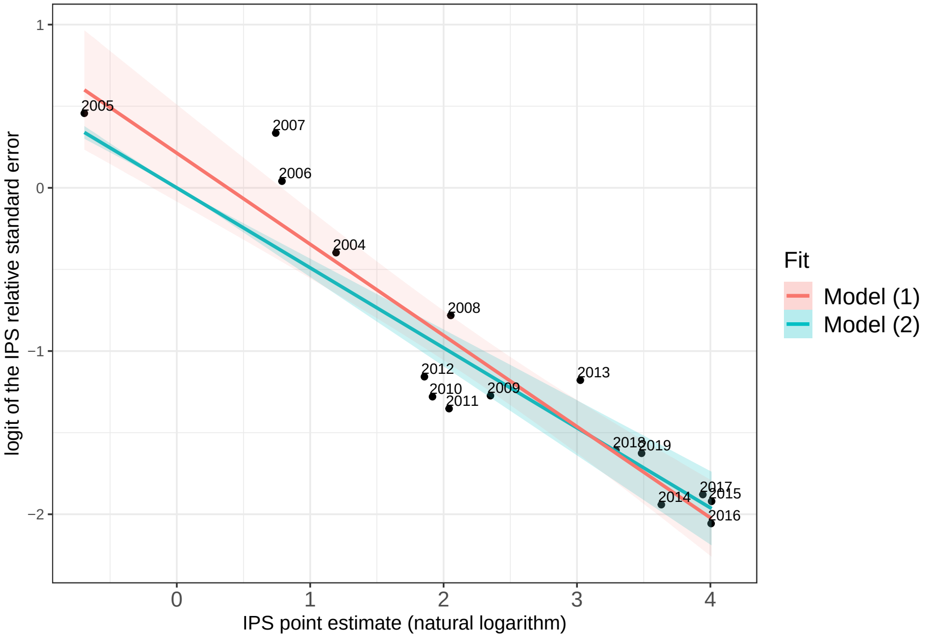
Source: Own calculations using Google Trends data.

Table A-3: Pearson's correlation coefficients ρ between the IPS data and the GTI clusters constructed using various monthly lags

lag	employment	education	currency	housing	control	all
0	0.55	0.32	0.53	0.55	0.58	0.76
1	0.58	0.62	0.51	0.50	0.58	0.74
2	0.60	0.79	0.49	0.60	0.55	0.72
3	0.67	0.87	0.48	0.64	0.44	0.73
4	0.71	0.92	0.44	0.63	0.40	0.71
5	0.74	0.96	0.38	0.63	0.35	0.68
6	0.70	0.95	0.36	0.83	0.37	0.65
7	0.72	0.98	0.32	0.78	0.27	0.61
8	0.74	0.94	0.31	0.69	-0.03	0.59
9	0.76	0.91	0.27	0.76	-0.12	0.54
10	0.76	0.95	0.27	0.81	-0.16	0.54
11	0.80	0.70	0.27	0.88	-0.14	0.52
12	0.77	0.27	0.26	0.86	-0.28	0.47

Note: Cluster “all” denotes average of all clusters except “control”.
Source: Own calculations using Google Trends data.

Figure A-1: Underdispersion in immigration data



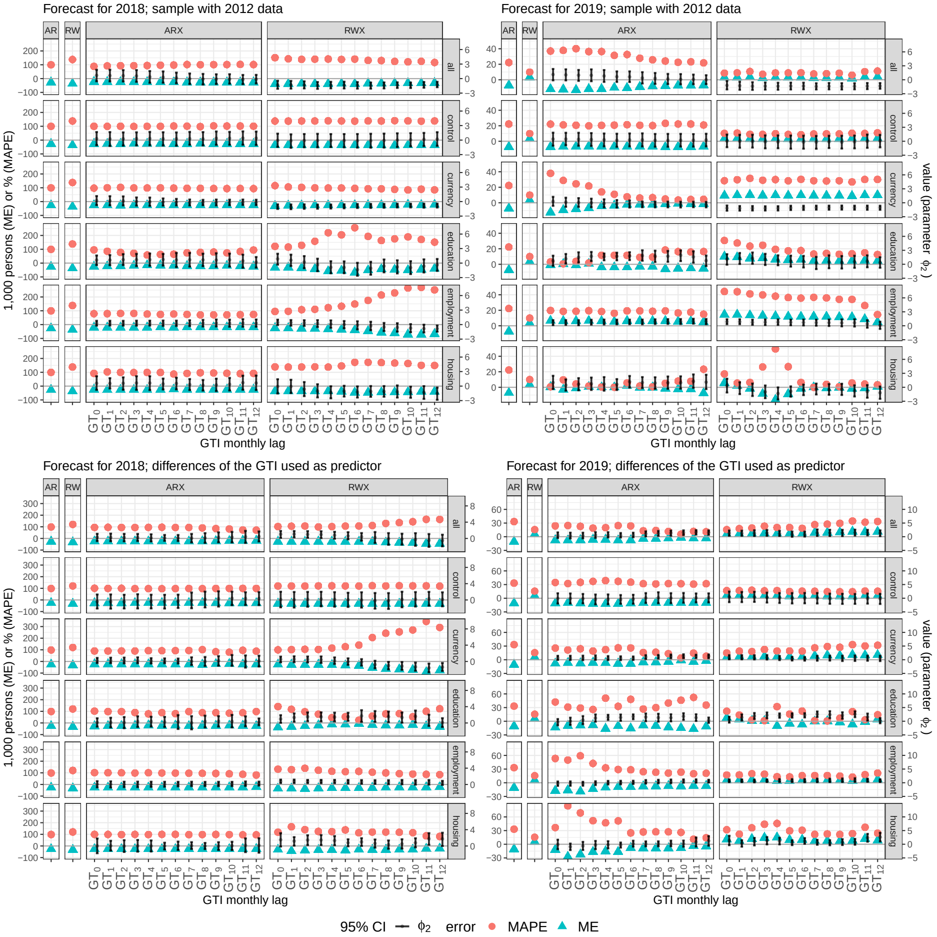
Source: Own calculations using International Passenger Survey (IPS) data.

Table A-4: Results of regressing the logit of the IPS relative standard errors (RSE_{IPS}) on the natural logarithm of IPS point estimates ($\log(IPS)$)

	Dependent variable	
	$logit(RSE_{IPS})$	
	Model (A)	Model (B)
$\log(IPS)$	-0.559 (0.051) t = -10.905 $p < 0.001$	-0.490 (0.026) t = -18.556 $p < 0.001$
Constant	0.213 (0.139) t = 1.537 $p = 0.147$	
Observations	16	16
R^2	0.895	0.958
Adjusted R^2	0.887	0.955
Residual Std. Error	0.274 (df = 14)	0.286 (df = 15)
F Statistic	118.9 (df = 1; 14) $p < 0.001$	344.3 (df = 1; 15) $p < 0.001$

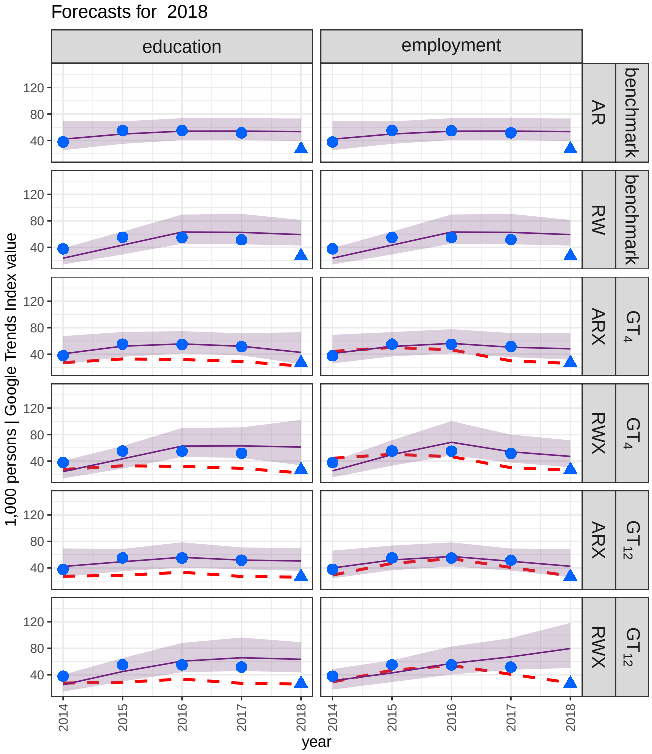
Note: The model relies on the assumption that the RSE is never larger than 100%, which is realistic for the data at hand. Model (A) is a model with a constant that does not differ from zero ($p = 0.147$); Model (B) has the constant fixed at zero, which implies that for the IPS estimate of 1,000 immigrants, the RSE is 50%.
Source: Own calculations using International Passenger Survey data, 2004–2019.

Figure A-2: Forecast errors and 95% credible intervals for 2018 and 2019, for benchmark models (AR and RW) and corresponding models with Google Trends Index (GTI) predictors (ARX and RWX)



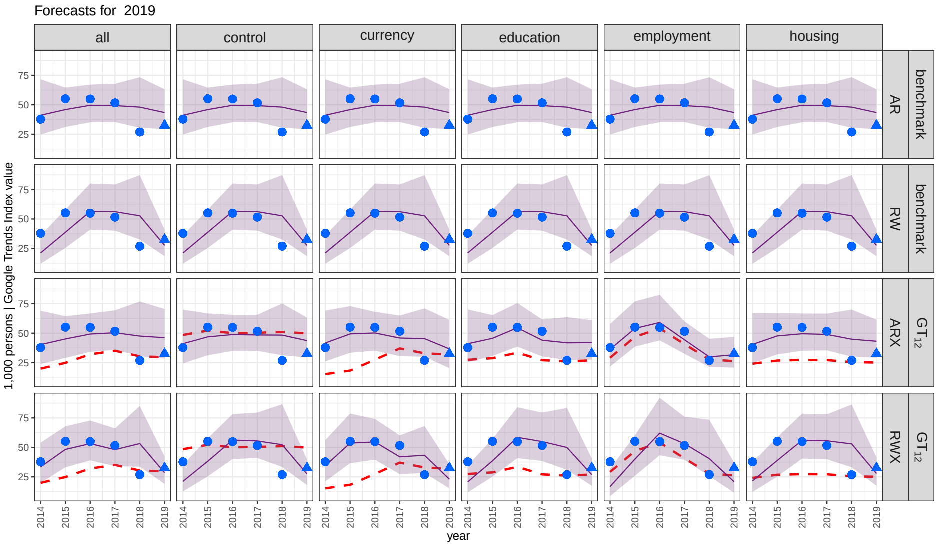
Note: Subscripts at GT denote a monthly lag in calculating the GTI variable. For example, for 2019, GT_0 refers to GTI calculated as an average over January–December 2019; GT_2 is an average over November 2018–October 2019.

Figure A-3: Forecasts of Romanian-born migration flows to the UK based on 2013–2017 data



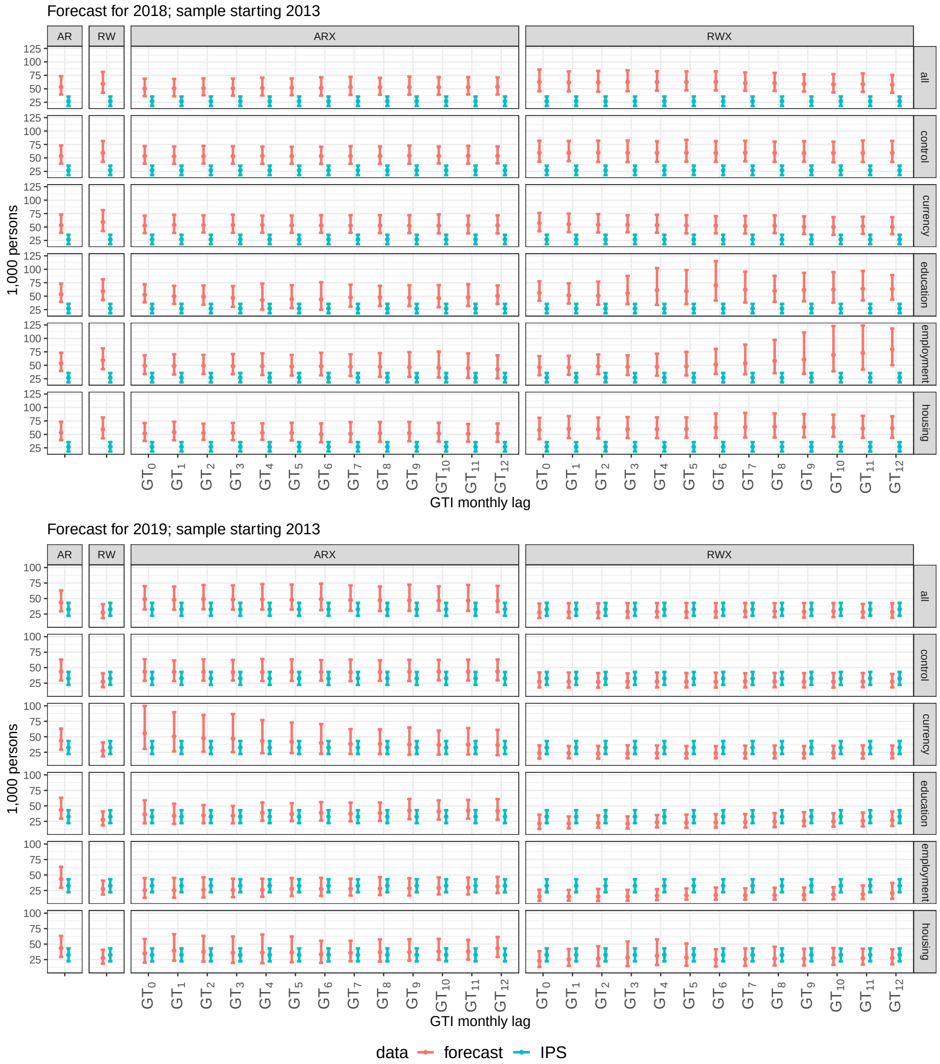
Note: Large blue dots denote observed data; triangles are observations that are being predicted by the models; dashed red lines are GTIs; lines and shading are posterior predictive medians and 95% predictive intervals, respectively; ARX and RWX denote models with a GTI predictor.
Source: Own calculations using International Passenger Survey and Google Trends data.

Figure A-4: Forecasts of Romanian-born migration flows to the UK based on 2013–2018 data



Note: Large blue dots denote observed data; triangles are observations being predicted by the models; dashed red lines are GTIs; solid lines and shading are posterior predictive medians and 95% predictive intervals, respectively.
Source: Own calculations using International Passenger Survey and Google Trends data.

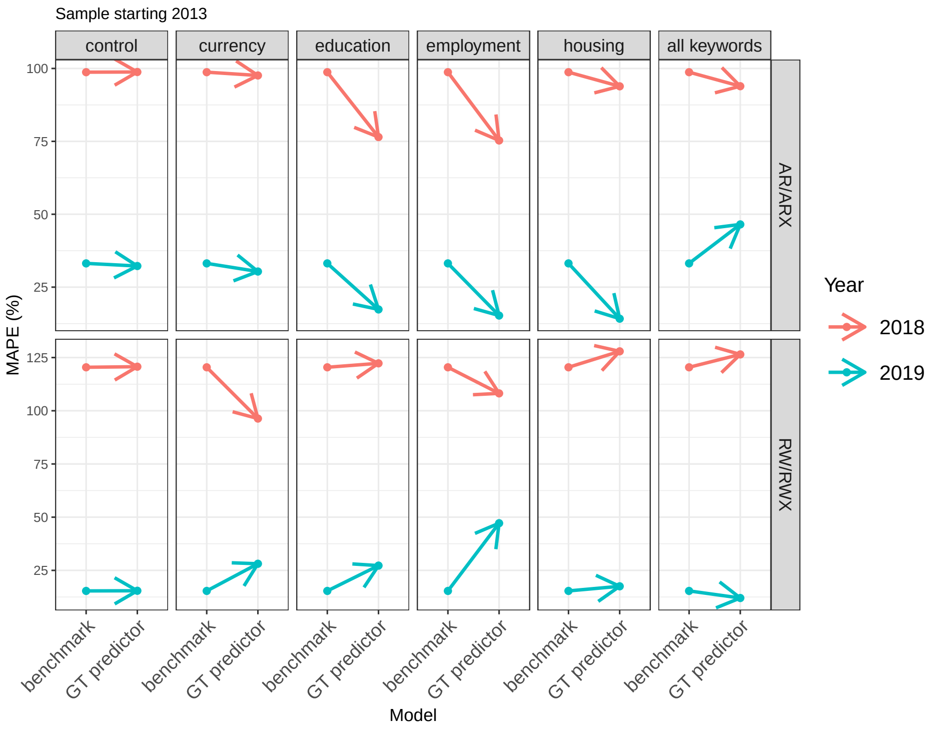
Figure A-5: Romanian-born migration flows: a comparison of means and 95% confidence intervals from the International Passenger Survey and medians with 95% predictive intervals for 2018 and 2019



Note: AR and RW denote benchmark models, and ARX and RWX are the corresponding models with Google Trends Index (GTI) predictors; subscripts at GT denote a monthly lag in calculating the GTI variable. For example, for 2019, GT_0 refers to GTI calculated as an average over January–December 2019 and GT_2 is an average over November 2018–October 2019.

Source: own calculations using International Passenger Survey and Google Trends data.

Figure A-6: Mean Absolute Percentage Errors (MAPE) for forecasts based on benchmark models (AR and RW) and corresponding models with Google Trends Index (GTI) keyword-specific predictors (ARX and RWX) averaged over all monthly lags



Source: Own calculations using International Passenger Survey and Google Trends data for 2013–2018.