Forecasting internal migration in Russia using Google Trends: Evidence from Moscow and Saint Petersburg

Fantazzini, Dean and Pushchelenko, Julia and Mironenkov, Alexey and Kurbatskii, Alexey

2021

Forecasting internal migration in Russia using Google Trends: Evidence from Moscow and Saint Petersburg

Dean Fantazzini [1,*], Julia Pushchelenko [2], Alexey Mironenkov [3], and Alexey Kurbatskii [4]

[1] Moscow School of Economics, Moscow State University.
[2] Higher School of Economics, Moscow.
[3] Moscow School of Economics, Moscow State University.
[4] Moscow School of Economics, Moscow State University.
* Correspondence: fantazzini@mse-msu.ru;

Abstract: This paper examines the suitability of Google Trends data for the modelling and forecasting of interregional migration in Russia. Monthly migration data, search volume data, and macro variables are used with a set of univariate and multivariate models to study the migration data of the two Russian cities with the largest migration inflows: Moscow and Saint Petersburg. The empirical analysis does not provide evidence that the more people search online, the more likely they relocate to other regions. However, the inclusion of Google Trends data in a model improves the forecasting of the migration flows because the forecasting errors are lower for models with internet search data than for models without them. These results also hold after a set of robustness checks that consider multivariate models able to deal with potential parameter instability and with a large number of regressors.

Keywords: Migration; Forecasting; Google Trends; VAR; Cointegration; ARIMA; Russia; Time-varying VAR; Multivariate Ridge regression.

1. Introduction

Google Trends (GT) is an online service launched in 2008, which provides an index that reflects the relative popularity of a particular keyword (or a topic) by calculating the share of users' searches for this keyword among the total Google searches. This tool has been used in various fields of research, including IT, communications, medicine, health, business, and economics, see the large survey by Jun et al. (2018) for a detailed review.

One of the latest advances in migration research proposed the inclusion of Google Trends data to forecast migration flows. In this regard, Böhme et al. (2019) stated that people acquire information about migration opportunities online before deciding to emigrate. Therefore, the online demand for information can serve as a proxy for future changes in the number of migrants: changes in online search intensity for specific keywords related to migration can indicate an increase in the demand for migration, and, thus, can help to predict migration flows. We remark that there is an increasing literature that shows that Google-based models significantly outperform most of the competitors in several economic and financial applications, see Chol and Varian (2012), Fantazzini (2014), D'Amuri and Marcucci (2017), Bulut (2017), Yu et al. (2019), and Borup and Schütte (2020). Jun et al. (2018) provide a useful review of the research studies using Google Trends in a wide range of areas, including IT, communications, medicine, health, business, and economics.

In this perspective, we propose to use online search data for forecasting the monthly aggregate migration inflows into Russian regions from all other regions. We justify this choice because the administrative

burden for registering in a new region is nontrivial and take some time[1], and searching the web for information is one of the main strategies a potential immigrant can do. Moreover, given that the most important requirement to register in a new region is having a place to stay, searching the web is needed to look for a house/flat to buy or rent. Furthermore, the official statistics on monthly migration are published with a lag of (usually) 6 months and are not available when a regional government start planning the social and labor policies in that region. Instead, internet search data are available on a weekly and monthly basis and they can help identify in advance the number of people that have an intention to move. Therefore, Internet data may provide precise migration forecasts much before the official statistics release, thus giving the regional governments more time and better information to plan their local policies. In this regard, Nikolopoulos et al. (2021a,b) recently highlighted that the lack of reliable hard data limits the possibility of policymakers making informed decisions, and they suggested employing auxiliary data from social media such as Google Trends. Our proposal in this paper goes in this direction[2].

We use monthly migration data, search volume data, and macro variables for the 2009-2018 time sample to analyze how these variables affect migration inflows for the two Russian cities with the largest migration inflows: Moscow and Saint Petersburg[3]. We consider both short- and long-term forecasts because in real-life the regional government has to plan the social and labor policy for at least a year in advance. ARIMA-class models are used to make 1-step ahead forecasts, while multivariate models are used for recursive long-term forecasting up to 24 months ahead.

The empirical analysis does not provide evidence that the more people search online, the more they relocate to other regions. Instead, we find that a one-time shock in internet search queries results in a negative migration inflow after approximately five months. However, the inclusion of Google Trends data in a model does improve the forecasting of the migration inflows because the forecasting errors are lower for models with internet search data than for models without them. These results also hold after a set of robustness checks that consider multivariate models able to deal with potential parameter instability and with a large number of regressors, potentially larger than the number of observations.

The use of Google search data represents an important leading indicator for migration dynamics, which can complement other instruments, such as data from other social media and telecommunications data, as

[1] See the official detailed requirements in Russian:
https://www.gosuslugi.ru/situation/residential_property/registration_of_citizens, and
http://www.consultant.ru/document/cons_doc_LAW_7271/2ab816e63f6cf336e7c992753d7a3c5c9a517997

[2] In August 2021, using the simple average of the market shares for search engines provided by the analytics services Yandex-Radar and StatCounter, Yandex was the top search engine in Russia with a share of 51%, while Google had a share of 45%. Unfortunately, Yandex provides only the last 24 months of search data, thus making any statistical analysis with monthly data unfeasible. It is for this reason that we used Google search data in place of Yandex data.

[3] The focus of this paper is on legal migrants. Of course, we are aware that there is a large number of illegal migrants in these two cities: unfortunately, the estimates of these immigrants vary widely and are not always available (see e.g. https://ru.wikipedia.org/wiki/%D0%93%D0%B0%D1%81%D1%82%D0%B0%D1%80%D0%B1%D0%B0%D0%B9%D1%82%D0%B5%D1%80%D1%8B_%D0%B2_%D0%A0%D0%BE%D1%81%D1%81%D0%B8%D0%B8 for a summary) so that it is difficult -if not impossible- to build a reliable model using these estimates. However, we are confident that both legal and illegal migration share the same temporal dynamics, as it was particularly evident during the Covid-19 pandemic in 2020, see e.g. https://en.wikipedia.org/wiki/Immigration_to_Russia.

recently discussed by Sirbu et al. (2021). The increasing availability to policymakers of a wide array of leading indicators can be useful to improve both the development and the implementation of migration policies[4].

The rest of this paper is organized as follows. Section 2 briefly reviews the literature devoted to migration research with Google Trends and online data, while the methods proposed for forecasting the migration flows in Moscow and Saint Petersburg are discussed in Section 3. The empirical results are reported in Section 4, while Section 5 briefly concludes. Robustness checks are discussed in the Appendix.

2. Literature review

2.1. Migration

The study of migration in Russia is based on different approaches. One of the oldest streams of migration research employed the spatial structure of data to explain migration flows between regions, see -just to name a few- Ravenstein (1885), Wilson (1970), Willekens (1980), and Alonso (1986).

Another strand of literature focuses on time series models and mainly employs two types of models: ARIMA class models and extrapolation of time series through the propagation of historical forecast errors, see Bijak et al. (2019) and references therein for a review. These models can also be extended using expert-based information through prior distributions and Bayesian methods. In this regard, Bijak et al. (2019) uses time series models with and without expert opinions and considers three types of models: ARIMA class models, autoregressive distributed lag (ADL) models, and historical propagation of forecast errors. They found that ARMA models of low orders showed better performances with stationary data, whereas ADL models worked better with non-stationary data.

In the last decade, there was a large set of works that focused on the main factors affecting migration like economic, institutional, and legal conditions, labor market performance measures, and numerous other factors, see, e.g., Mayda (2010), Constant and Zimmermann (2011), Bijak (2011), Ortega and Peri (2013), Chort (2014), Docquier et al. (2014), Dustmann and Okatenko (2014), Burkhauser et al. (2016), Ette et al. (2016), and Kuhlenkasper and Steinhardt (2017). We refer to Docquier and Rapoport (2012) and Fuchs et al. (2021) for an overview of this field of research.

There is also a smaller but increasing literature that uses social big data to measure migration dynamics and future patterns. These data come from social media, internet search services[5], mobile phones, supermarket transaction data, and other sources. They can contain detailed information about their users and can cover larger sets of the population than traditional data sources. Moreover, they can provide immigrants' movements in real-time and show the immigration trends even before the official statistics are published, see e.g. Hawelka et al. (2014). Zagheni et al. (2014) inferred migration patterns using Twitter data, while Moise et al. (2016) discovered the origins of immigrants from the language used in tweets. Skype ego networks[6] data can also be used to explain international migration patterns, see Kikas et al. (2015) for a detailed discussion. Furthermore, big data can be used to study the movements of individuals in the time of crisis, as suggested by Bengtsson et al. (2011), who proposed to improve the response to disasters and outbreaks by tracking population movements with mobile phone network data. Sirbu et al. (2021) provide a survey of this interesting new literature dealing with human migration and big data.

In the Russian literature, the focus has been to model interregional migration using econometric methods, moving from initial cross-sectional data, to panel data dealing with net migration rates, up to panel

[4] The research in this paper received financial support from a grant from the Russian Science Foundation. The policymakers' interest in using such instruments was indirectly confirmed by the request made to us by the grant reviewers to focus specifically on the possibility of forecasting migration flows using Google search data.

[5] A specific review of the literature dealing with internet search services is reported in section 2.2 .

[6] Ego-centric social networks (ego-networks) map the interactions that take place among the social contacts of individual people.

data models for interregional gross migration flows. Even though different datasets were used, the results of these studies are similar and they highlight that the overall migration flow is low compared to other countries of similar size (like the US or Canada), see Andrienko and Guriev (2004) and references therein. Besides, the main idea is that the Russian economy is in disequilibrium and the migration flows depend on economic fundamentals, such as the differences in the public services provisions, incomes, and the unemployment rates between regions. Vakulenko et al. (2011) and Korovkin et al. (2013) provided additional insights by showing that the main determinants of interregional migration are factors that reflect the situation in the labor and residential markets in the region of arrival. Finally, recent works employed time series methods for modelling migration data, like Pavlovskij (2017) who applied ARIMA models for the short-term forecasting of migration inflows and outflows in the Russian regions.

We remark that a large part of the migrants searching for work in Moscow and Saint Petersburg are from the former Soviet republics. Following the fall of the Soviet Union, Russia became a major destination country for international migrants, with officially almost 12 million foreign-born residents in 2017 (United Nations 2017). In the 1990s, most immigrants were ethnic Russians fleeing from the new post-Soviet republics, whereas the composition of migration flows changed in the 2000s to non-Russian labor migrants (Heleniak 2009; Chudinovskikh and Denisenko 2017). This shift was caused by two changes: more liberal policies to grant work permits to non-ethnic Russian citizens of the Commonwealth of Independent States (CIS), and better performance of Russia's economy compared to the other economies in the region, see Gerber and Zavisca (2020) and references therein for a large discussion. In this regard, we highlight that requirements for obtaining work permits have changed over time, both in policy and in implementation, see e.g. Ryazantsev (2016), and Schenk (2018). Moreover, several studies showed that most labor migrants from the CIS countries are illegal, due to government limits on the number of admitted migrants, complex procedures for obtaining legal status, and incentives for employers to hire undocumented migrants rather than follow those procedures, see Human Rights Watch (2009), and Schenk (2018). This lack of legal status has stimulated a business in fake documents and an array of methods to avoid deportation by the authorities, see Reeves (2013, 2015).

A large literature discussed how migrants from CIS countries learned of opportunities to migrate thanks to their connections with other migrants or family/friends in Russia (usually known as "migrant networks"), see Gerber and Zavisca (2020) and references therein. Demintseva and Peshkova (2014), Demintseva and Kashnitsky (2016), and Demintseva (2017) showed that social networking sites, such as Odnoklassniki.ru and VKontakte.ru, are among the most important means of communicating by foreign migrants, and they are actively used when looking for accommodation and work. Bedrina et al. (2018) recently provided a detailed econometric analysis of Uzbek migration networks in Russia. Timoshkin (2020) further analyzed the whole spectrum of digital migration networks, and he suggested that the success of these digital platforms is due to the complexity of official interfaces to communicate with state information nodes (regulations, job descriptions, normative acts), which make them unsuitable for communicating at a proper level. As a consequence, Timoshkin (2020) suggests that these "migrant" digital platforms such as social media and other information webpages have become an "instrument that compensates for the technological imperfection of the state information hubs". Abashin (2014), Chudinovskikh and Denisenko (2020), and Denisenko et al. (2020) provide large historical surveys and analyses about labor migration on the post-Soviet territory.

2.2. Google Trends and its applications in migration research

Ettredge et al. (2005) were among the first to discuss web-based search data to predict macroeconomic statistics. Since then, the research scope has expanded to a variety of other applications thanks to the seminal paper by Choi and Varian (2012), which proposed to use Google Trends data in several fields, like automobile sales, travel planning, consumer confidence, and many others. Several central banks analyzed the suitability of Google Trends for predicting economic fundamentals, see for example Artola and Galan (2012), and McLaren and Shanbhorge (2011).

Google Trends data have been widely used in the fields of fertility, mortality, and migration. As for fertility, Billari et al. (2013) found that online search queries could reveal the intention to have a child in the future months so that they can be used to increase in forecasting power of traditional demographic models. Mortality research in developing countries has benefited from using mobile phone data that stores information about causes of death across the country, see Tamgno et al. (2013) for more details. As for migration, Qin and Zhu (2018) studied the effect of an air pollution index on the intentions to emigrate using an online search index on "emigration" via Baidu, the largest Chinese search engine. They found that severe air pollution in the short run may significantly increase people's interest in emigration, but this effect varies across Chinese regions. Böhme et al. (2019), as far as we know, were the first to analyze the potential of online search data in predicting migration flows. They built a large set of fixed effects models for migration flows based on yearly migration data, Google Trends data from the origin countries, and several control variables, as suggested by Mayda (2010). This approach proved to be successful in providing real-time forecasts of current migration flows ahead of official statistics, and to improve the forecasting performances of conventional models of migration flow.

3. Materials and Methods

The goal of this paper is to verify whether Google Trends data can be useful for modelling and predicting internal migration in Russia. To this end, we will perform an out-of-sample forecasting analysis using a set of time series models: given that sufficiently long time-series data for migration in Russia have become available, time series analysis can now be used. Following Pavlovskij (2017), Böhme et al. (2019) and Bijak et al. (2019), we will use traditional ARIMA models with and without Google Trends to investigate the impact of this new data source for migration forecasting, as well as multivariate models for long-term forecasting. Moreover, as suggested by Keilman et al. (2001), for each class of models we will consider both a "standard" model with variables in levels and a model using logarithms.

Before presenting the results of the empirical analysis, we briefly review the forecasting models that we will use to predict the monthly migration data for the two Russian cities with the largest migration inflows: Moscow and Saint Petersburg.

3.1. Forecasting methods

The out-of-sample forecasting analysis will employ three classes of models: univariate time series models and Google-augmented univariate time series models for 1-step ahead forecasts, while multivariate models will be used for long-term forecasts. A brief description of each model is reported below.

3.1.1. Models for short-term forecasts

The first class of models employed in our analysis is the class of autoregressive integrated moving average (ARIMA) models based on migration data only. A non-seasonal ARIMA (p,d,q) model can be represented as follows,

$$(1-\phi_1 L-...-\phi_p L^p)(\Delta^d y_t-\mu)=(1+\theta_1 L+\ldots+\theta_1 L^q)\varepsilon_t$$

where $\Delta^d y_t=(1-L)^d$, μ is the mean of $\Delta^d y_t$, and L is the usual lag operator. ARIMA models represent a standard benchmark in time series analysis and we refer to Hamilton (1994) for more details. Following Keilman et al. (2001), we considered models with the variables in levels and in log-levels. In the case of seasonal data, a seasonal ARIMA (SARIMA) can be used:

$$(1-\Phi_1 L^S-...-\Phi_P L^{PS})(1-\phi_1 L-...-\phi_p L^p)(\Delta^d y_t-\mu)=(1+\Theta_1 L^S+\ldots+\Theta_Q L^{QS})(1+\theta_1 L+\ldots+\theta_1 L^q)\varepsilon_t$$

which can be written compactly as ARIMA (p,d,q)(P,D,Q)[S]. Information criteria can be used to find the optimal number of lags for the auto-regressive and moving average terms.

If we augment the previous class of models with Google search data, we obtain an autoregressive integrated moving average model with eXogenous variables (ARIMA-X),

$$(1-\phi_1 L-...-\phi_p L^p)(\Delta^d y_t-\mu)=\beta x_{t-1}+(1+\theta_1 L+\ldots+\theta_1 L^q)\varepsilon_t$$

where x_{t-1} is the lagged Google search index at time t -1 and β is a coefficient. Seasonal components may be added if needed.

3.1.2. Models for long-term forecasts

We used vector autoregression (VAR) models and vector error correction (VEC) models to consider the potential effects of both macroeconomic and search variables on migration flows, and to build long-term forecasts. A general VAR model of order p denoted as VAR(p) is given by

$$\mathbf{Y}_t=\mathbf{\Phi}_0+\sum_{i=1}^{p}\mathbf{\Phi}_i\mathbf{Y}_{t-i}+\boldsymbol{u}_t,\quad \boldsymbol{u}_t\sim WN\left(0,\mathbf{\Sigma}\right) \tag{1}$$

where $\mathbf{Y}_t$ is the (n × 1) vector of endogenous variables, $\mathbf{\Phi}_0$ is an intercept vector, while $\mathbf{\Phi}_i$ are the usual coefficient matrices with i=1,…,p. As the primary focus of this paper is forecasting, the VAR(p) model is estimated in levels and no differencing is applied to non-stationary data. The lag order p of the VAR is selected using the Akaike and Bayesian information criterion. The estimated VAR model is then analyzed by reporting its impulse response functions (IRF), and its forecast error variance decomposition (FEVD), see Lütkepohl (2005, chapters 2-5) for more details.

We decided to use a simple VAR(p) in levels following the suggestion by Gospodinov et al. (2013), who stated that the "*unrestricted VAR in levels appears to be the most robust specification when there is uncertainty about the magnitude of the largest roots and the co-movement between the variables*". This is definitely our case, given the moderate size of our dataset (120 observations): in this regard, we want to remark that Elliott (1998) was the first to show that cointegration methods may deliver large size distortions in the case of systems with near unit-roots. Similar distortions can take place when using sequential modeling and specification procedures based on pretests for unit roots. Moreover, it is possible to show that the estimates of the impulse responses using VAR in levels remain asymptotically valid under weak conditions, even when the underlying process contains a unit root (or is possibly cointegrated with other variables), and the same holds true for forecast error variance decompositions at any finite horizon, see Inoue and Kilian (2020) for more details. Instead, differencing the variables when they are stationary causes these estimates to be inconsistent and inference to be invalid. However, for sake of generality and interest, we will also consider a VEC model following the standard sequential specification procedure based on pretests for unit roots and cointegration, see Lütkepohl (2005, chapters 6-8) for more details.

Similar to univariate models for short-term forecasting, we will consider VAR and VEC models with and without Google search data to evaluate the impact of this new data source for migration forecasting.

3.2. Data

We used monthly migration, search volume data, and macro variables for the 2009-2018 period to analyze how search internet data and macro variables affect migration inflows into a region and to forecast migration. In case there were several alternative data sources for the same variables, we followed previous research in the field of migration and accepted standards among data sources.

3.2.1. Migration data and macroeconomic variables

We employed the monthly aggregate inflow into a region from all other regions using the dataset of interregional migration inflows inside Russia as reported by the Federal State Statistics Service (FSSS), all regions included, for the 2009-2018 period. The goal of this statistical service is to estimate the number of people living in each region when the census is not conducted, and the basis for this data collection is a change in the place of permanent registration. The FSSS is the primary source of information on migration for this work because other sources do not provide the same degree of reliability and they have smaller time samples: the latest population census was held in 2010, while the Russian Longitudinal Monitoring Survey and the Russian Sample Labor Force Survey are sample studies.

It is worth noting that in Russia there is currently freedom of movement within the country (except for some closed cities and territories related to state security), unlike in the Soviet era when migration to large cities was artificially hampered by a special type of registration known as "propiska". The so-called “propiska” was canceled on October 1, 1993. In its place, the Law of the Russian Federation No. 5242-1 of 25.06.1993 introduced the so-called "registration", which is applied following the “Rules for registration and removal of citizens of the Russian Federation from registration at the place of stay and the place of residence within the Russian Federation”, approved by the Decree of the Government of the Russian Federation No. 713 of 17.07.1995. This law is applied until now. Moreover, the right of movement is now enshrined in the Constitution (Article 27), and the current legislation provides only for the notification nature of the present-day registration. Therefore, if a citizen (or a foreigner) moves to a new place of residence for more than 90 days, he/she must notify the migration service within three days. The registration of the migration flows is handled by the Federal Migration Service, which was an independent federal service in 2012-2016, but it is currently a division of the Ministry of Internal Affairs (that is, the police). The registration procedure is regulated by the Government Decree No. 713 of 17.07.1995 with later amendments. The registration is carried out by the owner of the residential premises, and can take place with a personal visit to the office of the migration service, by mail, or using the state portal "Gosuslugi.ru". For further processing and use, the migration data are later transferred from the regional bodies of the Federal Migration Service to the Federal State Statistics Service.

The FSSS officially states that the migrants' statistical records are compiled upon registration and deregistration at their place of residence, as well as (since 2011) when registering at the place of stay for 9 months or more. The deregistration is carried out automatically when processing the migration data of the Russian citizens during their movements within the Russian Federation whereas, for foreign migrants, it takes place after the expiration of their period of stay, regardless of their place of former residence. Interestingly, the Federal State Statistics Service notes that the concepts of "arrivals" and "departures" affect migration data because the same person can change his place of permanent residence more than once during the year. See the official "*Methodological Explanations*" by the FSSS for more details[7]. We remark that there are two types of migration registration in Russia[8]: the *permanent registration* ("регистрация по месту жительства"), whose data are available on the Federal State Statistics Service website and are used in the paper; and the *temporary registration* for a predetermined period ("регистрация по месту пребывания"), which is requested by labor migrants.

Following the past Russian migration research discussed in the literature review, we used the following set of monthly variables dealing with the economic and social situation in Russia: the estimated Russian

[7] https://rosstat.gov.ru/storage/mediabank/%D0%9C%D0%95%D0%A2%D0%9E%D0%94%D0%9E%D0%9B%D0%9E%D0%93%D0%98%D0%A7%D0%95%D0%A1%D0%9A%D0%98%D0%95%20%D0%9F%D0%9E%D0%AF%D0%A1%D0%9D%D0%95%D0%9D%D0%98%D0%AF(1).html

[8] http://www.consultant.ru/document/cons_doc_LAW_2255/

GDP[9], the nominal wage of employees, the residential construction volume (in thousand square meters), the number of employed people in the 15-72 age class (in thousands), and the employers' need for employees (according to the Russian Federal Service for Labor and Employment). The descriptive statistics of these variables for Moscow and Saint Petersburg, respectively, are reported in Table 1, together with the FSSS sources from which they were collected.

Table 1. Descriptive statistics of the migration data and the macroeconomic variables

MOSCOW								
variable	***mean***	***min***	***Q1***	***median***	***Q3***	***max***	***st.dev***	***Source***
Migration Inflow	16252	4024	8455	16248	22962	38217	8534	https://rosstat.gov.ru/folder/12781
Number of employed	6612	5800	6064	6853	7047	7224	502	https://rosstat.gov.ru/labour_force
Nominal wage (per capita)	60666	29797	42719	59833	69791	361938	32509	https://rosstat.gov.ru/labour_costs
GDP (Russia)	44167	8483	23685	41540	62357	103627	23783	https://rosstat.gov.ru/compendium/document/50801
Employers' need	156347	97163	134390	153704	169585	272824	33380	https://rosstat.gov.ru/labour_force
Residential construction v.	242	1	95	171	294	1104	236	https://rosstat.gov.ru/folder/13706
SAINT PETERSBURG								
variable	***mean***	***min***	***Q1***	***median***	***Q3***	***max***	***st.dev***	***Source***
Migration Inflow	13655	3225	8735	14607	17291	25458	6061	https://rosstat.gov.ru/folder/12781
Number of employed	2800	2537	2630	2839	2967	3027	161	https://rosstat.gov.ru/labour_force
Nominal wage (per capita)	39923	21998	29623	38873	48426	72342	11698	https://rosstat.gov.ru/labour_costs
GDP (Russia)	44167	8483	23685	41540	62357	103627	23783	https://rosstat.gov.ru/compendium/document/50801
Employers' need	59404	35023	45548	57363	66519	113880	16912	https://rosstat.gov.ru/labour_force
Residential construction v.	248	21	97	160	250	2200	285	https://rosstat.gov.ru/folder/13706

3.2.2. Search volume data

Russia has two search engines that take most of the market: Yandex and Google. In this regard, we remark that the computation of market shares for search engines is not straightforward, it can be controversial[10], and different analytical services may provide different numbers. In the case of Russia, the two most well-known analytical services are Yandex Radar[11] and StatCounter[12]. We report in Figure 1 the market shares of Yandex and Google search engines since the beginning of 2015 for all platforms provided by these two services, together with their average (2015 is the first year when both analytical services are available).

[9] We are aware that the monthly estimates of the Russian GDP are sometimes considered disputable or doubtful statistical indicators. However, despite being potentially biased measures, they provide new (updated) information that is important for policymakers, and they can be useful to improve the efficiency of any model estimates. It is for these reasons that there are several efforts to estimate monthly GDP indicators: see, for example, the Eurocoin indicator for the euro area GDP growth rate developed by Altissimo et al. (2010), the Aruoba-Diebold-Scotti Business Conditions Index proposed by Aruoba et al. (2009) for the US, till the daily indicator of economic growth for the euro area proposed by Aprigliano et al. (2017).

[10] https://www.conductor.com/blog/2014/05/shouldnt-trust-comscores-numbers-search-engine-market-share-data

[11] https://radar.yandex.ru/search?period=all&group=month

[12] https://gs.statcounter.com/search-engine-market-share/all/russian-federation

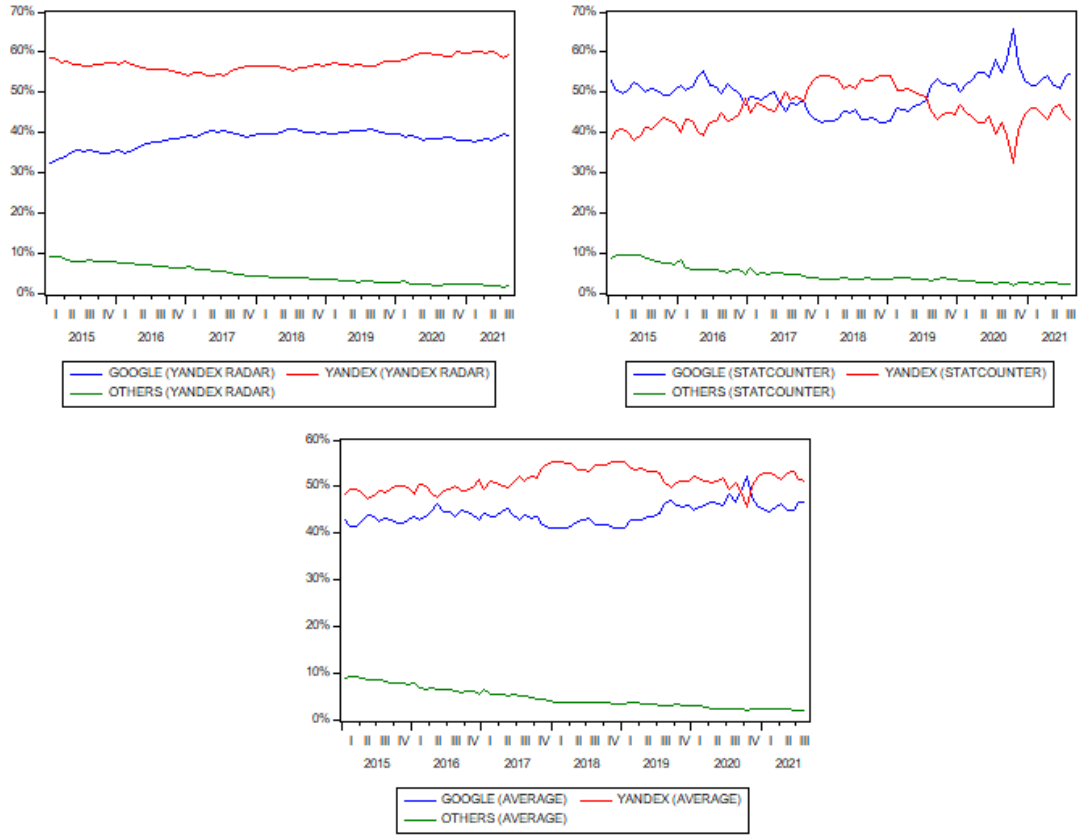

Figure 1. Market shares of Yandex and Google provided by Yandex Radar, Statcounter, and their average.

StatCounter shows that Google was the top search engine in Russia for most of the time, while the opposite is true for Yandex Radar. Given that investigating which online analytical service is more reliable goes beyond the scope of this work, we focused our attention on their simple average, and we observed that Google had a market share in the 40%-45% range, compared with a market share of 50-55% for Yandex. As we anticipated in the Introduction, Yandex provides only a limited amount of free monthly data, so that we had to use Google search data for our work. Even though the latter does not appear to be the main search engine in Russia, its high market share guarantees that its data can still provide useful insights for this research.

Google Trends is a website by Google that publishes a standardized index known as Google Index (GI), which estimates the popularity of a particular search query relative to the total number of searches in the same period in a specific region, and whose scale ranges from 0 to 100.

Although the general reach of Google Trends in Russia is wide, we found that the availability of online searches for our research purposes was quite limited, and search volumes were mostly available only from 2009 onwards. Therefore, we decided to focus only on the regions with the largest migration inflows, given that the online searches for the intentions to migrate were available only for these regions.

The top-10 regions by the total immigration flow in 2018 (see Table 2) represented the starting point that we used to look for online search queries.

Table 2. Top-10 Russian regions and cities for migrants inflows in 2018 (Federal State Statistics Service).

	2018 total inflow (in thousands)	Share from total inflow
Total migration inside Russia	4345.881	100%
Moscow oblast	343.373	7.9%
Moscow	314.868	7.2%
Saint-Petersburg	213.83	4.9%
Krasnodar krai	178.326	4.1%
Tyumen oblast	153.596	3.5%
Republic of Bashkortostan	135.867	3.1%
Krasnoyasrk krai	113.808	2.6%
Sverdolvsk oblast	113.222	2.6%
Leningrad oblast	110.254	2.5%
Rostov oblast	100.112	2.3%
Other regions and cities	2568.625	59.1%

After comparing the volumes of migration flows in Russian regions with the availability of online search queries, we decided to choose Moscow and Saint-Petersburg that account for 12% of the total migration inflow: even though the number of migrants in these cities is comparable to the migration inflows into other regions, the number of online searches for the other regions is almost insignificant compared to these two cities.

The choice of keywords for migration research is not predefined and clear cut unlike the studies dealing with unemployment (for example), where the set of keywords 'work' ("работа") and 'vacancies' ("вакансии") is generally enough to obtain a good estimate of the intentions to find a job, see D'Amuri and Marcucci (2017) and references therein for more details. It is for this reason that Böhme et al. (2019) used a wide range of words that could potentially reflect an intention to move, including indirect interest in economic and legal issues, using for example keywords such as "GDP" and "passport". According to the previously cited Russian studies dealing with migration, the main factors that explain the decision to emigrate are finding a job in the region of interest and finding an apartment. Therefore, we used not only the general query indicating the interest in emigrating ("переезд в «название региона»") but also queries on job and housing searches ("работа в «название региона»", "жилье в «название региона»"). This choice allows us to focus on capturing the intentions to move from one region to another, whether other queries may not indicate the direct intention to relocate. Moreover, we avoided the queries including the word 'migration' ("миграция") and its derivatives because they may be associated only with a general interest in migration policy. Furthermore, we specified the name of the region to exactly identify the direction of migration. We chose these three queries because they are the most popular search queries in each respective group of words concerning relocation, finding a job, and a place to live. As a result, compared to Böhme et al. (2019), our choice of keywords may provide an underestimated number of intentions to emigrate, but the willingness to move in our case is much more certain and it contains a specific geographical component.

We used the previous three queries separately for the in-sample analysis to examine the effect of each query on the migration flow. For forecasting purposes, we also considered the average of these three time series to reduce the number of variables involved and to improve the forecasting efficiency, see e.g. Fantazzini and Fomichev (2014) and Algan et al. (2019) for details.

4. Results

4.1. In-sample analysis

The monthly migrants' inflows in Moscow and Saint Petersburg, and the monthly averages for the three Google searches ("переезд в «название региона»","работа в «название региона»","жилье в «название региона»") are reported in Figure 2.

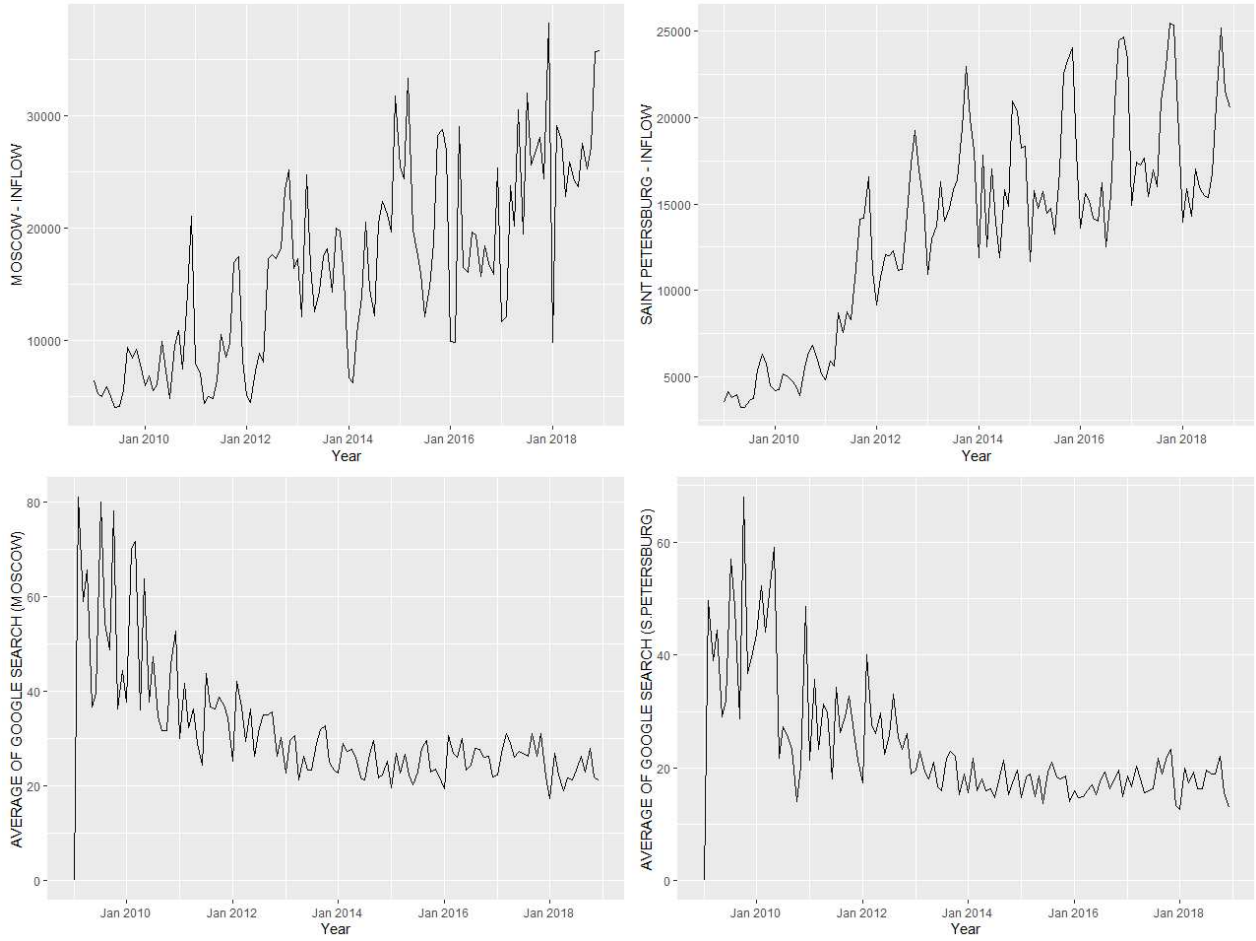

Figure 2. Monthly migrants' inflows in Moscow and Saint Petersburg, and monthly averages for the three Google searches ("переезд в «название региона»","работа в «название региона»","жилье в «название региона»") .

A first look at the data seems to show a certain degree of seasonality in the monthly inflows, particularly for Saint Petersburg. Therefore, we formally tested for seasonality using a battery of tests for the data in levels and in log-levels, which are reported in Table 3. More specifically, we used the F-test for seasonality based on the joint significance of seasonal dummies in a non-seasonal ARIMA model (where the latter is selected using the Hyndman-Khandakar (2008) algorithm), the Friedman (1937) test, the Kruskal and Wallis (1952) test, the QS test by Maravall (2011) that is a variant of the Ljung-Box test computed on seasonal lags, and the Welch (1951) test. We also implemented the Ollech-Webel (2020) test that is a machine learning (ML) classification approach, which first performs a recursive feature elimination algorithm using random forests to identify the most informative seasonality tests, and then uses their p-values as predictors within a single conditional inference tree to determine whether a time series has a significant seasonal component or not.

Table 3 Seasonality tests for the monthly migrants' inflows in Moscow and Saint Petersburg.

Seasonality test	P-values - Moscow		P-values - S. Petersburg	
	Levels	Log-levels	Levels	Log-levels
F-test on seasonal dummies	0.00	0.00	0.00	0.00
Friedman test	0.00	0.00	0.00	0.00
Kruskall-Wallis test	0.07	0.07	0.00	0.00
QS test	0.00	0.00	0.00	0.00
Welch test	0.08	0.04	0.05	0.25
Ollech-Webel ML test	Seasonal	Seasonal	Seasonal	Seasonal

The seasonality tests highlighted a significant seasonal component, so that we employed seasonal-ARIMA models and VAR/VEC models allowing for seasonality when modelling the monthly inflows data.

4.1.1. Univariate models

The best seasonal and non-seasonal ARIMA models, with and without Google search data, found using the Hyndman and Khandakar (2008) algorithm with the corrected Akaike criteria (AICC) proposed by Sugiura (1978) and Hurvich and Tsai (1989) are reported in Table 4 for both Moscow and Saint Petersburg. For sake of interest, Table 4 reports also the Bayesian Information Criteria (BIC) for each selected model.

Seasonal models have lower information criteria than non-seasonal models, and this is particularly true for Saint Petersburg, while the differences are much smaller for Moscow inflow data, thus confirming the previous seasonality tests. The Moscow data has a non-seasonal unit root, while the inflow data for Saint Petersburg displays both a seasonal and non-seasonal unit root. Interestingly, (S)ARIMA models augmented with Google search data as an exogenous regressor almost always show worse information criteria than the baseline models without Google data[13]. No qualitative differences are found when using data in levels and data in log-levels[14].

[13] The coefficients of the Google search data were never statistically significant across all models considered. These results are not reported for sake for space, but are available from the authors upon request.

[14] We remark that the information criteria for the data in levels and in log-levels cannot be directly compared because the datasets used are different, see section 2.11 in Burnham and Anderson (2004) for a detailed discussion of this issue at the textbook level.

Table 4. Best seasonal and non-seasonal ARIMA models, with and without Google search data for the Moscow and Saint Petersburg inflows data, selected using the AICC and the Khandakar and Hyndman (2008) algorithm.

Information criteria	**MOSCOW**			
	Data in levels		*Data in log-levels*	
	Best seasonal SARIMA	*Best non-seasonal ARIMA*	*Best seasonal SARIMA*	*Best non-seasonal ARIMA*
	ARIMA(0,1,1)(1,0,3)$_{[12]}$	ARIMA(1,1,1)	ARIMA(1,1,1)(2,0,0)$_{[12]}$	ARIMA(0,1,2)
AICC	2390	2399	83	92
BIC	2406	2408	97	103
	Best seasonal ARIMA-X	*Best non-seasonal ARIMA-X*	*Best seasonal ARIMA-X*	*Best non-seasonal ARIMA-X*
	ARIMA(0,1,1)(1,0,2)$_{[12]}$	ARIMA(1,1,1)	ARIMA(1,1,1)(0,0,2)$_{[12]}$	ARIMA(0,1,2)
AICC	2390	2401	89	95
BIC	2406	2412	105	108
Information criteria	**SAINT PETERSBURG**			
	Data in levels		*Data in log-levels*	
	Best seasonal SARIMA	*Best non-seasonal ARIMA*	*Best seasonal SARIMA*	*Best non-seasonal ARIMA*
	ARIMA(2,1,0)(0,1,1)$_{[12]}$	ARIMA(0,1,0)	ARIMA(0,1,2)(0,1,1)$_{[12]}$	ARIMA(0,1,0)
AICC	1910	2222	-156	-60
BIC	1920	2225	-146	-57
	Best seasonal ARIMA-X	*Best non-seasonal ARIMA-X*	*Best seasonal ARIMA-X*	*Best non-seasonal ARIMA-X*
	ARIMA(2,0,0)(0,1,1)$_{[12]}$	ARIMA(0,1,0)	ARIMA(0,1,2)(0,1,1)$_{[12]}$	ARIMA(1,1,1)
AICC	1929	2223	-154	-65
BIC	1944	2228	-141	-51

4.1.2. Multivariate models

Consistent with the past literature dealing with Russian migration research, we employed multivariate models for a set of variables including the migration inflows, the estimated Russian monthly GDP, the nominal wage of employees (per capita), the residential construction volume (in thousand square meters), the number of employed people in the 15-72 age class, the employers' need for employees (according to the Russian Federal Service for Labor and Employment), and the Google search data for the queries about moving in a certain region, about work and about and housing.

Information criteria selected a VAR(1) model for both Moscow and Saint Petersburg. Given the presence of seasonality, we estimated all multivariate models with centered seasonal dummies, which sum to zero over time and therefore do not affect the asymptotic distributions of testing procedures, see Johansen (1995, 2006) for more details. For ease of interpretation and sake of interest, we report the orthogonalised impulse responses[15] from a shock in Google searches on migration inflows in Moscow and Saint Petersburg in Figure 3 and 4, respectively, the forecast error variance decompositions[16] for the migration inflows are reported in Figure 5, while the full results are available from the authors upon request.

[15] The orthogonalised impulse responses are derived from a Choleski decomposition of the error variance-covariance matrix $\Sigma = PP'$, with P being lower triangular, see Lütkepohl (2005) for more details.

[16] The forecast error variance decomposition is based upon the orthogonalised impulse response coefficient matrices and shows the contribution of the variable j to the h-step forecast error variance of variable k, see Lütkepohl (2005) for more details.

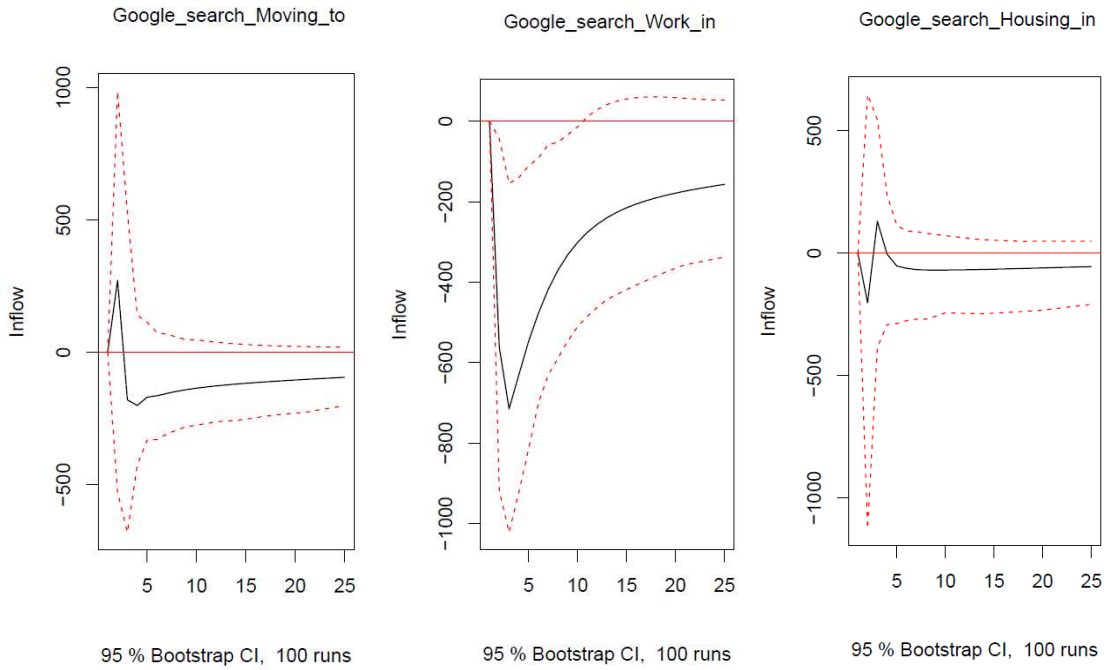

Figure 3. VAR(1) with centered seasonal dummies: orthogonalised impulse responses from a shock in Google searches on migration inflow in Moscow over 24 months.

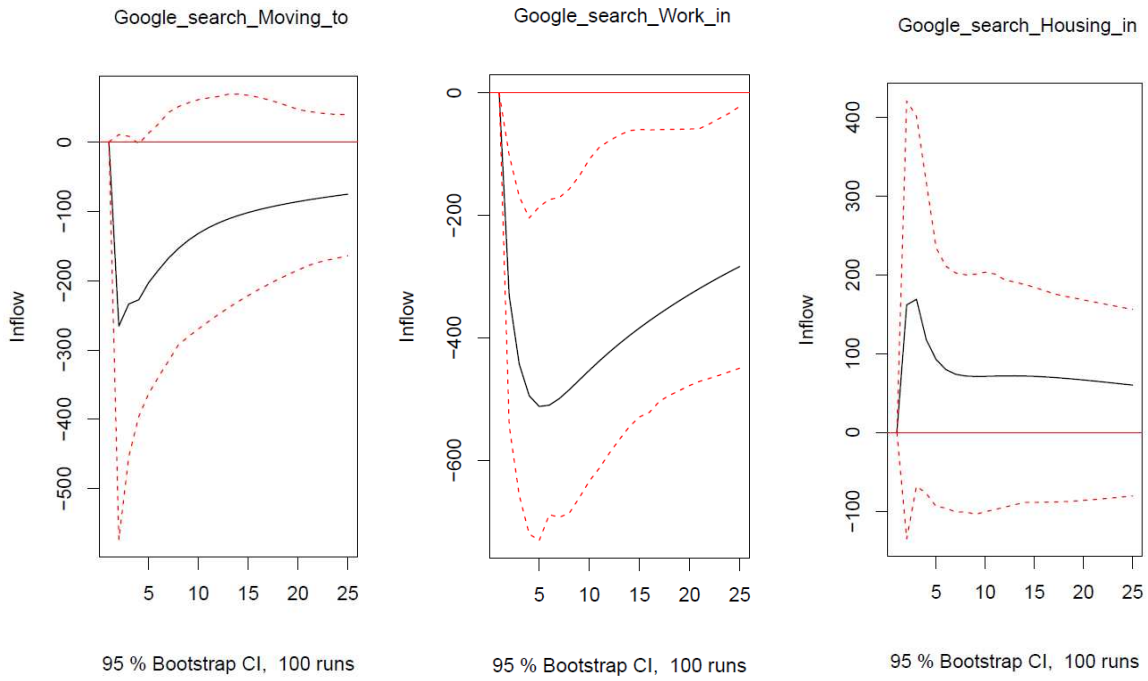

Figure 4. VAR(1) with centered seasonal dummies: orthogonalised impulse responses from a shock in Google searches on migration inflow in S.Petersburg over 24 months.

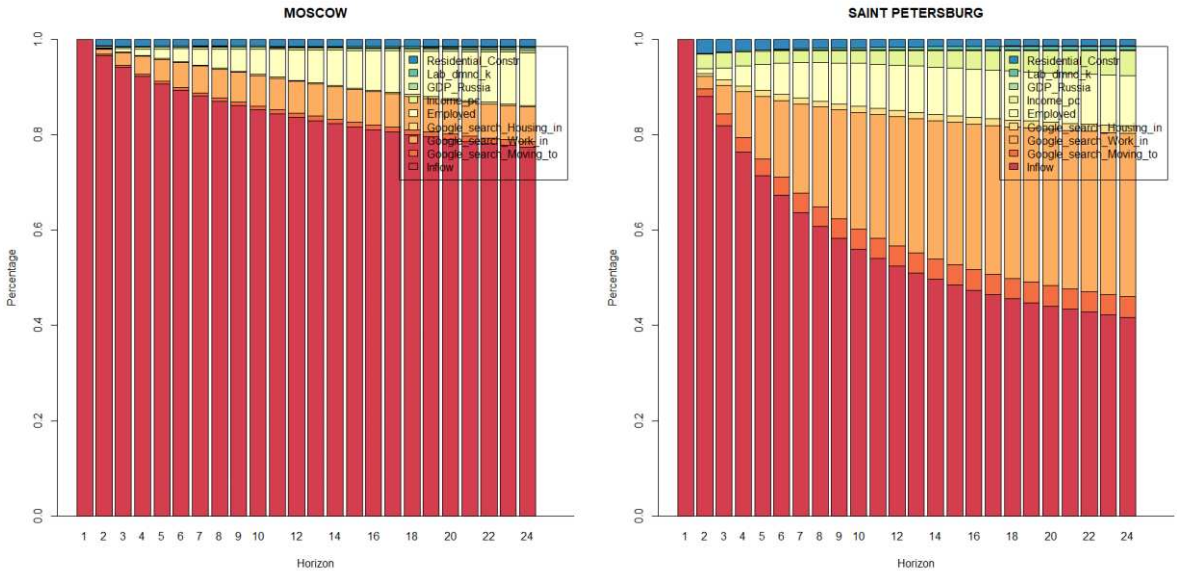

Figure 5. Forecast error variance decomposition of the VAR(1) with centered seasonal dummies: Moscow (left panel), Saint Petersburg (right panel).

Figure 3 and 4 show that the effects of shocks in internet searches on migration inflows are not significant for queries related to emigration and housing searches, while there are significant negative effects for queries related to job searches. In the latter case, it appears that a one-time shock in internet search queries results in a negative migration inflow after approximately five months. The forecast error variance decompositions in Figure 5 show that the variances of migration inflows are mostly affected by their own variances, but the effects of online job searches and the numbers of employed people become stronger in later periods, particularly for Saint Petersburg. The negative relationship between online job searches and migration inflows is probably due to immigrants moving to the regions bordering Moscow and Saint Petersburg because of the high cost of living and traffic congestion in these two metropolises, see e.g. Efimova and Mikhaltsov (2017), Pavlovskij (2017), Varaksin and Varaksina (2017), Demidova et al. (2020), and Vakulenko and Mkrtchyan (2020).

Given the evidence of non-stationarity that emerged from the previous univariate analysis, for sake of generality and interest, we also considered a VEC model following the standard sequential specification procedure based on pretests for unit roots and cointegration. We tested for cointegration using the Johansen trace test with centered seasonal dummies and we rejected the null hypothesis of no cointegration for both Moscow and Saint Petersburg. We estimated a VEC(1) model with six cointegration relationships and a constant term in the cointegration equations for both cities. The orthogonalised impulse responses from a shock in Google searches on migration inflows in Moscow and Saint Petersburg are reported in Figures A3 and A4 in the Appendix B, respectively, while the forecast error variance decompositions for the migration inflows are reported in Figure A5 and the full results are available from the authors upon request. The IRFs and the FEVDs obtained with VEC models are qualitatively similar to those estimated with VAR models in levels, confirming a significant negative effect of online job searches on migrants inflows (for Saint Petersburg), and a much larger importance of Google searches for Saint Petersburg than for Moscow.

4.2. Out-of-sample forecasting analysis

The last step to evaluate the ability of Google search data to predict internal migration in Russia was to perform an out-of-sample forecasting analysis for both Moscow and Saint Petersburg, to forecast the monthly inflows using several competing models with and without Google data, over different time horizons. The data in January 2009 – September 2015 were used as the first training sample for the models'

estimation, while the data for October 2015 - December 2018 was left for out-of-sample forecasting using an expanding estimation window.

4.2.1. Short-term forecasts: 1-step ahead forecasts

Three classes of models were considered for short-term forecasts for a total of 20 models:

1) *ARIMA models* with the dependent variable represented by the monthly inflows in levels or log-levels (2 models);

2) *Google-augmented ARIMA-X* models with the variables in levels or log-levels (8 models): we considered lagged Google search data for the query about moving in a certain region, queries about job and housing, as well as the average of these three queries.

3) *Seasonal-ARIMA (SARIMA) models with and without Google search data*, with the variables in levels or log-levels (10 models).

Additional models could surely be added, but this selection already gave important indications whether Google search data are useful for forecasting the monthly migration inflows in Moscow and Saint Petersburg. A summary of the models' performances according to the mean squared error (MSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE) is reported in Table 5[17].

Table 5. Models' performances according to the mean squared error (MSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). The smallest values are reported in bold font.

	MOSCOW			*SAINT PETERSBURG*		
	MSE	**MAE**	**MAPE (%)**	**MSE**	**MAE**	**MAPE (%)**
ARIMA	6.51E+09	5.79E+05	29.82	9.93E+08	2.59E+05	14.89
SARIMA	6.05E+09	5.50E+05	28.27	4.01E+08	1.69E+05	9.24
ARIMAX (Google: average)	6.44E+09	5.65E+05	29.22	8.94E+08	2.40E+05	13.65
SARIMAX (Google: average)	5.75E+09	5.14E+05	26.58	4.51E+08	1.76E+05	9.82
ARIMAX1 (Google: Moving)	6.49E+09	5.63E+05	29.11	9.82E+08	2.59E+05	14.95
SARIMAX1 (Google: Moving)	**5.37E+09**	**5.13E+05**	**26.17**	3.93E+08	1.67E+05	9.14
ARIMAX2 (Google: Work)	6.47E+09	5.69E+05	29.34	9.92E+08	2.65E+05	15.17
SARIMAX2 (Google: Work)	5.76E+09	5.31E+05	27.04	4.06E+08	1.71E+05	9.61
ARIMAX3 (Google: Housing)	6.51E+09	5.66E+05	29.54	1.04E+09	2.69E+05	15.58
SARIMAX3 (Google: Housing)	5.97E+09	5.33E+05	27.40	3.93E+08	1.67E+05	9.12
ARIMA.LOG	7.63E+09	6.16E+05	32.42	1.01E+09	2.45E+05	13.93
SARIMA.LOG	6.57E+09	5.74E+05	29.01	**3.52E+08**	**1.56E+05**	**8.46**
ARIMAX.LOG (Google: average)	7.64E+09	6.17E+05	32.48	9.72E+08	2.45E+05	14.20
SARIMAX.LOG (Google: average)	6.88E+09	5.84E+05	29.24	3.84E+08	1.63E+05	8.74
ARIMAX.LOG1 (Google: Moving)	8.63E+09	6.46E+05	34.34	1.06E+09	2.46E+05	14.11
SARIMAX.LOG1 (Google: Moving)	6.26E+09	5.83E+05	28.12	3.96E+08	1.70E+05	9.22
ARIMAX.LOG2 (Google: Work)	7.53E+09	6.13E+05	32.40	9.54E+08	2.46E+05	14.51
SARIMAX.LOG2 (Google: Work)	6.85E+09	5.85E+05	29.37	4.10E+08	1.67E+05	9.04
ARIMAX.LOG3 (Google: Housing)	7.55E+09	6.14E+05	32.48	9.87E+08	2.44E+05	13.91
SARIMAX.LOG3 (Google: Housing)	6.91E+09	5.87E+05	29.40	4.66E+08	1.87E+05	10.08

[17] The optimal seasonal and non-seasonal ARIMA models, with and without Google search data, were estimated using the Hyndman and Khandakar (2008) algorithm at each iteration of the forecasting procedure.

In general, Google-augmented time series models forecasted the monthly inflows better than models without Google data. However, the simple SARIMA model with data in logs turned out to be the best model for Saint Petersburg (even though Google-based models were close): this result was expected due to the strong local seasonality in monthly inflows, differently from Moscow where the seasonality was barely significant. This phenomenon may also explain why models with the variables in logs forecasted better than models with the variables in levels for Saint Petersburg, whereas the opposite was true for Moscow. Among Google search terms, queries about moving in a certain region or the averages of all three queries provided better forecasts than the other choices.

4.2.2. Long-term forecasts: 24-step ahead forecasts

The previous univariate models can also be used for long-term forecasting, but it is well known that their forecasting ability quickly degrades, see Hyndman and Athanasopoulos (2018) and references therein for more details. Moreover, if exogenous variables are present, multivariate models have to be used to build long-term forecasts.

More specifically, we used three classes of models to build long-term 24-step ahead forecasts:

1) *VAR models with centered seasonal dummies*, with and without Google data, with the variables in levels, log-levels, first differences, or log-returns (12 models).

2) *VEC models with centered seasonal dummies*, with and without Google data, with the variables in levels or log-levels (6 models).

3) *Seasonal-ARIMA models*, as simple univariate benchmark models, with the variables in levels or log-levels (2 models).

As for the Google search queries, we considered three possible variants: no Google data, the average of the three Google search queries, or all three Google search queries together. A summary of the models' performances according to the mean squared error (MSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE) is reported in Table 6.

In general, multivariate models with Google data forecasted better than multivariate models without Google data and much better than simple SARIMA models (as expected). In the case of Moscow, the VAR model with the variables in log levels and the average of the Google search queries was the best, while VAR models with the variables expressed in log returns (with and without Google data) provide the best forecasts: therefore, this forecasting evidence confirmed the initial in-sample analysis where the evidence of non-stationarity was much stronger for Saint Petersburg than for Moscow. Interestingly, the VEC models performed poorly, in some cases even worse than SARIMA models: these results were not a surprise because the large variance of the estimators for cointegrated models in small-medium samples is a well-known issue in the econometric literature, see Stock and Watson (1993), Maddala and Kim (1998) section 5.7 and Hayashi (2000) section 10.4, for more details. Moreover, Fantazzini and Toktamysova (2015) showed that the sampling noise of Google data can exacerbate this inference problem, and using the averages of Google data can solve this issue to some extent, but not completely. This is what we also found with our data, where models with the averages of Google data often performed better than models with the separate Google search queries.

These results are consistent with a large body of the forecasting literature that shows that Google-based models outperform their competitors, see -for example- Fantazzini (2014), D'Amuri and Marcucci (2017), Borup and Schütte (2020), Aaronson et al. (2021), and references therein.

Table 6. Models' performances according to the mean squared error (MSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). The smallest values are reported in bold font.

	MOSCOW			*SAINT PETERSBURG*		
	MSE	**MAE**	**MAPE (%)**	**MSE**	**MAE**	**MAPE (%)**
SARIMA	7.54E+07	7.21E+03	24.83	1.02E+07	2.70E+03	14.23
SARIMA.log	9.68E+07	7.84E+03	27.07	2.63E+07	3.89E+03	20.45
VAR (NO Google)	4.27E+07	5.70E+03	22.46	1.72E+07	3.27E+03	18.78
VAR.log (NO Google)	**3.30E+07**	4.52E+03	18.11	2.20E+07	3.34E+03	19.22
VAR.diff (NO Google)	7.44E+07	7.08E+03	26.32	1.09E+07	2.77E+03	14.81
VAR.dlog (NO Google)	9.89E+07	8.23E+03	28.73	**3.89E+06**	**1.64E+03**	**8.62**
VAR (all 3 Google queries)	5.23E+07	6.27E+03	23.81	8.24E+06	2.41E+03	13.55
VAR.log (all 3 Google queries)	4.90E+07	5.38E+03	19.72	6.59E+06	2.12E+03	11.54
VAR.diff (all 3 Google queries)	7.52E+07	6.91E+03	25.14	1.02E+07	2.67E+03	14.31
VAR.dlog (all 3 Google queries)	9.89E+07	8.23E+03	28.73	**3.89E+06**	**1.64E+03**	**8.62**
VAR (Google average)	4.52E+07	5.91E+03	23.17	1.69E+07	3.26E+03	18.79
VAR.log (Google average)	3.33E+07	**4.51E+03**	**18.09**	2.22E+07	3.38E+03	19.49
VAR.diff (Google average)	7.24E+07	6.95E+03	26.01	1.09E+07	2.77E+03	14.82
VAR.dlog (Google average)	9.89E+07	8.23E+03	28.73	**3.89E+06**	**1.64E+03**	**8.62**
VECM (NO Google)	6.94E+07	7.00E+03	27.12	1.07E+07	2.74E+03	14.33
VECM.log (NO Google)	7.46E+07	6.73E+03	25.82	7.00E+07	7.78E+03	40.25
VECM (all 3 Google queries)	5.95E+07	6.25E+03	24.21	1.12E+07	2.80E+03	14.65
VECM.log (all 3 Google queries)	5.69E+07	5.99E+03	21.91	8.01E+07	8.25E+03	42.62
VECM (Google average)	5.52E+07	5.94E+03	23.79	1.41E+07	3.22E+03	16.59
VECM.log (Google average)	5.63E+07	5.90E+03	23.28	6.93E+07	7.73E+03	40.02

5. Discussion and Conclusions

There is an increasing literature that shows that Google-based models significantly outperform most of the competitors in several economic and financial applications, see Jun et al. (2018) for a review. Böhme et al. (2019) analyzed the potential of online search data in predicting migration flows for the first time, and they showed that this approach improved the forecasting performances of conventional models of the migration flow. Moreover, it provided real-time forecasts ahead of official statistics.

Following this literature, this paper used monthly migration data, Google search volume data, and macroeconomic variables for the 2009-2018 time sample to analyze how these variables affected migration inflows for the two Russian cities with the largest migration inflows: Moscow and Saint Petersburg. The choice of keywords for migration research was not predefined and clear cut, unlike previous studies dealing with unemployment or financial and economic forecasting. We followed past Russian studies that showed that the main factors explaining the decision to emigrate are finding a job (in the region of interest) and finding an apartment. Therefore, we used not only the general query indicating the interest in emigrating ("переезд в «название региона»") but also queries on job and housing searches ("работа в «название региона»", "жилье в «название региона»"). We chose these three queries because they are the most popular search queries in each respective group of words concerning relocation, finding a job, and a place to live. As a result, compared to Böhme et al. (2019), our choice of keywords may provide an

underestimated number of intentions to emigrate but the willingness to move is more certain, and it contains a specific geographical component.

The empirical analysis did not provide evidence that the more people search online, the more they relocate to other regions, but we found that a one-time shock in internet search queries results in a negative migration inflow after approximately five months. We then performed an out-of-sample forecasting analysis to forecast the monthly inflows using several competing models with and without Google data, over different time horizons ranging from 1 month to 24 months ahead. In terms of short-term forecasting, Google-augmented time series models forecasted the monthly inflows usually better than models without Google data. However, the simple SARIMA model with data in logs turned out to be the best model for Saint Petersburg, thanks to the strong local seasonality in monthly inflows, whereas this was not the case for Moscow where the monthly seasonality was barely significant.

In terms of long-term forecasting, multivariate models with Google data forecasted better than multivariate models without Google data and much better than univariate models. Interestingly, the VEC models performed poorly, in some cases even worse than simple univariate models, thus confirming well-known estimation problems in small-medium samples that can be further exacerbated by the sampling noise of Google data. These results also held after a set of robustness checks that considered multivariate models able to deal with potential parameter instability and with a large number of regressors, potentially larger than the number of observations.

Our empirical evidence showed that Google Trends does help forecast migration inflows in the two Russian cities with the largest migration inflows (Moscow and Saint Petersburg). As recently highlighted by Nikolopoulos et al. (2021 a,b), the lack of reliable hard data limits the possibility of policymakers making informed decisions, and this is why they suggested employing auxiliary data from social media such as Google Trends. Given that migration inflows represent a sensitive social issue in Russia, the option to improve the modelling and forecasting of these flows via using auxiliary data such as Google Trends can be of great help to local policymakers. This improvement is even more important if we consider that a part of these migration inflows is represented by illegal immigrants, which are not included in official statistics but can be revealed by Google Trends.

The availability to policymakers of a wide array of leading indicators for migration dynamics, ranging from online search data to telecommunications data, can be useful to plan and implement more realistic migration policies that can significantly help the inclusion process of migrants, see Sirbu et al. (2021) for a large discussion.

The negative relationship between online job searches and migration inflows is probably due to immigrants moving to the regions bordering Moscow and Saint Petersburg because of the high cost of living and traffic congestion in these two metropolises, see e.g. Efimova and Mikhaltsov (2017), Pavlovskij (2017), Varaksin and Varaksina (2017), Demidova et al. (2020), and Vakulenko and Mkrtchyan (2020). An empirical analysis including also these bordering regions would require spatial econometric models able to deal with situations when the number of variables is larger than the number of time points for the data, see e.g. Ahrens and Bhattacharjee (2015), Lam and Souza (2020), and references therein. Given that this issue goes beyond the scope of this paper and the size of the paper is already quite substantial[18], we leave this issue as an avenue for further research.

Another possibility of future work will be to check how the empirical evidence found in this work will change when using Yandex search data in place of Google search data. To reach this aim, a direct agreement between Russian policymakers and Yandex will probably be necessary to have access to long time series of monthly search data, which currently are not available. The inclusion of such data will likely considerably improve the forecasting performances of the models proposed in this work, so we leave it as a compelling topic of further work.

[18] The authors want to thank an anonymous reviewer for highlighting the initial excessive length of the paper.

Acknowledgements: Dean Fantazzini, Alexey Mironenkov, and Alexey Kurbatskii gratefully acknowledges financial support from the grant of the Russian Science Foundation n. 20-68-47030.

Appendix A

Google Trends is a website (*https://trends.google.com*) that reports the standardized volume of Google searches for a keyword or a topic. Google Trends calculates the ratio of the number of online searches for a specific keyword (or topic) K in a given geographical region a, on a particular day t ($K_{a,t}$), to the total amount of searches for the same day and region ($T_{a,t}$): $R_{a,t} = K_{a,t} / T_{a,t}$. The obtained time series is then divided by the value of the day in which it reaches the maximum level, and multiplied by 100. The Google index (GI) for a specific keyword K on day t, and in the area a is thus given by, $GI_{Ka,t} = [100 \cdot R_{a,t} / \max_t(R_{a,t})]$. Google Trends tracks only queries with a minimum volume due to privacy considerations: if the search volume is too low, a value of zero is reported[19]. The data are available from an intraday time-frequency up to a monthly frequency (which was our case), depending on the selected time range. The longer is the time sample selected, the lower is the frequency provided by Google Trends (the lowest frequency possible is monthly data). Note that Google Trends allows comparing the search volumes of up to five search terms, or up to a maximum of 30 search terms grouped in a single entry using quotation marks (to return searches that match an exact expression), and using the + or - signs between the search terms to include or exclude search terms, respectively. The data are available since 2004, see *https://support.google.com/trends* for more details.

An example of the Google Trends interface to download the monthly data for the keywords "Работа в Москве" (="Job in Moscow") searched in Russia from 01/01/2009 until 31/12/2018, is reported in Figure A1:

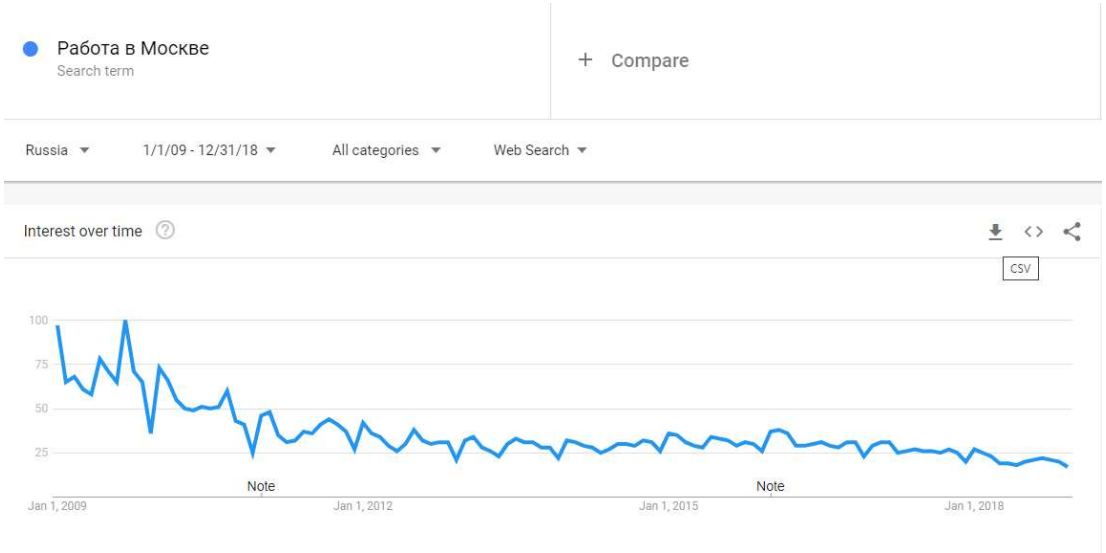

Figure A1. Google Trends data for the keywords "Работа в Москве", searched in Russia. Sample: 01/01/2009 - 31/12/2018.

[19] In the case of zero values, the GIs were linearly re-scaled using a small positive constant, following the approach proposed by Fantazzini and Toktamysova (2015).

The monthly GIs can be downloaded as a csv file by clicking on the arrow on the right, as highlighted in Figure A1. Given that the manual download of the GIs for several keywords can become too burdensome, it can be executed using an R script and the *gtrendsR* package as reported below:

```
library(gtrendsR)
dat=gtrends("Работа в Москве", geo = "RU", time = "2009-01-01 2018-12-31")
plot(dat)
```

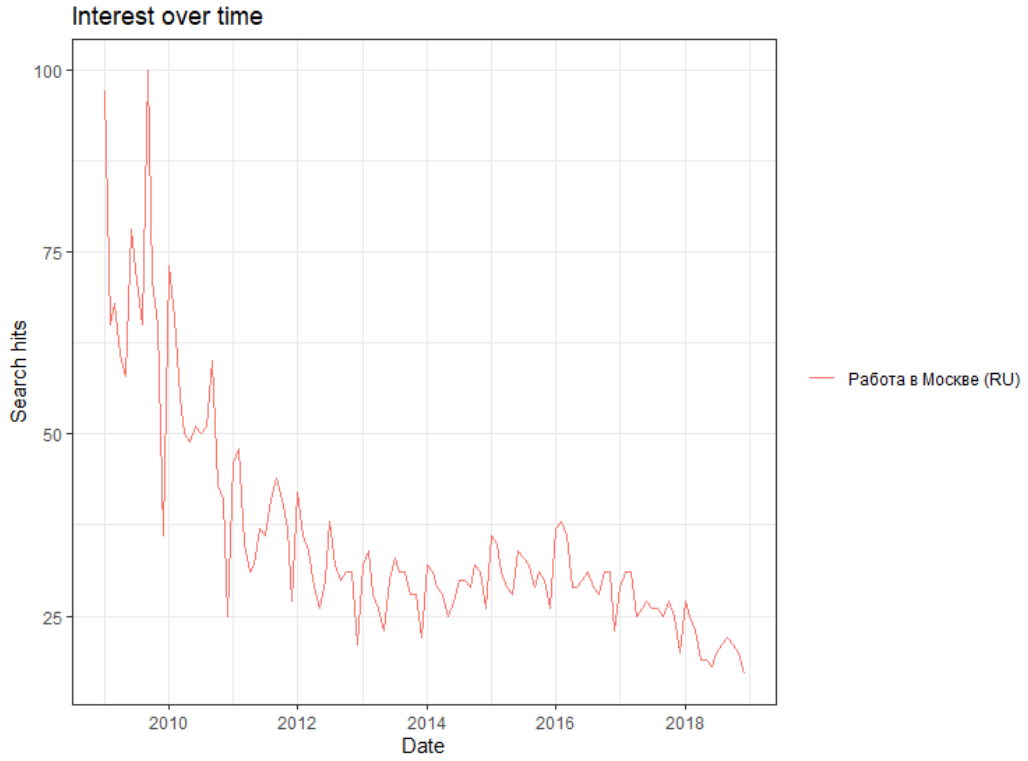

Figure A2. Google Trends data for the keywords "Работа в Москве", searched in Russia. Sample: 01/01/2009 - 31/12/2018. Data downloaded using the *gtrendsR* package.

We remark that Google Trends data are computed using a sampling method, so the results may be slightly different if the data are downloaded on different days. A possible way to decrease the sample variability is to compute the GIs as the simple average of different data downloads performed over different days. We also tried this approach as a robustness check, but we decided to use the original raw data coming from the single downloads because we found that using the raw data does not alter the final results, similarly to what found by Fantazzini and Toktamysova (2015) and D'Amuri and Marcucci (2017).

Google Trends has both advantages and limits when forecasting migration is of concern. In general, Google Trends has several advantages in terms of economy, coverage, and immediacy: they are free of charge, and they can cover larger sets of population than some of the traditional data sources, which may suffer from sample size limits. Moreover, they can allow researchers to monitor immigrants' intentions almost in real-time. In this regard, the main advantage of online search queries is the possibility to anticipate immigrants' movement, as highlighted by Böhme et al. (2019), who validated this proposition

by comparing the Gallup World Poll data about emigration[20] with the results obtained with Google Trends, and they found that Google Trends data can indeed now-cast the "*genuine migration intention*".

Yet, Google Trends data have also their limitations: for example, it is well known that online users may not represent the whole population, and these data may require significant cleaning, see Jun et al. (2018), Nikolopoulos et al. (2021a), and references therein. The impossibility to track specific categories of users may determine migration policies that perpetuate discrimination or neglect the needs of some groups. For these reasons, the latest research efforts try to combine online big data with more traditional data sources, see Salini et al. (2020) and Iacus and Porro (2021) for more details.

Despite these limitations, increasing literature showed that Google Trends and other online big data can still improve the understandings of migration patterns, see Hawelka et al. (2014), Zagheni et al. (2014), Moise et al. (2016), Iacus and Porro (2021), Sîrbu et al. (2021) for more details.

[20] This is a survey done over more than 160 countries and that wants to find whether the local individuals are planning to move to another country and, if so, whether the plan will take place within 12 months, see http://gallup.com for more details.

Appendix B

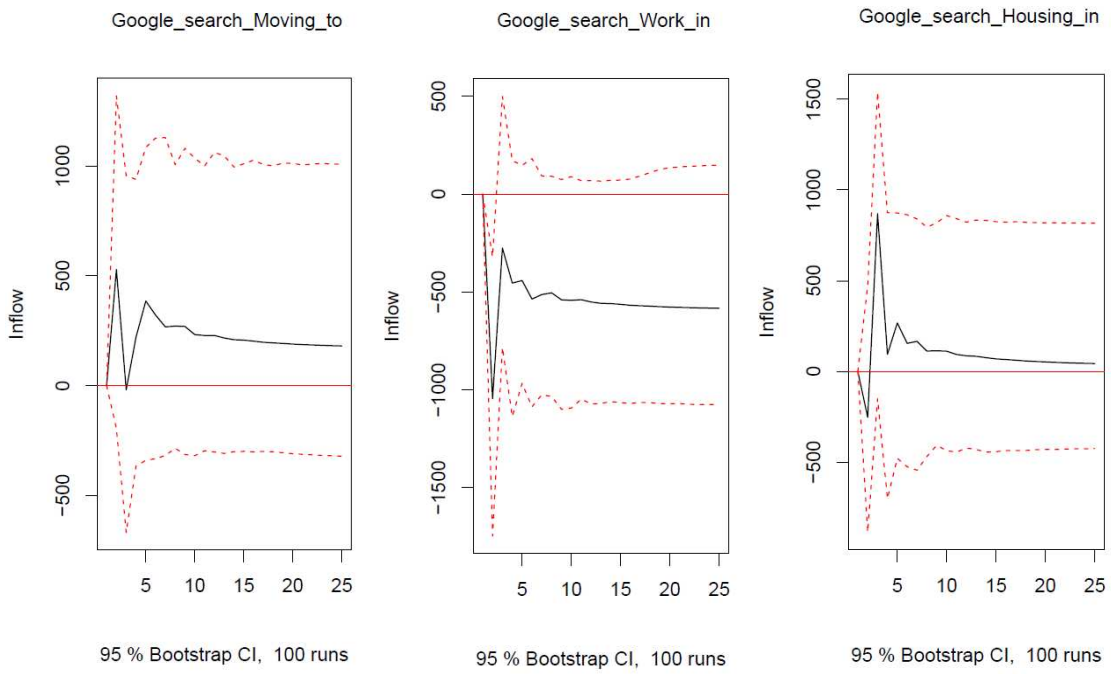

Figure A3. VECM(1) with centered seasonal dummies: orthogonalised impulse responses from a shock in Google searches on migration inflow in Moscow over 24 months.

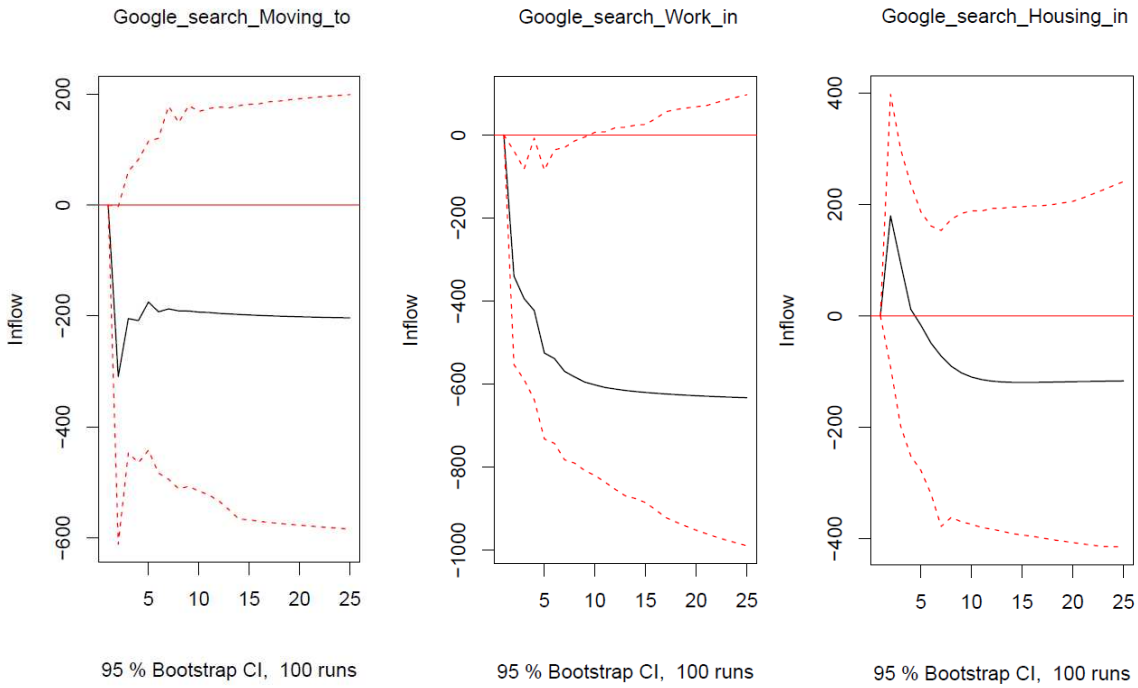

Figure A4. VECM(1) with centered seasonal dummies: orthogonalised impulse responses from a shock in Google searches on migration inflow in S.Petersburg over 24 months.

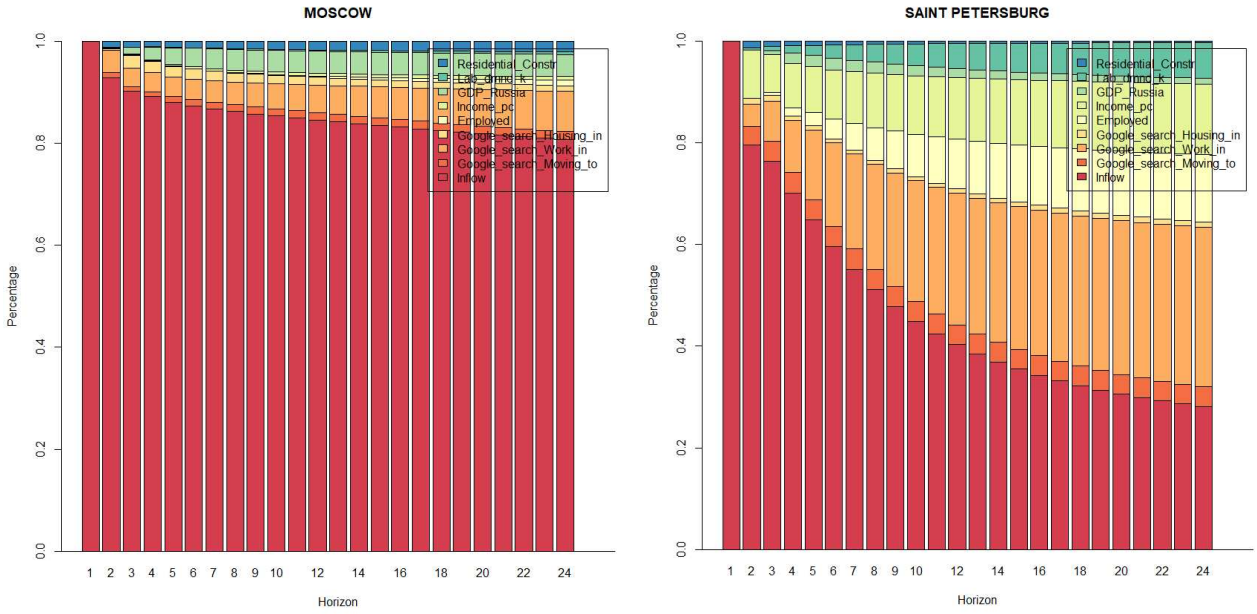

Figure A5. Forecast error variance decomposition of the VECM(1) with centered seasonal dummies: Moscow (left panel), Saint Petersburg (right panel).

Appendix C: Robustness Checks

We wanted to check how our previous results changed with models able to deal with potential parameter instability and with a large number of regressors, potentially larger than the number of observations. To achieve this goal, we employed the time-varying VAR model proposed by Casas and Fernandez-Casal (2018) and Casas et al. (2019), and a set of multivariate shrinkage estimation methods.

C.1. Parameter instability

We tested for the structural stability of our VAR(1) models using the generalized fluctuation tests discussed by Kuan and Hornik (1995), Zeileis et al. (2005), and Zeileis (2006). For sake of interest and space, we report below only the fluctuation test based on the moving OLS estimates for the VAR equation of the monthly migration flow in Moscow and Saint Petersburg, while the full results are available from the authors upon request[21].

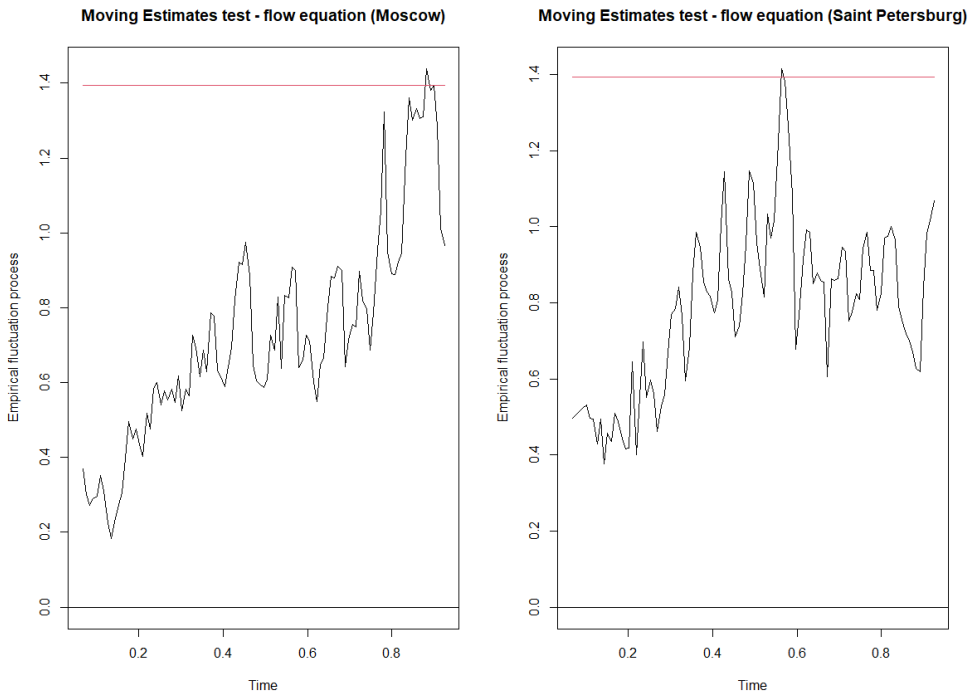

Figure 6. Fluctuation test based on the moving OLS estimates for the VAR equation of the monthly migration flow in Moscow and Saint Petersburg, with the boundary for the 5% confidence level (red line). The standardized sample cover the period 2009-2020.

Figure A6 and the battery of tests that we computed to test for structural stability highlighted that the evidence for parameter instability is mild or not significant. Nevertheless, we decided to implement the time-varying coefficient vector autoregressive model (TVVAR) proposed by Casas and Fernandez-Casal (2018) and Casas et al. (2019) to take any potential parameter instability into account:

[21] This (large) class of fluctuation tests for testing, monitoring and dating structural changes in linear regression models is implemented in the R package ***strucchange***.

$$\mathbf{Y}_t = \mathbf{\Phi}_{0,t} + \sum_{i=1}^{p} \mathbf{\Phi}_{i,t} \mathbf{Y}_{t-i} + \boldsymbol{u}_t, \quad \boldsymbol{u}_t \sim WN\left(0, \mathbf{\Sigma}_t\right) \tag{2}$$

where the elements of $\mathbf{\Phi}_{i,t}$ are unknown functions of either the rescaled time value $\tau = t / T$ with $\tau \in [0, 1]$, or of a random variable at time t. The variance-covariance matrix $\mathbf{\Sigma}_t$ can also be time varying. If the matrixes $\mathbf{\Phi}_{i,t}$ are a function of τ, then the TVVAR model is locally stationary in the sense of Dahlhaus (1997), which means that the functions in the matrices are constant or change smoothly over time. In this case, the TVVAR model (2) has a well-defined Wold representation given by

$$\overline{\mathbf{Y}}_t = \sum_{j=0}^{\infty} \mathbf{\Psi}_{j,t} \, \boldsymbol{u}_{t-j}$$

with $| \overline{\mathbf{Y}}_t - \mathbf{Y}_t | \to 0$ almost surely, $\mathbf{\Psi}_{0,t} = \boldsymbol{I}_n$, $\mathbf{\Psi}_{s,t} = \sum_{j=1}^{s} \mathbf{\Psi}_{s-j,t} \mathbf{\Phi}_{j,t}$ for horizons $s = 1, 2,...$, and where $\mathbf{\Psi}_{s,t}$ represent the time-varying coefficient matrices of the impulse response function (TVIRF), see Casas and Fernandez-Casal (2018) for more details. The orthogonal TVIRF can be computed using $\mathbf{\Psi}_{j,t} \boldsymbol{P}_t$ instead of $\mathbf{\Psi}_{j,t}$, where $\boldsymbol{P}_t$ is the lower triangular matrix obtained employing the Cholesky decomposition of $\mathbf{\Sigma}_t$ at time t given by $\mathbf{\Sigma}_t = \boldsymbol{P}_t \boldsymbol{P}_t'$.

The TVVAR model (2) can be estimated using a multivariate non-parametric Nadaraya–Watson estimator that minimizes a smoothed weighted sum of squared residuals, see Casas et al. (2019) for a detailed analysis of the asymptotic properties of this kernel estimator[22].

The orthogonal impulse responses from a shock in Google online searches on migration inflow in Moscow (left column) and Saint-Petersburg (right column) are reported in Figure A7, where the values reported are the means of the time-varying IRF over every time period.

[22] The TVVAR model is implemented in the R package ***tvReg***.

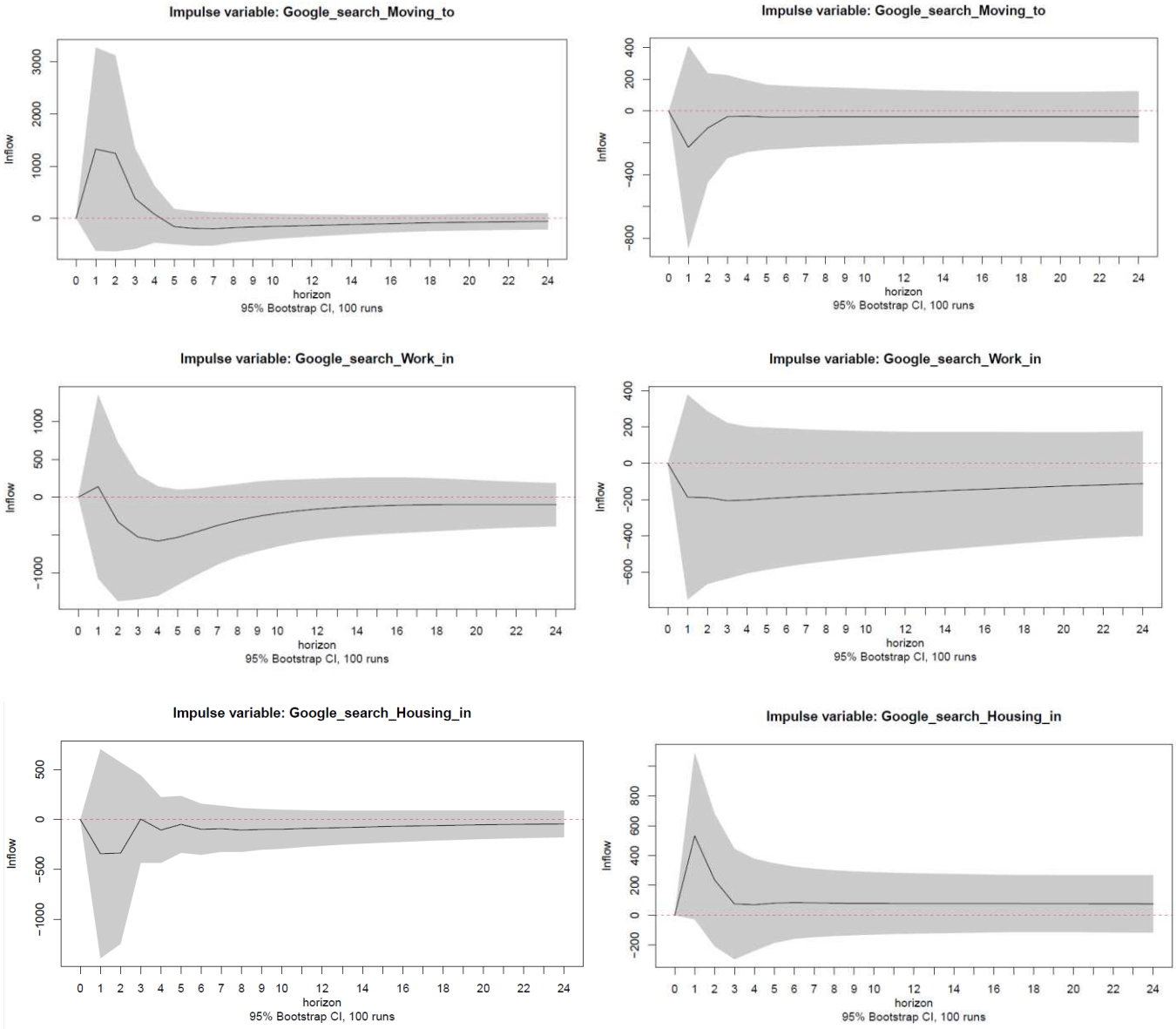

Figure A7. Orthogonal impulse responses from a shock in Google online searches on migration inflow in Moscow (left column) and Saint Petersburg (right column) using a TVVAR (1) model. The values reported are the means of the time-varying IRF over every period.

Similar to the baseline case, a one-time shock in online google searches related to emigration and job queries has a negative effect on migration inflows but, differently from the baseline case, these effects are no more significant.

The lack of significance of the IRFs can probably be explained by the larger variances of the TVVAR model estimates compared to traditional VAR models with constant parameters and by the weak evidence of model instability, which makes the TVVAR model more inefficient.

C.2. Additional lags

The simple VAR(1) model used in the baseline case can be an efficient way to deal with several variables, but it is hardly realistic, considering that the decision and the entire process to emigrate may take several months, at the very least[23]. Unfortunately, given the limited size of our dataset, VAR models with more than 6 lags were numerically unstable or simply impossible to estimate. Therefore, we resorted to multivariate shrinkage estimation methods that can be applied to high-dimensional VAR models with dimensionality potentially larger than the number of observations.

More specifically, we considered the multivariate ridge regression by Hoerl and Kennard (1970). If we rewrite the VAR model described in eq. (1) in a more compact form as follows,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}$$

where $\mathbf{Y}$ is a $(T - p) \times n$ matrix collecting the temporal observations of all endogenous variables, $\mathbf{X}$ is a $(T - p) \times (np+1)$ matrix collecting the lags of the endogenous variables and the constants, $\mathbf{B}$ is a $(np+1) \times n$ matrix of coefficients, while $\mathbf{U}$ is a $(T - p) \times n$ matrix of error terms, then the multivariate ridge regression estimator of $\mathbf{B}$ can be obtained by minimizing the following penalized sum of squared errors:

$$\mathbf{B}_{Ridge}(\lambda) = \arg\min_{\mathbf{B}} \frac{1}{T-p} \left\| \mathbf{Y} - \mathbf{XB} \right\|_F^2 + \lambda \left\| \mathbf{B} \right\|_F^2$$

where $\left\| \mathbf{A} \right\|_F = \sqrt{\sum_i \sum_j a_{ij}^2}$ is the Frobenius norm of a matrix $\mathbf{A}$, and $\lambda \geq 0$ is known as the regularization parameter or the shrinkage parameter. The ridge regression estimator $\mathbf{B}_{Ridge}(\lambda)$ has a closed form solution given by,

$$\mathbf{B}_{Ridge}(\lambda) = \left(\mathbf{X}'\mathbf{X} + (T-p)\lambda \mathbf{I} \right)^{-1} \mathbf{X}'\mathbf{Y}, \quad \lambda \geq 0$$

The shrinkage parameter λ can be automatically determined by minimizing the generalized cross-validation (GCV) score by Golub, Heath, and Wahba (1979):

$$GCV(\lambda) = \frac{1}{T-p} \left\| \mathbf{I} - \mathbf{H}(\lambda)\mathbf{Y} \right\|_F^2 \Bigg/ \left[\frac{1}{T-p} Trace(\mathbf{I} - \mathbf{H}(\lambda)) \right]^2$$

where $\mathbf{H}(\lambda) = \mathbf{X}'(\mathbf{X}'\mathbf{X} + (T-p)\lambda \mathbf{I})^{-1}\mathbf{X}'$.

Given our previous discussion, we considered a VAR(12) model estimated with the ridge regression estimator. The orthogonal impulse responses from a shock in Google online searches on migration inflow in Moscow (left column) and Saint Petersburg (right column) are reported in Figure A8.

[23] The first author of this paper immigrated to Moscow in August 2007: if the initial planning phase is considered, together with the time needed to satisfy all the administrative and migration requirements necessary for the physical transfer, the entire process took up to 1 year.

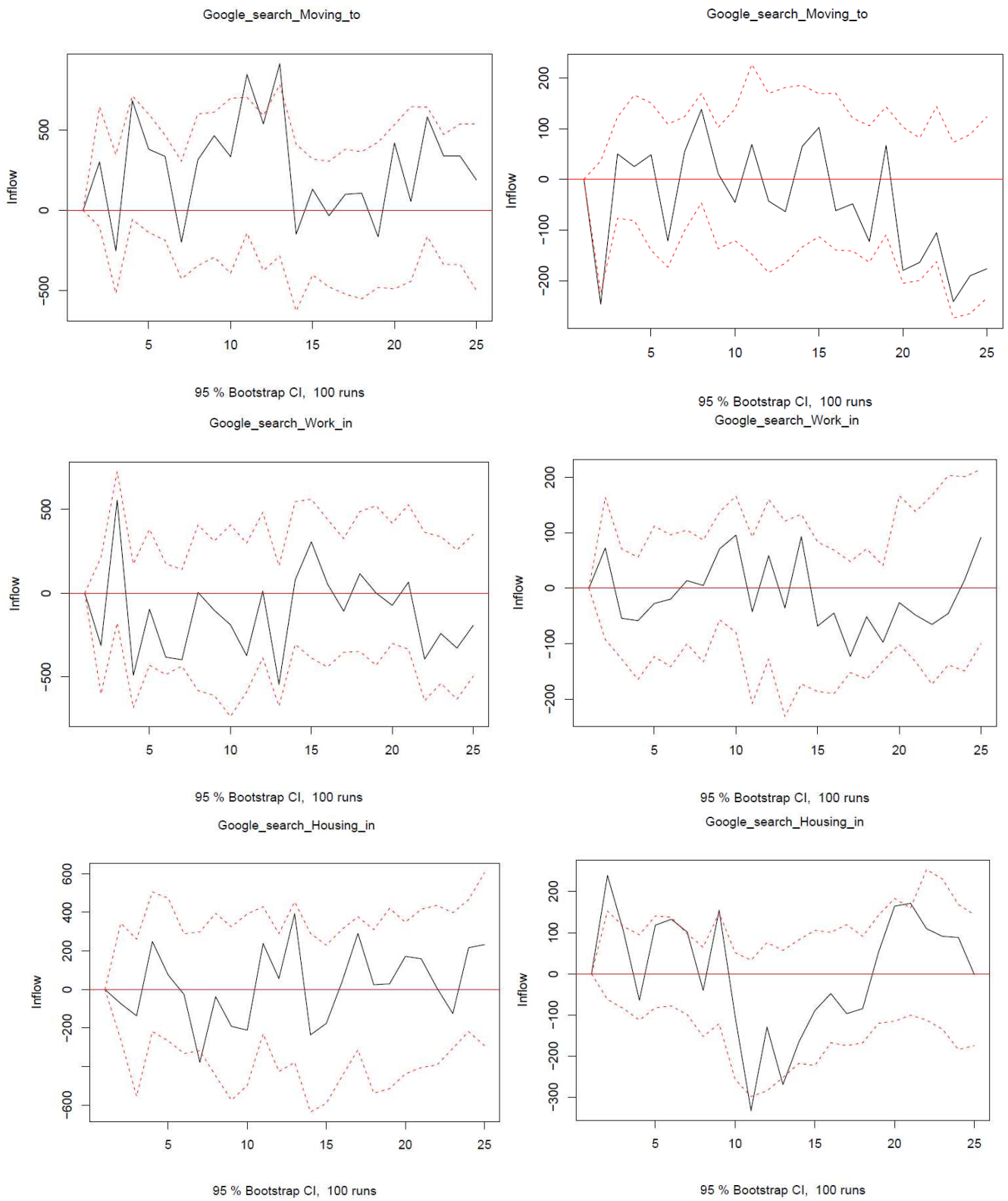

Figure 8. Orthogonal impulse responses from a shock in Google online searches on migration inflow in Moscow (left column) and Saint-Petersburg (right column) using a VAR(12) model estimated with the ridge regression estimator.

The estimated IRFs are similar to the baseline case, except for one-time shocks in online searches related to emigration that have a positive effect on migration inflows in Moscow, thus confirming similar evidence reported in Böhme et al. (2019). However, all these effects are no more statistically significant.

We remark that we also tried alternative multivariate shrinkage estimation methods for VAR models, like the nonparametric shrinkage estimation method proposed by Opgen-Rhein and Strimmer (2007), the full Bayesian shrinkage methods proposed by Sun and Ni (2004) and Ni and Sun (2005), and the semi-parametric Bayesian shrinkage method proposed by Lee et al. (2016): the results with these methods were qualitatively similar but their computational performance was much worse in several cases, so that we do not report them for sake of space and interest[24,25].

References

Aaronson, Daniel, Scott Brave, Andrew Butters, Michael Fogarty, Daniel Sacks, and Boyoung Seo. 2021. Forecasting unemployment insurance claims in realtime with Google Trends. *International Journal of Forecasting*, in press.

Abashin, Sergei 2014. Migration from Central Asia to Russia in the new model of world order. *Russian Politics & Law* 52(6): 8-23.

Ahrens, Achim, and Arnab Bhattacharjee. 2015. Two-step lasso estimation of the spatial weights matrix. *Econometrics* 3(1): 128-155.

Alonso, William. 1986. Systemic and log-linear models: from here to there then to now and this to that. Discussion Paper 86—10, Center for Population Studies. Harvard University, Cambridge, Massachusetts.

Algan, Yann, Fabrice Murtin, Elizabeth Beasley, Kazuhito Higa, and Claudia Senik. 2019. Well-being through the lens of the internet. *PloS one* 14(1): e0209562.

Altissimo, Filippo, Riccardo Cristadoro, Mario Forni, Marco Lippi, and Giovanni Veronese. 2010. New Eurocoin: Tracking economic growth in real time. *The review of economics and statistics* 92(4), 1024-1034.

Andrienko, Yuri, and Sergei Guriev. (2004). Determinants of interregional mobility in Russia. *Economics of transition* 12(1), 1-27.

Aprigliano, Valentina, Claudia Foroni, Massimiliano Marcellino, Gianluigi Mazzi, and Fabrizio Venditti. 2017. A daily indicator of economic growth for the euro area. *International Journal of Computational Economics and Econometrics* 7(1-2): 43-63.

Aruoba, S. Borağan, Francis X. Diebold, and Chiara Scotti. 2009. Real-time measurement of business conditions. *Journal of Business & Economic Statistics*, 27(4), 417-427.

Artola, Concha, and Enrique Martínez-Galán. 2012. Tracking the future on the web: construction of leading indicators using internet searches. *Banco de Espana Occasional Paper* 1203.

Bedrina, Elena, Yevgeniya Tukhtarova, and Natalia Neklyudova. 2018. Migration from Uzbekistan to Russia: Push-Pull Factor Analysis. In, *The International Science and Technology Conference "FarEastCon"*, pp. 283-296. Springer.

Bengtsson, Linus, Xin Lu, Anna Thorson, Richard Garfield, and Johan Von Schreeb. 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): e1001083.

Bijak, Jakub. 2011. *Forecasting international migration in Europe: A Bayesian view*. Springer Science & Business Media, vol. 24.

Bijak, Jakub, George Disney, Allan M. Findlay, Jonathan J. Forster, Peter WF Smith, and Arkadiusz Wiśniowski. 2019. Assessing time series models for forecasting international migration: Lessons from the United Kingdom. *Journal of Forecasting* 38(5): 470-487.

Billari, Francesco, Francesco D'Amuri, and Juri Marcucci. 2016. Forecasting births using Google. In *CARMA 2016: 1st International Conference on Advanced Research Methods in Analytics*. Editorial Universitat Politècnica de València.

Böhme, Marcus H., André Gröger, and Tobias Stöhr. 2020. Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics* 142: 102347.

Borup, Daniel, and Erik Christian Montes Schütte. 2020. In search of a job: Forecasting employment growth using google trends. *Journal of Business & Economic Statistics,* in press.

Burkhauser, Richard, Hahn, Markus, Hall, Matthew, and Nicole Watson. 2016. Australia Farewell: Predictors of emigration in the 2000s. *Population Research and Policy Review*, 35(2) 197-215.

Burnham, Kenneth, and Anderson, David. 2004. *Model selection and multi-model inference*. Second editionю NY: Springer-Verlag 63.

Casas, Isabel, and Ruben Fernandez-Casal. 2018. *tvreg*: Time-varying coefficients linear regression for single and multiple equations [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=tvReg (R package version 0.5.4)

Casas, Isabel, Eva Ferreira, and Susan Orbe. 2017. Time-varying coefficient estimation in SURE models. Application to portfolio management. *Journal of Financial Econometrics* nbz010.

[24] These additional results are available from the authors upon request.

[25] All the multivariate shrinkage estimation methods discussed in the text are implemented in the R package ***VARshrink.***

Choi, Hyunyoung, and Hal Varian. 2012. Predicting the present with Google Trends. *Economic Record* 88: 2-9.
Chort, Isabelle. 2014. Mexican migrants to the US: What do unrealized migration intentions tell us about gender inequalities? *World development* 59: 535-552.
Chudinovskikh, Olga., and Mikhail Denisenko. 2017. *Russia: A Migration System with Soviet Roots*. Washington, DC: Migration Policy Institute. https://www.migrationpolicy.org/print/15920
Chudinovskikh Olga, and Mikhail Denisenko. 2020. Labour Migration on the Post-Soviet Territory. In, *Migration from the Newly Independent States. Societies and Political Orders in Transition*. Springer: pp. 55-80.
Clemen, Robert T. 1989. Combining forecasts: A review and annotated bibliography. *International journal of forecasting* 5(4): 559-583.
Constant, Amelie. and Klaus Zimmermann, 2011. Circular and repeat migration: counts of exits and years away from the host country. *Population Research and Policy Review*, 30(4): 495-515.
Dahlhaus, Rainer. 1997. Fitting time series models to nonstationary processes. *Annals of Statistics* 25: 1–37
D'Amuri, Francesco, and Juri Marcucci. 2017. The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting* 33(4): 801-816.
Demidova, Anastasia, Olga Druzhinina, Olga Masina, and Alexey Petrov. 2020. Computer research of the controlled models with migration flows. In, *Proceedings of the 10th International Conference in Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems* (ITTMM-2020), volume 2639: 117–129.
Demintseva, Ekaterina, and Vera Peshkova. 2014. Migranty iz Srednei Azii v Moskve. *Demoscope Weekly*: 597–598. http://www.demoscope.ru/weekly/2014/0597/tema01.php
Demintseva, Ekaterina, and Daniel Kashnitsky. 2016. Contextualizing Migrants' Strategies of Seeking Medical Care in Russia. *International Migration* 54(5): 29–42.
Demintseva, Ekaterina. 2017. Labour migrants in post-Soviet Moscow: patterns of settlement. *Journal of ethnic and migration studies*, 43(15): 2556-2572.
Denisenko Mikhail, Mkrtchyan Nikita, and Olga Chudinovskikh. 2020. Permanent Migration in the Post-Soviet Countries. In, *Migration from the Newly Independent States. Societies and Political Orders in Transition*. Springer: pp. 23-53.
Docquier, Frédéric, and Hillel Rapoport. 2012. Globalization, brain drain, and development. *Journal of economic literature* 50(3): 681-730.
Docquier, Frédéric, Giovanni Peri, and Ilse Ruyssen. 2014. The cross-country determinants of potential and actual migration. *International Migration Review* 48(1): 37-99.
Elliott, Graham. 1998. On the robustness of cointegration methods when regressors almost have unit roots. *Econometrica* 66: 149-158.
Dustmann, Christian, and Anna Okatenko. 2014. Out-migration, wealth constraints, and the quality of local amenities. *Journal of Development Economics* 110: 52-63.
Efimova, Elena, and Semen Mikhaltsov. 2017. Road Traffic as a Factor of Regional Development: Case of Saint Petersburg Region, Russian Federation. *Procedia Engineering* 187: 135-142.
Ette, Andreas, Heß, Barbara, and Lenore Sauer. 2016. Tackling Germany's demographic skills shortage: permanent settlement intentions of the recent wave of labour migrants from non-European countries. *Journal of International Migration and Integration* 17(2): 429-448.
Ettredge, Michael, John Gerdes, and Gilbert Karuga. 2005. Using web-based search data to predict macroeconomic statistics. *Communications of the ACM 48*(11): 87-92.
Fantazzini, Dean, and Nikita Fomichev. 2014. Forecasting the real price of oil using online search data. *International Journal of Computational Economics and Econometrics* 4(1-2): 4-31.
Fantazzini, Dean, and Zhamal Toktamysova (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics* 170: 97-135.
Friedman, Milton. 1937. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association* 32(200): 675-701.
Fuchs, Johann, Söhnlein, Doris, and Patrizio Vanella. 2021. Migration Forecasting—Significance and Approaches. *Encyclopedia* 1(3): 689-709.
Gerber, Theodore, and Jane Zavisca. 2020. Experiences in Russia of Kyrgyz and Ukrainian labor migrants: ethnic hierarchies, geopolitical remittances, and the relevance of migration theory. *Post-Soviet Affairs* 36(1): 61-82.
Golub, Gene H., Michael Heath, and Grace Wahba. 1979. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics* 21(2): 215–23.
Gospodinov, Nikolay, Herrera, Ana María, and Elena Pesavento. 2013. Unit roots, cointegration, and pretesting in VAR models. *Advances in Econometrics* 32: 81-115.
Hawelka, Bartosz, Sitko, Izabela, Beinat, Euro, Sobolevsky, Stanislav, Kazakopoulos, Pavlos and Carlo Ratti. 2014. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* 41(3): 260-271.
Hayashi, Fumio. 2000. *Econometrics*. Princeton: Princeton University Press.
Heleniak, Timothy. 2009. Migration of the Russian Diaspora after the Breakup of the Soviet Union. *Journal of International Affairs* 57(2): 99–117.

Hoerl, Arthur E., and Robert W. Kennard. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12(1): 55–67
Hsiao, Cheng, and Shui Ki Wan. 2014. Is there an optimal forecast combination? *Journal of Econometrics* 178: 294-309.
Human Rights Watch. 2009. Are You Happy to Cheat Us? Exploitation of Migrant Construction Workers in Russia. https://www.hrw.org/report/2009/02/10/are-you-happy-cheat-us/exploitation-migrant-construction-workers-russia
Hyndman, Rob J., and George Athanasopoulos. 2018. *Forecasting: principles and practice*. OTexts.
Hyndman, Rob J., and Yeasmin Khandakar . 2008. Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software* 27(3): 1-22.
Iacus, Stefano, and Giuseppe Porro. *Subjective Well-being and Social Media*. CRC Press, 2021.
Inoue, Atsushi, and Lutz Kilian. 2020. The uniform validity of impulse response inference in autoregressions. *Journal of Econometrics* 215(2): 450-472.
Johansen, Soren. 1995. *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford: Oxford University Press.
Johansen, Soren. 2006. Cointegration: a survey. In: *Palgrave handbook of econometrics: Volume 1, Econometric theory*. Edited by Mills, T.C. and Patterson, K. Basingstoke (UK): Palgrave MacMillan, pp. 540–577.
Jun, Seung-Pyo, Hyoung Sun Yoo, and San Choi. 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.
Keilman, Nico, and Tomaš Kučera. 1991. The impact of forecasting methodology on the accuracy of national population forecasts: Evidence from the Netherlands and Czechoslovakia. *Journal of forecasting* 10(4): 371-398.
Keilman, Nico, Dinh Quang Pham, and Arve Hetland. 2001. Norway's uncertain demographic future. *Statistics Norway Social and Economic Studies* no. 105. Statistics Norway: Oslo
Kikas, Riivo, Dumas, Marlon and Ando Saabas. 2015. Explaining international migration in the skype network: The role of social network features. In *Proceedings of the 1st ACM Workshop on Social Media World Sensors*, pp. 17-22.
Korovkin, Andrei., Dolgova, Irina, and Ekaterina Edinak. 2013. Analysis of the relationship between internal migration and socio-economic differentiation of regions (on the example of the central Federal District). *Scientific works: Institute for Economic Forecasting, Russian Academy of Sciences,* pp 71-94.
Kruskal, William H., and W. Allen Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260): 583-62.
Kuan, Chung-Ming, and Kurt Hornik. 1995. The generalized fluctuation test: A unifying view, *Econometric Reviews,* 14, 135–161.
Kuhlenkasper, Torben, and Max Friedrich Steinhardt. 2017. Who leaves and when? Selective outmigration of immigrants from Germany. *Economic Systems* 41(4): 610-621.
Lam, Clifford, and Pedro CL Souza. 2020. Estimation and selection of spatial weight matrix in a spatial lag model. *Journal of Business & Economic Statistics* 38(3): 693-710.
Lee, Namgil, Hyemi Choi, and Sung-Ho Kim. 2016. Bayes Shrinkage Estimation for High-Dimensional VAR Models with Scale Mixture of Normal Distributions for Noise. *Computationl Statistics & Data Analysis* 101: 250–76.
Lütkepohl, Helmut. 2005. *New introduction to multiple time series analysis*. Berlin: Springer Science and Business Media.
Maddala, Gangadharrao S., and In-Moo Kim. 1998. *Unit Roots, cointegration, and structural change*. Cambridge University Press.
Maravall, Agustin. 2011. *Seasonality Tests and Automatic Model Identification in TRAMO-SEATS*. Bank of Spain.
Mayda, Anna Maria. 2010. International migration: A panel data analysis of the determinants of bilateral flows. *Journal of Population Economics 23*(4):1249-1274.
McLaren, Nick, and Rachana Shanbhogue. 2011. Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, (2011), Q2.
Moise, Izabela, Gaere, Edward , Merz, Ruben, Koch, Stefan and Evangelos Pournaras. 2016. Tracking language mobility in the Twitter landscape. In 2016 *IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pp. 663-670.
Nikolopoulos, Konstantinos, Christos Tsinopoulos, and Chrysovalantis Vasilakis. 2021a. Operational research in the time of COVID-19: The 'science for better'or worse in the absence of hard data. *Journal of the Operational Research Society*, 1-2.
Nikolopoulos, Konstantinos, Sushil Punia, Andreas Schäfers, Christos Tsinopoulos, and Chrysovalantis Vasilakis. 2021b. Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. *European journal of operational research* 290(1): 99-115.
Ni, Shawn, and Dongchu Sun. 2005. Bayesian Estimates for Vector Autoregressive Models, *Journal of Business and Economic Statistics* 23(1): 105–117.
Ollech, Daniel, and Karsten Webel. 2020. A random forest-based approach to identifying the most informative seasonality tests. Bundesbank Discussion Paper No. 55/2020.
Opgen-Rhein, Rainer, and Korbinian Strimmer. 2007. Learning Causal Networks from Systems Biology Time Course Data: An Effective Model Selection Procedure for the Vector Autoregressive Process. *BMC Bioinformatics* 8(2): 1-7.
Ortega, Francesco, and Giovanni Peri. 2013. The effect of income and immigration policies on international migration. *Migration Studies* 1(1): 47-74.
Pavlovskij, Egor. 2017. Arima Models in the Short-Term Forecasting of Internal Migration in Russia. *Voprosy Statistiki*, 1(10): 53-63.
Qin, Yu, and Hongjia Zhu. 2018. Run away? Air pollution and emigration interests in China. *Journal of Population Economics* 31(1): 235-266.

Ravenstein, Ernest George. 1885. The laws of migration. *Journal of the statistical society of London* 48(2): 167-235.
Reeves, Madeleine. 2013. Clean Fake: Authenticating Documents and Persons in Migrant Moscow. *American Ethnologist* 40(3):508–524.
Reeves, Madeleine. 2015. Living from the Nerves: Deportability, Indeterminacy, and the 'feel of Law' in Migrant Moscow. *Social Analysis* 59(4): 119–136.
Ryazantsev, Sergey. 2016. Labour Migration from Central Asia to Russia in the Context of the Economic Crisis. *Russia in Global Affairs*, August 31. http://eng.globalaffairs.ru/valday/Labour-Migration-from-Central-Asia-to-Russia-in-the-Context-of-the-Economic-Crisis-18334
Salini, Silvia, Siletti Elena, and Porro Giuseppe. 2020. Controlling for Selection Bias in Social Media Indicators through Official Statistics: a Proposal. *Journal of Official Statistics* 36(2): 315-338.
Schenk, Caress. 2018. *Why Control Immigration? Strategic Uses of Migration Management in Russia*. Toronto: University of Toronto Press.
Sîrbu, Alina, Gennady Andrienko, Natalia Andrienko, Chiara Boldrini, Marco Conti, Fosca Giannotti, Riccardo Guidotti et al. 2021. Human migration: the big data perspective. *International Journal of Data Science and Analytics* 11(4): 341-360.
Stock, James H., and Mark W. Watson. 1993. A simple estimator of cointegrating vectors in higher order integrated systems. *Econometrica* 61(4): 783–820.
Sun, Dongchu, and Shawn Ni. 2004. Bayesian Analysis of Vector-Autoregressive Models with Noninformative Priors. *Journal of Statistical Planning and Inference* 121(2): 291–309.
Tamgno, James K., Roger M. Faye, and Claude Lishou. 2013. Verbal autopsies, mobile data collection for monitoring and warning causes of deaths. In *2013 - 15th International Conference on Advanced Communications Technology (ICACT),* pp. 495-501. IEEE.
Timmermann, Allan. 2006. Forecast combinations. *Handbook of economic forecasting* 1: 135-196.
Timoshkin, Dmitry. 2020. Construction of Horizontal Networks on "Migrant" Russian-Language Digital Platforms, *Journal of Siberian Federal University. Humanities & Social Sciences*, 13(5): 688-699.
United Nations. 2017. *International Migration Report 2017*. New York: United Nations Population Division
Varaksin, Sergei, and Natal'ya Varaksina. 2017. Application of fuzzy linear regression for modeling the migration process in Russia. In, *Economic and Social Development: Book of Proceedings*: 332-340.
Vakulenko, Elena, Nikita Mkrtchyan, and Kirill Furmanov. 2011. Modeling registered migration flows between regions of the Russian Federation. *Applied Econometrics* 21(1): 35-55.
Vakulenko, Elena, and Nikita Mkrtchyan. 2020. Factors of Interregional Migration in Russia Disaggregated by Age. *Applied Spatial Analysis and Policy* 13(3): 609-630.
Zagheni, Emilio, Venkata Rama Kiran Garimella, Ingmar Weber, and Bogdan State. 2014. Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*, pp. 439-444.
Zeileis, Achim. 2006. Implementing a class of structural change tests: An econometric computing approach. *Computational Statistics & Data Analysis* 50(11):2987–3008.
Zeileis, Achim, Friedrich Leisch, Christian Kleiber, and Kurt Hornik. 2005. Monitoring structural change in dynamic econometric models. *Journal of Applied Econometrics* 20(1): 99–121.
Welch, Bernard Lewis. 1951. On the Comparison of Several Mean Values: An Alternative Approach. *Biometrika* 38(3/4): 330-336.
Willekens, Frans. 1980. Entropy, multiproportional adjustment and the analysis of contingency tables. *Systemi Urbani 2*(3):171-201.
Wilson, Alan. (1970). *Entropy in urban and regional modelling.* Volume 1. London: Routledge.