DOI: 10.15439/2023F3141
ISSN 2300-5963 ACSIS, Vol. 37

Forecasting migration of EU citizens to Germany using Google Trends

Nicholas Steinbrink
0000-0002-6163-0212
Bertelsmann Stiftung
Carl-Bertelsmann-Str. 256, 33335 Gütersloh, Germany
Email: nicholas.steinbrink@bertelsmann-stiftung.de

***Abstract*—The study examines the potential of Google Trends data as an additional data source for forecasting EU migration to Germany. For that aim, candidate search queries with relation to migration intent are proposed. The resulting Google Trends Indices (GTI) are combined with macroeconomic and past migration data and used to build a machine learning regression model. It is shown that GTI predictors can moderately reduce the forecast error and enable a slight expansion of the forecast horizon. However, the presence of outliers emphasizes the need for continuous improvement in data quality to increase the robustness of the approach.**

I. Introduction

MIGRATION policy plays a pivotal role in shaping the labor market policies of OECD countries, offering potential solutions to mitigate labor market rigidities. Notably, in the context of Germany, the labor market has greatly benefited from the free internal mobility of EU nationals, which serves as a crucial source of skilled labor migration. However, the effectiveness of migration policy faces a central challenge arising from the uncertainty surrounding the goals and scale of future migration. This uncertainty is driven by a diverse array of political and socio-economic push and pull factors. Although intra-EU labor mobility is subject to regular monitoring [1], there is currently a lack of substantial efforts towards forecasting, despite the occurrence of significant mobility shifts in the past, such as those witnessed in the aftermath of the Eurozone financial crisis.

Regarding external migration, forecasts primarily rely on three methodologies: (a) extrapolation of past migration patterns using time series methods such as ARIMA models, (b) explanatory econometric models incorporating variables like GDP and unemployment, which are presumed to be linked to migration, or (c) spatial interaction models like gravity models, connecting origins and destinations. [2] These models often integrate expert opinions within a Bayesian framework. Despite their methodological sophistication, these approaches often exhibit considerable forecast uncertainties, leading to potential over- or underestimation of actual migration.

An innovative approach to migration forecasting involves focusing on the individual planning behaviors of individuals who have made the decision to migrate, as opposed to relying solely on macroscopic factors. This can be achieved, for example, by incorporating data from migration intention surveys into the forecasting process [3]. A promising alternative lies in leveraging digital trace data, which has the potential to identify individual migration intentions earlier by capturing active behaviors of individuals seeking information about emigration and migration planning. One suitable data source for this purpose is Google Trends, which measures the temporal and regional search intensity associated with specific keywords in the Google search engine, thanks to Google's high market share.

The aim of this project is to explore the applicability of a Google Trends as a predictor in a novel forecast of migration for EU nationals to Germany. Specifically, the research seeks to determine whether Google Trends data can enhance the accuracy of a forecast method based purely on past migration patterns and macroeconomic variables, particularly within a short- to medium-term timeframe (3 to 12 months).

II. Related work

Although no specific work regarding intra-EU mobility can be found, various attempts have been made to utilize digital trace data for migration forecasts. Data sources include for example Facebook's advertising platform [4], geolocalized IP addresses from e-mails [5], as well as Twitter messages [6], [7]. Moreover, the potential of Google Trends data as a predictive data source for migration has been examined in previous studies. Boehme et al. [8] demonstrated the correlation between search activity related to migration and migration intention, as well as between migration intention and successful migration, establishing the viability of using search engine data for predictive purposes. Carammia et al. [9] have developed an early warning and forecast system based on data from Google Trends, applied to monthly asylum data for EU destination countries. Closely related to that, ongoing research by Boss et al. [10] highlights the particular usefulness of Google Trends data for forecasting bilateral refugee flows at scale across multiple corridors. In contrast, Wanner [11] presented mixed results using a minimal model based on a single Google Trends keyword as predictor for work-related regular migration from EU countries to Switzerland, emphasizing the need for further validation and research.

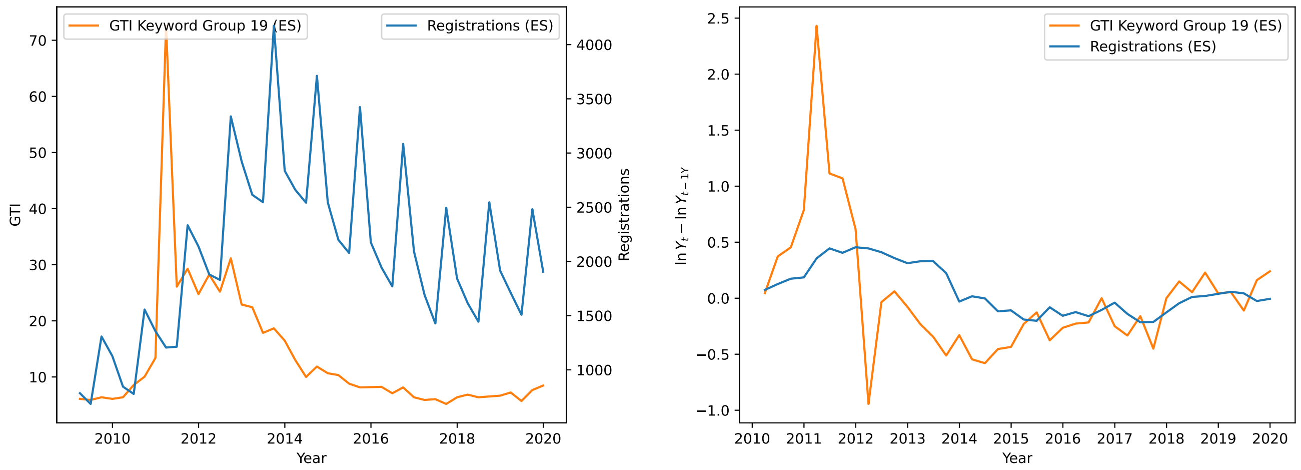

Fig. 1. Left panel: Number of registrations (blue) and GTI for keyword group 19 (related to work and jobs in Germany) for the example of Spain. The peak during 2011 coincides with a period of high unemployment in Spain during the Eurozone financial crisis and the search for jobs in Germany has possibly gone "viral" for a short period of time. Right panel: transformed values of registrations and GTI according to (1).

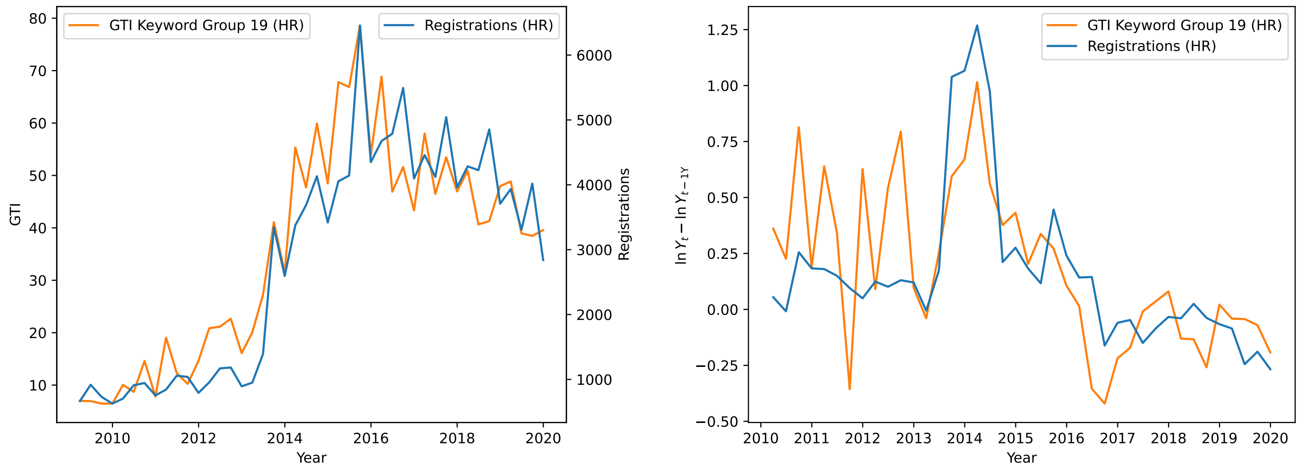

Fig. 2. Left panel: Number of registrations (blue) and GTI for keyword group 19 (related to work and jobs in Germany) for the example of Croatia. Registrations increase continuously after the admission of Croatia to the EU in July 2013 and the full access to free movement to Germany in July 2015, with the GTI following the trend. Right panel: transformed values of registrations and GTI according to (1).

III. Methods

A. Prediction target

Since EU nationals are required to register within 3 months after their arrival in Germany, the official number of registrations by country of origin, which is provided by the federal office of statistics (DESTATIS), is taken as a proxy of the number of arrivals. The number of registrations exhibits a clear seasonal pattern with a peak during the summer months and a drop during the winter months consistently for all EU origin countries (for an example, see figs. 1 and 2, left panels, blue line). A naive forecast could in most situations produce reasonable results by taking the number of registrations from the same period of the previous year as prediction. Therefore, the interesting quantity to predict is not the absolute number of registrations, but the change compared the previous year. The target variable is then stipulated as:

$$Y_{t,c} = \ln R_{t,c} - \ln R_{t-1\mathrm{y},c}, \tag{1}$$

where $R_{t,c}$ is the number of registrations at time t of nationals with country of origin c and $R_{t-1\mathrm{y},c}$ the corresponding number lagged by one year. The log transformation ensures that all countries are given equal weight regardless of the absolute number of registrations if the forecast accuracy is determined by conventional metrics. For an example, see figs. 1 and 2 (right panels, blue line).

B. Countries of origin

The selected countries of origin encompass the 28 member countries of the EU prior to the Brexit, except Malta, Cyprus and Germany, which is the destination country. Additionally, Switzerland has been added due to the shared border with Germany and the membership in the Schengen area. Countries with only small number of registrations have been grouped according to similar migration behavior, similar size and regional proximity. These groups comprise: Austria and Switzerland; Belgium, the Netherlands and Luxembourg; the Czech Republic and Slovakia; Lithuania, Latvia and Estonia; Sweden, Finland and Denmark.

C. *Google Trends index*

The Google Trends index (GTI) measures the relative search interest, given a certain keyword, over a specified amount of time. The index is based on a sample of total search queries, and is normalized in such a way that the maximum value for a given index time series is set to 100. If the sample becomes too small, no data is returned by Google, which results in a flat zero response when using the API. No detailed information is specified by Google regarding the sampling process or the cutoff.

The keywords of interest have been taken from [8] and been professionally translated to the languages of the countries of origin. A crucial challenge in keyword selection is to achieve an optimal balance between specificity and generality. The keywords must be specific enough to avoid any confusion with search activity unrelated to migration. However, they should not be excessively specific, as this is essential to ensure sufficient data quality and comprehensive coverage of relevant search activity. Especially for smaller countries, the Google Trends API requests often resulted in either a flat zero response when a keyword lacked sufficient search frequency, or an overly noisy and unusable response. To balance both aims, the Google Trends queries have been enriched in the following way:

- Semantically related keywords have been grouped.
- Each keyword is considered in the language(s) of the country of origin, as well as German and English.
- To every keyword, the postfix "Germany" is added in the corresponding language.
- Multiple spellings (for instance with and without accents) have been considered.
- The keywords in one semantic group have been connected with a logical "or" (+).

An exemplary query is given in appendix V-B, while an example for a resulting GTI time series can be found in fig. 1 (left panel, orange line). Additionally, a range of keywords without the postfix "Germany" has been included as well to consider push-effects. In total, GTI time series of 48 keyword group candidates have been generated that way (see appendix V-A, table II for a complete list). To mitigate statistical variance, each query has been performed ten times and averaged.[1]

D. *Data preparation and modeling*

All time series have been discretized by 3-month intervals.[2] Data have been taken from 2010 to 2019. The lower limit has been set due to insufficient data quality of Google Trends prior to 2010 while the upper limit was set to avoid effects of the COVID-19 pandemic.[3] In addition to Google Trends, GDP per capita and unemployment of the countries of origins have been selected as explanatory variables. Google Trends indices, GDP and unemployment have as well been transformed according to (1) (for an example of a transformed GTI, see fig. 1, right panel, orange line). The full set of features is then given by lagged values of Google Trends indices, GDP, unemployment and autoregressive lags of the numbers of registrations themselves. Only for Google Trends a minimum lag of 3 months has been used, while for the other variables the minimum lag was 6 months due to the publication delay, which forbids smaller lags in a forecast situation.

An array of both linear and ensemble-based machine learning models has been tested with different feature-sets.[4] The choice of models is guided by similar motivations as in [10]: no prior theoretical knowledge is utilized, and the algorithms are suitable for a combined model across a multitude of origin countries with a relatively large number of features. To reduce dimensionality and mitigate multicollinearity, a feature selection step has been added, which is described in section III-E.

For each configuration, a mean out-of-sample R^2, given by $R^2_{OOS} = 1 - \sum(y_i - \hat{y}_i)^2 / \sum y_i^2$ [12], and MAE have been determined via n-fold cross-validation (CV). Each year corresponds to a CV fold, while the year 2019 has been additionally set apart as hold-out set for a sanity-check of the CV results. While in principle future information is used as training data in this CV scheme, it has been shown [13] that such a method is valid as long as residuals are uncorrelated, which is typically only the case in severe underfitting. The advantage, on the other hand, is a maximum use of the available training data and comparability across folds in contrast to time-series CV methods.

As there are no known comparable forecasts for intra-EU mobility, a range of benchmarks has been chosen to assess the forecast accuracy. The const(0) benchmark is the simplest baseline, setting the target variable constantly zero, $\hat{Y}_{t,c} = 0$, corresponding to no annual change in the registration rate. A model performing below the const(0) benchmark corresponds to $R^2_{OOS} < 0$.

[1] To force Google Trends to draw a new sample for each request, a random, sufficiently long character string can be added to each query.

[2] A finer discretization has been tested as well, but did not lead to any significant improvement of the forecast accuracy in the modeling stage.

[3] While the German border has been officially closed only for a few months, it is reasonable to assume that individual mobility has been reduced for a longer period of time during the pandemic. As a stable relationship between online search activity and registrations is a necessary assumption of the methodology, the cutoff has been set to 2020 for this principle study.

[4] The tested ensemble models include: Random Forest, XGBoost and AdaBoost. The tested linear models include: OLS, ElasticNet, Bayesian Ridge Regression, Automatic Relevance Determination, Huber Regression and Theil-Sen Regression.

The benchmark previous(1) corresponds to a random walk, where the previous lag of the target variable is set as predictor, $\hat{Y}_{t,c} = Y_{t-1,c}$. This is not a realistic scenario due to the aforementioned publication delay of registration data. Therefore, the benchmark previous(2), $\hat{Y}_{t,c} = Y_{t-2,c}$, provides a realistic benchmark based on previous lags of the target variable.

Eventually, a non-naive realistic benchmark is given by a comparison of the best models with and without GTI variables, which accounts for the benefit of adding Google Trends data itself.

E. Feature selection

In total, a maximum number of features of roughly $P \sim 51 \cdot L$, where L denotes the number of lags used for the prediction, are available. The maximum lag has been set to $L = 8$ to account for possible correlations between search activity and immigration for up to 24 months. In effect, we end up with a relatively large number of features compared to the number of data points of $N = 576$. While some models with internal variable selection mechanisms are in principle able to handle large-P-small-N-problems, a reduction of the feature dimensionality is nonetheless advisable to maximize model performance and increase interpretability.

In a pre-selection step, the complete set of features is filtered to ensure that only features which are sufficiently correlated with the target are included. To that aim, the p value of the Spearman lag-correlation of all variables with the target is determined. Only those variables with a minimum p value below a cut-off given by a conservative Bonferroni correction, $p < 0.05/P$, are kept. This conservative limit both minimizes the likelihood of spurious correlation and maximizes the robustness of the subsequent feature selection step. The remaining 10 GTI variables include both pull- and push-related queries and are indicated in table II.

Finally, the selection of the input features for the regression model has been performed separately for the linear and ensemble models, respectively, as well as for feature configurations both including and excluding the GTI and autoregressive lags of the prediction target. To estimate the optimum feature sets for the linear models, a Sequential Forward Floating Selection (SFFS) has been performed with an ordinary least squares linear regression model, which has shown to be an efficient search technique [14]. The selected features have been checked for multicollinearity using the variable inflation factor (VIF). It could be shown that by the selection procedure multicollinearity could be reduced to a moderate level of VIF < 10. Since SFFS is a rather costly greedy method, for the ensemble models Recursive Feature Elimination (RFE) with a tuned Random Forest regressor has been chosen. For both methods, the CV scheme as outlined in section III-D with MAE as optimization metric has been used to determine the optimum number of features. For all feature configurations, a stable optimum could be found. An illustration of the variables and the selection procedure is shown in fig. 3.

IV. Results

A. Model performance

Table I shows a performance comparison of the best optimized models of each class (linear and ensemble) after hyperparameter tuning with different feature configurations (all features, without GTI, without autoregressive lags of the target variable), compared to naive benchmarks, as outlined in the previous section. For the ensemble models, a Random Forest has been found to perform best, while for the linear models, a linear regression with Huber loss has shown the optimum performance.

If the cross validation results of the ML models are compared to the benchmarks, it can be observed that most models to surpass the const(0) and the previous(2) benchmark, but only the linear model with all features is on par with the previous(1) benchmark. Since the latter is not realistic due to the publication delay of registration data, this observation implies that a real-time monitoring of registrations alone would be sufficient to provide a competitive short-term forecast.

If the models are compared among themselves, it can be seen that the linear model has a clear advantage over the Random Forest model, which is prone to overfitting due to the small number of training examples. For both model classes, it can be observed that the models with GTI exhibit slight but clear reduction of the MAE by up to 10 %, compared to the models without GTI. A mean R^2_{OOS} of up to 0.54 can be achieved. The models without autoregressive lags provide, while not being competitive, still a reasonable forecast accuracy on average and could in principle be used as an alternative if past registration data are not available.

Compared to the mean CV scores, the MAE for the 2019 holdout set are smaller. However, only for the linear models with included autoregressive lags, R^2 is noticeably above zero. A plausible explanation is that the trend in registrations (1) is largely flat in 2019, suggesting that Google Trends data provide additional predictive power only if there are shifts in registrations which can not be predicted by other variables.

B. Comparison by country of origin

Fig. 4 (left panel) shows strong heterogeneity regarding the model performance for the linear model by country of origin. It can be observed that the model produces largely reasonable forecasts in terms of R^2 especially for Southern European and some Eastern European countries. Coincidentally, most EU citizens moving to Germany are native to these regions and registrations from these countries have been subject to greater variability over the last decade. However, even for countries which perform well on average, some outliers are present, corresponding to periods for which the forecast performs poorly. If the forecast errors of the best model with and without GTI, respectively, are compared (right panel), it can be seen that for some countries the model benefits more clearly from the GTI predictors. These are especially Spain (see example, fig. 5, left panel), Portugal, Greece and Italy, which have been particularly hit by the Eurozone financial crisis of the early

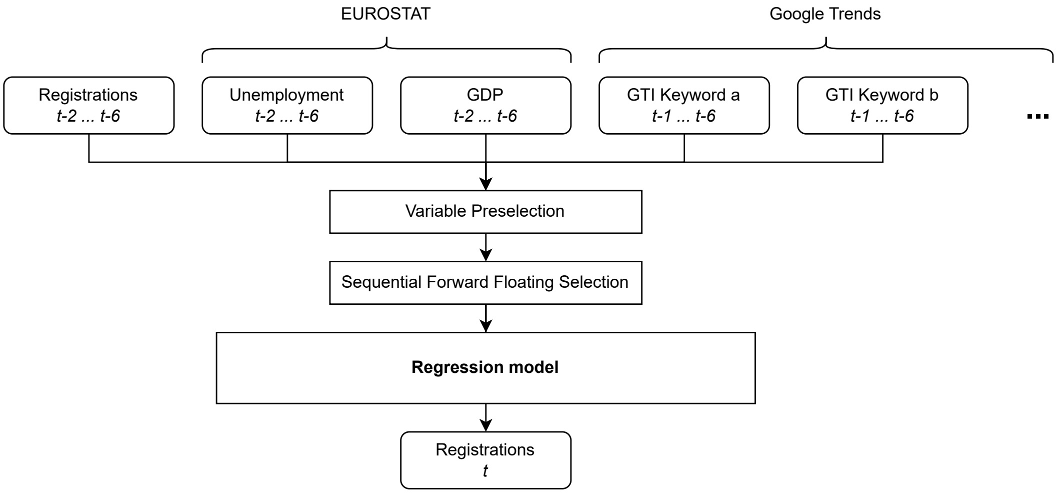

Fig. 3. Illustration of the input features and selection procedure.

TABLE I
Comparison of model performance metrics

Model	Feature Sets	Cross Validation		Holdout	
		MAE	R^2_{OOS}	MAE	R^2_{OOS}
Benchmark cost(0)	-	0.127 (0.025)	0.00 (0.00)	0.070	0.00
Benchmark previous(1)	-	0.070 (0.005)	0.51 (0.10)	0.061	0.22
Benchmark previous(2)	-	0.092 (0.007)	0.23 (0.11)	0.068	0.07
Random Forest	all	0.082 (0.025)	0.44 (0.07)	0.068	0.07
Random Forest	without autoregression	0.092 (0.011)	0.23 (0.17)	0.072	0.05
Random Forest	without GTI	0.089 (0.010)	0.38 (0.07)	0.068	-0.05
Linear Regression	all	0.071 (0.007)	0.54 (0.08)	0.063	0.30
Linear Regression	without autoregression	0.088 (0.010)	0.34 (0.11)	0.062	0.30
Linear Regression	without GTI	0.078 (0.009)	0.47 (0.09)	0.070	-0.00

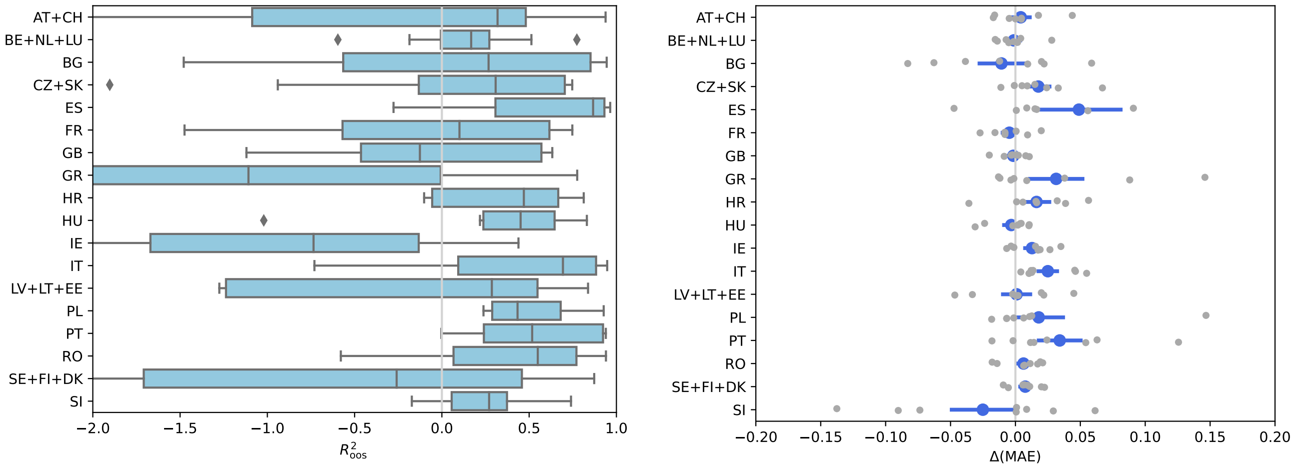

Fig. 4. Left panel: Distribution of cross validation out-of-sample R^2 by country of origin for linear model with autoregression and GTI. Right panel: Cross validation distribution of difference of mean absolute error between linear autoregressive model with and without GTI by country of origin. Positive values indicate lower prediction errors with GTI. Blue circles correspond to mean with standard errors.

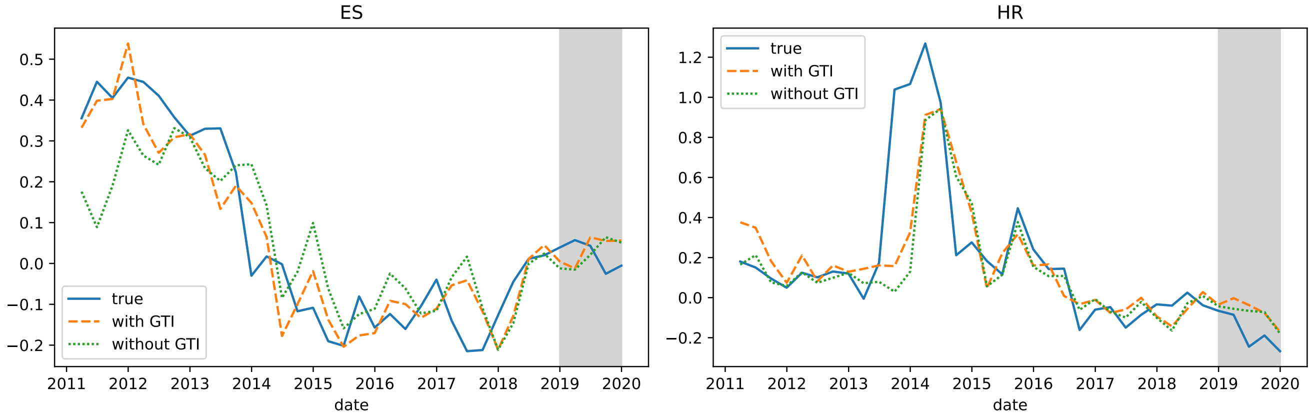

Fig. 5. Transformed number of registrations (1) (blue solid line), as well as prediction for the best linear model with GTI (orange dashed line) and without GTI (green dotted line) for Spain (left panel) and Croatia (right panel). The prediction is composed of the individual test sets of the cross validation folds. The shaded area corresponds to the holdout set of 2019.

2010s. In addition, a very subtle performance improvement can be observed for Poland, Croatia (see example, fig. 5, right panel), the Czech Republic, Slovakia and Ireland. For other countries, the benefit is negligible or even negative, such as in case of Bulgaria and Slovenia.

The existence of some forecast errors can partly be attributed to the data quality of the GTI predictors. In general, the lag correlation between GTI predictors and target is not stable in time, which can be observed in the examples in figs. 1 and 2. Especially for smaller countries, the GTI predictors can be noisy, which becomes even more pronounced in terms of relative changes, as after transformation (1). Moreover, while noise and outliers can be accommodated by using a diverse array of search queries, many of the corresponding GTI variables are unusable or missing (zero) for smaller countries. In the example of Croatia (fig. 5, right panel) this causes the forecast, for instance, to predict the sharp increase of registrations during 2013 and 2014 too late and to erroneously predict an increase in early 2011.

C. Forecast Horizons

Fig. 6 compares the best linear model with and without GTI, respectively, for different forecast horizons. Due to the machine learning setup, the results have been simulated by shifting the lags of all features $n-1$ periods to the past, with n denoting the number of forecast periods ahead, except for those features which would still be available at $t = t_0 - n$. For all forecast horizons from 3 to 12 months, the model with GTI predictors consistently outperforms the model without GTI moderately. Whereas for $n = 3$ and $n = 4$ the performance of the model without GTI becomes nearly indistinguishable from that of the const(0) baseline, the model with GTI exhibits at least some predictive power.

V. Conclusion

Google Trends can in principle be seen as a viable additional data source for a forecast of EU migration to Germany,

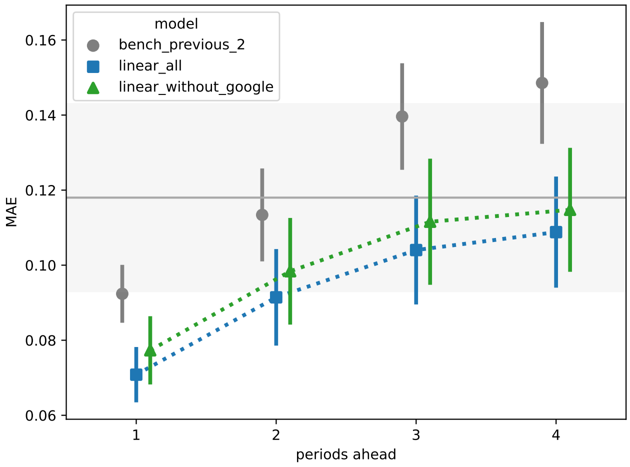

Fig. 6. Mean absolute error of the best model with GTI (blue) and without GTI (green), respectively, for forecast horizons of 1 to 4 periods corresponding to 3 to 12 months. The horizontal line represents the mean score for the const(0) benchmark with the shared area representing the standard error, while the grey circles represent the mean score for the previous(2) benchmark. All error bars are standard errors.

especially as long as a real-time monitoring of registrations is not available yet. For an existing forecast based on past registrations and macroeconomic variables, the addition of Google Trends data can on average reduce the forecast error moderately and enable some expansion of the forecast horizon.

The benefit of Google Trends data is especially given for scenarios and origin countries with greater shifts in migration behavior, where these can not always reliably predicted by other variables. For countries where the seasonality-adjusted migration to Germany is largely stationary, no improvement is gained by Google Trends, as naive forecasts are sufficient in these cases. For a small set of origin countries, however, the Google Trends based forecast performs weakly. Even for bigger countries the forecast occasionally produces outliers. The lack of robustness can be partly attributed to insufficient data quality of the GTI predictors. It is unclear if that is a purely statistical effect. Even if multiple samples are drawn from Google Trends and averaged, the resulting time series are still noisy in many cases. Unfortunately, details about the sampling procedure are not made public by Google. As long as data quality is still an issue, a possible workaround might be a more sophisticated modeling of the relationship between GTI variables and migration intent using Bayesian inference techniques or microsimulations.

Nonetheless, it is reasonable to assume that role of Google Trends in migration forecasts will become more prominent in the future, given that Google claims to continuously improve the data quality, that usage of the Google search engine grows in some countries and that simply more data will become available.

Appendix

A. List of keyword groups

The complete list of candidate keywords groups can be found in table II.

B. Examplary query

Below, the full Google Trends search query corresponding to keyword group 19 for Spain is given as an example.

> contrato de trabajo alemania + contrato laboral alemania + contrato de empleo alemania + trabajo alemania + empleo alemania + ocupación alemania + ocupacion alemania + trabajar alemania + empleo alemania + empleos alemania + trabajo alemania + trabajos alemania + arbeitsvertrag deutschland + arbeit deutschland + arbeiten deutschland + job deutschland + jobs deutschland + work contract germany + employment germany + working germany + job germany + jobs germany

C. Source code

The source code for the study, including the data and queries, can be found online at https://github.com/bertelsmannstift/eu-migration-forecast.

Acknowledgment

This study has been part of the project "Migration Forecast EU" of the Bertelsmann Stiftung. Special thanks goes to Orkan Kösemen, who initiated and managed the project. I would like to thank Ingmar Weber, André Gröger, Tobias Heidland, Johannes Müller and Leonard Mandtler for important and valuable input to the project, as well as Carsten Große Starmann, Ulrich Kober and Jörg Dräger for making the project possible.

References

[1] European Commission (2022). "Annual report on intra-EU labour mobility." https://ec.europa.eu/social/BlobServlet?docId=26778&langId=en.

[2] Sohst, R., Tjaden, J., de Valk, H., & Melde, S. (2020). "The future of migration to Europe: A systematic review of the literature on migration scenarios and forecasts." International Organization for Migration, Geneva, and the Netherlands Interdisciplinary Demographic Institute, the Hague, https://publications.iom.int/system/files/pdf/the-future-of-migration-to-europe.pdf.

[3] Tjaden, J., Auer, D., & Laczko, F. (2018). "Linking migration intentions with flows: Evidence and potential use." *International Migration*, 57(1), 36–57, https://doi.org/10.1111/imig.12502.

[4] Zagheni, E., Weber, I., & Gummadi, K. (2017). "Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants." *Population and Development Review*, 43(4), 721–734, https://doi.org/10.1111/padr.12102.

[5] Zagheni, E., & 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*, 348–351, https://doi.org/10.1145/2380718.2380764.

[6] Zagheni, E., Garimella, V.K.R., Weber, I., & State, B. (2014). "Inferring international and internal migration patterns from Twitter data" *Proceedings of the 23rd International Conference on World Wide Web*, 439–444, https://doi.org/10.1145/2567948.2576930.

[7] Hawelka, B. et al (2014). "Geo-located Twitter as proxy for global mobility patterns" *Cartography and Geographic Information Science*, 41(3), 260–271, https://doi.org/10.1080/15230406.2014.890072.

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

[9] Carammia, M., Iacus, S.M., & Wilkin, T. (2022). Forecasting asylum-related migration flows with machine learning and data at scale. *Sci Rep* 12, 1457. https://doi.org/10.1038/s41598-022-05241-8.

[10] Boss, K., Gröger, A., Heidland, T., Krüger, F., & Zheng, C. (2023). "Forecasting Bilateral Refugee Flows with High-dimensional Data and Machine Learning Techniques." BSE Working Paper 1387, https://www.itflows.eu/wp-content/uploads/2023/03/1387.pdf.

[11] Wanner, P. (2021). "How well can we estimate immigration trends using Google data?" *Qual Quant* 55, 1181–1202, https://doi.org/10.1007/s11135-020-01047-w.

[12] S. Hawinkel, W. Waegeman, & S. Maere (2023). "Out-of-Sample R2: Estimation and Inference." *The American Statistician*, https://doi.org/10.1080/00031305.2023.2216252.

[13] Bergmeir, C., Hyndman, R.J., & Koo, B. (2018). "A note on the validity of cross-validation for evaluating autoregressive time series prediction." *Computational Statistics & Data Analysis*, 120, 70–83, https://doi.org/10.1016/j.csda.2017.11.003.

[14] Ferri, F. J., Pudil P., Hatef, M., $ Kittler, J. (1994). "Comparative study of techniques for large-scale feature selection." *Machine Intelligence and Pattern Recognition*, 16, 403–413, https://doi.org/10.1016/B978-0-444-81892-8.50040-7.

TABLE II
LIST OF CANDIDATE KEYWORDS, BASED ON [8]

ID	Keywords	With postfix "Germany"	Included in forecast
2	passport, passport office	yes	
10	immigrant, emigrant, immigrate, emigrate, immigration, emigration	yes	
11	visa, entry requirements, required documents	yes	
12	minimum wage	yes	
14	pension	yes	
15	unemployment	yes	
16	internship	yes	
17	inflation, living expenses	yes	
18	social benefits, unemployment benefits	yes	
19	work contract, employment, working, job, jobs	yes	x
20	employment agency, employer, hiring, recruitment	yes	
21	income, tax	yes	
22	GDP, prosperity	yes	
24	wage, salary	yes	x
26	economy, German economy	partially	
28	vacancies, job offers	yes	x
32	job application, application letter, job interview, resume	yes	
33	insurance premium, health insurance, social insurance	yes	
37	university qualification, university	yes	
38	credentials, diploma, certificate	yes	
39	language school, German language school, Goethe Institut	partially	x
41	language test, German language test, German certificate	partially	
42	studies, study, Bachelor, Master, phd	yes	
44	trainee, vocational training, apprenticeship, German apprenticeship	partially	
48	bank account, account	yes	
49	apartment, flat, room	yes	
51	spouse, marry, marriage, intermarriage	yes	
52	rent, utilities, rent deposit	yes	
54	move, moving, relocation	yes	
55	Germany	no	
56	customs	yes	
57	business	yes	
58	migrant, migration, foreigner	yes	
59	nationality, citizenship	yes	
60	arrival, tourist, visit	yes	
112	minimum wage	no	
113	welfare	no	
114	pension	no	
115	unemployment	no	x
117	inflation, living expenses	no	x
118	social benefits, unemployment benefits	on	x
119	work contract, employment, working	no	x
121	income, tax	no	
122	GDP, prosperity	no	
123	job, jobs	no	x
124	wage, salary	no	x
125	gross net, allowances	no	