National Research University Higher School of Economics

As a manuscript

Bronitsky Georgy Timurovich

MIGRATION FORECASTING USING GOOGLE TRENDS

PhD Dissertation Summary

for the purpose of obtaining academic degree

Doctor of Philosophy in Economics

Academic supervisor:
Doctor of Economic Sciences
Vakulenko Elena Sergeevna

JEL: J61, C53, C81

Moscow – 2024

Migration has a profound impact on demographic, economic, social and other indicators of the countries of migration origin and destination. Migration indicators are subject to external shocks, such as pandemics, armed conflicts, natural disasters and others. Applying modern technological methods and tools for the analysis of migratory flows as well as migration intentions, can provide insights for better informing demographic policies of the (migrants' receiving and sending) states. Such evidence would assist in timely addressing the external shocks with minimum adverse effects for the economy of the country. However, the analysis of migration is often complicated due to the limitations of the available data (not all movements are covered by official statistics), changes in the methodology of data collection (Chudinovskikh & Stepanova, 2020), lack of reliable data, and also release lags of the official migration statistics (in some cases, for more than one year).

With information technology advancement, a new strand of migratory flows research emerged in the literature, based on the analysis of migrants' digital traces on the internet. These data open up more opportunities for migration researchers, allowing for shorter lags in economic indicators projections, and sometimes, even for real-time projections (now-casting). In 2006, there appeared Google Trends Index (GTI), a tool that allows to extract data on users' interests in various areas (medicine, telecommunication, business and economy). Choi and Varian (2012) were among the first scholars who published a paper based on the analysis of such data, where they indicate a huge potential of using the data on users' search queries produced by Google Trends Index to measure the interests of economic agents in rel-time (demand for housing, cars, etc.). The results of the meta-analysis of 657 academic papers (Jun et al., 2018) show that the Google Trends popularity is progressively growing among researchers, just within the first 10 years of its existence.

In total, 10 papers (main parameters of which are given in Table 1). The authors of the analysed papers calculated migration estimates using Google Trends Index, including two papers focusing on migration flows in Russia – internal migration from Moscow to St. Petersburg (Fantazzini et al., 2021), and international migration flows from Tajikistan to Russia (Tsapenko & Yurevich, 2022). Seven papers out of eight - except (Wladyk, 2017) - were published in the past three years, that is after 2020.

According to the migration theory, actual resettlement happens only at the third out of five stages of migration to a new place of residence (Benson-Rea, Rawlinson, 2003). In some studies (Wladyka, 2017; Böhme, 2020; Tsapenko & Yurevich, 2022), explanatory variables include not only GTI values per se, but also the lags, most often, not exceeding one year. At the same time, the authors point out certain challenges when estimating migratory flows using GTI: these are due to the unavailability of monthly data, and limitations of the models. For example, the author (Wladyka, 2017) relies on alternative data sources (such as municipal population registry) to overcome the difficulties of determining the lags of search queries based on the yearly data from official statistical agencies. Additionally, in most studies under analysis, search queries are selected by experts, that might negatively impact the quality of the resulting estimates (Ormerod et al., 2014). In most cases, the sets of search queries are too large to include several lags at a time, so that the authors (Wladyka, 2017; Wanner, 2021) have to estimate models for each lag separately.

This dissertation research contributes to migration studies by making use of the Google Trends data. The methods developed in the research serve to overcome analytical challenges when nowcasting migration flows, and novel approaches are proposed for estimating migration using Google Trends Index data.

Table 1. Overview of the studies where Google Trends was used for estimating migration

Authors	Countries	Years	Search queries	Lags, choosing the lag	Selection of search queries	Models

Wladyka (2017)	From Latin America (Argentina, Peru, Columbia) to Spain	2005 – 2010	Embassy of Spain, visa to Spain, work in Spain	Columbia: work and embassy (9th lag);
				Argentina: work (4th lag), embassy (8th lag);
				Peru: embassy (9th lag).
				Cross-correlation with migration data
					Cross-correlation with migration data, three separate searches	Paired time-series regression on variables in differentiation

Böhme et al. (2020)	In 101 migration origin countries in relation to 35 migration destination countries of OECD	2004–2015	Migration and economy (work, wages, visa) + country	Yearly data, 1 lag	Individual searches	FE (fixed effects), a panel of countries, gravity model
Golenvaux et al. (2020)	In 101 migration origin countries in relation to 35 migration destination countries of OECD (from Böhme, Gröger, Stöhr, 2020)	2004 – 2015	Migration and economy (work, wages, visa) + country	Yearly data, 1 lag	Individual searches	Neural networks (LSTM) better predict than ANN, linear gravity model as in (Böhme, Gröger, Stöhr, 2020)
Wanner (2021)	To Switzerland from France, Italy, Germany and Spain	2006 – 2016	Work in Switzerland	Up to 3 years. R^2 Criterion	Expert judgement	Linear regression with GTI lags
Fantazzini et al. (2021)	Russia, Moscow and St. Petersburg	2009 – 2018	Migration to ‘name of the region’, work in ‘name of the region’, housing in ‘name of the region’	1 month	Expert judgement	ARIMAX, SARIMAX, VECM
Avramescu, Wiśniowski (2021)	From Romania to the United Kingdom of Great Britain and Northern Ireland	2012 – 2019	Clusters of keywords: employment, education, currency, housing. WordNet corpus to identify set of synonyms for keywords.	No lags, moving average for 12 months	Correlation with migration data, averages for clusters	ARIMAX
Jurić (2022)	From Croatia to Austria and Germany	2004 – 2020	Work, housing, education, general words (passport, visa, citizenship)	no lags	Most popular search term, individual searches	Paired time-series regression
Tsapenko, Yurevich (2022)	From Tajikistan, Kyrgyzstan, Uzbekistan to Russia	2015 – 2020	(work OR vacancies) AND (Moscow OR Russia)	7 or 11 months. Information criteria, coefficients significance, backward elimination algorithm	Expert judgement	ARIMAX

Source: (Bronitsky, Vakulenko, 2024).

Subject, object of the research

The **subject** of the dissertation research is international migration. The **object** of the research is forecasting of migration to Germany from various countries, using Google Trends data.

Objective and aims of the research

The main **objective of the dissertation research** is to forecast migration of the population ahead of the release of the official migration statistics (now-casting). Data on search queries from Google Trends Index (GTI) can be used for migration modeling. Such models allow to calculate migration flows estimates that would precede the publication of the official statistics. Such an approach has been used earlier to forecast both internal and international migration (flows) (Böhme et al., 2020), including in Russia (Fantazzini et al., 2021; Tsapenko & Yurevich, 2022).

In those papers, the authors used time series models (similar to SARIMAX) with migration flow as a dependent variable, while explanatory variables were GTI values for a number of search queries selected, as a rule, by experts. This dissertation research improves the described models and methods of selecting search queries. Additionally, the research tests the hypothesis that migration occurs not at the time of the search request but with a certain lag since it takes time to make a decision and arrange the necessary documents.

Main aims of the research are the following:

1. To develop a method of selecting keywords and combination of words to predict the fact of migration with highest probability; to create an algorithm of automatic identification of relevant search queries substituting the selection based on expert judgement;
2. Develop and test of a method for increasing the frequency of initial data, allowing to obtain an estimate of migration in monthly frequency;
3. To explore different means of reducing dimension of the set of search queries in order to reduce the number of explanatory variables and corresponding lags;
4. To estimate migration models including GTI lags ranging from one to 12 months. To assess the performance of forecasting models estimating migration from Russia to Germany, to estimate average lag length within the themes ‘embassy’, ‘study’, and ‘work’, to compare the findings with results of previous similar studies;
5. To test asymmetry of reaction of the migration flows to the increase and decrease of popularity of search requests, as well as to investigate lags weight during the shock periods;
6. To verify the sustainability of the findings against migration estimates for other countries (migration from Russia, Poland, Italy, Romania, Spain and Bulgaria to Germany), to develop algorithms for selecting search queries of migrants who do not use the Russian language when searching information on the internet.

Data sources

The dissertation research relies on the search query data such as Google Trends Index and Yandex Wordstat. A set of search queries was obtained using a database of pre-trained embeddings (pre-trained machine learning models) based on the National Dictionary as the initial lexical database, as well as a database of Wikipedia text files.

Additionally, migration statistics was used: for the analysis of migration from Russia to Germany, the official international migration statistics published by the national statistical agencies in Russia (Rosstat), as the country of origin, and in Germany (German Federal Statistical Office), as the country of destination. Moreover, the analysis also relied on the OECD data on annual migration flows from the European Union countries, and on Eurostat data. Besides open access official statistics, monthly data were obtained upon request from the Federal Statistical Office of Germany. The monthly data were converted into yearly values which coincided with the publicly available data.

Theoretical framework and methodology of the research

The analysis of migration from Russia to other countries conducted by (Chudinovskikh & Stepanova, 2020; Denisenko, 2012) focused on the dynamics of migration flows, as well as on the strengths and drawbacks of the official international migration statistics in Russia. Due to considerable release lags of the official statistics (up to five years for developing countries), as well as to the differences in the approaches to calculations, the researchers would often employ alternative data sources (Cesare et al., 2018), that allow for estimating migration flows ahead of the publication of the official statistics. There emerge more studies relying on new types of exogeneous data, facilitating faster estimations of migration (now-casting). These include:

- GPS coordinates of mobile devices (Bengtsson et al., 2011);
- Social networks data (Kim et al., 2020);
- IP addresses of devices (Zagheni, Weber, 2012);
- Search query statistics (Jun et al., 2018);
- Other data sources (international flights data, news reports, etc.) (Gabriel et al., 2019).

This dissertation research makes use of GTI search queries as an alternative data source, due to the availability of data in the open access, and to relatively less bias among users covered by the statistics (Tjaden, 2021). When analysing existing papers devoted to migration estimations based on GTI, it is important to consider the following key elements of such research: (1) Search query sets, i.e. which key search queries authors select to include in migration models, as well as methods of preparing and selecting the queries; (2) Lag structure, i.e. which search queries have lags, and what is the count of those lags; (3) Migration modeling, i.e. which econometric models are used for migration estimations, and the properties of those models.

Considering these elements of the research, Bronitsky and Vakulenko (2024), compiled a table (Table 1) comparing the papers most relevant to the subject of this dissertation research. The study (Bronitsky and Vakulenko, 2024) also noted that "in most cases, the authors of the papers just take such Google Trends search queries as embassy, visa, job, accommodation, and add a destination region. Quite often, the authors choose to identify individual queries based on expert judgment and include only those queries into their models (Böhme et al., 2020; Golenvaux et al., 2020; Wanner, 2021; Fantazzini et al., 2021; Jurić, 2022; Tsapenko & Yurevich, 2022). Only a few studies start by exploring the correlation between migration flows data and search queries (Wladyk, 2017; Avramescu, Wiśniowski, 2021), and only afterwards consider the models for individual search queries." Importantly, the mentioned papers, except (Wladyk, 2017), were published in the past four years, thus, indicating the relevance of the chosen research direction.

Selecting a set of search queries. To use GTI values as explanatory variables, a set of search words is to be defined. This could be done in different ways:

- Relying on expert judgment, including on the opinion of other authors who conducted similar research (Golenvaux et al., 2020; Wanner, 2021; Jurić, 2022);
- Using machine learning methods to identify words and phrases most closely associated with the word 'migration' (Avramescu, Wiśniowski, 2021);
- Relying on statistics of internet searches (for example, Yandex Wordstat data).

This dissertation research uses a combination of the second and third approaches. Initially, using machine learning methods, the word combinations in Russian were identified, which corresponded to the meaning of the word ‘migration’. Next, these phrases were used to search in the Yandex Wordstat system. This approach yields best predictability of migration compared to similar models that rely on other algorithms of building a search query set.

Also, when a set of search queries was being formed, the queries were grouped by most popular themes. Key themes included queries related to job search, studying opportunities, and embassies (Böhme et al., 2020; Fantazzini et al., 2021). This ‘multi-query’ method is used in this dissertation research to estimate migration flows from Russia to Germany. Apart from this method, the dissertation develops also a ‘one-query’ approach, used for migration estimations for Poland, Italy, Romania, Spain, Bulgaria, Russia and Germany. One-query approach consists of using only one search query – ‘work in Germany’, in the language of the country of migration origin – and aims to simplify the process of data collection used for migration estimations.

Preparing the GTI search query data. The Google Trends Index (GTI) reflects the dynamics of search popularity $S_{d,r}$ among users for specific keywords over time (*d*) within a particular region (*r*). However, the index $S_{d,r}$ does not show the absolute number of queries for the chosen search term $V_{d,r}$, rather it shows the number of queries relative to all search queries in that region on that day $T_{d,r}$. Thus, the index $S_{d,r} = \frac{V_{d,r}}{T_{d,r}}$ indicates the share of queries related to a specific search word relative to the total number of queries in the chosen geographic region at a given moment in time.

Using GTI data is associated with two challenges: first, the data are relative, thus it is problematic to use it in models without making additional transformations; second, the data may change when time windows are changed. One way to address this issue is through the standardization of Google Trends data (Fantazzini et al., 2021):

$$Z = \frac{X - \bar{X}}{\hat{\sigma}_X} \tag{1}$$

,where $\bar{X}, \hat{\sigma}_X$ are a mean value and a standard deviation of a random variable, respectively.

Standardization (1) of data allows to compare the search queries and assess the relative popularity of themes over time. Additionally, this approach enables the application of statistical tools, such as Principal Component Analysis (PCA), to the transformed time series, as all series have the same scale — zero mean and unit variance.

Reducing dimensionality. When assessing migration models, limitations arise regarding the number of variables in use. The reason is that the number of observations in the migration dataset is small: only 132 observations on migration from Russia to Germany since the beginning of 2011 (data before 2011 cannot be used due to changes in GTI methodology). Additionally, a ‘control’ group is to be identified to evaluate the predictive power of the models, and a switch to seasonal differences with a 12-month period is to be done.

The study tests the hypothesis regarding the necessity of including lagged search queries (from 1 to 12 lags) in the models, which is motivated by the assumption that "some time" passes between the moment of the search and actual migration (estimating this time period for different search queries is of special interest): when lags are included, the number of explanatory variables increases by 12 (times). These limitations refer mostly to ‘multi-query’ models assessing migration from Russia to Germany, due to the large number of explanatory variables used. This approach requires additional efforts to select explanatory variables and to reduce dimensionality of data before using the data in the models.

1. The first step is to perform R^2 decomposition for the multiple regression model including all search queries, using Shapley method (Israeli, 2007). This approach allows for selecting regressors that contribute the most to the explained variance of the migration model with GTI.
2. The second step is to apply the principal component analysis (PCA) method only to those search queries that had been identified in step 1, and were grouped by themes (education, employment, embassy are the main themes contributing the most to the explained variance (step 1)). Principal components are selected for each group in such a way that the coefficients yield the smallest p-values in migration models (Aivazyan, 2012).

The Shapley method in combination with the Principal Component Analysis, allows to obtain three vectors of queries grouped by the following themes: 'work', 'study', 'embassy'. Next step is to assess the models using these data, as well as incorporating lags from 1 to 12 months, for each theme (Table 2). Importantly, when following this algorithm further, only test sample is used to evaluate the eigenvalues and eigenvectors.

Distributed Lag Models. The models used for forecasting migration flows between different countries include: SARIMA – Seasonal Autoregressive Integrated Moving Average, –a model to forecast migration flows without using exogenous data, the model also serves the 'baseline' migration estimate; SARIMAX - a variant of SARIMA models, where GTI and its lags from 1 to 12 months are included as explanatory variables; and a distributed lag model is proposed - multiple regression, where the number of migrants serves as the dependent variable, and GTI values and the corresponding time lags from 1 to 12 months are used as explanatory variables. To estimate the time lags of search queries, GTI lags are added, i.e., a distributed lag model ($l = 1 \dots 12$ months) is estimated for variables in seasonal adjustment[1]:

$$Y_t - Y_{t-12} = \beta_0 + \sum_{k=1}^{m} \sum_{l=0}^{12} \beta_{k,l} \left(X_{k,l-1} - X_{k,l-12-1} \right) + \varepsilon_t \qquad (2)$$

where 'arrivals of foreigners' to Germany Y_t is used as a dependent variable of the model estimating migration flows, $t = 1,...,T$ - the year, $X_1,...,X_k$ are explanatory variables (GTI search queries), $\varepsilon_t \sim iid\left(0,\sigma^2\right)$ - regression errors. For the analysis of migration flows from Russia to Germany, a total of 36 GTI variables, selected to be used as explanatory variables, were transformed into three vectors – 'study', 'work', 'embassy', using principle component analysis (PCA).

[1] A larger number of lags is not considered for several reasons. On the one hand, with more lags, the dimensionality of the model increases significantly, complicating the estimation, as there are limitations on the length of the series under study. On the other hand, migration intentions are typically realized within a year.

Table 2. Description of the models in seasonal adjustment.

Model	Initial data	Description
PCA by themes without lags	Standardized GTI-indices in differences	Three themes were identified: ‘study’, ‘work’, ‘embassy’. PCA-vectors were identified for each category. Only the first principal component is used.
PCA themes + dummy variables	PCA by themes	Two dummy variables are used on the model: 1) a binary variable corresponding to the 5th upper and lower percentiles for the ‘embassy’ GTI 2) multiplication of this variable by the ‘embassy’ PCA vector These account for modeling rapid changes in search activity in response to shocks.
PCA-vector 'study', ‘work’, ‘embassy’ with lags	PCA by themes	*For each theme, the first principal component with lag of 1-12 months was taken.* All possible combinations of lags are tested, and the best model is determined based on the AIC criterion.
PCA-vectors ‘study’, ‘work’, ‘embassy’ with lags	PCA by themes	*For all themes simultaneously,* PCA-vectors were used with lags, determined in models by separate themes. The best lags and model are determined based on the AIC criterion.

Source: (Bronitsy &Vakulenko, 2024)

To determine the required number of time lags in the distributed lag model, an algorithm was developed testing all possible models including time lags from 1 to 12 months (a total of 8,192 models were estimated for each country). The best model was chosen using the Akaike Information Criterion (AIC). When applying AIC to compare the models, it is necessary to use the same number of observations as was used for estimating the parameters of the model. For SARIMAX models, the AIC was also employed to determine the model parameters p, d, q, where *p* is the order of autoregression, *d* is the order of integration, *q* is the moving average order, and the parameter s=12 is chosen based on the Auto-Correlation Function (ACF). These parameters are also selected using the AIC by testing parameters on the ‘test’ group data. In SARIMAX model, the exogenous variables are the search query itself and its time lags, determined for the distributed lag model. Best lags are not explored further due to the larger number of model parameters to be tested and limited computing capabilities.

Evaluation metrics. The performance of the models is assessed by dividing the initial dataset into ‘test’ sample and ‘control’ sample: in terms of methodology, it is an important step enabling out-of-sample validation of the models (MAPE and MAE metrics are used in the research). In all these evaluations, the test and control samples do not overlap, because the ‘test’ sample is taken for the period starting 01.01.2011 (the beginning of the period under study) and up to the "control" sample.

Validation of the models and migration flows forecasting, is conducted using the time-series in seasonal differences. When further comparing the results of the model validation to identify interpretability of the models, the initial set of variables is used, so that to see migration flows forecast errors rather than seasonality differences of migration. The pairs of test and control time-series were used corresponding to the 2nd and 3rd forecasting years, respectively. The first pair of test and control time-series is intended to assess the model performance over the 2-year forecast period: from June1, 2021 to June 1, 2023 (with the corresponding test sample covering the period from January1, 2011 to June 1, 2021).

It is important to note that during this period, most Covid-19 pandemic restrictions were lifted, meaning that model performance is not subject to significant external shocks. The second pair of time-series is used to investigate a 3-year forecast period from June 1, 2020 to June 1, 2023 (with the corresponding test sample from January 1, 2011 to June 1, 2020). At the beginning of this period, COVID-19 pandemic restrictions were in force which is reflected in the decline of migration activity across all countries due to travel challenges. Yet studying such a period allows to assess how well the models perform in the context of the shocks caused by pandemic. The obtained findings are generalized for other contexts involving external factors, such as armed conflicts, natural disasters, etc.

Average lag length. For the distributed lag models (2), the contribution of each lag can be calculated using the average time lag L_k estimation, within the thematic frame of search queries k:

$$L_k = \sum_{l=0}^{12} l\beta_{k,l}^2 \Big/ \sum_{l=0}^{12} \beta_{k,l}^2 \,. \qquad (3)$$

The value of the average time lag L_k is calculated using the squares of the coefficients for the lags $l = 1,...,12$ from model (2), so that to deal with the cases of negative coefficients $\beta_{k,l}$. The average lag is calculated separately for each theme ('work', 'study', 'embassy"). The coefficient values in the estimated model are random variables, and so is the value of the average lag (3). To estimate the confidence interval, the distribution of the average lag is evaluated using the Monte Carlo simulation, and then the mean value of the obtained distribution is determined using bootstrapping. Larger values of the average lag L are indicative of the greater contribution of time lags closer to 12 months, implying a longer interval between the moment of internet search and actual migration. Similarly, smaller values of L correspond to 'faster reaction' of migration to changes in GTI data.

Roadmap of the dissertation

The dissertation has 105 pages and consists of an introduction, three chapters, a conclusion, a list of references, and appendices. Chapter 1 is based on the research by Bronitsky and Vakulenko (2022), where authors describe major analytical challenges when using the official statistics on migration flows from Russia. These challenges include release lags of the statistics and underestimation of Russian migrants in various official statistical datasets. The chapter also clarifies the data on migration used in the dissertation research, including the statistics of the migration destination and migration origin countries. Additionally, the chapter describes the methods for preparing data for further modeling, including algorithms developed for increasing the frequency of the initial migration data. The second half of the chapter contains an overview of digital footprint data sources that could be used in migration modeling and provides analysis of previous research where Google Trends Index data was used for modeling both international and internal migration. Finally, the chapter discusses the main advantages and disadvantages of migration modeling based on the digital footprint data.

Chapter 2 is based on the paper by Bronitsky and Vakulenko (2024). The chapter examines various methods of modeling migration using Google Trends Index data, focusing on migration flows from Russia to Germany. The chapter describes techniques for reducing the dimension of the initial time series, so that to consider not only the current values of search indexes but also the corresponding time lags ranging from 1 to 12 months. Several types of models are compared: without lags; with lags; for separate themes (such as 'work', 'study', 'embassy'); and for all themes simultaneously. Additionally, the chapter includes evaluation of the length of the average lag, indicating the time interval between the moment of internet search query and actual migration. The chapter concludes by comparing performance of the models used, with the baseline SARIMA models where migration data from external sources are not used.

Chapter 3 is based on the paper by Bronitsky (2024), which applies the previous research findings for the analysis of international migration flows to Germany. The chapter describes methods of building a set of search terms when dealing with several languages. The hypothesis tested in this chapter is that the predictive power of migration models increases when GTI data is included into the models, especially in the context of external shocks, where SARIMA models without exogenous variables show poorer performance. Additionally, the chapter compares the performance of the two types of models: when search queries used in the model refer to several themes and when a single search query and its lags are included in the model.

Scientific contribution of the dissertation research

The existing research on forecasting migration using Google Trends data, has not offered yet the methods for automatic selection and aggregation of search queries, as well as for exploring the lag structure of the search queries:

1. This dissertation develops a method of selecting search queries to improve the existing approach which relies mostly on expert judgement (Wanner, 2021; Jurić, 2022). Machine learning methods (NLP) are used allowing for automatic selection of search queries associated with migration intentions;
2. The algorithm that allows to increase the frequency of initial migration data from annual to monthly values. Unlike MIDAS (Mixed Data Sampling) models, which are used to work with mixed frequency data, this approach allows us to study the lag structure of search queries;
3. A method of aggregating key research queries is offered which groups search queries into indexes by themes that describe migration more accurately. Considering search queries by thematic groups associated with various migration purposes (to work, to study, etc.), allows to reduce dimensionality of the data to consider several types of search queries at a time and to explore the lag structure of these queries;
4. The dissertation research improves existing methods of working with search queries time lags, including the method described by Wanner (2021), by means of adding several GTI time lags into the model (distributed lag model).
5. Average time lag is assessed for search queries by themes, indicating the time interval between the moment of internet search and actual migration, as reflected in the corresponding migration statistics data.
6. Algorithm for selecting search queries is developed, which can be used for the purposes of the analysis of international migration from several migration origin countries (the dissertation studies international migration to Germany from six countries); the results of the analysis show that predictability of the models using GTI data is better; also, distributed lag models demonstrate higher predictability compared to the models with only one time lag (Wanner, 2021).

Findings to be defended

The dissertation proposes a methodology for migration modeling with minimal time lag, termed as ‘migration statistics nowcasting’. Although a similar methodology was developed elsewhere, it can hardly be directly applicable to the Rosstat data and statistical services of many other countries.

1. The proposed algorithm of selecting search queries allows for automatic identification of relevant search terms to be further used for forecasting.
2. The proposed method for increasing the frequency of data by extracting the seasonal component from migration data enables estimation of monthly indicators from yearly values.
3. Various econometric models for forecasting migration (SARIMA, SARIMAX, distributed lag models) were assessed in the dissertation, and the results show that the use of exogenous variables such as GTI data for different thematic categories improves the predictive power of models for all countries studied in the dissertation.
4. Using several GTI lags simultaneously in the model reduces prediction errors, compared to models where each lag is used separately. For five out of six countries under study, models with multiple lags demonstrate the best performance.
5. Methodology for estimating the time interval between the moment of change in search query dynamics and the change in migration flows data is developed, based on the average lag length. The average lag L for the ‘embassy’ thematic category is L=5.6, with a 95% confidence interval [3.64;7.92], while for ‘study’ and ‘work’ categories L=8.0 [5.36;10.8], and L=6.5 [4.72;8.21] respectively.
6. After assessing the performance of models for forecasting international migration to Germany from several countries, it is concluded that in the context of external shocks (such as the Covid-19 epidemic, the launch of the Special Military Operation by Russia), distributed lag models demonstrate superior predictive power compared to SARIMAX models for all countries covered in the dissertation research. In the case of migration from Poland, Italy, Romania, and Spain to Germany, the model error metric MAPE over a 3-year period is more than 2 times lower compared to SARIMAX models.

Theoretical and empirical contribution of the dissertation research

As a theoretical contribution, the dissertation research further develops methods of migration forecasting with minimal time lag (nowcasting) using digital traces left by migrants on the internet. The proposed approaches can be applied for forecasting not only migration flows but also other economic indicators where minimal time lag is of importance. Additionally, these approaches can also be used when considering alternative methods for calculating migration flows estimates, for example, the proposed methods allow to estimate discrepancies in migration data caused by undocumented migration.

The methodological approaches proposed in the dissertation can be used for lecturing on nowcasting. The lectures based on this dissertation will help students gain knowledge about the basic methods of collecting and processing the data for modeling of economic processes. While the sources of the digital footprint are numerous, the dissertation focuses on Google Trends Index.

As an empirical contribution, the dissertation addresses the issue of underreporting in Russia's official migration statistics, confirming the necessity of using ‘mirror analysis’ so that to estimate migration outflows based on the migration inflows data collected in destination countries. When estimating migration flows using GTI data, it is important to include not only current query data but also the lagged values. Thematical categories ‘embassy,’ ‘work,’ and ‘study’ can be used when modeling international migration.

The findings of the dissertation research could be considered by the Main Directorate for Migration Affairs of the Ministry of Internal Affairs of the Russian Federation, for the purpose of developing an alternative methodology of migration data collection and statistics compilation, as well as by the Government of Russia and relevant ministries for formulating policies in the areas of demography and migration.

Limitations of the research

It should be noted that the proposed approaches do have some limitations associated with migration data as well as search queries data from Google Trends Index and the utilized models. First, the dissertation research relies on aggregated migration flows data, not distinguishing between the purposes of migration (to work, to study, to reunite with the family, etc.); second, official statistical data used for migration modeling, could provide an underestimated number of migrants, for example, not covering migrants with double citizenship and undocumented migrants; third, when using GTI data for migration modeling, the results are biased towards those people who use internet to search for required information, and in some countries Google search is not used at all; fourth, while the dissertation research explores the asymmetry of responses to the external shocks (such as pandemics and other crises), the potentially different speed of responses to external shocks is not considered for most models, and as a result, the obtained data can significantly diverge from the actual migration flows.

Additionally, the asymmetry effect may be also caused by unknown true motivations of internet search, for example, search queries may be motivated by a debatable news item or any event which created much noise and attracted public attention, thus not related to migration intentions; finally, when applying the proposed algorithms for the analysis of migration from several European Union (EU) countries, the search term ‘work in Germany’ was used since there is no need in visa for EU nationals, however, an alternative search term ‘embassy of Germany’ could be used to study the countries where citizens have to obtain visa to enter Germany.

Verification and approbation of the research findings

The findings of the dissertation research were presented and discussed at the following **conferences:**

- XXIV Yasin (April) International Scientific Conference on Economic and Social Development Issues (presentation date is April 4, 2023). Demography and Labor Markets Section. Title of the presentation: Forecasting International Migration Using Google Trends.

- Fifth Russian Economic Congress 2023 (presentation date is September 11, 2023). Applied Econometrics Section. Title of the presentation: Investigating Lag Structure of Google Trends Indices for Forecasting Migration from Russia.

- Seventh Consortium of Journals Conference (presentation on October 25, 2023). Session on Modern Methods and Data in Demographic Analysis. Presentation: Application of Google Trends for Forecasting Migration from Russia.

- Eleventh International Conference “Multivariate statistical analysis, econometrics and simulation of real processes” named after S.A. Ayvazyan (presentation date is June 25, 2024). Multivariate statistical analysis and econometrics section. Title of the presentation: Migration nowcasting using Google Trends Index.

The results of the dissertation research have been **published in the following articles**:

- Bronitsky, G.T. (2024). Migration Nowcasting Using Google Trends: Application to Different Countries. *Population and Economics*. (in press). Scopus Q3.

- Bronitsky, G.T., & Vakulenko, E.S. (2024). Application of Google Trends for Forecasting Migration from Russia: Aggregation of Search Queries and Consideration of Lag Structure. *Applied Econometrics*, 73, 78–101. DOI:10.22394/1993-7601-2024-73-78-101. Scopus Q4.

- Bronitsky, G.T., & Vakulenko, E.S. (2022). Using Google Trends for external migration prediction. *Demographic Review*, 9 (3), 75–92. DOI: 10.17323/demreview.v9i3.16471.

References

Ayvazyan S.A. (2012). Analysis of the quality and lifestyle of the population (econometric approach). M: Nauka, 432 p. (in Russ.)

Bronitsky, G.T. (2024). Migration Nowcasting Using Google Trends: Application to Different Countries. *Population and Economics*. (in press, in Russ.).

Bronitsky, G.T., & Vakulenko, E.S. (2022). Using Google Trends for external migration prediction. *Demographic Review*, 9 (3), 75–92. DOI: 10.17323/demreview.v9i3.16471. (In Russ.)

Bronitsky, G.T., & Vakulenko, E.S. (2024). Application of Google Trends for Forecasting Migration from Russia: Aggregation of Search Queries and Consideration of Lag Structure. *Applied Econometrics*, 73, 78–101. DOI:10.22394/1993-7601-2024-73-78-101. (In Russ.)

Vorob’eva O.D., & Grebenyuk A.A. (2017). Comparative analysis of domestic and foreign statistics on the emigration of russian citizens. *Voprosy statistiki*. 1(9), 64-73. (In Russ.)

Denisenko M.B. (2012). Emigration from Russia to foreign countries. *Demoscope Weekly*. No. 513-514.

Tsapenko I.P., & Yurevich M.A. (2022). Nowcasting migration using statistics of online queries. Economic and Social Changes: Facts, Trends, Forecast, 15(1), 74–89. DOI: 10.15838/esc.2022.1.79.4 (in Russ.)

Chudinovskikh O.S. (2018). Big data and statistics on migration. Voprosy statistiki. 25(2), 48-56. (In Russ.)

Chudinovskikh O.S., & Stepanova A. V. (2020). On the quality of the federal statistical Observation of migration processes. *Demographic Review*, 7 (1), 54–82. DOI: 10.17323/demreview.v7i1.10820. (in Russ.)

Akaike H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723. DOI:10.1109/TAC.1974.1100705

Avramescu A., Wiśniowski A. (2021). Now-casting Romanian migration into the United Kingdom by using Google Search engine data. *Demographic Research*, *45*, 1219–1254. DOI: 10.4054/DemRes.2021.45.40.

Bengtsson L., Lu X., Thorson A., Garfield R., von Schreeb J. (2011). Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. PLoS Medicine, 8 (8). DOI: 10.1371/journal.pmed.1001083.

Benson-Rea M., Rawlinson S. (2003). Highly skilled and business migrants: Information processes and settlement outcomes. International Migration, 41 (2), 59–79. DOI: 10.1111/1468-2435.00235.

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. DOI:10.1016/j.jdeveco.2019.04.002.

Chi G., State B., Blumenstock J.E., Adamic L. (2020). Who Ties the World Together? Evidence from a Large Online Social Network. In: Cherifi, H., Gaito, S., Mendes, J., Moro, E., Rocha, L. (eds) Complex Networks and Their Applications VIII. COMPLEX NETWORKS 2019. Studies in Computational Intelligence, 882. DOI: 10.1007/978-3-030-36683-4_37.

Choi H., Varian H. (2009). Predicting the Present with Google Trends. Technical report, Google. [Cited 1 April 2012.] Available from: http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf.

Choi H., Varian H. (2012). Predicting the present with Google Trends. Economic record, 88, 2–9. DOI: 10.5018/economics-ejournal.ja.2018-34.

Fantazzini D., Pushchelenko J., Mironenkov A., Kurbatskii, A. (2021). Forecasting internal migration in Russia using Google Trends: evidence from Moscow and Saint Petersburg. Forecasting, 3 (4), 774–803. DOI:10.3390/forecast3040048.

Gabrielli, L., Deutschmann, E., Natale, F., Recchi, E., Vespe, M. (2019). Dissecting global air traffic data to discern different types and trends of transnational human mobility. EPJ Data Science, 8(1), 26.

Golenvaux N., Alvarez P. G., Kiossou H. S., Schaus P. (2020). An LSTM approach to Forecast Migration using Google Trends. DOI: 10.1145/1122445.1122456.

Israeli O. (2007). A Shapley-based decomposition of the R-square of a linear regression. *Journal of Economic Inequality*, 5, 199 –212. DOI: 10.1007/s10888-006-9036-6.

Jun S. P., Yoo H. S., Choi S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological forecasting and social change*, 130, 69–87. DOI: 10.1016/j.techfore.2017.11.009.

Jurić T. (2022). Facebook and Google as an Empirical Basis for the Development of a Method for Monitoring External Migration of Croatian Citizens. *Ekonomski pregled*, *73* (2), 186–214. DOI: 10.32910/ep.73.2.2.

Kim J., Sîrbu A., Giannotti F., Gabrielli L. (2020). Digital Footprints of International Migration on Twitter (pp. 274–286). DOI: 10.1007/978-3-030-44584-3_22.

Ormerod P., Nyman R., Bentley A. (2014). Nowcasting economic and social data: when and why search engine data fails, an illustration using Google Flu Trends. DOI:10.48550/arXiv.1408.0699

Tjaden J. (2021). Measuring migration 2.0: A review of digital data sources. Comparative Migration Studies, 9 (1), 59. DOI: 10.1186/s40878-021-00273-x.

Wanner, P. (2021). How well can we estimate immigration trends using Google data?. Quality & Quantity, 55 (4), 1181–1202. DOI: 10.1007/s11135-020-01047-w.

Wladyka D. K. (2017). Queries to Google Search as predictors of migration flows from Latin America to Spain. *Journal of Population and Social Studies [JPSS]*, *25* (4), 312–327.

Zagheni E., Weber I. (2012). You are where you E-mail: Using E-mail data to estimate international migration rates. Proceedings of the 3rd Annual ACM Web Science Conference, WebSci'12. DOI: 10.1145/2380718.2380764.